Tournament Theory Applied to Professional Tennis Players

Nuno Vicente Amado e Vasconcelos Duarte Silva

Thesis to obtain the Master of Science Degree in

Industrial Engineering and Management

Supervisor: Prof. António Sérgio Constantino Folgado Ribeiro

Examination Committee

Chairperson: Prof. Miguel Simões Torre Preto
Supervisor: Prof. António Sérgio Constantino Folgado Ribeiro
Member of the committee: Prof. Hugo Miguel Fragoso de Castro Silva

November 2021

Abstract

Tournament theory has been used to understand what motivates employees to advance in their careers. Tournament models imply that the introduction of a differentiated reward system, where the biggest prize is awarded to the best competitor, increases the average level of effort, success and performance of players (Lazear and Rosen, 1981). The goal of this dissertation is to study how professional tennis players respond to the two main incentives they can get while playing a professional tennis match, prize money and ATP ranking points, if the players exert bigger effort while facing both these incentives and, if one of the incentives, prize money or ranking points, is more appealing than the other. We used data from the 2019 ATP Tour and created two regressions applying OLS method using the rStudio software for this purpose: one of the regressions presents a model to assess the most adjusted prize money for match while the other regression does the same, but for ranking points. It was found that, although the structure of the reward system for both incentives in professional tennis is not structured accordingly to tournament theory, with no extra reward in the final round to ensure the optimization of effort from the players (Rosen, 1986), players will exert bigger effort while facing these incentives and that, prize money is more appealing than ranking points to professional tennis players.

Key words: tournaments, incentives, tennis players, effort, prize money, ranking points, reward system

Resumo

A Teoria dos Torneios tem sido usada para perceber o que motiva os funcionários a progredir nas suas carreiras. Os modelos de torneio implicam que a introdução de um sistema de recompensa diferenciado, onde o maior prémio é atribuído ao melhor competidor, aumenta o nível médio de esforço, sucesso e desempenho dos jogadores (Lazear e Rosen, 1981). O objetivo desta dissertação é estudar de que forma os jogadores de ténis profissionais respondem aos dois principais incentivos que podem ganhar num jogo de ténis profissional, prémio em dinheiro e pontos de ranking ATP, se os jogadores exercem maior esforço devido aos dois incentivos e, se algum dos incentivos, prémio em dinheiro ou pontos de ranking, é mais atraente do que o outro. Usámos dados do ATP Tour 2019 e criámos duas regressões aplicando o método OLS, usando o software rStudio para esse propósito: uma das regressões apresenta um modelo para avaliar o prémio em dinheiro mais ajustado para um certo jogo, enquanto a outra regressão faz o mesmo, mas para os pontos de ranking. Verificou-se que, embora a estrutura do sistema de recompensa para os dois incentivos no ténis profissional não esteja estruturada de acordo com a Teoria dos Torneios, por não haver uma recompensa extra na ronda final que garanta a otimização do esforço dos jogadores (Rosen, 1986), os jogadores exercem um esforço maior devido a esses incentivos e que, prémio em dinheiro é mais atraente do que os pontos de ranking para jogadores de ténis profissionais.

Palavras chave: torneios, incentivos, jogadores de ténis, esforço, prémio em dinheiro, pontos de ranking, sistema de recompensa

Table of contents

ABSTRACT .... **I**

RESUMO .... **III**

LIST OF ABBREVIATIONS .... **VI**

LIST OF FIGURES .... **VII**

LIST OF TABLES .... **VIII**

1. INTRODUCTION **1**

1.1 Introduction 1
1.2 Problem Definition and Objectives 3
1.3. Organization of Chapters 3

2. LITERATURE REVIEW **4**

2.1. Tournament Theory and its applications 4
2.1.1. Tournament Theory 4
2.1.2. Tournament Theory applied to Industry 6
2.1.3. Tournament Theory applied to Sports 8
2.2. Tournament Theory applied to Tennis 11
2.3. Why tennis fits Tournament Theory 13

3. RESEARCH METHODOLOGY AND DATA **15**

3.1. Dataset 15
3.1.1. Variables 15
3.1.1.1. Dependent Variables 15
3.1.1.2. Independent Variables 16
3.2. Research Methodology 19
3.2.1. Least squares Model 21
3.3. ATP Tour 22

4. DESCRIPTIVE ANALYSIS **27**

4.1. Descriptive Analysis to the Dependent Variables 27
4.2. Descriptive Analysis to the Independent Variables 33
4.2.1. Match Difficulty Variables 33
4.2.2. Effort Variables: 35
4.2.3. Players' Characteristics Variables: 36

5. RESULTS **37**

5.1. Chosen Models 37
5.1.1. Prize Money 41
5.1.2. Ranking Points 44
5.2. Answering the questions 49

6. CONCLUSION **50**

REFERENCES **52**

APPENDIX **55**

List of Abbreviations

ATP	*The Association of Tennis Professionals*
IMSA	*International Motor Sports Association*
NASCAR	*National Association for Stock Car Auto Racing*
PGA	*Professional Golfers' Association*

List of Figures

Figure 3.1 - ATP Tour Ranking Points Distribution

Figure 3.2 - 2019 US Open Prize Money Distribution $

Figure 3.2 - 2019 ATP 250 Pune Prize Money Distribution $

Figure 4.1 – Matches by Tournament Level

Figure 4.2 – Matches by Surface

List of Tables

Table 3.1 – Dependent Variables

Table 3.2 – Independent Variables

Table 3.3 – Surface Dummy Variables

Table 3.4 – Draw Size Dummy Variables

Table 3.5 – Tournament Level Dummy Variables

Table 3.6 – Tournament Round Dummy Variables

Table 3.7 - 2019 US Open Prize Money Distribution

Table 3.8 - 2019 ATP 250 Pune Prize Money Distribution

Table 4.1 – Dependent Variables Descriptive Statistics

Table 4.2 – Prize Money Frequency Table

Table 4.3 – Ranking Points Frequency Table

Table 4.4 – Prize Money per Surface

Table 4.5 – Prize Money by Round and Tournament Level

Table 4.6 – Correlation Matrix

Table 4.7 – Match Difficulty Descriptive Statistics

Table 4.8 – Match Difficulty per Tournament Level

Table 4.9 – Match Difficulty per Round

Table 4.10 - Players' Effort Descriptive Statistics

Table 4.11 – Players' Characteristics Descriptive Statistics

Table 5.1 – Prize Money Models

Table 5.2 – Ranking Points Models

Table 5.3 – Estimated Prize Money Regression

Table 5.4 – Estimated Ranking Points Regression

Table A1 -2019 Grand Slams Total Prize Money

Table A2 - 2019 ATP Masters 1000 Total Prize Money

Table A3 - 2019 ATP Tour 500 Total Prize Money

Table A4 - 2019 ATP Tour 250 Tournaments Prize Money

1. Introduction

1.1 Introduction

In conventional business operations, many companies and organizations use payment systems based on performance or rank results in order to encourage the effort of the average worker. Tournament theory has been used to understand what motivates employees/professional players to advance in their careers/get promoted. It was developed based on the observation that workers in a firm are sometimes not payed based on their output level but instead based on their relative position in the firm. Given an ideal pay dispersion scheme, instead of output compensation, compensation based on their position can be really effective and it is less costly to observe the relative position than the worker's output. Tournament models imply that the introduction of a differentiated reward system, where the biggest prize is awarded to the best competitor, increases the average level of effort, success and performance of players (Lazear and Rosen, 1981).

A tournament model is typically defined by three main features: the prize structure is set ahead of time and, regardless of the margin of victory, the winner earns the same prize; relative output is significant, not absolute output (even if all the rivals have outstanding talent, what counts is that a player performs well enough to win the game) and the effort of a competitor depends on the size of the price differential between winning and losing. The greater the gap, the greater the effort that the contestants make to win it (Lazear and Rosen, 1981; Rosen, 1986).
While writing about elimination tournaments with a finite/restricted number of players, just like tennis tournaments, Rosen (1986) defends that the ideal prize structure should provide an additional incentive in the final stage to ensure that the players optimize their efforts.

Tournament theory says that when the contestants of a tournament are homogeneous, i.e. no contestant is known to be stronger than the others, the total amount of effort exerted by the contestants in the tournament is higher (Lazear and Rosen, 1981; Rosen, 1986; Prendergast, 1999; Sunde, 2003). Tournaments with uneven contestants, where one contestant is *ex ante* stronger than the others, result in a lower effort for both of the competitors. In other words, the bigger the initial gap of a contestant, the more costly is for them to compensate this disadvantage gap and a strong player knowing *ex ante* their advantage will difficultly maintain a high effort and can reduce it without jeopardizing their probable success (Sunde, 2003).
Sunde (2003**)** finds that the key reason for companies to design promotion tournaments is to provide employees with rewards for effort and given that tournaments between heterogeneous contestants are the standard in practice, whereas homogeneous contestants are the exception, the absence of proof for the implications of heterogeneity among contestants calls for an empirical analysis of the theoretical predictions.

Various sports related studies, papers and journals have already analyzed elimination-style tournaments and tournament theory in professional bowling (Abrevaya, 2002), golf (Ehrenberg and Bognanno, 1990a, 1990b), tennis (Sunde, 2003), for example, with different results obtained among their studies.

Using data from professional bowling, Abrevaya (2002) analyzes the determinants of match winning probabilities of *ex ante* unequal players in ladder tournaments. He finds no financial reward impacts on the absolute score of a player or the possibility of an upset; the number of previous wins in the championship round (" hot hand effect ") and the player’s ability to maintain this momentum are the main variables in the explanation of performance. However, he does not consider effort exertion.

Ehrenberg and Bognanno (1990a) were the first using data from golf tournaments (PGA tour) in order to understand if incentives matter. Overall, as calculated by the lower scores required to complete a round, they found that the prize money at stake has a dramatically positive effect on the concentration and effort players exert.

Sunde (2003, 2009) tests the effects of incentives by using data regarding the two final stages (semi-final and final) of the major tennis tournaments, Grand Slams and the Masters Series. He finds a positive relation between number of victories and stage prizes.

Professional tennis is probably one of the best sports to be studied in the eyes of tournament theory as said by Sunde (2003) since it uses almost exclusively single elimination tournaments in which the pay gaps between the successive stages of the tournament are increasingly and the losers go home with their prize money.

In theoretical models of single elimination tournaments, the structure of information available in tennis data resembles the structure and requirements of information. At the same time, tennis tournaments mimic the atmosphere of a business tournament, and contain major stakes for the players. For each tournament and at each stage of it, prizes are known *ex ante*. The strength of the players can be inferred from their rank prior to the respective tournament, allowing a measure of the heterogeneity of the participants to be constructed, which is not commonly available in other data. Before the match starts, rankings and prizes are already predetermined and well known, a fact which satisfies the theory's criteria and prevents endogeneity issues such as sabotage. In addition, tennis tournaments end with the final, they have a finite horizon unlike intermediate-level firm data.

In this dissertation, professional tennis players will be the ‘employees’ and the stages of the tournaments will be their relative position in the firm in order to study and comprehend tournament theory, focusing on professional tennis.

While other papers related to tennis applied to tournament theory mainly focus on two questions: “Does the effort of a competitor depend on the size of the prize differential between winning and losing?” and “When the tournament is heterogeneous and both players know *ex ante* that one is stronger than the other, do they both exert less effort?”, this thesis approaches tournament theory in a different way by showing how professional tennis players respond to the two main incentives they can get while playing a professional tennis match: prize money and ATP ranking points.

1.2 Problem Definition and Objectives

Annually, more than 750 tennis tournaments and more than 20 thousand games are played all around the world. These tournaments are divided into six main categories, played on different surfaces, either outdoor or indoor[1] and offer different rewards, ranking points and prize money. To reach the top, a player needs to progress in the ranking in order to have direct entry in the bigger tournaments which award bigger prize money. A top 50 player has access to tournaments that a top 500 player does not. Plus, to get there, they have to get ranking points every week.

The effort, intensity, game and stroke placed in every tournament, vary every week and from player to player. This leads us to ask some questions and try to understand why this happens. What motivates professional tennis players to perform better?

- Are these two incentives appealing to the players?
- If so, is one of the incentives, prize money or ranking points, more appealing than the other?
- For each of these two incentives do players exert greater effort?

This project aims to answer these questions and contribute to the number of empirical analyses that already exist about compensation models and effort exertion, using data from professional tennis.

1.3. Organization of Chapters

This dissertation contains six different chapters.
The Introduction contextualizes the main topics that will be addressed, clarifies the motivation for the study and describes the problem and the objectives. Chapter 2, the Literature Review, provides a theoretical framework, discussing the main concepts and studies related to tournament theory, and explains the ATP Tour. Chapter 3, Research Methodology and Data, presents the dataset's structure, gives an introduction to the variables that will be used and explains the research methodology developed in the dissertation stage, while Chapter 4, Descriptive Analysis, performs a statistical analysis and characterization of the data.
Chapter 5, Results, presents the results founds and discusses them and, in the last chapter of this project, the Conclusion, the main ideas concluded from this dissertation and its concepts are summarized.

[1] All the information about tennis tournaments and how the ATP Tour works will be further explained in chapter 3.3 (see pages 22 to 26).

2. Literature Review

This chapter will provide the theoretical framework to address the topic under study, tournament theory applied to professional tennis players. It begins by introducing tournament theory, its main concepts and why it started to be used. Next, we talk about tournament theory different applications, in industry, sports and more specifically in professional tennis, the focus of this thesis and the advantages of studying tournament theory using tennis data. Finally, we introduce the ATP Tour, the professional tennis tour.

2.1. Tournament Theory and its applications

2.1.1. Tournament Theory

Tournament theory has been used for almost 40 years to understand what motivates employees/professional players to advance in their careers/get promoted, getting more attention between 1980-2000. It was developed on the observation that workers in a firm sometimes are not payed based on their output level but instead on their relative position in that firm.

Firms face a maximization dilemma in the labor market in which, with a fixed wage purse, they strive to optimize the efficiency of their human resources. "Piece rate" compensation, or compensation in the form of a commission, and "salary-based" compensation are the two most common payment types. In order to calculate overall payment methods, "piece rate" compensation relies solely on the productivity of each worker individually, while "salary-based" compensation relies on each individual's relative role in the firm. "Piece rate" compensation can seem to be the most beneficial choice for the organization as it explicitly encourages maximization of effort but controlling outputs may be an expensive undertaking (Lazear and Rosen, 1981).

A combination of “salary-based” and "piece rate" pay is often used by most compensation systems (i.e. base salary plus bonus). However, in practice, measuring the commitment of an employee and therefore the quality of the pay structure of an organization is extremely hard.

Lazear and Rosen (1981) come with three conclusions:

First, given an ideal pay dispersion scheme, instead of output compensation, compensation based on workers position can be really effective and it is less costly to observe the relative position than the worker’s output. Tournament models imply that the introduction of a differentiated reward system, where the biggest prize is awarded to the best competitor, increases the average level of effort, success and performance of workers/players.

Second, the prize salary structure no longer duplicates the resource distribution caused by the ideal piece rate when risk aversion is implemented. Contests will appear to rule events with a high level of inherent riskiness. The competitive piece rate appears to rule for reasonably low levels of inherent riskiness. Under risk aversion, a highly skewed total income distribution is the inevitable result of combining the income distribution for the paid prizes with that for those paid piece rates.

Third, workers/competitors are allowed to be heterogeneous because this adds an important result. Competitive contests would not sort employees generally in a way that results in effective resource distribution. Low-quality employees, in particular, would try to “contaminate” a company with high-quality employees, even though there are no collaborations in production. If low-quality employees are not stopped from joining, this “contamination” leads to a general breakdown of the efficient solution. As a result, one justification for high-quality companies using initial qualifications when assigning jobs to applicants is this.

A tournament model is typically defined by three main features. First of all, the prize structure is set ahead of time and, regardless of the margin of victory, the winner earns the same prize. Secondly, relative output is significant, not absolute output. Even if all the rivals can have outstanding talent, what counts is that a player performs well enough to win the game. Thirdly, the effort of a competitor depends on the size of the price differential between winning and losing. The greater the gap, the greater the effort that the contestants make to win it (Lazear and Rosen, 1981; Rosen, 1986). While talking about elimination tournaments with a finite/restricted number of players, Rosen (1986) defends that the ideal prize structure should provide an additional incentive in the final stage to ensure that the players optimize their efforts.

In 1981, Lazear and Rosen demonstrated that the prize money leads to the decision of entering a tournament, but not to the decision of how much effort will be exerted. Instead, the degree to which the worker's effort affects the worker's probability of winning or improving a position is determined by the gap between the rewards given to the different places and the degree to which the worker's effort affects the worker's probability of winning or improving a place. In 1986, Rosen defends that the difference in the payment of a vice-chief to a chief gap is bigger than the difference in their abilities and outputs. This gap is important to understand how much effort an employee is willing to do, given their talents, ability and the will they have to get that extra money.

Players/workers who succeed in attaining regularly high ranks in elimination models will not make the extra effort unless the top-ranking prizes are given a disproportionate weight in their purse. They will keep fighting if the gap gets bigger. High wage differentials are a must motivating extra effort from the workers since the start of their careers until the end. The same applies to tennis players since the beginning of a new tournament until the end of it.

The other main hypothesis of tournament theory is that when the contestants of a tournament are homogeneous, i.e. no contestant is known to be stronger than the others, the total amount of effort exerted by them is higher. (Lazear and Rosen, 1981; Rosen, 1986; Prendergast, 1999; Sunde, 2003). On balance, the more uneven the chances of winning the match *ex ante*, the less effort both players, the favorite and the underdog, will expend (Rosen, 1986). Sunde (2003) writes about tournaments with uneven contestants, where one contestant is *ex ante* stronger than the others, resulting in less effort from both of the competitors, defending that the bigger the initial gap of a contestant strength in relation to the other is, the more costly it is for them to compensate this disadvantage gap and second, a strong player knowing *ex ante* his advantage, will be less likely to maintain a high effort and can reduce it without jeopardizing their probable success. The author finds that the key reason for companies to design promotion tournaments is to provide employees with rewards for effort. Given that tournaments between heterogeneous contestants are standard practice, whereas homogeneous contestants are the exception, the absence of proof for the implications of heterogeneity among contestants calls for an empirical analysis of the theoretical predictions. Furthermore, for heterogeneous matches, knowledge of the reward effects of heterogeneity is critical for the ideal design of tournaments in terms of disadvantages.

Empirical analysis of tournament theory and tournament models are few and the main reason is the difficulty faced when applying tournament theory models in a firm's environment and being able to produce meaningful results and conclusions due to the difficulty of measuring the worker's effort level in the firm (Ivankovic, 2007).

2.1.2. Tournament Theory applied to Industry

Main, O'Reilly and Wade (1993) were pioneers in testing tournament models in a firm setting. They provided results of empirical tests of executive compensation, analyzing more than two hundred corporations. These results are consistent with the operation of tournaments and the concepts of ever larger rewards in order to motivate the top positions. They found out that between 1980-84, the CEO's total earnings (salary plus bonus) were about 141% higher than their immediate subordinates, or the people one-step down the hierarchy. The authors also confirmed that the ratio of payment between the different levels seems to increase more and more as one moves up in the hierarchy, as said by Rosen (1981). Their study explored two contradictory explanations of structure of compensation among top executives. The first being that tournament theory suggests the need for meaningful variations in compensation among top executives in a firm. The second follows the idea of Milgrom and Roberts (1988) and Lazear (1989) who say that compressed wages could be more efficient since they reduce sabotage and promote teamwork. They find a positive evidence for the operation of tournament models. Assuming the metaphor of a firm working as a tournament and the need of forever larger prizes for top positions in order to continuously motivate workers, a substantial dispersion in compensation is expected. However, their empirical evidence showed that once winners are identified and promoted, they pass from the high end of payment to the low of the next level.
Hambrick (1988) showed evidence that the senior management in most firms operates as a team. Main, O'Reilly and Wade (1993) concluded that compensation in form of independent accomplishments may be inappropriate. If a tournament were to operate at senior levels, firms would find really difficult to motivate the losers. Tournament theory pays close attention in how to keep winners motivated but largely ignores losers (Dye, 1984).

Main, O'Reilly and Wade (1993) proposed the possibility of having ongoing, sequential tournaments in order to maintain incentives to everyone. Rather than having a single elimination tournament, top management compensation seems to rely on continuous small promotions and raises rather than large and uncertain playoffs. This suggests that any tournaments are most likely to be sequential. This way, winning leads to the payment of prizes and rewards and enhanced promotion prospects but losing is not the end because it still allows the loser to try again.

Another study with great relevance to this paper is Knober and Thurman (1994) where they use data from production of broiler chickens and got really close to a firm setting. Broiler chickens are raised by contract growers whose rewards depend explicitly upon relative performance. They found out that the growers respond to different payment structures with different levels of effort exerted. The evidence displays that the factor which determines effort is the incremental rewards and not the absolute levels; this means that changes in the level of prizes which leave unchanged prize differentials will not affect the performance. Their second conclusion regards the relation between the player's ability and the risk they are able to take in heterogeneous tournaments. They find out that**,** in these tournaments, players with lowest ability will adopt riskier strategies, which is consistent with the theory. The last empirical results imply that tournament organizers will want to handicap players of unequal ability or try to homogenize tournaments in order to prevent the disincentive impact of mixed tournaments.

Prendergast (1999) also examined the incentive effects implied by the design of promotion tournaments in firms. According to the basic theoretical framework of tournament theory, the payment for a certain level in the hierarchy is already fix *ex ante* by the firm and it increases as the levels in the hierarchy get higher. A group of candidates competes for the promotion which means the prize of a higher salary is associated with the position in the hierarchy. The candidate with the best relative performance wins the tournament and the prize associated.

The essential assumption for Prendergast is that agents can affect the probability of winning the prize through some actions. The possibility of winning the prize (rise in the salary) and participate in another round of the "tournament" provides incentive to exert effort on the current job (level of hierarchy) – the higher the prize to be won, the larger the incentive.

2.1.3. Tournament Theory applied to Sports

More support of tournament model is found in sport-related papers. Ehrenberg and Bognanno (1990a, 1990b) were pioneers in using empirical data from professional golf tournaments in the United States and Europe in order to test if and how incentives matter. The prize distribution is the incentive, the players' scores are the outputs, and other variables such as player quality, opponent quality, tournament complexity, and weather conditions are also observed. The hypothesis used for both papers is based on the work of Lazear and Rosen (1981) and Rosen (1986) in relation to prize structures.

Their conclusions are the following: higher prize money leads to lower scores (which in golf represents better scores), although they find that this effect is stronger in later rounds of tournaments where such factors as fatigue have more impact in players' concentration. Given a player's performance on the first rounds of the tournament, their performance on the last round seems to rely on the marginal returns to effort they faced. If a player faces larger marginal returns, they achieve better scores. In addition, the level of prize money of the tournament influences who enters the tournament, with higher prize money attracting better players.

The influence of tournament prizes on performance is their last conclusion. The results show that better players are more responsive to financial incentives or that the tournament's prize level doesn't reflect in the proper way the reward system faced by the nonexempt players since their main problem is how their finish in the tournament will influence the probability of qualifying as exempt players on the tour the following year.

Ehrenberg and Bognanno (1990b) also give some additional notes for future studies of tournament theory. Firstly, they defend that it would be desirable to use data from more than a year since the one year's analysis focus on how players allocate their effort within that year show no evidence at all on how the increasing prize levels over time influences players who wish to join the tour or how much effort and resources they put into their off-season practices. The other remark is about analyzing the tour and the sponsors. Their main concern is related to the level and structure of prizes that exist in tournaments, leaving some questions for others to analyze.

Orszag (1994) challenges the robustness of these findings by Ehrenberg and Bognanno, (1990a, 1990b) while using golf data from the 1992 PGA tour. He finds that the effect of total prize money on final scores to be insignificant giving two different explanations: the first is called the "choke factor" and is explained by the increasing of the amount of prize money and television exposure between these two studies[3]. Although larger prize differentials increase effort and concentration, they also make the golfers more nervous which makes them "choke"; the second is that there is a big random component to performance in golf that weakens the relationship between performance and effort exerted.

[3] Time between the studies of Ehrenberg and Bognanno (1990a, 1990b) and Orszag (1994).

Becker and Huselid (1992) use data from the National Association for Stock Car Auto Racing (NASCAR) and International Motor Sports Association (IMSA) to show that increasing the absolute prize spread awarded to the top finishing positions increases the driver's performance. They also show that the driver's safety is negatively affected by the increasing spread of prizes since drivers take more risks.

Maloney and McCormick (1994) applied the tournament theory to foot races. They used data of 1462 individual runners, from 115 foot races held in the Southeastern United States of America between 1987 and 1991, in order to understand the overall market response of higher prizes in two components of the basic tournament model. The first component is that an entry effect exists so that higher wages attract better runners. The second one is that the spread between prizes makes runners work harder and put more effort into the race. They test this component by measuring individuals' results against their own average performance. They describe the racing tournament as a labor market in which the supply curve is represented as the sum of two unrelated components, potential responses by the existing participants and by the new who enter. The results found strongly demonstrate a market-sorting effect that generates a positive price elasticity of supply, the bigger the pay the better the performance. They also demonstrate that the more concentrated is the prize money in a race, the higher the runners' effort.
Continuing with foot races, Lynch and Zacx (2000) examine the incentive effects of tournaments and confirm Maloney and McCormick's (1994) findings that times are faster in races which offer higher prize money because competitors with greater ability are more likely to choose tournaments with the higher prize spread.

For Lynch and Zacx (2000), road racing differs from other sport related studies about tournament theory. Firstly, there are important questions about the utility of participation and the worth of effort as a good. Secondly, competitors can choose amongst the available tournaments since their number of choices is high. Thirdly, the only major factor of production is the runner. And fourthly, the effort in road racing is mainly physical.
They defend that participation as utility is important in road racing because, in some races, tens of thousands of runners compete with zero hope of getting prize money. In road racing, there may be increasing returns from participation, this meaning that in larger races, runners can derive utility from the large number of spectators and increased media attention.
Road racing is a repeated-play tournament in which the total number of tournaments is bigger than the number of tournaments that a competitor can possibly enter, so each participant can choose a subset of tournaments to participate. Remaining everything equal, it is expected that competitors will choose tournaments in which the expected prizes are bigger, but they consider other factors as well, such as the depth of the field and their long-run training goals.
If non-prize compensation, depending on the competitors' performance across tournaments rather than single performance, is received by the competitors of the repeated-play tournament, prize money in a single race may elicit little additional effort.

In the scenario of road-racing, the best runners receive appearance fees, housing, transportation, equipment and contracts from sponsors. According to Lazear and Rosen (1981), theoretically appearance fees sort more efficiently than prizes do. This situation varies from runner to runner and depends on the consistency of their performances across multiples races instead of a single race. Thus, the incentive is to perform well in all races instead of dedicating all the effort in the races with the highest prize money.

While studying the reward system in NASCAR, Von Allmen (2001) defends that the need for sponsorships creates incentives for consistency over victory. He says that there is a possibility of NASCAR being just a vehicle for advertising what leads to the value of sponsorships being far bigger than any prize money that might be received by winning races. This means that drivers cannot compete for prize money without having a sponsor and that turns into their main job, where they concentrate their effort. In addition, the only way to maintain sponsorship is by keeping the car on track, regardless the position. Being this true, it's better for any driver to maintain a risk-averse strategy for driving instead of racing to win.

For the implication that incentives and effort should be lower in uneven tournaments, (Abrevaya, 2002) examined performance and the determinants of match winning probabilities of *ex ante* heterogeneous players in ladder tournaments using data from professional bowling. His results indicate that although there is an advantage of starting with a higher hand, players with lower ranking win more often than expected. He explains these findings with two theories consistent with the structure of a ladder tournament, regression-to-the-mean and hot-hand-theory.

Ladder tournaments are described by having the players pre-ranked and placed in a "ladder". The two lowest ranked players compete against each other in the first round and the winner advances to the next round, which is competing against the next player ranked in the ladder. Competition continues until the top of the ladder, at which point the winning player competes against the highest ranked player. In this kind of tournaments, the higher ranked players do not compete in earlier rounds; they have to wait until the victorious lower ranked players climb the ladder until they reach their position.
There is a clear advantage to the higher-rank players because they do not have the hypothesis of being eliminated in the first rounds. There is, however, a possible drawback of not competing in the early rounds and this is the hot-hand-theory: future success is positively related with the success of the recent past, so lower ranked players with a "hot-hand" will have a greater chance of beating the higher ranked opponent in a later round.
The pre-ranking mechanism can also have an effect of outcomes. This pre-ranking, in bowling, is determined by performance in earlier rounds of a certain tournament, which means based on short-run performance. If a player does better or worse than predicted in earlier rounds, the regression-to-the-mean theory can be applied: performance in ladder tournaments is partly

determined by the competitor's long-run ability. He finds no financial reward impacts on the absolute score of a player or the possibility of an upset and he does not consider effort exertion.

Lynch (2002) tests two hypothesis of tournament theory using data relative to Arabian horse racing. The hypotheses are: the worker's efforts will be greater when the difference between the winning and losing prizes is greater and the worker's efforts increase have a positive impact on the probability of winning. The author finds that jockeys increase their efforts/reduce their times in the second half of races when the amount of prize money lost by dropping a place is greater and if the distance between them and their closest competitors is shorter. Both of their findings support tournament models.

2.2. Tournament Theory applied to Tennis

In one of the first studies applying professional tennis to tournament theory, Sunde (2003) tested two of the main hypotheses of the theory: first that uneven tournaments, where contestants are *ex ante* known to be heterogeneous, lead to less effort exertion and second, if incentives through prizes do matter for effort exertion.

The author used data from 156 male's single tournaments between 1990 and 2002. By considering only the two final rounds of each tournament, he ruled out all sorting or selection problems that could exist by the initial seeding rules in tennis[4]. On the other hand, this data set allowed Sunde to investigate the effects of the stage of the tournaments in terms of performance incentives.

Sunde found two major findings while testing both hypotheses. First, the findings show that monetary rewards improve performance in terms of individual player game wins and overall effort exerted during a match, as measured by the total number of games played. Higher stage prizes, in fact, and the significantly higher prizes to be won in finals compared to semifinals, significantly encourage effort exertion. The second finding is that the degree to which a match between two opponents is biased, in terms of their initial relative advantage, does also influence match outcomes by influencing effort exertion. The contestants' ranking prior to the match influence the number of games won by a player, the total number of games played, and the difference in games won by the favorite and underdog, through a capability and an opportunity effects. Both of them can be separated by looking at various types of players independently. When it comes to underdogs, both effects strengthen each other but, when it comes to favorites, both effects work against each other. The remarkably different results for underdogs and favorites show that player heterogeneity does have a negative impact on their incentives.

[4] Every tournament on the ATP Tour use a seeding system to help balance the bracket. Players with the best ranking (between 8 and 32, depending on the tournament) are seeded so the best matches occur later in the tournament, making it impossible for the two players with the best ranking to compete before the final.

Furthermore, the more important the heterogeneity information is, the greater the incentive effects of heterogeneity, as calculated by relative rankings prior to the match. Rankings provide a clear measure of their opponents' present shape in matches with fairly similar experienced opponents who have already been professionals for some time. Consequently, in these matches, the heterogeneity impact measured by ranking differences in individual and total games won, is bigger than in matches in which both players have less experience against each other.

Ivankovic (2007), used data from the Association of Tennis Professionals, ATP Tour to investigate marginal pay spreads in professional tennis tournaments and the effects it has on the effort exerted by the players. He chose the total time of the match as the main variable for effort, which means the longer the match lasts, the more effort the player is exerting.
Ivankovic's objective was to test three hypotheses: first, spreads increase at the linear or increasing rate as the number of rounds left to play decrease but, in the final round, spreads should increase; second, the top four ranked players (both semi-finalists, finalist and winner) should receive at least 50% of the total prize and third, that effort should increase as the spread level increases.
The author concludes that the structures of marginal payoffs (spread level) in professional tennis do not correspond to the tournament theory. Marginal payoffs rise, but at a slower pace, until the final round, where the final marginal payoff falls. In the semi-finals, the percentage change in marginal payoff is greater than in the finals. In a similar manner, the top four finishers earn less than half of the overall prize money, roughly 40%. However, changes in marginal payoffs do cause direct changes in the effort exerted by players, as it increases, players try harder.

Gilsdorf and Sukhatme, (2008a) too, used data from the ATP tour to test Rosen's (1986) sequential elimination tournament model, more specifically, they investigated the effects of increasing prize money differentials between rounds (spread) on the stronger player's probability of winning a match. They estimated a probit model using data from 2632 individual matches over 68 tournaments during the 2001 ATP season. Their results support Rosen's theory that increases in prize money differentials have a positive effect on the stronger player's probability of winning a match.
Passing from men's professional tennis to women's professional tennis, (Gilsdorf and Sukhatme, 2008b) analyzed a data set of 2098 matches over 58 tournaments during the 2004 WTA tour to examine, again, the effect of prize money differentials on the stronger player's probability of winning a match. They estimated, again, a probit model using detailed information about prize money, player's characteristics and match play outcomes. Their results showed, again, that larger prize money differentials have a positive and statistically significant effect on the strongest player's probability of winning a match but this time they added that the number of stages remaining in a tournament have a negative effect on the probability of winning.

Lallemand, Plasman and Rycx (2008) applied the same methodology used by Sunde (2003) to examine how female tennis players react to prize incentives and *ex ante* heterogeneity in player's abilities. They analyzed a data set that covers all professional tennis players who played in the two final rounds (semi-finals and final) of all WTA tournaments between 2002 and 2004, containing information about players (age, ranking, head-to-head matches against opponent...) and tournaments (size, prize, points…) in a total of 502 matches. They used as an indicator of heterogeneity, the difference in the players' WTA ranking; as a sensitivity test, they relied on an indicator built by Klaassen and Magnus (2003) to control the pyramid structure of player's quality; they measured the prize spread through the difference in prizes between winning and losing a match; and they measured the player's performance by the individual number of games won at the end of the match. On the one hand, the results show a positive, significant relation between prize spread and players' performance, consistent with tournament theory. On the other hand, they came to the conclusion that the final outcome of a match is more connected with the player's ability than player's incentive to adjust their effort; uneven tournaments lead to more games win by the favorites and less performance by the underdogs, that means that as the players' ranking differential increases, the difference in terms of number of games won by the favorite and the underdog increases.

2.3. Why tennis fits Tournament Theory

The information available in tennis data and tennis tournaments makes it a perfect sport to analyze tournaments models because they are based on the concept of single elimination, two-person tournament, where winners continue to play with a chance of winning win a bigger reward, since the pay gap between stages increases, while losers receive their share of the total prize money (Sunde, 2003; Ivankovic, 2007; Sunde, 2009). There are several advantages in analyzing a single elimination, two-person tournament. There are no dynamic issues to consider, such as strategic support of certain contestants and abatement of other contestants, endogenous alliance forming, for example. This type of tournaments, usually allows for better control than tournaments with more competitors, reduce sabotage, collusion, etc…

Tennis tournaments for being a two-person, single elimination tournament, mimic the atmosphere of a corporate tournament in a way that frequently there are two candidates battling for a promotion in an economic context, for example for top-level management positions, acceptance of a bid in procurement auctions (Sunde, 2003). Also, most of the times tournaments are uneven, with contestants being stronger than their opponents. That is why there is general interest in knowing the effects of prizes and the heterogeneity of candidates on incentives.

In theoretical models of single elimination tournaments, the structure of information available in tennis data resembles the structure and requirements of information. Players' strength can be deduced from their rank prior to the respective tournament, allowing a measure of the heterogeneity of the participants to be constructed. Other measures such as their history in certain type of tournaments or surfaces can also be used to deduce their strength in a specific tournament. These measures of strength of the players/contestants, mainly the world rank prior to a tournament, make tennis data particularly adequate for testing hypotheses about the effects of heterogeneity which is not commonly available in other data. For each tournament, and for each stage of it, prizes are known *ex ante* and do not change regardless the margin of the victory.

So, both players' strength (measured by their ranking) and prizes are known *ex ante*, before the match or even when the tournament starts, which corresponds to the tournament theory theoretical requirements and reduces problems of endogeneity. In addition, tennis tournaments end with the final. In other words, they have a finite horizon as opposed to intermediate-level firm data.

3. Research Methodology and Data

3.1. Dataset

For the aim of the dissertation, we used data provided by the ATP, with main offices in Ponte Vedra Beach, Florida. The website (https://www.atptour.com/) has been the home of the ATP since its inauguration in 1990 and has available data about players' current ranking and prize money, ranking history, head-to-head against other players, tournament winners, players' activity (wins/losses). It also shows players' statistics and information, such as age, height and weight, number of years in the circuit and if they are right or left-handed. The website also has data about Grand Slams and Davis Cup (Davis Cup data will not be used in the future dissertation since our main focus are the main tournaments of the ATP Tour and the Grand Slams).

The dataset chosen will be about the 2019 ATP Tour (because of the pandemic, the 2020 ATP Tour had suffered many changes and many tournaments weren't played) and will comprise information for the main tournaments of the season. ATP Challenger Tour, Futures Series, ATP Next Gen Finals and ATP Finals will not be taken in consideration as well. We will analyze 65 tournaments, 4 Grand Slams, 9 ATP Masters 1000, 13 ATP Tour 500 and 39 ATP Tour 250.

Data about prize money in a more detailed way will be used from the website (https://www.perfect-tennis.com/prize-money/) which has information about prize money in each stage of every tournament.

3.1.1. Variables

In order to study the performance/effort exerted of professional tennis players and how they react to certain incentives through the spread of prize money/ranking points of the different stages of the tournaments, with the information available from (https://www.perfect-tennis.com/prize-money/) and (https://www.atptour.com/), we will try to identify and define a set of variables that can impact the probability of the possible outcomes that will be studied in the future dissertation.

3.1.1.1. Dependent Variables

As addressed before, we aim to study a fair value for the prize money of each match and a fair value for the ranking points. For that, the dependent variables refer to the different outcomes of the study: For each of the incentives is the player playing: prize money or ATP ranking points? The outcomes will be an estimate of what the fair prize money/ATP ranking points for the effort they put into each match should be. These variables are described in Table 3.1.

Table 3.1 - Dependent Variables

Dependent variables		**Description**
Ranking Points	Continuous	ATP ranking points awarded to the winner of the match
Prize Money	Continuous	Prize money awarded to the winner of the match

Since both the prize money values and ranking points have a wide range (from 6 500$ until 3 850 000$ and from 10 until 2000 points, respectively) we have turned their scale into a logarithm base, in order to flatten the gap between the lowest and the highest values and what becomes of interest are the scales increases.

3.1.1.2. Independent Variables

Within the regression model context, the variables can be classified either as independent or dependent. In a simple way, the variables that we want to predict their values are considered as dependent variables and the other ones which we use to build the model are the independent ones. For this dissertation the independent variables are associated with the players' characteristics, the players' effort through matches, match difficulty and tournaments. They are enumerated in Table 3.2.

Regarding players' characteristics, the variables are always divided into winners and losers. These variables are age, height, if the player is right or left handed, if he is a seeded player in the game in question and their rank. These variables are used to capture ability effects.

The players' effort variables analyze their effort and concentration in each match, in order to see if these variables explain the different incentives they face in each match. The assumptions for effort in this paper are explained in the Research Methodology, in Chapter 3.

Both variables sets percentage and match time per set should all be bigger if the effort put by both the players is bigger as well, since the match turns out to be more balanced. The variable total game difference should be lower if the effort put by both players is bigger, for the same reason listed before.

All the variables that are "per Set" and the variable sets percentage were metrics that we decided to put on a base of "average per set" to even out games with different number of sets (as stats on a best of 5 set match would be considerably bigger than on a best of 3 set match, for example).

Table 3.2 - Independent Variables

Independent variables		**Description[10]**
Sets Percentage	Continuous	Percentage of sets played per total number of sets (3 or 5)
Total Game Difference	Discrete	Total difference of games won and lost between the winner and the loser
Match Time per Set	Continuous	Match time played per set played
Winner Aces per Set	Continuous	Number of aces per set by the winner
Winner Double Faults per Set	Continuous	Number of double faults per set by the winner
Winner First Service In	Continuous	Percentage of first services in by the winner
Winner First Service Points Won Percentage	Continuous	Percentage of first service points won by the winner
Winner Second Service Points Won Percentage	Continuous	Percentage of second service points won by the winner
Winner Break Points Saved Percentage	Continuous	Percentage of break points saved by the winner
Loser Aces per Set	Continuous	Number of aces per set by the loser
Loser Double Faults per Set	Continuous	Number of double faults per set by the loser
Loser First Service In	Continuous	Percentage of first services in by the loser
Loser First Service Points Won Percentage	Continuous	Percentage of first service points won by the loser
Loser Second Service Points Won Percentage	Continuous	Percentage of second service points won by the loser
Loser Break Points Saved Percentage	Continuous	Percentage of break points saved by the loser
Winner Rank	Discrete	Winner ranking
Loser Rank	Discrete	Loser ranking
Winner Height	Continuous	Winner’s height
Winner Age	Continuous	Winner’s age
Loser Height	Continuous	Loser’s height
Loser Age	Continuous	Loser’s age
Dummy Winner Seeded	Categorical-Binary	The variable assumes value equals to 1 if winner is seeded, otherwise 0
Dummy Loser Seeded	Categorical-Binary	The variable assumes value equals to 1 if loser is seeded, otherwise 0
Dummy Winner Right Handed	Categorical-Binary	The variable assumes value equals to 1 if winner is right handed, otherwise 0
Dummy Loser Right Handed	Categorical-Binary	The variable assumes value equals to 1 if loser is right handed, otherwise 0

[10] Both discrete and continuous variables are quantitative.

All the categorical variables were transformed into dummies variables in order to incorporate them into the regressions.

Regarding tournaments, we used different sets of dummy variables.

First, there are the dummies associated with the surface of the tournament, in other words, if it is played in grass, clay or hard court. There were 317 matches played on grass, 764 on clay and 1434 on hard court. These variables are described in Table 3.3 below.

Table 3.3 - Surface Dummy Variables

Surface	
Dummy Grass	The variable assumes value equals to 1 if played on grass, otherwise 0
Dummy Clay	The variable assumes value equals to 1 if played on clay, otherwise 0
	If both 0, then the match is played on hard court

Second, there are the dummies associated with the draw size of each tournament. They can be 32, 64 or 128 players entering the tournament. The bigger the draw size, more matches are played in the tournament, meaning more ranking points and more prize money. These variables are described in the table below.

Table 3.4 - Draw Size Dummy Variables

Draw Size	
Dummy Draw 64	The variable assumes value equals to 1 if tournament draw size is 64, otherwise 0
Dummy Draw 128	The variable assumes value equals to 1 if tournament draw size is 128, otherwise 0
	If both 0, tournament draw size is 32

Last, there are the dummies associated with the level of the tournament, if it is an ATP250, ATP500, Masters 1000 or a Grand Slam. There were 4 Grand Slams, 9 Masters 1000, 13 ATP500 and 39 ATP250 tournaments, a total of 65 tournaments. The prestige of Grand Slam and Masters 1000 tournaments, in addition to having a larger field of stronger players, they could lead to an increase of effort by the players. These variables are described in the following table.

Table 3.5 - Tournament Level Dummy Variables

Tournament Level	
Dummy ATP500	The variable assumes value equals to 1 if tournament is ATP500, otherwise 0
Dummy Masters1000	The variable assumes value equals to 1 if tournament is Masters1000, otherwise 0
Dummy GrandSlam	The variable assumes value equals to 1 if tournament is Grand Slam, otherwise 0
	If all 0, tournament is ATP250

There is also a last set of dummy variables to be used in this project, regarding the round of each game. Different tournaments have a different number of rounds, depending on the draw size number. But rounds go since the round of 128 until the final, each round awarding a different level of ATP ranking points, depending on the type of tournament (same type tournaments award the same number of ATP ranking points in the same rounds) and different prize money. These variables are described in Table 3.6.

Table 3.6 - Tournament Round Dummy Variables

Round	
Dummy Semi Final	The variable assumes value equals to 1 if it is a semi-final game, otherwise 0
Dummy Quarter Final	The variable assumes value equals to 1 if it is a quarter final game, otherwise 0
Dummy Round of 16	The variable assumes value equals to 1 if it is a round of 16 game, otherwise 0
Dummy Round of 32	The variable assumes value equals to 1 if it is a round of 32 game, otherwise 0
Dummy Round of 64	The variable assumes value equals to 1 if it is a round of 64 game, otherwise 0
Dummy Round of 128	The variable assumes value equals to 1 if it is a round of 128 game, otherwise 0 **If all 0, the game is a final**

3.2. Research Methodology

The main objective is to study how professional tennis players respond to the two main retributions they can get while playing a professional tennis match: prize money and ATP ranking points and to study their response to aforementioned incentives, using for this purpose some regression models by the OLS method.

We will focus on Lazear and Rosen’s (1981) theory that the introduction of a differentiated reward system, where the biggest prize is awarded to the best competitor, increases the average level of effort, success and performance of players.

In order to do this, we will test if players exert bigger effort while facing increasingly higher prize money or when they have the chance of winning more ATP ranking points.

In opposition to previous papers about tournament theory applied to professional tennis, in which the dependent variable is the probability of a player winning a match, in this thesis the main focus was in ranking a match in relation to the incentives accordingly to the effort variables, using the incentives as dependent variables.

We wanted to analyze which of the independent variables that characterize the players’ performance and effort through the game and that characterize a match difficulty have an influence and explain each of the hypotheses. To do this, we started by doing a descriptive analysis of the variables, both dependent and independent, through various scenarios to test if they were correlated, if they had influence in each other etc. in order to contextualize our dataset.

We analyzed all the data, checking their averages and standard deviations and for the dependent variables, ranking points and prize money, we also analyzed the kurtosis and skewness of both.
Step two of the dissertation was the creation of two regressions applying OLS method using the rStudio software for this purpose. One of the regressions presents a model to assess the most adjusted prize money for match while the other regression does the same, but for ranking points. These models were fine-tuned as we plugged in and out variables that may or may not be significant in each model, in order to get the best regressions out of our dataset.
In the end we aim to answer to the following questions:

- What is a fair value for a match?
- Are these two incentives appealing to the players?
- If so, is one of the incentives, prize money or ranking points, more appealing than the other? For each of these two incentives do players exert bigger effort?

Effort is not easily observed in tennis matches and in order to analyze and quantify effort in this paper, some assumptions were made.
Effort can be seen in two different ways, by individual stats or by match difficulty. In this thesis, the assumptions for effort are about match difficulty:

- Matches get more difficult/even while facing bigger rewards;
- Players exert bigger effort while facing bigger rewards;
- The match is more difficult/even when both players exert bigger effort:
    - By the winner: an increase in the percentages of both aces and double faults means that the winner takes a more riskier strategy, meaning that the match is easier so they don't need to make that much effort -> bigger effort means a more conservative strategy while facing the opportunity of a bigger reward, because the match is harder; percentages of 1st and 2nd service points won and break points saved also decrease when the match is more difficult because the opponent/loser plays better;
    - By the loser: more difficult/even match means that the loser plays better, making it harder for the winner -> bigger percentages for everything, except for number of double faults that are lower.

3.2.1. Least squares Model

In order to relate the decisions being studied with the variables present in the database, we use the least squares model, a linear regression model.

A linear regression is an approach to study the relationship between a scalar response (dependent variable) and explanatory (independent) variables.
If the model is used with only one explanatory variable, it is called a simple linear regression and if it has more than one explanatory variables, it is called multiple linear regression.

Given a certain dataset $\{yi, xi1, \dots, xip\}_{i=1}^{n}$ the linear regression model assumes that the relationship between the dependent variable y and the explanatory variable/variables x is linear. In order to model this relationship, it is used an extra variable – ε, the error variable, an unobserved random variable which captures all the factors that affect y other than the regressors and its relationship with these regressors is crucial to formulate the appropriate model. Thus, the model for linear regression takes the following form seen in equation 1:

$$yi = \beta 0 + \beta 1xi1 + \cdots + \beta pxip + \varepsilon 1, i = 1, \dots, n$$

Equation 1 - Linear Regression Model

In this thesis the method used was the Least-Squares Method, which calculates the best fitting regression line for the observed dataset through the minimization of the sum of the squares of the residuals, being these the vertical deviations from each data point to the line – difference between the observed value and the fitted value proposed by the model.
Since deviations are first squared then summed, there are no cancellations between negative and positive values.

3.3. ATP Tour

The Association of Tennis Professionals (ATP) was formed in 1972 when the leading professionals joined forces with the idea of discussing the need for a player's association and it is currently the official organizer of the men's worldwide tennis tour. The ATP Tour debuted in 1990.

There are four main categories of tournaments played all around the world: Grand Slams, ATP tournaments, ATP Challenger tournaments and Futures tournaments. The ATP Tour comprises ATP Masters 1000, ATP 500 and ATP 250 tournaments and, also, oversees the ATP Challenger Tour, the level below ATP Tour. Both Grand Slams and Futures tournaments are overseen by the International Federation of Tennis (ITF) but, contrary to other tournaments such as Laver Cup, Hopman Cup and even the Olympics, ATP ranking points are awarded in these tournaments.

In each tournament, players compete for both monetary prizes and for ATP points, which determine a player's ranking. The ATP rankings are the only which both accurately identifies a player's current level of play while providing the benefits of seeding players in tournaments. Tournaments have different point weights according to the type of tournament a player has entered. The higher the category, the more points scored.

Tournament Level	W	F	SF	QF	R16	R32	R64	R128	Q	Q3	Q2
Grand Slam	2000	1200	720	360	180	90	45	10	25	16	8
ATP 1000 - 96 Draw	1000	600	360	180	90	45	25	10	16		8
ATP 1000 - 48/56 Draw	1000	600	360	180	90	45	10		25		16
ATP 500 - 48 Draw	500	300	180	90	45	20			10		4
ATP 500 - 32 Draw	500	300	180	90	45				20		10
ATP 250 - 48 Draw	250	150	90	45	20	10			5		3
ATP 250 - 32 Draw	250	150	90	45	20				12		6

[11]

Figure 3.1 - ATP Tour Ranking Points Distribution

All tournaments are counted immediately after they finish (usually on the Monday after the tournament is finished, with the exception of the tournaments that last for more than a week: Grand Slams, ATP Masters 1000 of Miami and Indian Wells). Once played, tournament points will be counted for the next 52 weeks. Ranking points are awarded according to the round the player loses in and the tournament's category.

A player's ranking is computed over the past 52 weeks, and it is based on the total points a player accumulates in their best 18 tournaments (19 if the player qualifies for the ATP Finals, only the 8 best players of the year qualify). These tournaments are:

- the four Grand Slams: Australia Open, Roland Garros, Wimbledon and US Open
- eight mandatory ATP Masters 1000 (there are nine ATP Masters 1000 events, but Monte Carlo does not count as a mandatory event)
- the six best results from all the other tournaments (Monte Carlo, ATP Tour 500's, ATP Tour 250's, ATP Challenger Tour, Futures Series)

There are certain exceptions to miss the mandatory events: being 30 years old or more in January 1st of the year makes you able to skip 1 ATP Masters 1000; having 12 years of career makes you able to skip another and having played more than 600 matches makes you able to

[11] Image from (https://mytennishq.com/how-do-tennis-players-earn-ranking-points/).

skip another. If a player meets these three criteria (e.g., Federer or Nadal), they can skip the events they want.
If a commitment player (top 30 player) skips a mandatory event and plays another in the same week (e.g., missing an ATP Masters 1000 to play an ATP Tour 250), they will not receive the points gained in that tournament. There are no money penalties for skipping mandatory events.

Tournaments range in size from 32 to 128 competitors in the main draw (there are also qualifying for lower ranked players), depending on the type of tournament (e.g., Grand Slams have 128 competitors, in which 32 are seeded players) and in three different surfaces: clay, hard court and grass.
Higher ranked players automatically qualify to participate in better tournaments (therefore tournaments with higher prize money) and, depending on their ranking, they can enter as a seeded player, which gives them an advantage in the position of the draw, thus, increasing their opportunity to win more points and more money.
Lower ranked players have different types of opportunities to participate in ATP tournaments and Grand Slams. Most of the times they have to play qualifier rounds in order to access the first round of the tournament. Grand Slams have three rounds of qualifying and ATP tournaments usually just have two. Just like in the tournaments, the qualifying is seeded.
If a player drops from the main draw before the start of the tournament, they will be replaced by the highest ranked player who lost in the last round of qualifying – this is called a lucky loser.
Another way of entering a tournament is by being attributed a wild card. Tournaments hand out a certain amount of vacancies to players who do not have sufficient ranking points to enter a tournament directly.

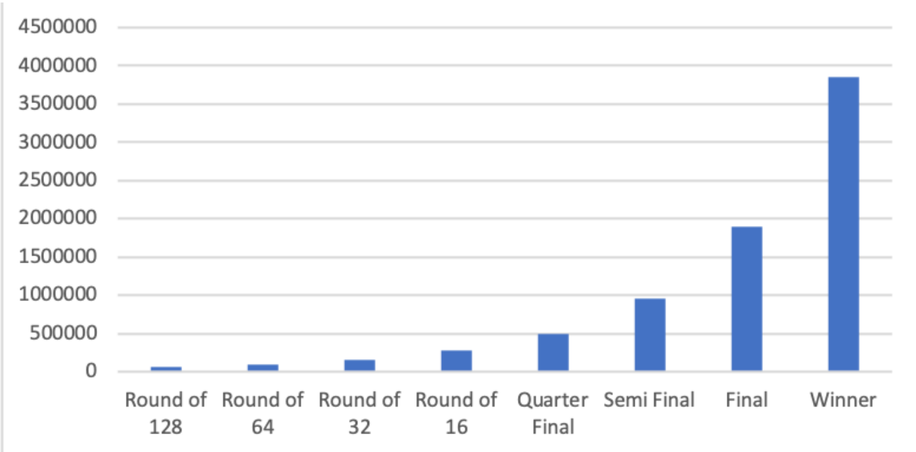
The 2019 ATP Tour Calendar [12] comprised 66 tournaments in total, 4 Grand Slams, 9 ATP Masters 1000, 13 ATP Tour 500's, 39 ATP Tour 250's and the ATP Finals. The Next Gen ATP Finals were also played in this year but will not be counted in this study. Neither will the ATP Finals. For this study, only main draw games will be counted (no qualifiers).
The 2019 tour started on 31 December 2018 with three ATP Tour 250 tournaments (Doha, Brisbane and Pune) and it ended on 17 November 2019 with the ATP Finals.

In this season, the US Open was the tournament that handled the biggest prize money, a total of 28,619,350$ with the winner receiving a total of 3,850,000$ (13.45% of the total prize money). The Antalya Open, an ATP Tour 250 tournament handled the smallest prize money, 568,389$ with the winner receiving a total of 87,517$ (15.40% of the total prize money).

[12] Due to the pandemic, the 2020 tour is not fit to analyze because a lot of tournaments were cancelled, and the rankings were changed

Table 3.7 - 2019 US Open Prize Money Distribution

Position	Prize Money $
Round of 128	58,000
Round of 64	100,000
Round of 32	163,000
Round of 16	280,000
Quarter Final	500,000
Semi Final	960,000
Final	1,900,000
Winner	3,850,000



[13]

Figure 3.2 - 2019 US Open Prize Money Distribution $

Table 3.8 - 2019 ATP 250 Antalya Prize Money Distribution

Position	Prize Money $
Round of 32	5,114
Round of 16	8,531
Quarter Final	14,846
Semi Final	26,139
Final	47,312
Winner	87,517

[14]

[13] Both information of the table and image come from (https://www.perfect-tennis.com/prize-money/us-open/).
[14] Both information of the table and image come from (https://www.perfect-tennis.com/prize-money/antalya-open/).

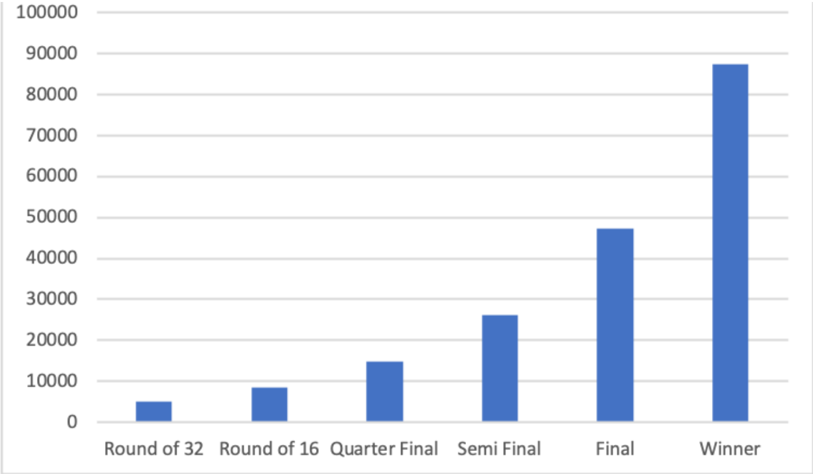


Figure 3.3 - 2019 ATP 250 Antalya Prize Money Distribution $

Payouts are awarded on a sliding scale, where the prizes keep increasing in every round. Points too are awarded on a sliding scale, increasing in every round, although not so drastically as the prize money.

Both these reward systems, points and prize money, are consistent with Rosen's theory that the higher the position, the higher the rewards.

4. Descriptive Analysis

This chapter is focused on describing the data used for this project, focusing mainly on describing the data set and preforming some descriptive analysis, in order to contextualize tennis statistics in this dissertation.

This analysis and project used a data set recording match and player statistics of 2594 matches of the 2019 ATP Tour Season, of which 79 instances were not used, as all mid-match retirements were disregarded, leaving a total of 2515 instances to analyze.

44% of the matches analyzed were played in ATP250 tournaments (lower rank tournaments), 15% correspond to ATP500 tournaments, 21% to Masters 1000 tournaments and 20% to matches played in Grand Slams tournaments. Of a total of 2515 matches, more than half of them were played in hard court surface (57%), 30% in clay terrain and 13% in grass, being the latter the least usual surface on professional tennis matches. This occurs because hard court can be played any time of the year and in every Continent, while both clay and grass depend more on favorable weather conditions.

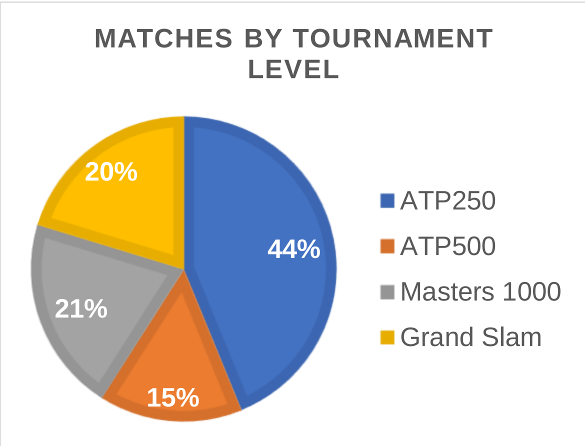


Figure 4.1 - Matches by Tournament Level

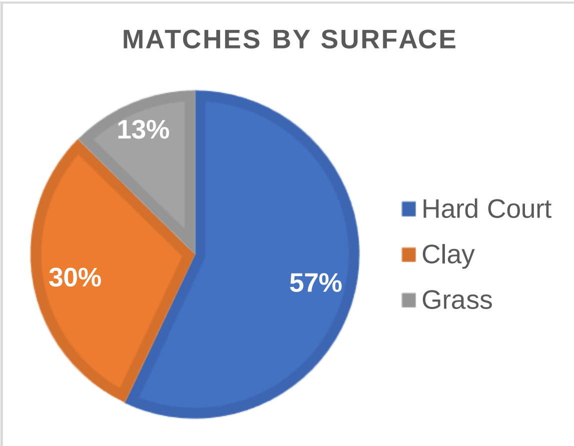


Figure 3.2 - Matches by Surface

4.1. Descriptive Analysis to the Dependent Variables

For this project, as addressed before, regression models will be used to estimate what could be some fair values for the prize money and the ATP ranking points of each match, taking into consideration the characteristics of some matches and tournaments. As so, prize money and ranking points were defined as being the two dependent variables. In Table 4.1, the main descriptive measurements of these variables are exposed:

The two dependent variables are described in the table below.

Table 4.1 - Dependent Variables Descriptive Statistics

	Ranking Points	Prize Money
Average	95	89879
SD	151	194327
Skewness	6,14	9,11
Kurtosis	54,58	121,15
Max	2000	3850000
Min	10	6500

The average prize money of a game is 89 879 $, with a standard deviation of 194 327 $, which gives us a variation coefficient of 216%. This allow us to show that the average, *per se,* does not provide a much relevant piece of information as there is a high degree of variation between prize money. This high variability on the prize money of each game can also be shown on the difference between the lowest value paid (6 500$) and the highest value paid (3 850 000$, corresponding to the amount due to the winner of the US Open, one of the Grand Slams).

Both skewness and kurtosis results help the understanding the aforementioned conclusions: Skewness coefficient is of 9,11 (way above 0), which shows that there is a high concentration of values below the average (low prize moneys) and that there are some few matches with very high prize moneys that tend to inflate the average. Kurtosis of 121 (also way above zero) shows that there is a big concentration of instances around a value (in this case between 15 000 $ and 50 000$). In fact, around 65% of the instances are values below the average, as it can be seen in Table 4.2 below:

Table 4.2 - Prize Money Frequency Table

Prize Money Awarded	Frequency	Relative Frequency	Cumulate Frequency	Cumulate Relative Frequency
[6 500 ; 15 000 [	464	18,4%	464	18,4%
[15 000 ; 30 000 [	509	20,2%	973	38,7%
[30 000 ; 50 000 [	507	20,2%	1480	58,8%
[50 000 ; 100 000 [	434	17,3%	1914	76,1%
[100 000 ; 300 000 [	490	19,5%	2404	95,6%
[300 000 ; 500 000 [	49	1,9%	2453	97,5%
[500 000 ; 1 000 000 [	41	1,6%	2494	99,2%
[1 000 000 ; 2 000 000 [	17	0,7%	2511	99,8%
[2 000 000 ; 3 850 000]	4	0,2%	2515	100,0%
TOTAL	**2515**	**100,0%**		

With regards to the ATP ranking points, the points attributed to each game depend on the rank of the tournament and the stage of the game within the tournament. The average points per game are 95, with a standard deviation of 151 points, showing that this is also a highly variability variable. There are way more games awarding lesser ranking points, as seen in Table 4.3, having a variation coefficient of 159%. Maximum points awarded tare 2000 which happens 4 times to the winner of each Grand Slam final.

Both skewness and kurtosis values give the same answer as the prize money: there is a high concentration of values below the average, as shown in Table 4.3 below:

Table 4.3 - Ranking Points Frequency Table

Ranking Points Awarded	Frequency	Relative Frequency	Cumulate Frequency	Cumulate Relative Frequency
10	15	0,6%	15	0,6%
20	569	22,6%	584	23,2%
45	969	38,5%	1553	61,7%
90	512	20,4%	2065	82,1%
150	76	3,0%	2141	85,1%
180	183	7,3%	2324	92,4%
250	39	1,6%	2363	94,0%
300	25	1,0%	2388	95,0%
360	63	2,5%	2451	97,5%
500	13	0,5%	2464	98,0%
600	15	0,6%	2479	98,6%
720	15	0,6%	2494	99,2%
1000	9	0,4%	2503	99,5%
1200	8	0,3%	2511	99,8%
2000	4	0,2%	2515	100,0%
TOTAL	**2515**	**100%**		

Due to the analysis shown above, some additional cross analyses were done to better contextualize both dependent variables. As so, we will cross tab the aforementioned two with the variables that most affect their values: stage of the tournament, tournament ranking and surface. Starting by comparing the average and standard deviation of the prize money by surface type, it is noticeable that the standard deviation values are way lower than the value of the variable as whole, lowering to a variation coefficient 20-25% of the value of the average which makes it possible to get a more detailed analysis. As so, we can see that grass tournaments are the best paying ones, awarding an average of 104 907 $ per game, while clay games pay the lowest on average (82 898 $ per game). (Table 4.4)

Table 4.4 - Prize Money per Surface

	Hard Court Obs 1434	**Clay** Obs 764	**Grass** Obs 317
Prize Money	Av 90276 SD 199040	Av 82899 SD 170670	Av 104908 SD 223830

In Table 4.5, it is noticeable that the average prize money increases as increases the rank of the tournament (ATP 250, ATP 500, Masters 1000 and Grand Slams). The average prize money won per round is always increasing, from 1,57 times to 2 times the value before, and generally increasing round after round, until it reaches the final (except in ATP250 tournaments) which goes against tournament theory that in an ideal prize structure there must be an extra incentive in the final round to ensure that players optimize their efforts. Looking into Figure 2.1, it can be seen

that the same happens to ranking points – in all tournaments, they duplicate from round to round, until they reach the semi-finals and final, where the value only increases 1,67 times.
Looking at the final round payouts by type of tournament, the average prize money for the winner of each tournament is increasing with the tournament rank. The prize money variation is 4,14x from an ATP 500 to an ATP 250, 2,53x from a Masters 1000 to an ATP 500 and 2,59 from a Grand Slam to a Masters 1000. This does not happen on the points systems as the ATP ranking points awarded in each type of tournament double from tournament to tournament (ATP 500 awards double the points of an ATP 250, a Masters 1000 awards double that of an ATP 500 and a Grand Slams awards twice that of a Masters).

Table 4.5 - Prize Money per Round and Tournament Level

Prize Money $	ATP 250	ATP 500	Masters 1000	Grand Slam
Final	Av 114793 SD 30981 Obs 39	Av 475505 SD 105608 Obs 13	Av 1204723 SD 147176 Obs 9	Av 3114595 SD 503037 Obs 4
Semi Final	Av 62105 SD 16648 Obs 76	Av 240672 SD 51026 Obs 25	Av 613726 SD 72732 Obs 15	Av 1559885 SD 213110 Obs 8
Quarter Final	Av 34461 SD 9137 Obs 151	Av 120101 SD 26287 Obs 49	Av 312921 SD 37279 Obs 32	Av 770631 SD 121440 Obs 15
R16	Av 19412 SD 4956 Obs 301	Av 63132 SD 13074 Obs 101	Av 161194 SD 18801 Obs 71	Av 419719 SD 74754 Obs 31
R32	Av 11235 SD 2977 Obs 475	Av 31494 SD 6487 Obs 205	Av 80946 SD 9396 Obs 139	Av 242892 SD 41815 Obs 63
R64	Av 6500 SD 0 Obs 15	Av 16283 SD 3493 Obs 32	Av 43171 SD 5262 Obs 216	Av 145348 SD 22847 Obs 121
R128			Av 26430 SD 0 Obs 62	Av 92355 SD 10549 Obs 247

Another point of interest would be to assess the dependent variables and some independent variables in order to identify if there were any match/player data in specific that would be highly impactful on the ranking points and prize moneys won. Of the total set of correlations between dependent and independent variables, the stronger ones (with absolute values closer to 1) are the ones presented below:

- Correlation between winner ranking and the prize money (logarithm) = -38%

- Correlation between winner ranking and ranking points (logarithm) = -36%

Both correlations have negative values, as expected: the bigger the incentive, the lower the rank of the winner, which was the premise expected. In both cases (for each dependent variable), since the values have an enormous gap, if we utilize the logarithmic function to this variable, the value of the correlation is much bigger.

If we look only for the final rounds of every tournament (finals, semi-finals and quarter finals), the correlation values get bigger for both incentives, as expected:

- Correlation between winner ranking and the prize money (logarithm) = -42%
- Correlation between winner ranking and ranking points (logarithm) = -38%

Both of these values are quite low and don't show a big correlation due to some reasons: first, the best players (ranked 1 to 10) don't play a lot of tournaments, even if the prize money or the ranking points awarded to the winner is big, because of their full schedule. There is more than 1 tournament a week which means that the same player can't be competing in both. And for a player to be the winner of the tournament, they have to win every round until there, so it is normal that in lower rounds with smaller incentives, the winner rank is low as well as in the statistics.

Correlation between the incentives and the explanatory variables of players' effort (for both winners and losers) and game difficulty were also done, as shown in Table 4.6 below, with no interesting values taken from there. All the values where too close to 0, ranging from -0,082 until 0,102. This can be explained with the fact that the players are professionals, so they play each game probably with the same ambition, no matter the incentives.

Looking again into Table 4.6, of the total set of correlations made, the stronger ones (with absolute values closer to 1) are the ones presented below:

- Correlation between loser first service points won % and loser aces per set = 55%
- Correlation between total game difference and sets percentage = -53%
- Correlation between match time per set and total game difference = -50%
- Correlation between loser first service points won % and total game difference = -49%
- Correlation between winner first service points won % and winner aces per set = 44%
- Correlation between loser second service points won % and total game difference = -41%

By observing these correlation values, some interpretations can be made. Correlation between loser first service points won % and loser aces per set, and between winner first service points won % and winner aces per set are both positive with high values comparing with the rest of the values presented in the Correlation Table. These values are expected since aces, in tennis, are services without answer from the opposite player. So, it is normal that if a player gets more first serves in and wins more points with their first serve, that the number of aces will be bigger.

The other four correlation values presented above are all related with the variable total game difference and they have all negative values. As said before in Chapter 3.1.1.2, the more difficult/even the match is, the lower is the difference of games won and lost between the winner and the loser, the lower the value of this variable. Again, as described before in Chapter 3.1.1.2, the more difficult/even the match is, the biggest the percentage of sets played out of the possible, the biggest the value of sets percentage, so their correlation value of -53% confirms the assumptions made before.

Correlation value between match time per set and total game difference confirms once again the same assumptions made before: the biggest the game time played per set, the more difficult/even the match is, so, as expected, this correlation value has a negative value.

Lastly, the correlation negative values between total game difference and loser first service points won % and loser second service points won % are the final prove needed to describe effort: matches where players exert bigger effort are the more difficult/even and this happens when losers' percentages get higher, meaning that they played better. By knowing already that bigger effort in a match leads to a lower value of total game difference, these negative correlation values of -0,488 and -0,414, respectively, are as expected.

Table 4.6 - Correlation Matrix

	Prize $	Sets %	Total Game Diff.	Match Time p/Set	Win. Aces p/Set	Win. 2x Flts p/Set	Win. 1st Serv %	Win. 1st Serv Won %	Win. 2nd Serv Pts Won %	Win. Brk Pts Svd %	Los Aces p/Set	Los. 2x Flts p/Set	Los. 1st Serv %	Los. 1st Serv Pts Won %	Los. 2nd Serv Pts Won %	Los. Brk Pts Saved %	Win. Rank	Los. Rank
Rank. Pts	95%	-2%	6%	7%	-4%	-8%	6%	2%	5%	-1%	2%	-5%	8%	-1%	0%	1%	-25%	-24%
Prize Money		-4%	10%	5%	-5%	-6%	5%	1%	4%	0%	1%	-5%	7%	-2%	0%	2%	-22%	-19%
Sets %			-53%	14%	-6%	9%	-5%	-28%	-25%	3%	11%	-3%	7%	24%	21%	12%	9%	-8%
Total Game Diff.				-50%	-14%	-17%	6%	18%	22%	1%	-30%	0%	-13%	-49%	-41%	-23%	-10%	12%
Match Time p/Set					2%	18%	-10%	-29%	-21%	15%	17%	7%	4%	22%	22%	33%	4%	-7%
Win. Aces p/Set						15%	2%	44%	-3%	2%	15%	-5%	11%	26%	26%	12%	-6%	0%
Win. 2x Flts p/Set							-34%	-1%	-38%	8%	2%	9%	-6%	6%	4%	4%	2%	7%
Win. 1st Serv %								-6%	4%	-1%	4%	-8%	9%	0%	4%	-2%	-6%	-2%
Win. 1st Serv Won %									10%	-10%	15%	-6%	7%	19%	14%	5%	-10%	1%
Win. 2nd Serv Pts Won %										-10%	12%	-3%	0%	7%	3%	0%	-6%	0%
Win. Brk Pts Svd %											6%	2%	-1%	12%	9%	2%	4%	-2%
Los Aces p/Set												12%	2%	55%	10%	17%	1%	-11%
Los. 2x Flts p/Set													-33%	10%	-23%	10%	8%	-3%
Los. 1st Serv %														-7%	6%	4%	-3%	-4%
Los. 1st Serv Pts Won %															14%	15%	2%	-14%
Los. 2nd Serv Pts Won %																10%	1%	-4%
Los. Brk Pts Saved %																	-1%	-2%
Win. Rank																		14%

4.2. Descriptive Analysis to the Independent Variables

Now, independent variables will be subject to some descriptive analysis other than correlations, in order to better understand the data that will be supporting the regression models of this project.

4.2.1. Match Difficulty Variables

The variables sets percentage, total game difference and match time per set provide information about how competitive the match is, the longer a match is, the bigger the percentage of sets is, and the lower the game difference, the more challenging and difficult the match is, meaning that the effort is bigger from both players, as seen before with the correlation values.

On average, the percentage of sets played is of 77,96% with a standard deviation of 15,9%.
Total game difference has an average of 4,846 but with a big standard deviation of 2,689 almost 55% of the average which can be explained with the fact that the final result of a match can be highly different (a match can finish 6-0/6-0 or it can end 7-6/5-7/7-5). The lower this value is, the most difficult/even the match was.
Match time per set is on average of 42,562 minutes, with a standard deviation of 8,38 minutes.
All these results are shown in Table 4.6.

Table 4.7 - Match Difficulty Descriptive Statistics

	Average	SD
Sets Percentage	77,96%	0,16
Total Game Difference	4,85	2,69
Match Time per Set	42,56	8,38

As a way to see how these variables behave in different circumstances, we analyzed them with other variables:

Table 4.8 - Match Difficulty per Tournament Level

	ATP250 Obs 1057	**ATP500** Obs 366	**Masters1000** Obs 499	**Grand Slam** Obs 489
Sets Percentage	Av 79,44% SD 0,16	Av 78,87% SD 0,16	Av 77,69% SD 0,16	Av 74,52% SD 0,15
Total Game Difference	Av 4,41 SD 2,34	Av 4,35 SD 2,20	Av 4,54 SD 2,36	Av 6,58 SD 3,37
Match Time per Set	Av 42,33 SD 8,32	Av 43,24 SD 8,48	Av 43,80 SD 9,36	Av 41,48 SD 7,16

While looking at these three variables in a more detailed way, in Table 4.7, there are some interpretations to be made:

Concerning sets percentage, the bigger the tournament, the smaller the value. This can be explained because in bigger tournaments, there are way more uneven matches than in smaller tournaments, since there will be more “better” players because of the better incentives and the prestige that these tournaments have, and there will be more “worse” players since the draw size gets bigger with the level of the tournaments (example given, all Grand Slams have 128 players in their draws, while normally ATP 250 have 32 players in their draw) so they have the opportunity of playing these tournaments as well.
In total game difference, the only big change in value is while playing Grand Slams, since matches are played best of 5 sets and again, because there are more uneven matches, so it is expected that this value gets bigger. As the match gets longer, players will get more exhausted, make more errors and don’t perform as well.
Game time played per set increases as well as the tournament difficulty increases, as expected, until Grand Slams, where the value is lower than all the other tournaments. Again, this happens because of the assumptions made before about Grand Slams: the more uneven matches and the exhaustion because of the 5 sets.

Table 4.9 - Match Difficulty per Round

	Match Time per Set	**Sets Percentage**	**Total Game Difference**	**Obs**
Final	Av 44,52 SD 8,90	Av 79,1% SD 0,16	Av 4,52 SD 2,57	65
Semi Final	Av 43,32 SD 8,43	Av 79,8% SD 0,17	Av 4,5 SD 2,66	124
Quarter Final	Av 43,93 SD 8,70	Av 80,2% SD 0,16	Av 4,39 SD 2,55	243
R16	Av 43,10 SD 8,87	Av 78,7% SD 0,16	Av 4,51 SD 2,54	504
R32	Av 42,18 SD 8,37	Av 77,4% SD 0,16	Av 4,65 SD 2,37	882
R64	Av 42,67 SD 7,98	Av 78,5% SD 0,16	Av 5,01 SD 2,76	384
R128	Av 40,90 SD 7,31	Av 75,2% SD 0,15	Av 6,29 SD 3,31	309

[15]

[15] In this table, only rounds are taken into consideration and not the level of the tournament, so the variables do not always take the values expected since, for example, a quarter final match in a Grand Slam gives better incentives than a final in an ATP 250 tournament.

Looking again into these 3 variables in Table 4.8, we can see that:

- Match time per set value is bigger in finals, matches where theoretically players exert biggest effort;
- Sets percentage values do not show any linearity, only from quarter finals until finals, but here the value decreases, as opposite to what was expected with the assumptions made;
- Total game difference values decrease from round of 128 until quarter finals, as expected, but then increase in semifinals and finals, contrarily to what was expected. Here, an explanation can be made, about the exhaustion. Players who get to the final rounds, played the rounds before, so it is normal that they are more tired which leads them into making more errors and making the game more uneven.

4.2.2. Effort Variables:

Looking for both winner and loser, every variable in average is bigger for the winners than for the losers, unless double faults, which is the expected.
The biggest differences are between first service points won % and second service points won %, as it is normal in a tennis match, where winners tend to win more points than the losers.
Aces per set and double faults per set have big standard deviations – from 70% to 80%. This happens because these variables depend on the players' game type. Some players win more points with aces while others have worse serves.
The percentage of break points saved being higher for winners contextualizes an important assumption of tennis matches – break points are points where the player who is not serving has the opportunity of winning the game from the one who is serving (in tennis, service games are an advantage), so this makes break points more important than "regular" points.
The fact that the winner saves more of these "important" points is a good indicator for winning a match.

Table 4.10 - Players' Effort Descriptive Statistics

	Winner		**Loser**	
	Average	**SD**	**Average**	**SD**
Aces per Set	2,91	2,24	2,18	1,86
Double Faults per Set	1,06	0,82	1,28	0,91
First Service In %	62,9%	0,07	61,1%	0,07
First Service Points Won %	77,2%	0,08	67,0%	0,09
Second Service Points Won %	56,6%	0,10	46,2%	0,10
Break Points Saved %	59,7%	0,33	51,7%	0,20

4.2.3. Players' Characteristics Variables:

Table 4.11 - Players' Characteristics Descriptive Statistics

	Winner		Loser	
	Average	**SD**	**Average**	**SD**
Age	27,81	4,88	27,63	4,69
Height	187,27	7,83	186,64	7,36
Rank	57,04	52,77	76,64	62,85
Seeded	0,42	0,49	0,26	0,44
Right Handed	0,87	0,34	0,86	0,35

Looking into the last set of variables, as shown in the table above, some interpretations were made:

- In terms of age and height, the values are very similar for both winners and losers, so no interesting interpretations can be taken;
- In terms of ranking, although the big value of standard deviation is of almost 90%, winners' rank is slightly superior to losers;
- Concerning seeded players, "better" players who start with an advantage position in the tournament, as expected, the amount of times they win is way bigger than when they lose – by being seeded they have an "easier" path in the tournament than non-seeded players.

5. Results

In this chapter, the results incoming from the OLS regression models used to study both the incentives/rewards present in a professional tennis match and trying to answer to our questions, are presented and discussed.

Several regressions were run with the purpose of selecting the most accurate ones for each one of the incentives: prize money and ranking points, with the objective of explaining the impact that some of the independent variables have on the dependent variables (the only that are this study's subject).

Afterwards, the estimated values provided by each of the models were cross-correlated with the real values of the prize money and ranking points (real prize money with estimated ranking points and real ranking points with estimated prize money) in order to answer this project questions: are these incentives appealing to players? And, if so, which of them is more appealing?

As the dependent variables for both models are in a logarithmic scale, the intercept ($\beta 0$) is explained by its exponential value, while the remaining βs are explained through the percentage impact on the dependent variables caused by percentage variations in those variables (independent variables).

Dummy variables were included in every model to capture fixed effects that in fact influenced the regressions. In order to quantify and analyze effort, the assumptions made in Chapter 3 – Methodology were used.

5.1. Chosen Models

Tables 5.1 and 5.2 show the results of the models obtained from each of the incentives: prize money and ranking points, respectively.

We started by testing various different regressions, with several combinations of independent variables in order to make possible the explanation of each dependent variable.

For both cases, we started by testing a regression with all the independent variables, (regressions P2 and R2) and then, by adding and withdrawing other different variables, other regressions were calculated. In both tables, 5 illustrative regressions of the work that was done are represented.

Focusing on the information in the prize money table, it can be seen that all five regressions are significant but regressions P2, P3 and P5 get excluded for having non-significant variables (at least one variable in each regression has a p-value > 5%). Regression P4, as well as many other regressions that are not presented in the table, have all the independent variables significant but got excluded in relation to regression P1 because it has a lower explanatory power (R^2 = 29% against R^2 = 78%).

Regarding ranking points, it can be observed in the table that not all the regressions are statistically significant – regressions R4 and R5. In this case, regressions R2 gets excluded for the same reason as regressions P2, P3 and P5 – for having non-significant variables in the model. Regressions R4 and R5 got excluded for not being significant. Regression R3, as regression P4, has all the independent variables significant but a low explanatory power, with a R^2 = 65% against regression P1's R^2 of 86%.

That said, we chose to use both regressions P1 and R1, for prize money and ranking points, respectively.

Table 5.1 - Prize Money Models

	Regression Results									
	Regression P1		**Regression P2**		**Regression P3**		**Regression P4**		**Regression P5**	
	Coef.	Sig.	Coef.	Sig.	Coef.	Sig.	Coef.	Sig.	Coef.	Sig.
Sets Percentage			0,032	<1	0,192*	<0,10			0,212	<1
Total Game Difference	0,052***	<0,01	0,0002	<1	0,015*	<0,10	0,174***	<0,01	0,181***	<0,01
Match Time per Set	0,004***	<0,01	-0,0004	<1	0,011***	<0,01	0,019***	<0,01	0,021***	<0,01
Winner Height			0,002***	<0,01	0,007***	<0,01	0,008***	<0,01	0,008***	<0,01
Winner Age	0,004**	<0,05	0,001	<1	0,002	<1	0,007**	<0,05	0,010**	<0,05
Loser Height			0,001	<1	0,004*	<0,10	0,007***	<0,01	0,008***	<0,01
Loser Age			0,0001	<1	-0,001	<1			-0,002	<1
Winner Aces per Set	-0,009*	<0,10	-0,007***	<0,01	-0,015**	<0,05	-0,033***	<0,01	-0,024**	<0,05
Winner Double Faults per Set	-0,038***	<0,01	-0,004	<1	-0,06***	<0,01	-0,070***	<0,01	-0,065**	<0,05
Winner First Service In %			0,08	<1	0,478**	<0,05			0,077	<1
Winner First Service Points Won %	-0,331*	<0,10	-0,06	<1	0,514**	<0,05	-1,420***	<0,01	-1,125***	<0,01
Winner Second Service Points Won %	-0,334***	<0,01	0,059	<1	0,191	<1	-0,995***	<0,01	-0,915***	<0,01
Winner Break Points Saved %	-0,075**	<0,05	0,007	<1	-0,019	<1	-0,195***	<0,01	-0,195***	<0,01
Loser Aces per Set			-0,007**	<0,05	-0,01	<1			-0,011	<1
Loser Double Faults per Set			-0,007	<1	-0,032**	<0,05	-0,070***	<0,01	-0,055**	<0,05
Loser First Service In %	0,749***	<0,01	0,077	<1	0,356*	<0,10	1,382***	<0,01	1,486***	<0,01
Loser First Service Points Won %	0,781***	<0,01	-0,067	<1	-0,226	<1	1,922***	<0,01	2,046***	<0,01
Loser Second Service Points Won %	0,512***	<0,01	0,003	<1	-0,053	<1	1,480***	<0,01	1,446***	<0,01
Loser Break Points Saved %	0,108*	<0,10	0,035*	<0,10	0,032	<1	0,325***	<0,01	0,285***	<0,01
Winner Rank	-0,002***	<0,01	-0,0003*	<0,10	-0,001***	<0,01	-0,004***	<0,01	-0,006***	<0,01
Loser Rank	-0,002***	<0,01	-0,0001*	<0,10	-0,001***	<0,01	-0,003***	<0,01	-0,004***	<0,01
Grass	0,154***	<0,01	-0,034***	<0,01	-0,011	<1	0,104*	<0,10		
Clay			-0,087***	<0,01	-0,076**	<0,05	-0,184***	<0,01		
Draw 64	1,239***	<0,01	-0,017	<1	-0,417***	<0,01				
Draw 128	2,290***	<0,01	0,043	<1	-0,732***	<0,01				
ATP500			1,114***	<0,01	0,979***	<0,01				
Masters 1000			2,014***	<0,01	1,684***	<0,01				
Grand Slam			3,128***	<0,01	2,640***	<0,01				
Semi Final	-0,626***	<0,01	-0,641***	<0,01						
Quarter Final	-1,243***	<0,01	-1,264***	<0,01						
Round of 16	-1,803***	<0,01	-1,858***	<0,01						
Round of 32	-2,304***	<0,01	-2,451***	<0,01						
Round of 64	-2,846***	<0,01	-3,068***	<0,01						
Round of 128	-3,110***	<0,01	-3,503***	<0,01						
Winner Seeded			0,007	<1	0,458***	<0,01	0,409***	<0,01		
Winner Right Handed	0,081***	<0,01	0,015	<1	-0,048	<1				
Loser Seeded	-0,091***	<0,01	0,0001	<1	0,539***	<0,01	0,397***	<0,01		
Loser Right Handed	0,077***	<0,01	0,011	<1	0,060*	<0,10				
Constant	10,828***	<0,01	11,311***	<0,01	6,416***	<0,01	5,348***	<0,01	4,807***	<0,01
Observations	2,515		2,515		2,515		2,515		2,515	
R2	0,783		0,972		0,692		0,337		0,294	
Adjusted R2	0,781		0,971		0,688		0,332		0,288	
Residual Std. Error	0,514 (df = 2488)		0,187 (df = 2476)		0,613 (df = 2482)		0,898 (df = 2493)		0,926 (df = 2493)	
F Statistic	345,233*** (df = 26; 2488)		2,221,611*** (df = 38; 2476)		174,265*** (df = 32; 2482)		60,452*** (df = 21; 2493)		49,516*** (df = 21; 2493)	

Table 5.2 - Ranking Points Models

	Regression Results									
	Regression R1		Regression R2		Regression R3		Regression R4		Regression R5	
	Coef.	Sig.	Coef.	Sig.	Coef.	Sig.	Coef.	Sig.	Coef.	Sig.
Sets Percentage			-0,007	<1			0,168	<1	0,257**	<0,05
Total Game Difference	0,032***	<0,01	-0,0003	<1	0,093***	<0,01	0,017*	<0,10	0,067***	<0,01
Match Time per Set	0,004***	<0,01	-0,0001	<1	0,011***	<0,01	0,013***	<0,01	0,020***	<0,01
Winner Height	-0,002**	<0,05	-0,0001		0,003**	<0,05	0,006***	<0,01	0,007***	<0,01
Winner Age			-0,0002*	<0,10	0,006***	<0,01	0,001	<1	0,005*	<0,10
Loser Height			-0,0001	<1	0,003*	<0,10	0,003	<1	0,005**	<0,05
Loser Age			-0,0002	<1			-0,001	<1	-0,002	<1
Winner Aces per Set			0,001**	<0,05	-0,016**	<0,05	-0,008	<1	-0,018**	<0,05
Winner Double Faults per Set	-0,022**	<0,05	-0,0001	<1	-0,033**	<0,05	-0,059***	<0,01	-0,065***	<0,01
Winner First Service In %			0,003	<1			0,497**	<0,05	0,510**	<0,05
Winner First Service Points Won %			-0,008	<1	-0,666***	<0,01	0,623**	<0,05	0,349	<1
Winner Second Service Points Won %	-0,222***	<0,01	-0,006	<1	-0,569***	<0,01	0,136	<1	-0,159	<1
Winner Break Points Saved %	-0,049**	<0,05	-0,0002	<1	-0,106***	<0,01	-0,035	<1	-0,090*	<0,10
Loser Aces per Set			0,0003	<1			-0,003	<1	-0,007	<1
Loser Double Faults per Set	-0,028***	<0,01	-0,0003	<1	-0,026*	<0,10	-0,025	<1	-0,037**	<0,05
Loser First Service In %	0,305***	<0,01	-0,011	<1	0,739***	<0,01	0,317	<1	0,791***	<0,01
Loser First Service Points Won %	0,556***	<0,01	0,006	<1	1,131***	<0,01	-0,142	<1	0,439*	<0,10
Loser Second Service Points Won %	0,216***	<0,01	0,0002	<1	0,801***	<0,01	-0,064	<1	0,287	<1
Loser Break Points Saved %			0,005	<1	0,114**	<0,05	0,014	<1	0,003	<1
Winner Rank	-0,002***	<0,01	-0,00002	<1	-0,003***	<0,01	-0,001***	<0,01	-0,005***	<0,01
Loser Rank	-0,001***	<0,01	-0,00001	<1	-0,003***	<0,01	-0,001***	<0,01	-0,004***	<0,01
Semi Final	-0,508***	<0,01	-0,510***	<0,01	-0,497***	<0,01				
Quarter Final	-1,001***	<0,01	-1,021***	<0,01	-0,962***	<0,01				
Round of 16	-1,662***	<0,01	-1,713***	<0,01	-1,579***	<0,01				
Round of 32	-2,351***	<0,01	-2,471***	<0,01	-2,140***	<0,01				
Round of 64	-2,956***	<0,01	-3,143***	<0,01	-1,935***	<0,01				
Round of 128	-3,552***	<0,01	-3,856***	<0,01	-2,186***	<0,01				
Winner Seeded			0,006***	<0,01	0,222***	<0,01	0,541***	<0,01		
Winner Right Handed			0,001	<1			-0,073*	<0,10		
Loser Seeded			0,005***	<0,01			0,636***	<0,01		
Loser Right Handed			0,0002	<1			0,068*	<0,10		
Draw 64	0,889***	<0,01	-0,032***	<0,01			-0,484***	<0,01		
Draw 128	1,548***	<0,01	-0,061***	<0,01			-1,008***	<0,01		
ATP500			0,748***	<0,01			0,590***	<0,01		
Masters 1000			1,478***	<0,01			1,096***	<0,01		
Grand Slam			2,214***	<0,01			1,615***	<0,01		
Grass	0,048***	<0,01	0,001	<1	-0,042*	<0,10	0,017	<1		
Clay	0,131***	<0,01	0,0001	<1	0,073**	<0,05	0,025	<1		
Constant	5,338***	<0,01	5,560***	<0,01	3,186***	<0,01	0,401	<1	-0,221	<1
Observations	2,515		2,515		2,515		2,515		2,515	
R2	0,858		0,999		0,645		0,448		0,256	
Adjusted R2	0,856		0,999		0,641		0,441		0,25	
Residual Std. Error	0,334 (df = 2492)		0,030 (df=2476)		0,528 (df=2488)		0,659 (df=2482)		0,763 (df=2493)	
F Statistic	681,866*** (df = 22; 2492)		57,585,470*** (df=38;2476)		173,821*** (df=26;2488)		62,996*** (df=32;2482)		40,945*** (df=21;2493)	
Note:	*p<0,1;		**p<0,05;		***p<0,01					

5.1.1. Prize Money

Table 5.3 - Estimated Prize Money Regression

Variable	Coefficient	$e^{\beta}-1$	Significance level	$e^{\beta 0}$
(Intercept)	10,83		< 2e-16	50421,276
Total Game Difference	0,052	5,39%	<0,01	
Match Time per Set	0,004	0,43%	<0,01	
Winner Aces per Set	-0,009	-0,91%	0.097	
Winner First Service Points Won %	-0,331	-28,16%	0.061	
Winner Second Service Points Won %	-0,334	-28,40%	<0,01	
Winner Break Points Saved %	-0,075	-7,22%	0.024	
Winner Double Faults per Set	-0,038	-3,73%	<0,01	
Loser First Service In %	0,749	111,42%	<0,01	
Loser First Service Points Won %	0,781	118,40%	<0,01	
Loser Second Service Points Won %	0,512	66,93%	<0,01	
Loser Break Points Saved %	0,108	11,37%	0.052	
Winner Age	0,004	0,44%	0.041	
Winner Rank	-0,002	-0,24%	< 2e-16	
Loser Rank	-0,002	-0,17%	< 2e-16	
Loser Seeded	-0,091	-8,72%	<0,01	
Draw 64	1,239	245,23%	< 2e-16	
Draw 128	2,291	887,95%	< 2e-16	
Semi Final	-0,626	-46,52%	< 2e-16	
Quarter Final	-1,243	-71,15%	< 2e-16	
Round of 16	-1,243	-71,15%	< 2e-16	
Round of 32	-1,243	-71,15%	< 2e-16	
Round of 64	-1,243	-71,15%	< 2e-16	
Round of 128	-1,243	-71,15%	< 2e-16	
Grass	0,154	16,67%	<0,01	
Winner Right Handed	0,081	8,39%	<0,01	
Loser Right Handed	0,077	7,99%	<0,01	

Taking into consideration the information present in Table 5.1, the best model chosen, regression P1, is not the one with the best determination coefficient (R^2) but the one with more significant variables, still with a good value of R^2, with a p-value (< 2.2e-16) lower than the statistical significance (α = 0.05). This regression has a very high significance level, since 21 out of the 26 independent variables used have a p-value lower than 1%.

Table 5.3 shows the results of the estimated prize money regression.
While analyzing the fair prize money regression, it can be noticed that the R^2 coefficient of 78% provides security on the explaining power of this model as it explains approximately 78% of the dependent variable, leaving only 22% to be affected by stochastic events. This was the best regression found after plugging in and out independent variables to see which of them were significant in the model.
This model is explained by 26 significant independent variables, of each:

- 9 are dummy variables about tournament characteristics – 6 about the round of the tournament, 2 about the draw size and 1 about the surface;
- 11 are variables regarding match difficulty and the players' effort in the match – for both winners and losers;
- 6 are related to the players' characteristics – 3 of them are dummy variables and they regard both winners and losers as well.

The Intercept or $\beta 0$ value is 10,83, which corresponds to $e^{10,83}$ = 50421,28$. Mathematically speaking, this value represents the regression estimated value when all the independent variables are all equal to 0:

- In this case, this value is an estimate for the prize money of a final of a 32 player tournament, played in either clay or hard court, where the loser is not seeded, both winner and loser are left handed, the winner is 0 years old, both winner and loser have a rank of 0, the total game difference is 0, match time per set is 0 minutes and all players percentages are 0.[16]

Starting with match difficulty variables, both variables, total game difference and match time per set, have positive influence in the prize money. Match time per set value was expected, meaning that an increase in the effort by both players leads to an increase of the time of the game. In this case, the coefficient value for total game difference 0,052 means that a percentage increase on this variable results in an increase of 5,39% on the prize money. This goes against the premise and the assumptions previously assumed, the bigger is the game difference, the less balanced is the match, and so, easier it gets to the winner, which should lead to a decrease on the prize money, instead of the modelled increase. Two matches played exactly under the same

[16] This is a base value for the regression with no sport value but used anyway to set the base for the remainder values within this model.

circumstances, if a match is won by 7-5 6-3 and the other 7-5 6-2, the second deserves 5,39% more prize money than the first, since the difference in terms of number of games is 1.

The variable match time per set has a smaller impact on the prize money value – one more minute played per set translates in +0,43% of prize money, but is still significant and has a positive signal.

Analyzing winner's effort variables, the results are in agreement with the assumptions made before. Both number of double faults and aces per sets lead to a decrease in the prize money. The increase of number of aces per set leads to a decrease of 0,91% for extra ace per set, in the prize money of the match.
Taking these assumptions in mind, the premise for double faults is the same. The bigger the number of double faults by the winner translates into less prize money awarded (-3,73% for extra double fault per set in a match). This result goes to the encounter of the theory if we think that the biggest the effort, the less the number of double faults – winners exert biggest effort by risking less in their second service in more difficult matches, avoiding double faults while facing more prize money.

Looking into winners' percentages of first and second service points won, we can see that both have a big decrease impact on the prize money – 28,16% and 28,40% respectively. Again, following the assumptions above (more difficulty in a game translates into biggest effort from both players), these values make sense in the regression.
Winners' percentage of break points saved follows the same direction as the other percentages. A bigger percentage translates into smaller prize money – this can also be because biggest percentage of break points saved can mean more break points faced.
While playing a more difficult match, winners' percentages will be lower because the loser will play better.

Opposite to winners' percentages, losers' percentages have a positive impact on the prize money. Taking the assumption that a more difficult/even match means bigger effort in a match and that a match is more even when the loser plays better, these values are the expected.
Looking to the percentages of first service in and first service points won, we can see that an increase in these percentages by 1% leads to a 111,42% and 118,40% respectively, increase in the prize money, being consistent with the theory.
These big values make sense if we think that losers take a riskier strategy when games are more difficult, so the effort by both players is way bigger, having such a big impact in the prize money.
The percentages of second service games won and break points saved have also a positive impact into the prize money, again, being consistent with the theory and the assumptions.

Now, observing the variables that explain the players' characteristics, we can see that an increase of 1 year in the age of the winner leads to +0,44% of the prize money and that if both winner and loser are right handed, the prize money also increases by 8,39% and 7,99% respectively.
If the loser is seeded, this leads to a decrease of 8,72% in the prize money value.
Both winner and loser rank have a negative impact in the prize money, which is consistent with the theory. The lower the value of the rank, the better the player is, so an increase in the rank means that players are worse, and the game doesn't pay that much.
The dummy grass variable has a positive impact in the prize money by 16,67%. This means that if a match is played in grass instead of clay or hard court, it should pay +16,67%.
Grass is known to be the most difficult surface in tennis, since it is the surface where the least number of tournaments happen and that it is played in only 2 months of the year, making it even more difficult for players to prepare for it. By the assumption that a more difficult match deserves better incentives, this value goes accordingly to the theory.

Dummy variables related to the draw size have the biggest impact in the prize money. If the tournament has a draw size of 64 instead of 32, the increase in the prize money is of 245,23% and if the draw is of 128 players the increase is of 887,95%. This big impact is easy explained by the fact that the bigger the tournament size, the bigger the reward.

Finally, the model controls the round that is being played because the prize money depends on it. The latest the round, the bigger the prize money awarded, the bigger the effort by the players. As expected, every time the round decreases, the prize money decreases as well. These values are consistent with the theory.

5.1.2. Ranking Points

The same methodology was used for ranking points. Table 5.4 shows the results of the ranking points regression.
Evidentially, the R^2 coefficient of 86% provides security on the explaining power of this model as it explains approximately 86% of the dependent variable, ranking points, leaving only 14% to be affected by stochastic events. Again, this was the best regression found after plugging in and out independent variables to see which of them were significant in the model.
This model is explained by 22 significant independent variables, of each:

- 10 are dummy variables about tournament characteristics – 6 about the round of the tournament, 2 about the draw size and 2 about the surface;
- 9 are variables regarding the players' effort in the match – for both winners and losers, and match difficulty;
- 3 are related to the players' characteristics – winner and loser rank and winner height.

The Intercept or $\beta 0$ value, 5,338 corresponds to approximately 208 ranking points. Mathematically speaking, this value represents the regression estimated value when all the independent variables are all equal to 0:

- In the case of this model, this value is an estimate of the ranking points that should be awarded for a match that is a final, played in hard court, in a tournament with a draw of 32 players, where both winner and loser have a rank of 0, the winner height is 0, the total game difference is 0, match time per set is 0 minutes and all players' percentages are 0.[17]

Between the two models, certain independent variables were present in both models: winner and loser rank, total game difference and match time per set, dummy variables about the round of the tournament and the draw size, grass dummy variable, loser's percentage of first service in, first and second service points won and winner's percentage of double faults per set, break points saved and second service points won. These variables shared by both models have the same signal (positive or negative) in both of them. Stating the obvious, it can be noticed that motivation works the same way for both incentives.

[17] As explained before in Chapter 5, this is a base value for the regression with no sport value but used anyways to set the base for the remainder values within this model.

Table 5.4 - Estimated Ranking Points Regression

Variable	Coefficient	$e^{\beta}-1$	Significance level	$e^{\beta 0}$
(Intercept)	5,338		< 2e-16	208,148
Total Game Difference	0,032	3,25%	<0,01	
Match Time per Set	0,004	0,45%	<0,01	
Winner Height	-0,002	-0,19%	0.037	
Winner Double Faults per Set	-0,022	-2,16%	0.016	
Winner Second Service Points Won %	-0,222	-19,92%	<0,01	
Winner Break Points Saved %	-0,049	-4,80%	0.020	
Loser Double Faults per Set	-0,028	-2,79%	<0,01	
Loser First Service In %	0,305	35,66%	<0,01	
Loser First Service Points Won %	0,556	74,44%	<0,01	
Loser Second Service Points Won %	0,216	24,13%	<0,01	
Winner Rank	-0,002	-0,16%	< 2e-16	
Loser Rank	-0,001	-0,10%	< 2e-16	
Semi Final	-0,508	-39,83%	< 2e-16	
Quarter Final	-1,001	-63,25%	< 2e-16	
Round of 16	-1,662	-81,02%	< 2e-16	
Round of 32	-2,351	-90,47%	< 2e-16	
Round of 64	-2,956	-94,80%	< 2e-16	
Round of 128	-3,552	-97,13%	< 2e-16	
Draw 64	0,889	143,34%	< 2e-16	
Draw 128	1,548	370,03%	< 2e-16	
Clay	0,048	4,93%	<0,01	
Grass	0,131	13,95%	<0,01	

Again, by analyzing the match difficulty variables, both variables total game difference and match time per set have positive influence in the dependent variable ranking points. Equals to what happened with the prize money model described in Chapter 5.1, match time per set value was as expected (an increase in effort by both players leads to an increase of the time of the match), having a positive effect in the incentive and the value of the total game difference goes against the expected signal of this variable.
Match time per set has a positive impact of 0,45% per extra minute played of the match per set, a value really close to the one saw in the prize money and the total game difference variable has an impact of +3,25% in the dependent variable, for each game of difference in the match.

Regarding the effort variables for the winners, as seen in the prize money model, the results are in agreement with the assumptions about effort. Winner's percentage of double faults leads to a decrease of the ranking points. For an extra 1% of double faults per set, there is a decrease of 2,16% in the fair value of ranking points that should be awarded.
The percentage of break points saved by the winner of a match also leads to a decrease of the dependent variable, such as all the other winner's percentages. In this case, a 1% increase in the number of break points saved leads to a decrease of 4,8% of the ranking points.
Lastly, the winner's percentage of second service games won has a bigger impact on the ranking points than any other winner's percentage – an increase of 1% leads to a decrease of 19,92%, a value slightly lower than in the prize money model.
The increase of all the winner's percentages leads to a decrease on the ranking points, such as expected with the assumptions made.

Such as the prize money model, the losers' percentages have a big positive impact on the ranking points, except for the percentage of double faults (again, as expected with the premise that a more difficult/even game means bigger effort in a match and that a game is more even when the loser plays better).
The same happens with the loser's percentage of double faults. For the losers, the expected signal for this variable is to be negative – while facing a bigger reward, the match gets more even, meaning that both players play better/exert bigger effort and, in the case of double faults, this number decreases. In this model, an increase on the percentage of double faults per set leads to a decrease of 2,79%.
Analyzing the percentages of the first service in and the first and second service points won, an increase of percentages of all of these variables lead to an increase of the fair ranking points.
So, as what happened in the prize money model, these positive impacts on the dependent variable go accordingly to the theory.
In terms of the players' characteristics variables, there are only three that are present in the model: both winner's and loser's rank and winner's height.
Winner's height has a negative small impact of -0,19% on the ranking points awarded.

Same as in the prize money model, both winner's and loser's rank have a negative impact in the prize money, which is consistent with the theory - The lower the ranking number of the player, the better the rank, the better the players are (leading to a more difficult match), so an increase in the rank means that the match doesn't pay that much.

Both dummy grass and clay variables have positive impacts on the ranking points by 13,95% and 4,93%, respectively. This means that matches played in grass or clay, instead of hard court, reward more ranking points than hard court matches. Again, this may happen because there are much more matches played in hard court rather than on the other two surfaces.
Dummy variables related to the draw size have, same as in the prize money model, the biggest impact in dependent variable. If the tournament has a draw size of 64 instead of 32, the increase of ranking points is of 143,34% and if the draw is of 128 players the increase is of 370,03%. This big impact is easy explained by the fact that the bigger the tournament size, the bigger the reward.

Finally, the model controls the round that is being played because ranking points depends on it. The latest the round, the more ranking points awarded, the bigger the effort by the players. As expected, every time the round decreases, ranking points awarded decrease as well. These values are consistent with the theory and similar to what happens in the case of prize money.

5.2. Answering the questions

As we used these regressions to calculate the effort level, measured in a fair value of prize money and ranking points, we tried to understand how this effort evolves with the existing incentives.

For this, it was calculated the so-called fair value of a match in relation to ranking points with the real prize money, and vice-versa, the so-called fair value of prize money with the ranking points. In order to analyze this, the linear correlation coefficients between these variables were calculated:

- Correlation between Fair Value of Prize Money and Ranking Points = 42%
- Correlation between Fair Value of Ranking Points and Prize Money = 77%

While talking about correlation coefficients, the closer they are to 1 or -1, he greater the intensity of correlation between the variables, the more effort is associated with the incentive variables. By observing that both the coefficients have positive values, we can see that effort and incentives (both prize money and ranking points) evolve in the same direction: a greater value of reward leads to a bigger exertion of effort by the players, answering the first question. It is shown that the incentives are appealing to the players, leaving them to exert more effort while facing better rewards since both correlation coefficients are positive.

Looking at the correlation values, we came to the conclusion, that in both cases, the incentive leads to a bigger effort (coefficient > 0) and taking into account that the value for prize money is 77% against the ranking points' value of 42%, it can be concluded that the prize money is the strongest incentive of both since it has a stronger value of correlation coefficient, making it possible to answer both the questions made.

6. Conclusion

The main focus of this dissertation was the place of incentives in professional tennis and if there is one incentive more appealing than the other. The two main incentives, while talking about professional tennis, are prize money and ATP ranking points.
Our goal was to analyze if players exert bigger effort while facing both these incentives. To test this, two questions were asked and answered in this dissertation:

- Are these two incentives appealing to the players?
- If so, is one of the incentives, prize money or ranking points, more appealing than the other? For each of these two incentives do players exert bigger effort?

In order to answer these questions, we used data from the (https://www.atptour.com/) and (https://www.perfect-tennis.com/prize-money/). We obtained a total of 2515 matches to analyze, with statistics about players' characteristics, players' effort, match difficulty, tournaments' characteristics and prize money and ranking points awarded to the winner of each match.

While doing descriptive analyzes of all the variables used in this dissertation, it was found that the reward system for both incentives, prize money and ranking points, in professional tennis tournaments in the 2019 ATP Tour season, are not structured in the most efficient way, as defended by tournament theory, just in ATP250 tournaments. Briefly, in all tournaments, other than ATP250 tournaments, the spread level of prize money increases at a decreasing rate, instead of having an extra incentive in the final round to ensure that players optimize their efforts. If we talk about ranking points, this happens in all tournaments.
Still, even though the reward system is not structured accordingly to tournament theory, both the incentives are appealing to players. Players will exert bigger effort while facing bigger incentives.

To answer these questions, two OLS regressions were made. One of the regressions presents a model to assess the most adjusted prize money for match while the other regression does the same, but for ranking points. These models were fine-tuned as variables, that might or might not be significant in each model, were plugged in and out to get the best regressions out of our dataset.
Then we used these regressions to calculate the effort level, measured in a fair value of prize money and ranking points, to understand how this effort evolves with the existing incentives.

For this, it was calculated the so-called fair value of a match in relation to ranking points with the real prize money, and vice-versa, the so-called fair value of prize money with the ranking points. In order to analyze this, the linear correlation coefficients between these variables were calculated.

Through the analyses of both of these models, by the signal and impact that the explanatory variables have on the two incentives, prize money and ranking points, and by studying the correlation coefficients, it was possible to understand that, in fact, players deserve better incentives when the match is more difficult and when both of them, winner and loser, exert bigger effort.

Looking at the correlation values shown in Chapter 5 - Results, we came to the conclusion that, in both cases, the incentive leads to a bigger effort (coefficient > 0) and taking into account that the value for prize money is 77% against the ranking points' value of 42%, it can be concluded that the prize money is the strongest incentive of both since it has a stronger value of correlation coefficient, making it possible to answer both the questions made.

Some limitations of this work include the difficulty in measuring effort in terms of professional tennis players statistics since all psychological effects and other incentives, such as prestige of winning certain tournaments, or playing in certain stadiums are not included in this study. Other big limitation was the lack of recent related papers, both for tournament theory that had its biggest focus between 1980-1990 and for tournament theory applied to Sports.

A potential improvement for the work done and for papers related to tournament theory applied to Professional Tennis would be focusing the analysis in a single player over the years, instead of looking to the whole Tour in a year space and trying to capture other statistics related to players' effort.

References

Abrevaya, Jason. "Ladder tournaments and underdogs: lessons from professional bowling." *Journal of Economic Behavior & Organization* 47.1 (2002): 87-101.

Becker, Brian E., and Mark A. Huselid. "The incentive effects of tournament compensation systems." *Administrative Science Quarterly* (1992): 336-350.

Dye, Ronald A. "The trouble with tournaments." *Economic Inquiry* 22.1 (1984): 147.

Ehrenberg, Ronald G., and Michael L. Bognanno. "Do tournaments have incentive effects?." *Journal of political Economy* 98.6 (1990a): 1307-1324.

Ehrenberg, Ronald G., and Michael L. Bognanno. "The incentive effects of tournaments revisited: Evidence from the European PGA tour." *ILR Review* 43.3 (1990b): 74-S.

Gilsdorf, Keith F., and Vasant A. Sukhatme. "Testing Rosen's sequential elimination tournament model: Incentives and player performance in professional tennis." *Journal of Sports Economics* 9.3 (2008a): 287-303.

Gilsdorf, Keith F., and Vasant A. Sukhatme. "Tournament incentives and match outcomes in women's professional tennis." *Applied Economics* 40.18 (2008b): 2405-2412.

Lynch, James G. "The effort effects of prizes in the second half of tournaments." *Journal of Economic Behavior & Organization* 57.1 (2002): 115-129.

Hambrick, Donald C., and Gerard L. Brandon. *Executive values*. Elsevier Science/JAI Press, 1988.

Ivankovic, Miren. "The tournament model: an empirical investigation of the ATP Tour." *Zbornik radova Ekonomskog fakulteta u Rijeci: časopis za ekonomsku teoriju i praksu* 25.1 (2007): 83-111.

Klaassen, Franc JGM, and Jan R. Magnus. "Forecasting the winner of a tennis match." *European Journal of Operational Research* 148.2 (2003): 257-267.

Knoeber, Charles R., and Walter N. Thurman. "Testing the theory of tournaments: An empirical analysis of broiler production." *Journal of labor economics* 12.2 (1994): 155-179.

Lallemand, Thierry, Robert Plasman, and François Rycx. "Women and competition in elimination tournaments: evidence from professional tennis data." *Journal of sports economics* 9.1 (2008): 3-19.

Lazear, Edward P., and Sherwin Rosen. "Rank-order tournaments as optimum labor contracts." *Journal of political Economy* 89.5 (1981): 841-864.

Lazear, Edward P. "Pay equality and industrial politics." *Journal of political economy* 97.3 (1989): 561-580.

Llorca, Matthieu, Eric Barget, and Thierry Teste. "The Effects of Financial Incentives on Women's Performance: The Tournament Theory Applied to Female Tennis Players." *Management and Economics Review* 2.1 (2017): 1-13.

Lynch, James G., and Jeffrey S. Zax. "The rewards to running: prize structure and performance in professional road racing." *Journal of Sports Economics* 1.4 (2000): 323-340.

Main, Brian GM, Charles A. O'Reilly III, and James Wade. "Top executive pay: Tournament or teamwork?." *Journal of labor economics* 11.4 (1993): 606-628.

Maloney, Michael T., and Robert E. McCormick. "The response of workers to wages in tournaments: Evidence from foot races." *Journal of Sports Economics* 1.2 (2000): 99-123.

Milgrom, Paul, and John Roberts. "An economic approach to influence activities in organizations." *American Journal of sociology* 94 (1988): S154-S179.

Orszag, Jonathan M. "A new look at incentive effects and golf tournaments." *Economics Letters* 46.1 (1994): 77-88.

Prendergast, Canice. "The provision of incentives in firms." *Journal of economic literature* 37.1 (1999): 7-63.

Rosen, Sherwin. "Prizes and incentives in elimination tournaments." *The American Economic Review* Vol. 76, No. 4 (1986): 701-715.

Sunde, Uwe. "Potential, prizes and performance: Testing tournament theory with professional tennis data." *Available at SSRN 477442* (2003).

Sunde, Uwe. "Heterogeneity and performance in tournaments: a test for incentive effects using professional tennis data." *Applied Economics* 41.25 (2009): 3199-3208.

Von Allmen, Peter. "Is the reward system in NASCAR efficient?." *Journal of Sports Economics* 2.1 (2001): 62-79.

Appendix

Table A1 -2019 Grand Slams Total Prize Money

Grand Slam	Prize Money
Australia Open	$20,644,340
Roland Garros	$22,467,200
Wimbledon	$22,694,567
US Open	$28,619,350

Table A2 - 2019 ATP Masters 1000 Total Prize Money

ATP Masters 1000	Prize Money
Indian Wells	$9,314,875
Miami	$9,314,875
Monte Carlo	$6,255,234
Madrid	$8,152,782
Rome	$6,486,234
Canada	$6,338,885
Cincinnati	$6,735,690
Shanghai	$8,322,885
Paris	$6,486,234

Table A3 - 2019 ATP Tour 500 Total Prize Money

ATP Tour 500	Prize Money
Rotterdam	$2,350,298
Rio de Janeiro	$1,937,740
Dubai	$2,887,895
Acapulco	$1,931,110
Barcelona	$3,076,030
Halle	$2,485,448
Queen's Club	$2,485,448
Hamburg	$2,078,149
Washington	$2,046,340
Beijing	$3,666,275
Tokyo	$2,046,340
Vienna	$2,725,867
Basel	$2,486,372

Information about prize money of Tables A1 to A4 come from (https://www.atptour.com/en/scores/results-archive?year=2019&tournamentType=atpgs). In order to get all the Prize Money values in $, we used the 2019 average price of exchange: 1£ = 1,2772$; 1€ = 1,12$; 1A$ = 0,6954$.

Table A4 - 2019 ATP Tour 250 Tournaments Prize

ATP Tour 250	Prize Money
Doha	$1,416,205
Brisbane	$589,680
Pune	$589,680
Sidney	$589,680
Auckland	$589,680
Cordoba	$589,680
Montpellier	$656,477
Sofia	$656,477
New York	$777,385
Buenos Aires	$673,135
Marseille	$833,291
Delray Beach	$651,215
São Paulo	$618,810
Houston	$652,245
Marrakech	$656,477
Budapest	$656,477
Estoril	$656,477
Munich	$656,477
Geneva	$656,477
Lyon	$656,477
Stuttgart	$845,085
s-Hertogenbosch	$796,628
Antalya	$568,389
Eastbourne	$835,386
Newport	$652,245
Bastad	$656,477
Umag	$656,477
Atlanta	$777,385
Gstaad	$656,477
Los Cabos	$858,565
Kitzbuhel	$656,477
Winston-Salem	$807,210
St. Petersburg	$1,248,665
Metz	$656,477
Chengdu	$1,213,295
Zhuhai	$1,000,000
Moscow	$922,520
Antwerp	$796,628
Stockholm	$796,628

Information about prize money of Tables A1 to A4 come from (https://www.atptour.com/en/scores/results-archive?year=2019&tournamentType=atpgs). In order to get all the Prize Money values in $, we used the 2019 average price of exchange: 1£ = 1,2772$; 1€ = 1,12$; 1A$ = 0,6954$.