PROJECT REPORT

The Role of Data Analytics in Professional Tennis

Abstract

The ATP tour has been one of the few professional sports leagues to avoid the use of data analytics to advance the sport. This project analyses the untapped potential of data analytics within tennis. This report outlines and describes the goals, process, and outcome of our project.

Andrew Meyer, Pegah Karimi, Naveen Kuman, Dikshant Joshi, Rajasheker

Contents

List of figures

List of tables

1 Introduction

Tennis has long been viewed as “gentleman’s” game, where players show their utmost respect for both their opponents and the traditions of the game. This may be one of the main reasons tennis has taken longer than nearly all other professional sports leagues to embrace the potential of new technologies which enhance the fan experience and allow players to gain a competitive edge in matches. Despite having used an advanced camera system called “Hawkeye” to assist line judges in making in and out calls since the Miami Open in 2006, the data captured by the Hawkeye system has mainly remained used for this purpose (Newcomb 2006). In recent years the system has been used to enhance the television experience for viewers, breaking down players service location and hit points but it is unclear how widely available the data is for players and coaches. This is in sharp contrast to the use of data analytics in other major sports leagues such as baseball and football, where teams are constructed based on rigorous analysis of player’s strengths and weaknesses. The use of data analytics to construct and improve a sports team’s performance has been coined Moneyball after the popular movie of the same name which told the story of the Oakland A’s use of data analytics to find undervalued players and create a dominate baseball team, ringing in the era of data analytics in professional sports (Dargis 2011).

The goal of this project is to evaluate whether the use of data analytics is viable for the sport of tennis by developing an understanding of the available data and what insights into a player’s performance it may provide. The existence of informative data and analysis may signal that tennis is ready to leap forward into the Moneyball era. Our group focused on two main areas where we felt data analytics had the greatest potential to enhance the sport of tennis. First, we wanted to understand if the data could provide insight into what metrics were most important in determining player success which would be valuable to tennis coaches. Second, we sought to understand if the available data could help expand the betting industry surrounding tennis, both through helping set more accurate betting odds for betting agencies and providing bettors more information on which to base their bet. These two areas would help improve competitiveness in a sport dominated by its top 4 players, as well as provide an influx of money by enhancing the legitimacy of tennis betting.

To guide our project, we consulted our team member Andrew Meyer who played highschool tennis. Andrew mentioned the importance of a players serve in determining match outcomes. Given this information we focused on finding data containing serve statistics, as its importance to the match outcomes makes it a strong candidate to test whether there truly are insights to be found from data analysis for Tennis. The remainder of the report will cover our methodology in identifying data sources, the process for creating a relational database to store the data, reports on initial analysis performed on, and reflection on the future potential of the database as well as the process undergone to this point.

2 Methodology

2.1 Identifying Potential Data Sources

The first step towards performing informative analysis was finding a data set with the appropriate variables encapsulating a players match performance. Given our focus on a players serve we sought data sets which contained information relevant to a players serve. Additionally, our end goal of improving betting lines and outcomes meant we needed to find data which contained betting odds for matches. With the above guiding criteria, we understood that match level data was required, as betting odds are for individual matches, and understanding how a players serve effects their ability to win would require looking at individual matches and their result. Furthermore, we wanted to have an understanding of more general factors which may affect a player's performance; therefore, we sought data which contained information about the tournament a match was a part of in addition to data which had characteristics about the players involved. Potential tournament variables we sought were tournament level (Grand Slam, Masters 1000/500/200), court surface (grass, clay, hardcourt) and time of year. Player variables we desired were age, height, and handedness. With the above desired variables in mind, we found 4 data sets which we believed were best suited to the end goals of our project. The following table shows the comparison between these datasets.

Table 1 Evaluations of the data sets

	Pros	Cons
ATP tour results and betting data 2000-2019 (*ATP and WTA Tennis Results*)	– Match venue, winners, and losers – Data from 2000 onwards. – Contains betting odds history – Columns references are available.	– Does not have the player statistics. – Betting data starts in 2010
WTA matches and ranking data 2000-2017 (*ATP and WTA Tennis Results*)	– 17 years of match data as well as ranking and player information – Match Details: games won, sets played, court surface, tournament level, opponent ranking, age	– Null values for some variables with greater frequency before 2006 – Data hasn't been scrapped for past 4 years – WTA matches only
Tennis match-charting project (*JeffSackmann*)	– Detailed shot and return data – Comprehensive and current (monthly updates) – Opensource, user updated scores and matches	– A large pool of data, excessive and repeated data points – The data only serves as a reference from 1960 - Now – And not every file begins recording in 1960
ATP world tour tennis data 1877-2017 (*ATP World Tour Tennis Data*)	– Tournament information: name, court surface, time of year – Player data: height, age, name, player id, handedness – Match data: winner and loser stats for first serve percentage, breakpoints save, etc.	– Dataset is not indexed. – Most recent year of data is 2017 – Not a clean dataset as there are null values for some of the attributes in dataset.

2.2 Data Set Decision

We chose two data sources from the initial selection of four to be included in our database based upon their fit with our initial criteria and project goals. The first data source we selected was "ATP World Tennis data 1877-2017". This data source was chosen because of the breadth of variables it contained. The data set contained match level data for each tournament on the ATP tour, as well as player information for both the winner and loser of each match. Furthermore, the data set contained the most variables which reflected the effectiveness of a players serve. These variables include, first serve points won, second serve points won, total number of services, break points faced, breakpoints saved, and number of aces for both the winner and loser of each match. With these variables we could also create other important variables such as first serve percentage for both the winner and loser of a match, as well as the winner and losers break point save percentage, both which will be helpful for understanding the impact of a players serve on the match outcome. Additionally, the data set had a unique player id column and a unique tournament id column, both of which would be useful for normalization of the database.

The second data source chosen was the "ATP tour results and betting results 2000-2019". This data source was important to include to accomplish the projects second goal of improving the betting industry surrounding tennis. To accomplish this, we need to understand how historical betting odds compared with match outcomes and player statistics. This data source included betting odds for the winner and loser of a match from multiple betting agencies. Additionally, the data contained in match betting odds for after each set, which could potentially allow for an understanding on how betting odds change based on in match performance. Lastly, the betting odds data when in the database with the match level player statistics from the first data set will provide insight into what player statistics influence a specific matches betting line, which will help accomplish the goal of improving the legitimacy of tennis betting.

While our dataset contains over 100 variables allowing for a variety of different analyses, there is the potential to further increase potential insights by expanding the scope of match level variables. Unfortunately, from the publicly available data we found, our data set contains nearly all the key variables. Potential variables which would be valuable to include if made available would be first serve and second serve speed, rally length, forehand and backhand winners, net winners (points won at net during a match), average first and second serve return position (distance from baseline), and racket brand. These variables would provide a comprehensive picture of a players match performance allowing for much more precise analysis of the determinants of player and match success, as well as allow an individual to understand a player's style of play. This data could help us understand if a player is aggressive or conservative by looking at their return position, number of winners hit, and points won at the net. These data points could also show whether a player prefers to serve and volley (follows serve to net to volley return), or rally (baseline hitting) by looking at serve speed, net points won, forehand and backhand winners.

Despite not having access to those more specific data points, we felt that the two data sources chosen in combination would provide the most informative data for our project goal given the data available. The large number of player and match statistics will allow us to answer questions not only related to the outcome of a match, but also understand what makes a more successful player. With the presence of winner and loser data for each match as well as winner and loser id's to identify the winner and loser, we can aggregate a players data across all their matches to create a profile of each players career statistics. This will allow us to see whether the variables such as first serve percentage, break point save percentage for a player effect their match win percentage. With this information we can understand what factors make for the most successful tennis players, and therefore better inform coaches. Additionally with the match level data we can see what match specific factors contribute to winning a match by comparing the different value a variable (serve percentage, height, aces) between the winner and loser. Lastly these match insights can be combined with the betting data and player career information to better understand what variables are most important to betting odds, or potentially create insights into finding matchups where a better is more likely to place a winning bet.

3 Database Design and Migration

Data migration to AWS was done as instructed on Canvas. The database was accessed through MySQL Workbench to normalize data and create ERD. For the testing to capabilities of our data base we used only data from 2000-2016, a time period which both data bases covered. The primary data from "ATP World Tennis data 1877-2017" was normalized into three different tables. The original data table had multiple functional dependencies. Tourney_id determined the tournament level, tournament name, court surface, and match data. Therefore, a new table called tournament_info with tourney_id as the primary ley was created. Furthermore, the winner_id and loser_ id column determined the winner and losers name, height, handedness, country or origin, rank, and seed. A new table was created called player_info which contained the related player information with player_id as the primary key. The primary key, player_id, can appear in the main atp_match table in both the loser_id and winner_id column, however, the player id cannot appear in both the loser_id and winner_id in one match. Lastly, all redundancies were removed from the main atpMatch table.

The "ATP tour results and betting results 2000-2019" was then added to the database. Given this was additional data from a different source the table was not completely normalized, however, it was connected to the other three tables with redundancies removed. The primary key for the bettingData table was a composite key comprised of the winner id, loser id, and match date. The table was joined to the tournament_info with tournament id(TID) as the foreign key. The bettingData table was then joined with player_info table with winner id(WID) and loser id(LID) both as foreign keys connected with player_id in the player_info table. Lastly, the bettingData table was joined with the atpMatch table by the composite foreign key containing tournament id and match id. Redundancies of variables already existing in the joined tables were then removed. Figure 1 shows the related ERD and the Appendix contains the data dictionary for each table within the database.

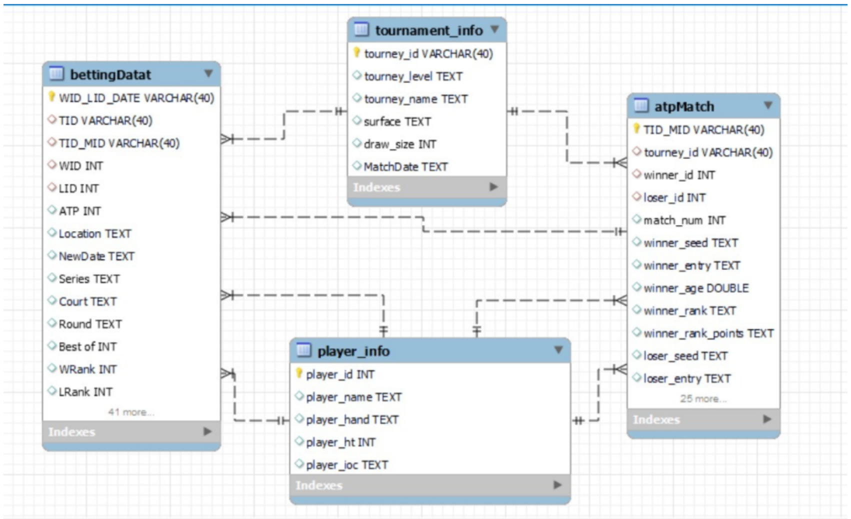


Figure 1 Normalized ERD

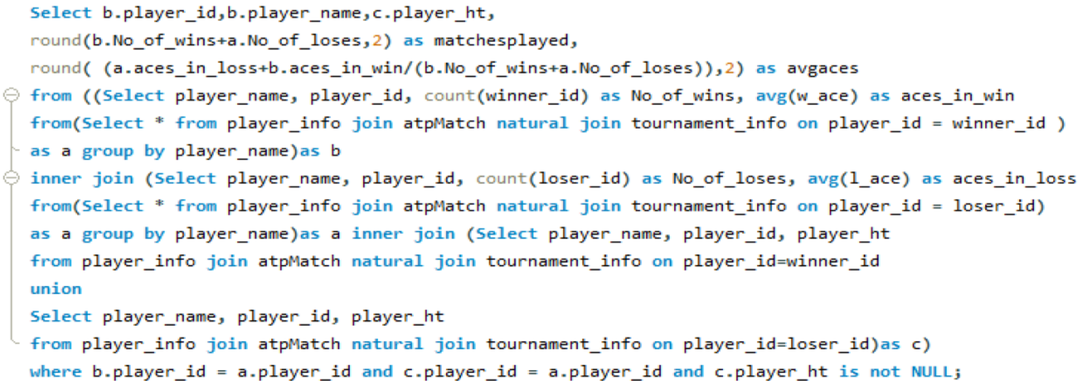
4 Data Analysis

The main goal of our project was to design a database which could help test whether tennis was ready to move into an era of data analytics. With data sources we compiled and migrated to the database, we ran multiple analyses to evaluate if there were insights to be found within the data and therefore show that tennis could benefit from the implementation data analytics. Our test analyses focused on answering our specific project goals of using the database to understand key metrics of player success, and the potential to increase the legitimacy of betting within tennis. Three main research questions were created to see if our database could achieve these goals:

1. Is there a relationship between player’s winning probability and his height?
2. Is there a relationship between player’s winning probability and the breakpoints saved?
3. Why did in the semi final match of BNP Paribas 2013, Paris Masters 1000 title tournament, David Ferrer won Rafael Nadal, although the betting agencies predicted a very high chance of him losing that match? What are some elements that those agencies did not consider but we can conclude from our data?

4.1 Relationship between player height and win percentage

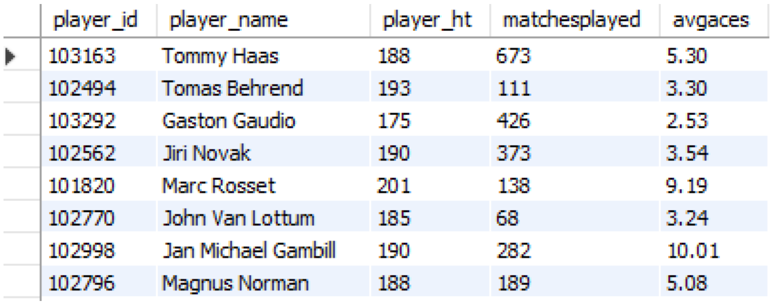
The first question we sought to answer was whether taller players had a higher win percentage. Given the importance of the serve in tennis, our team thought a taller player may have a faster serve allowing them to win more serve points therefore holding more service games and increasing their chances of winning the match. The following query was used to pull each players height and the average number of aces they hit per match.



```
Select b.player_id,b.player_name,c.player_ht,
round(b.No_of_wins+a.No_of_loses,2) as matchesplayed,
round( (a.aces_in_loss+b.aces_in_win/(b.No_of_wins+a.No_of_loses)),2) as avgaces
from ((Select player_name, player_id, count(winner_id) as No_of_wins, avg(w_ace) as aces_in_win
from(Select * from player_info join atpMatch natural join tournament_info on player_id = winner_id )
as a group by player_name)as b
inner join (Select player_name, player_id, count(loser_id) as No_of_loses, avg(l_ace) as aces_in_loss
from(Select * from player_info join atpMatch natural join tournament_info on player_id = loser_id)
as a group by player_name)as a inner join (Select player_name, player_id, player_ht
from player_info join atpMatch natural join tournament_info on player_id=winner_id
union
Select player_name, player_id, player_ht
from player_info join atpMatch natural join tournament_info on player_id=loser_id)as c)
where b.player_id = a.player_id and c.player_id = a.player_id and c.player_ht is not NULL;
```

Table 2 was produced and exported to be used in R Studio.

Table 2 Query for each players height and the average number of aces



player_id	player_name	player_ht	matchesplayed	avgaces
103163	Tommy Haas	188	673	5.30
102494	Tomas Behrend	193	111	3.30
103292	Gaston Gaudio	175	426	2.53
102562	Jiri Novak	190	373	3.54
101820	Marc Rosset	201	138	9.19
102770	John Van Lottum	185	68	3.24
102998	Jan Michael Gambill	190	282	10.01
102796	Magnus Norman	188	189	5.08

With the data in R studio, a scatter plot (Figure 2) with a fitted linear regression line was created.

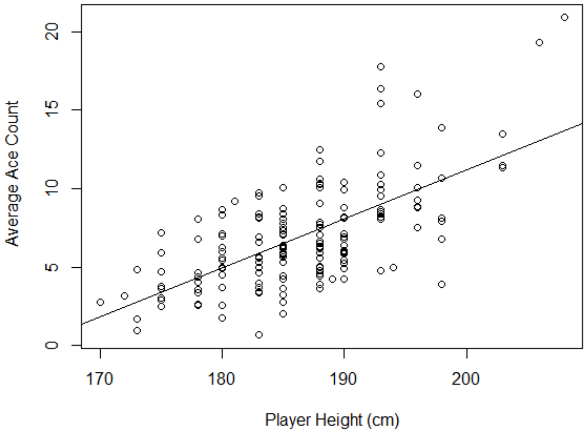


Figure 2 Height Vs. Ace Count

The plot and fitted regression line show that as a player's height increases, they will serve a higher number of aces on average in a match. This is likely a result of taller players being able to hit faster serves, increasing the probability of a hitting an ace. To further understand the effect of height on a players serve, we wanted to see if taller players not only hit more aces but if that translated to winning a greater percentage of first serve points. Using a similar query of the database which was then exported to R the plot below (Figure 3) was created.

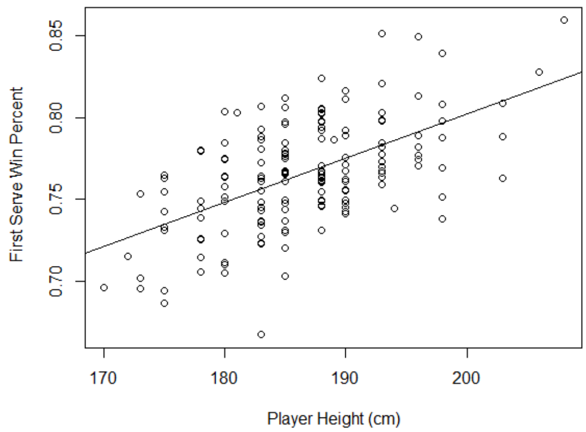
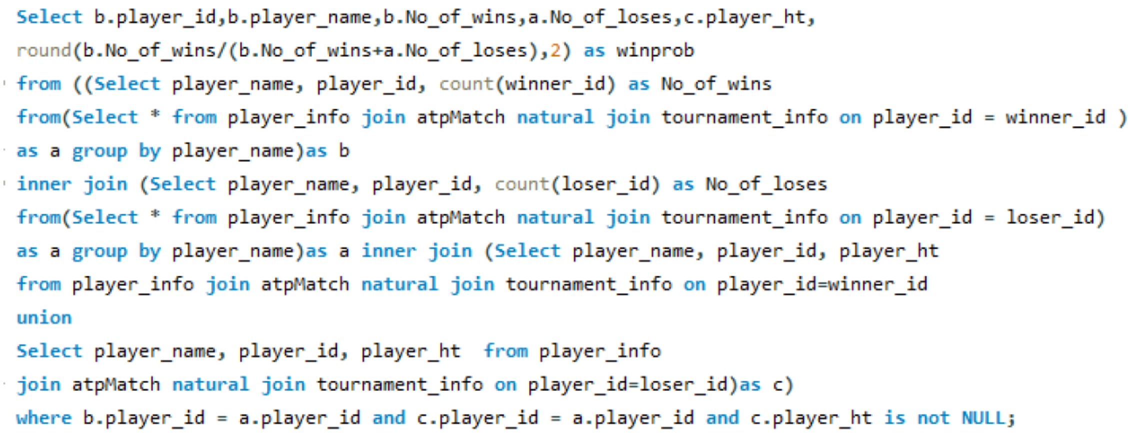


Figure 3 Height Vs. First Serve Win Percentage

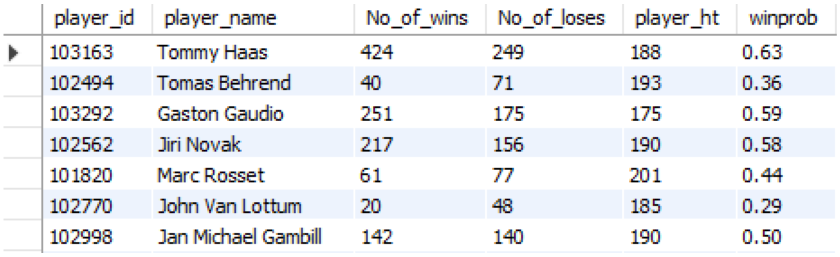
Plot two shows that taller players win a higher percentage of first serve points in comparison to shorter players. However, to understand if the benefits of a stronger serve from increased height translated into winning a plot with a correlation line between height and win percentage was created in R. The following SQL code was run to pull the necessary data from the database:



```
Select b.player_id,b.player_name,b.No_of_wins,a.No_of_loses,c.player_ht,
round(b.No_of_wins/(b.No_of_wins+a.No_of_loses),2) as winprob
from ((Select player_name, player_id, count(winner_id) as No_of_wins
from(Select * from player_info join atpMatch natural join tournament_info on player_id = winner_id )
as a group by player_name)as b
inner join (Select player_name, player_id, count(loser_id) as No_of_loses
from(Select * from player_info join atpMatch natural join tournament_info on player_id = loser_id)
as a group by player_name)as a inner join (Select player_name, player_id, player_ht
from player_info join atpMatch natural join tournament_info on player_id=winner_id
union
Select player_name, player_id, player_ht  from player_info
join atpMatch natural join tournament_info on player_id=loser_id)as c)
where b.player_id = a.player_id and c.player_id = a.player_id and c.player_ht is not NULL;
```

The resulting table (Table 3) was then exported to R Studio for the creation of a scatter plot with a fitted correlation line (Figure 4).

Table 3: Win Probability Vs. Player Height



player_id	player_name	No_of_wins	No_of_loses	player_ht	winprob
103163	Tommy Haas	424	249	188	0.63
102494	Tomas Behrend	40	71	193	0.36
103292	Gaston Gaudio	251	175	175	0.59
102562	Jiri Novak	217	156	190	0.58
101820	Marc Rosset	61	77	201	0.44
102770	John Van Lottum	20	48	185	0.29
102998	Jan Michael Gambill	142	140	190	0.50

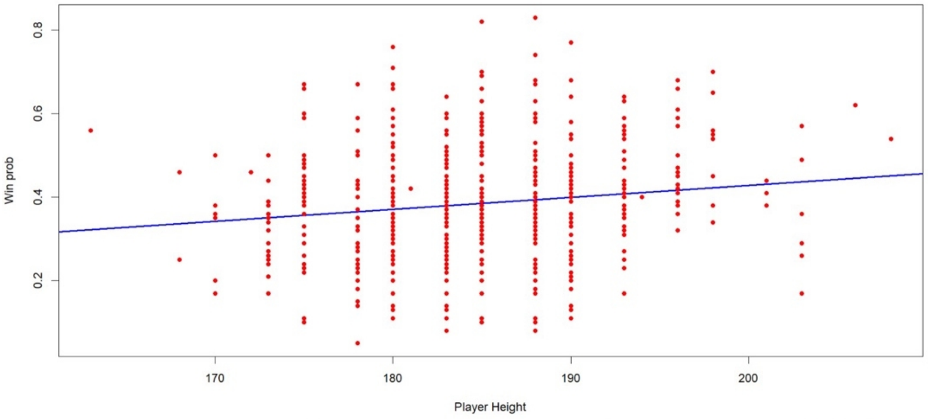
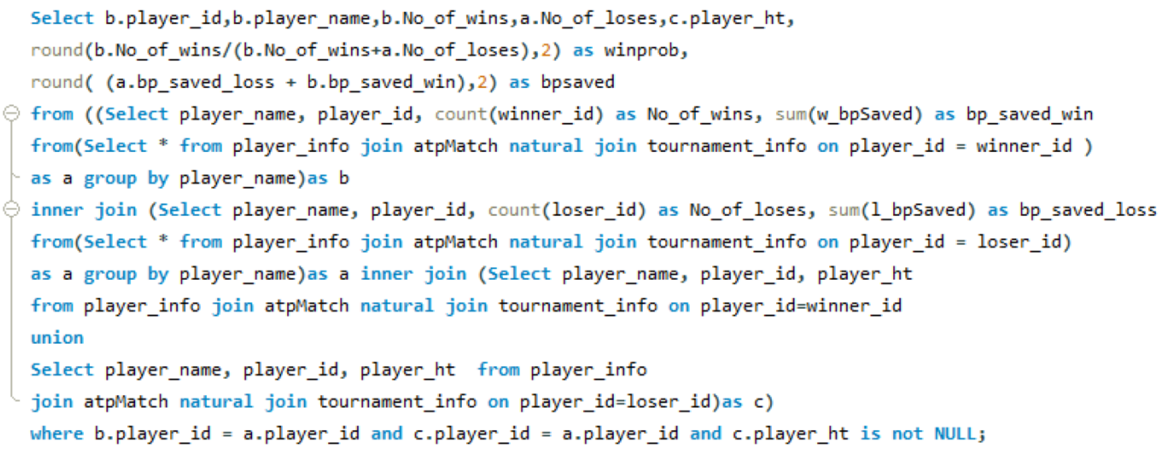


Figure 4 Height vs. Win Probability Regression

The plot of player height and win percentage shows that there is a weak correlation, however not as strong as would be expected from the advantage height provides on the serve. This suggest that there are other factors that may hold back a taller player from capitalizing on their serve. A possible explanation is slower movement around the court, making it harder to win points on their second serve, or when returning their opponents serve. Ultimately, the data successfully showed that taller players do have a stronger performing serve, but also revealed that it does not have a significant impact on improving their chances of winning matches.

4.2 Players Win Percentage and Break Points Saved

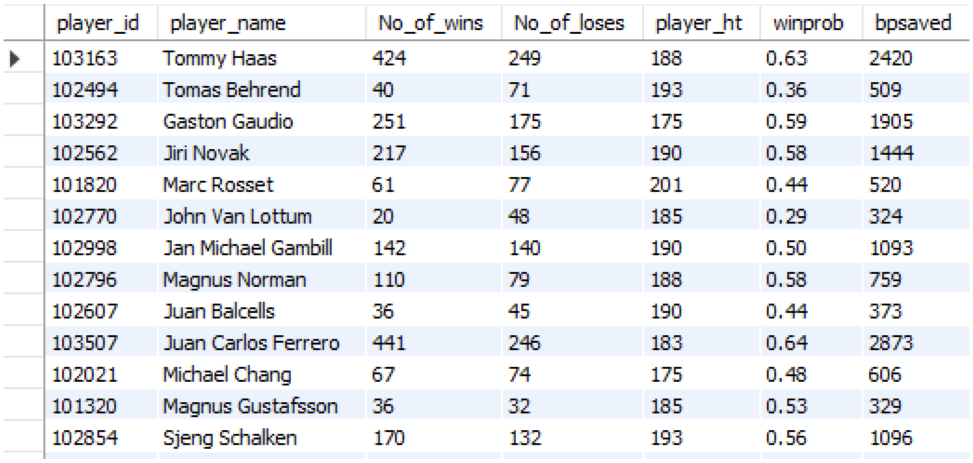
The second research question we sought to answer was whether the number of breakpoints a player saved has an impact on their match winning percentage. A breakpoint is when a server must win the point to avoid losing the game. When a player is broken (loses game while serving), it means they must win a game when their opponent is serving to avoid losing the set. Winning a point while serving is easier than when returning because the server can dictate the point based of their serve or hit an ace. Therefore, players who save a high number of breakpoints are less likely to have their serve broken which should have a favorable effect on their outcome in the match. The query below was run on the database to pull the necessary data.



```
Select b.player_id,b.player_name,b.No_of_wins,a.No_of_loses,c.player_ht,
round(b.No_of_wins/(b.No_of_wins+a.No_of_loses),2) as winprob,
round( (a.bp_saved_loss + b.bp_saved_win),2) as bpsaved
from ((Select player_name, player_id, count(winner_id) as No_of_wins, sum(w_bpSaved) as bp_saved_win
from(Select * from player_info join atpMatch natural join tournament_info on player_id = winner_id )
as a group by player_name)as b
inner join (Select player_name, player_id, count(loser_id) as No_of_loses, sum(l_bpSaved) as bp_saved_loss
from(Select * from player_info join atpMatch natural join tournament_info on player_id = loser_id)
as a group by player_name)as a inner join (Select player_name, player_id, player_ht
from player_info join atpMatch natural join tournament_info on player_id=winner_id
union
Select player_name, player_id, player_ht  from player_info
join atpMatch natural join tournament_info on player_id=loser_id)as c)
where b.player_id = a.player_id and c.player_id = a.player_id and c.player_ht is not NULL;
```

The following table was produced.

Table 4: Win Probabality Vs. Breakpoints



	player_id	player_name	No_of_wins	No_of_loses	player_ht	winprob	bpsaved
▶	103163	Tommy Haas	424	249	188	0.63	2420
	102494	Tomas Behrend	40	71	193	0.36	509
	103292	Gaston Gaudio	251	175	175	0.59	1905
	102562	Jiri Novak	217	156	190	0.58	1444
	101820	Marc Rosset	61	77	201	0.44	520
	102770	John Van Lottum	20	48	185	0.29	324
	102998	Jan Michael Gambill	142	140	190	0.50	1093
	102796	Magnus Norman	110	79	188	0.58	759
	102607	Juan Balcells	36	45	190	0.44	373
	103507	Juan Carlos Ferrero	441	246	183	0.64	2873
	102021	Michael Chang	67	74	175	0.48	606
	101320	Magnus Gustafsson	36	32	185	0.53	329
	102854	Sjeng Schalken	170	132	193	0.56	1096

The table was exported into R for regression analysis, after analyzing preliminary regressions it was decided to focus the 100 players with the most wins. The simple linear regression of breakpoints saved on win percentage produced an R squared of .65, meaning the 65% of the variation in a players win percentage can be attributed to the number of breakpoints they saved. The coefficient for the break points saved variable was .0005, suggesting that for every 100 breakpoints saved a players win percentage increases by 5%. The coefficient was statistically significant above the 99% confidence level which allows us to say that the data shows that saving break points increases a player's winning percentage, highlighting that the best players are able to serve under pressure and save break points. The associated scatter plot and fitted regression are below.

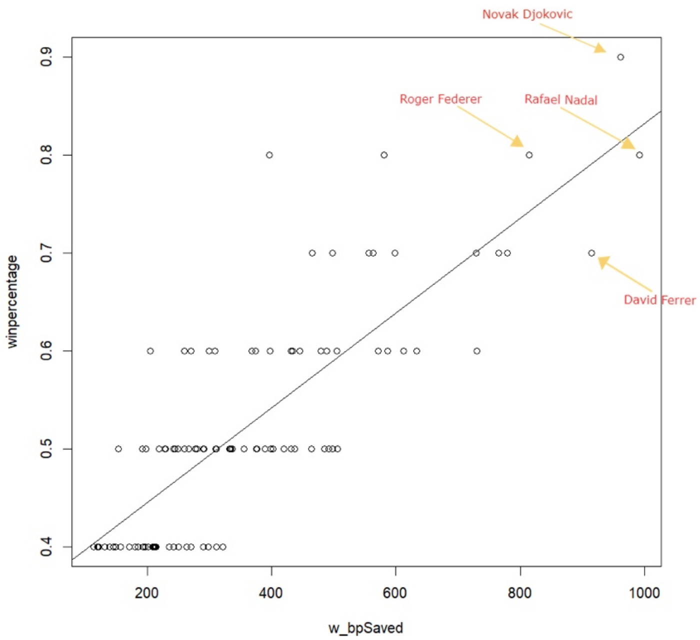


Figure 5 Scatter plot and regression line for win percentage vs. breakpoints saved

4.3 Betting Odds Case study: Rafael Nadal vs. David Ferrer

The analysis for the first two research questions showed that our database, using querying and R, was able to accomplish the first goal of the project of discovering insights into the critical attributes of a successful player. The last research question is designed to test our second goal which is to understand if the match data combined with the betting data can help predict upsets, or alternatively provide information betting agencies could use to set more accurate betting odds. To answer this question, we analyzed a tennis match between David Ferrer and Rafael Nadal during the 2013 Paris Paribas Masters 1000 Tournament. Rafael Nadal was the number one seed at the tournament and was the heavy favorite to win the match. David Ferrer had been a consistent player on the ATP tour and was the defending champion of the tournament, however, has only beat Rafael Nadal 4 times in their 24 meetings.

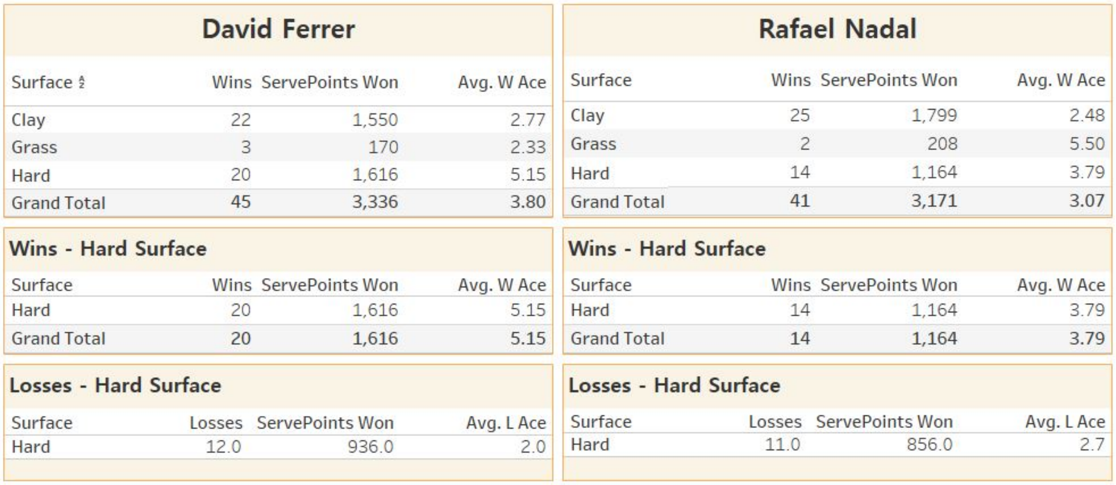
The following betting odds data for the match was pulled from the database.

Table 5 Betting Agency Odds

	David Ferre	Rafael Nadal
Pinnacles Sports	6.65	1.15
Bet365	6.5	1.1
Ladbrokes	6.5	1.1

Table 5 shows that Rafael Nadal was the heavy favorite for the match, however, David Ferrer won the match. To evaluate if our data base could uncover potential indicators that these odds were miscalculated multiple queries (attached in appendix C) looking at win percentage and serve percentage grouped by surface type were ran on the previous two years of data (2010-2012). Table 6 were created from the queries after export into excel.

Table 6: Nadel Vs. Ferrer Match Statistics



David Ferrer

Surface	Wins	ServePoints Won	Avg. W Ace
Clay	22	1,550	2.77
Grass	3	170	2.33
Hard	20	1,616	5.15
Grand Total	45	3,336	3.80

Rafael Nadal

Surface	Wins	ServePoints Won	Avg. W Ace
Clay	25	1,799	2.48
Grass	2	208	5.50
Hard	14	1,164	3.79
Grand Total	41	3,171	3.07

Wins - Hard Surface

Surface	Wins	ServePoints Won	Avg. W Ace
Hard	20	1,616	5.15
Grand Total	20	1,616	5.15

Wins - Hard Surface

Surface	Wins	ServePoints Won	Avg. W Ace
Hard	14	1,164	3.79
Grand Total	14	1,164	3.79

Losses - Hard Surface

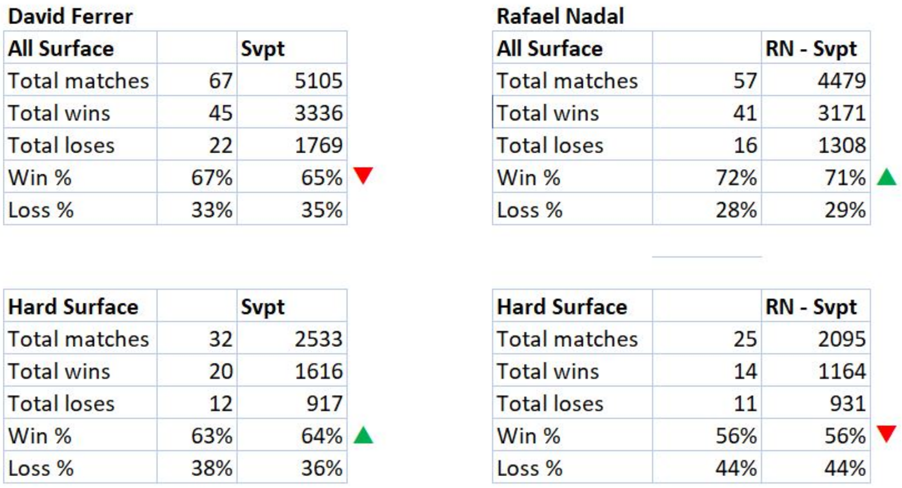
Surface	Losses	ServePoints Won	Avg. L Ace
Hard	12.0	936.0	2.0

Losses - Hard Surface

Surface	Losses	ServePoints Won	Avg. L Ace
Hard	11.0	856.0	2.7

Table 6 shows that that David Ferrer had a higher winning percentage on hard courts over the previous two season, in addition to having a higher average ace count in hard court matches. Furthermore, the two tables below show that when looking at all surfaces Rafael Nadal has a higher win percentage on his service points and match win percentage than David Ferrer. However, when looking at hard courts only over the two years before the 2013 Paris Paribas 1000 tournament, David Ferrer had both a first serve point win percentage and match win percentage.

Table 7 : Rafael Nadal Vs. David Ferrer Hard Court Vs. All Surfaces



David Ferrer

All Surface		Svpt	
Total matches	67	5105	
Total wins	45	3336	
Total loses	22	1769	
Win %	67%	65%	▼
Loss %	33%	35%	

Rafael Nadal

All Surface		RN - Svpt	
Total matches	57	4479	
Total wins	41	3171	
Total loses	16	1308	
Win %	72%	71%	▲
Loss %	28%	29%	

Hard Surface		Svpt	
Total matches	32	2533	
Total wins	20	1616	
Total loses	12	917	
Win %	63%	64%	▲
Loss %	38%	36%	

Hard Surface		RN - Svpt	
Total matches	25	2095	
Total wins	14	1164	
Total loses	11	931	
Win %	56%	56%	▼
Loss %	44%	44%	

Given that the Paris Paribas Masters 1000 tournament is held on hard courts, the data suggest that David Ferrer should have been less of an underdog in the match, considering his higher average ace count, service point win percent, and overall match win percentage on hard courts. Furthermore, given that Paris tournament takes place in October each year, we wanted to understand if the time of year might have been an indicator of the outcome. A query was run on the previous two years of matches to find match wins by month for Rafael Nadal and David Ferrer.

The resulting table was exported to excel, and Tableau was used to create the graphic below.

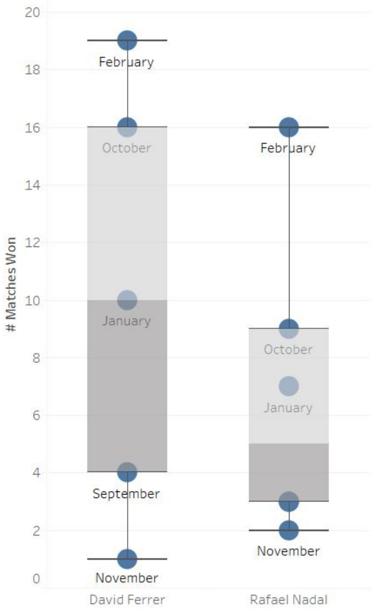


Figure 6 Ferrer vs. Nadal Wins by Month

The graphic shows the Ferrer has produced much stronger results in terms of match wins in October in comparison to Nadal. With the tournament taking place in October the data suggest that Ferrer should be expected to perform above his normal level. This information combined with the hard-court match statistics above reveal that the odd makers over favored Nadal in this match up.

The above analysis in this situation reveals how our database could help improve betting in the sport of tennis. First, if used by the odds makers, it could have allowed them to set more accurate odds by allowing them to evaluate each player based on their performance by court type and time of year, factors that appeared to be overlooked in actual odds for the match. Additionally, this data in the hands of bettors could have allowed them to see the upset potential given the favorable player statistics for Ferrer in comparison to the match odds which favored Nadal, giving the better a strong chance at receiving a bigger pay out. This example illustrates that out data base can achieve the second project goal of helping advance the betting industry surrounding tennis.

5 Project Reflection

Throughout the project, our team developed an understanding of the data surrounding tennis, discovered the most useful tools to interpret the available data, and most importantly learned each group members strengths to allocate assignments efficiently. The most interesting part of the project was understanding the availability of data for tennis, and the insights which could be found from it. Our group was surprised to find that there was limited publicly available data which would describe a players match performance, such as serve speed, ground stroke speed, number of backhands vs. forehands hit, returns in play percentage, and serve location. Given the presence of the Hawkeye system, which tracks ball movement, all of the statistics should be trackable for each player in a match and would provide an abundance of information about what factors determine a tennis match winner. However, given the data we were able to find, it was interesting and exciting to see that there were insights to be derived from the data that highlighted critical aspects of a players game. Specifically, understanding the relationship between a players height and serve success, and how the positive relationship between the two did not result in significantly better match outcomes. Furthermore, it was exciting to see that our prediction that saving more breakpoints would lead to a higher win percentage was validated by the data analysis, showing the potential of data analytics within tennis.

To derive these insights from the data the most helpful tools we used were SQL and R. The use of a normalized, relational database allowed us to easily query the relevant attributes from the dataset to run analysis in R. Considering the number of variables our data set contained, having normalized tables allowed for easier viewing of the data, and using SQL to query only the selected variables for a given analysis significantly reduced to amount of time filtering and sorting the data in R. R specifically helped our group run higher level analysis on the data which helped show the potential of our data to provide meaningful analytics for tennis players. While SQL can filter and organize the data to make analysis easier, R allowed us to expand beyond descriptive analysis methods with the ability to run regressions. Regression analysis allowed our group to understand the actual relationships between player attributes, such as height or break points save, and outcomes such a match winning percentage and serve statistics, a necessary feature of tennis is to become an analytical sport.

While our group was able to accomplish and show proof of concept of our original goals there are various areas for improvement. The biggest area for potential improvement would be to expand the breadth of variables relating to a players match play, many of which were discussed above. Access to these variables would greatly expand the analytical capabilities of our data base and may allow for the creation of a predictive model which predicts match winners. The largest hurdle to accomplishing this goal is sourcing the data as it was not found in our initial data source search. The second area for improvement would be to further normalize our database, specifically the betting data table which was not completely normalized. The last large area for potential improvement would be to increase the complexity of our analyses, while we ran multiple simple linear regression models, the use of interaction terms and multiple regressors in the models may provide more detailed and relevant information regarding player outcomes in a match.

Throughout the process our team learned how to best delegate task to make the project run more efficiently. Originally, we attempted to each participate in all aspects of the project, however, this often led to repeated work and an inefficient use of time. By taking time to understand everyone's strengths we were able to effectively split up tasks resulting in higher quality work.

For example, we leveraged Andrew’s experience playing tennis to help us isolate the most relevant data points and understand the game of tennis so we could target our research and analysis appropriately. For creating and normalizing the database we used Dishkant and Rajashekar’s experience to guide to the process, and for visualization we relied on Naveen’s Tableau experience. Pegah was critical to managing our workflow process and developing our ideas into a report and presentation. Learning how to capitalize on each person’s strengths allowed us to complete the project on time and improve its quality.

While the project overall went smoothly there were some difficulties encountered, the process of migrating data from two different sources into a relational database required the creation of composite keys to ensure a matches betting data aligned with the same matches player statistics. Furthermore, simply understanding what variables might be relevant to tennis analytics required background research and discussions with Andrew to ensure we were targeting the appropriate information and running useful analysis. Both difficulties were addressed by leveraging members of the group who had experience in each specific area. Ultimately this collaboration, allowed us to overcome these difficulties and develop a database which met our project goals of developing a database capable of driving insights into the key components of a successful tennis player, and evaluating the favorability of betting odds for tennis matches.

6 Appendices

6.1 Appendix A: Data Dictionary ATP Matches

Table Name	Object Type	is_key	Column Name	Data Type	Nullable	Column Description
atpMatch	TBL	PK	TID_MID	varchar(40)	NOT NULL	composite primary key/(Tournament Id + matchId)
atpMatch	TBL	FK	tourney_id	varchar(40)	NULL	Tournament ID
atpMatch	TBL	FK	winner_id	int(10)	NULL	winner_id
atpMatch	TBL	FK	loser_id	int(10)	NULL	loser_id
atpMatch	TBL		match_num	int(10)	NULL	match_num
atpMatch	TBL		winner_seed	text(65535)	NULL	Winner_seed position
atpMatch	TBL		winner_entry	text(65535)	NULL	
atpMatch	TBL		winner_age	double(22)	NULL	winner_age
atpMatch	TBL		winner_rank	text(65535)	NULL	winner_rank
atpMatch	TBL		winner_rank_points	text(65535)	NULL	winner_rank_points
atpMatch	TBL		loser_seed	text(65535)	NULL	Loser seed position
atpMatch	TBL		loser_entry	text(65535)	NULL	
atpMatch	TBL		loser_age	double(22)	NULL	loser_age
atpMatch	TBL		loser_rank	text(65535)	NULL	loser_rank
atpMatch	TBL		loser_rank_points	text(65535)	NULL	loser_rank_points
atpMatch	TBL		score	text(65535)	NULL	score
atpMatch	TBL		best_of	int(10)	NULL	best_of
atpMatch	TBL		round	text(65535)	NULL	round
atpMatch	TBL		minutes	int(10)	NULL	minutes
atpMatch	TBL		w_ace	int(10)	NULL	winner_Ace Point
atpMatch	TBL		w_df	int(10)	NULL	Winner_#DoubleFault
atpMatch	TBL		w_svpt	int(10)	NULL	Winner_Total Serve Points
atpMatch	TBL		w_1stIn	int(10)	NULL	winner # 1st serves in
atpMatch	TBL		w_1stWon	int(10)	NULL	Winner # points won on 1st serve
atpMatch	TBL		w_2ndWon	int(10)	NULL	Winner # points won on 2nd serve
atpMatch	TBL		w_SvGms	int(10)	NULL	Winner # Serve games
atpMatch	TBL		w_bpSaved	int(10)	NULL	Winner break point saved
atpMatch	TBL		w_bpFaced	int(10)	NULL	Winner break point faced
atpMatch	TBL		l_ace	int(10)	NULL	Loser_Ace Point
atpMatch	TBL		l_df	int(10)	NULL	Loser_#DoubleFault
atpMatch	TBL		l_svpt	int(10)	NULL	Loser Total Serve Points
atpMatch	TBL		l_1stIn	int(10)	NULL	Loser # 1st serves in
atpMatch	TBL		l_1stWon	int(10)	NULL	Loser # points won on 1st serve
atpMatch	TBL		l_2ndWon	int(10)	NULL	Loser # points won on 2nd serve
atpMatch	TBL		l_SvGms	int(10)	NULL	Loser # serve games
atpMatch	TBL		l_bpSaved	int(10)	NULL	Loserbreak point saved
atpMatch	TBL		l_bpFaced	int(10)	NULL	Loser break point faced

6.2 Appendix B: Data Dictionary Betting Data

Table Name	Object Type	is_key	Column Name	Data Type	Nullable	Column Description
bettingDatat	TBL	PK	WID_LID_DATE	varchar(40)	NOT NULL	composite primary key/(Winner Id + Loser Id + date)
bettingDatat	TBL	FK	TID	varchar(40)	NULL	Tournament ID
bettingDatat	TBL	FK	TID_MID	varchar(40)	NULL	Composite foreignkey(Tournament Id+ Match Id)
bettingDatat	TBL	FK	WID	int(10)	NULL	Winner ID
bettingDatat	TBL	FK	LID	int(10)	NULL	Loser ID
bettingDatat	TBL		ATP	int(10)	NULL	Tournament number (men)
bettingDatat	TBL		Location	text(65535)	NULL	Venue of tournament
bettingDatat	TBL		NewDate	text(65535)	NULL	Match Date
bettingDatat	TBL		Series	text(65535)	NULL	Name of ATP tennis series (Grand Slam, Masters, International or International Gold)
bettingDatat	TBL		Court	text(65535)	NULL	Type of court (outdoors or indoors)
bettingDatat	TBL		Round	text(65535)	NULL	Round of match
bettingDatat	TBL		Best of	int(10)	NULL	Maximum number of sets playable in match
bettingDatat	TBL		Wrank	int(10)	NULL	ATP Entry ranking of the match winner as of the start of the tournament
bettingDatat	TBL		Lrank	int(10)	NULL	ATP Entry ranking of the match loser as of the start of the tournament
bettingDatat	TBL		W1	int(10)	NULL	Number of games won in 1st set by match winner
bettingDatat	TBL		L1	int(10)	NULL	Number of games won in 1st set by match loser
bettingDatat	TBL		W2	int(10)	NULL	Number of games won in 2nd set by match winner
bettingDatat	TBL		L2	int(10)	NULL	Number of games won in 2nd set by match loser
bettingDatat	TBL		W3	text(65535)	NULL	Number of games won in 3rd set by match winner
bettingDatat	TBL		L3	text(65535)	NULL	Number of games won in 3rd set by match loser
bettingDatat	TBL		W4	text(65535)	NULL	Number of games won in 4th set by match winner
bettingDatat	TBL		L4	text(65535)	NULL	Number of games won in 4th set by match loser
bettingDatat	TBL		W5	text(65535)	NULL	Number of games won in 5th set by match winner
bettingDatat	TBL		L5	text(65535)	NULL	Number of games won in 5th set by match loser
bettingDatat	TBL		Wsets	int(10)	NULL	Number of sets won by match winner
bettingDatat	TBL		Lsets	int(10)	NULL	Number of sets won by match loser
bettingDatat	TBL		Comment	text(65535)	NULL	Comment on the match (Completed, won through retirement of loser, or via Walkover)
bettingDatat	TBL		CBW	text(65535)	NULL	Centrebet odds of match winner
bettingDatat	TBL		CBL	text(65535)	NULL	Centrebet odds of match loser
bettingDatat	TBL		GBW	text(65535)	NULL	Gamebookers odds of match winner
bettingDatat	TBL		GBL	text(65535)	NULL	Gamebookers odds of match loser
bettingDatat	TBL		IWW	text(65535)	NULL	Interwetten odds of match winner
bettingDatat	TBL		IWL	text(65535)	NULL	Interwetten odds of match loser
bettingDatat	TBL		SBW	text(65535)	NULL	Sportingbet odds of match winner
bettingDatat	TBL		SBL	text(65535)	NULL	Sportingbet odds of match loser
bettingDatat	TBL		B365W	double(22)	NULL	Bet365 odds of match winner
bettingDatat	TBL		B365L	double(22)	NULL	Bet365 odds of match loser
bettingDatat	TBL		B&WW	text(65535)	NULL	Bet&Win odds of match winner
bettingDatat	TBL		B&WL	text(65535)	NULL	Bet&Win odds of match loser
bettingDatat	TBL		EXW	double(22)	NULL	Expekt odds of match winner
bettingDatat	TBL		EXL	double(22)	NULL	Expekt odds of match loser
bettingDatat	TBL		PSW	double(22)	NULL	Pinnacles Sports odds of match winner
bettingDatat	TBL		PSL	double(22)	NULL	Pinnacles Sports odds of match loser
bettingDatat	TBL		WPts	int(10)	NULL	ATP Entry points of the match winner as of the start of the tournament
bettingDatat	TBL		LPts	int(10)	NULL	ATP Entry points of the match loser as of the start of the tournament
bettingDatat	TBL		UBW	text(65535)	NULL	Unibet odds of match winner
bettingDatat	TBL		UBL	text(65535)	NULL	Unibet odds of match loser
bettingDatat	TBL		LBW	double(22)	NULL	Ladbrokes odds of match winner
bettingDatat	TBL		LBL	double(22)	NULL	Ladbrokes odds of match loser
bettingDatat	TBL		SJW	double(22)	NULL	Stan James odds of match winner
bettingDatat	TBL		SJL	double(22)	NULL	Stan James odds of match loser
bettingDatat	TBL		MaxW	double(22)	NULL	Maximum odds of match winner (as shown by Oddsportal.com)
bettingDatat	TBL		MaxL	double(22)	NULL	Maximum odds of match loser (as shown by Oddsportal.com)
bettingDatat	TBL		AvgW	double(22)	NULL	Average odds of match winner (as shown by Oddsportal.com)
bettingDatat	TBL		AvgL	double(22)	NULL	Average odds of match loser (as shown by Oddsportal.com)

6.3 Appendix C: Data Dictionary Tournament and Payer Info

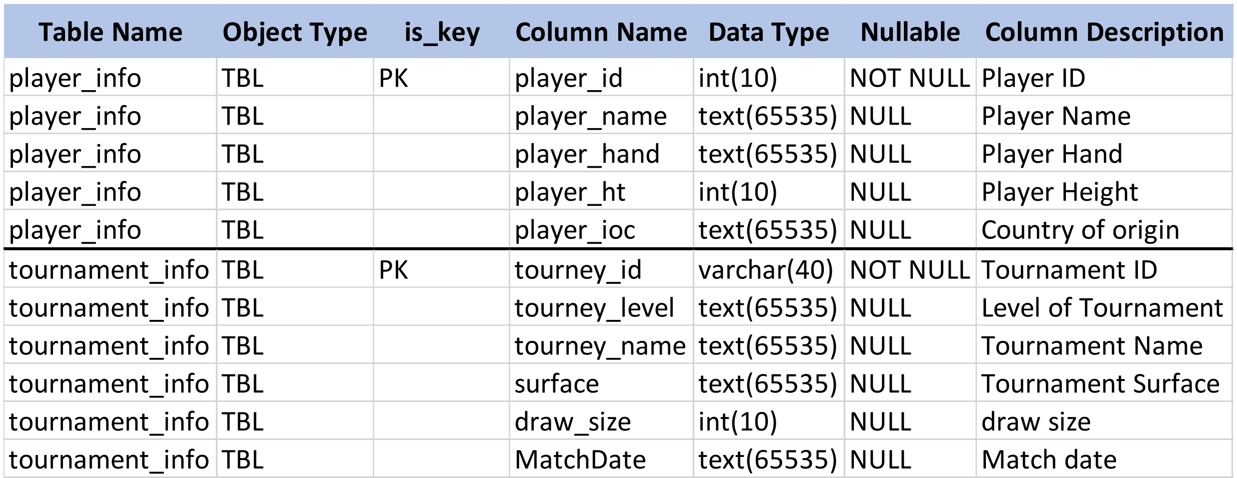
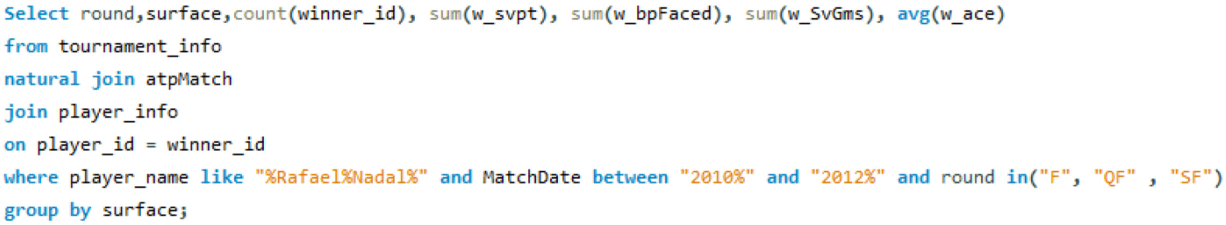


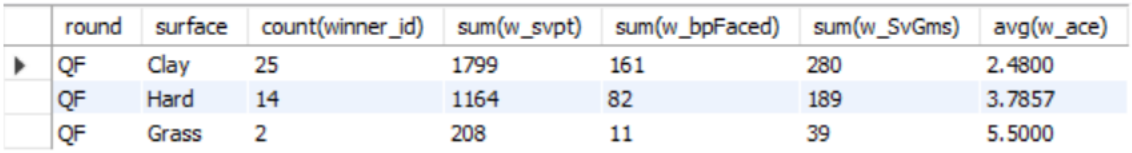
Table Name	Object Type	is_key	Column Name	Data Type	Nullable	Column Description
player_info	TBL	PK	player_id	int(10)	NOT NULL	Player ID
player_info	TBL		player_name	text(65535)	NULL	Player Name
player_info	TBL		player_hand	text(65535)	NULL	Player Hand
player_info	TBL		player_ht	int(10)	NULL	Player Height
player_info	TBL		player_ioc	text(65535)	NULL	Country of origin
tournament_info	TBL	PK	tourney_id	varchar(40)	NOT NULL	Tournament ID
tournament_info	TBL		tourney_level	text(65535)	NULL	Level of Tournament
tournament_info	TBL		tourney_name	text(65535)	NULL	Tournament Name
tournament_info	TBL		surface	text(65535)	NULL	Tournament Surface
tournament_info	TBL		draw_size	int(10)	NULL	draw size
tournament_info	TBL		MatchDate	text(65535)	NULL	Match date

6.4 Appendix D: Nadal Vs. Ferrer Query's

- Query and Result for Nadal Statistics by Surface for Matches Won:

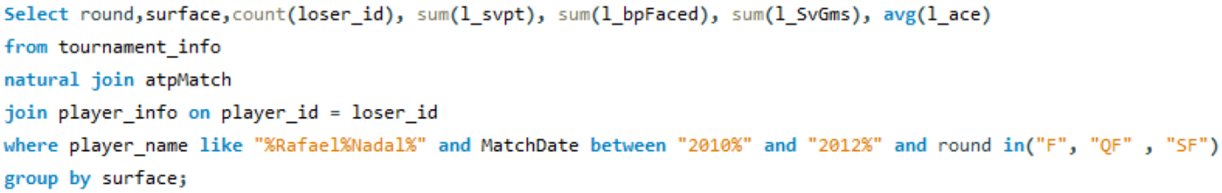


```
Select round,surface,count(winner_id), sum(w_svpt), sum(w_bpFaced), sum(w_SvGms), avg(w_ace)
from tournament_info
natural join atpMatch
join player_info
on player_id = winner_id
where player_name like "%Rafael%Nadal%" and MatchDate between "2010%" and "2012%" and round in("F", "QF" , "SF")
group by surface;
```

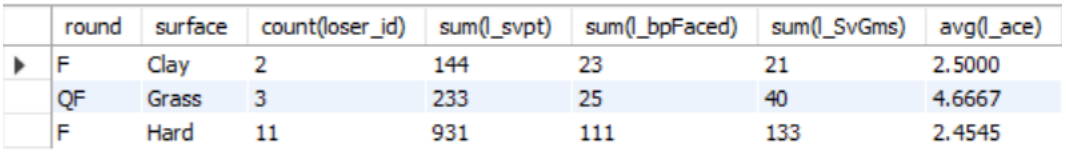


	round	surface	count(winner_id)	sum(w_svpt)	sum(w_bpFaced)	sum(w_SvGms)	avg(w_ace)
▶	QF	Clay	25	1799	161	280	2.4800
	QF	Hard	14	1164	82	189	3.7857
	QF	Grass	2	208	11	39	5.5000

- Query and Result for Nadal Statistics by Surface for Matches Lost:

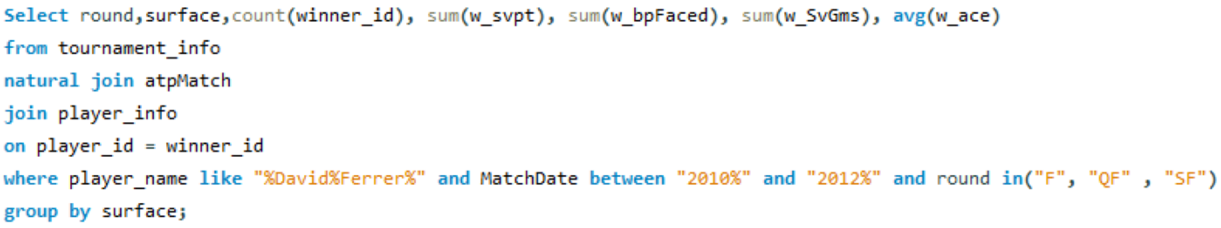


```
Select round,surface,count(loser_id), sum(l_svpt), sum(l_bpFaced), sum(l_SvGms), avg(l_ace)
from tournament_info
natural join atpMatch
join player_info on player_id = loser_id
where player_name like "%Rafael%Nadal%" and MatchDate between "2010%" and "2012%" and round in("F", "QF" , "SF")
group by surface;
```

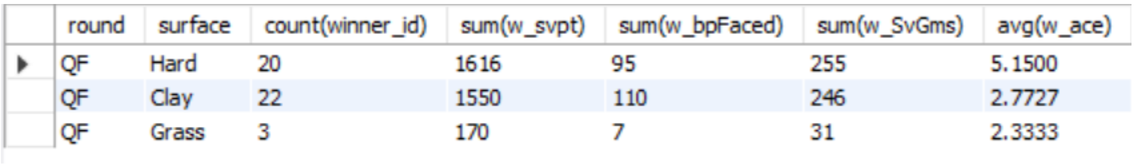


	round	surface	count(loser_id)	sum(l_svpt)	sum(l_bpFaced)	sum(l_SvGms)	avg(l_ace)
▶	F	Clay	2	144	23	21	2.5000
	QF	Grass	3	233	25	40	4.6667
	F	Hard	11	931	111	133	2.4545

- Query and Result for Ferrer Statistics by Surface for Matches Won:

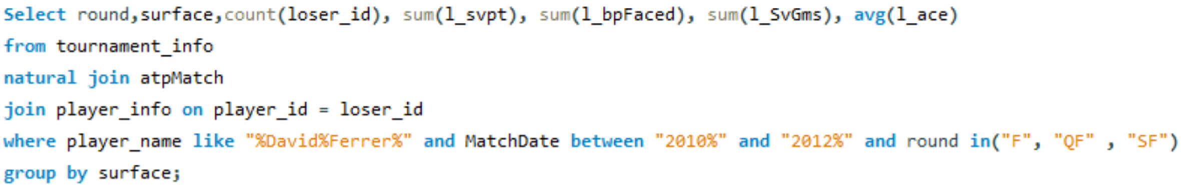


```
Select round,surface,count(winner_id), sum(w_svpt), sum(w_bpFaced), sum(w_SvGms), avg(w_ace)
from tournament_info
natural join atpMatch
join player_info
on player_id = winner_id
where player_name like "%David%Ferrer%" and MatchDate between "2010%" and "2012%" and round in("F", "QF" , "SF")
group by surface;
```

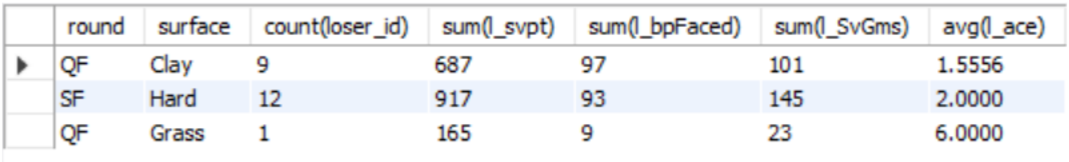


	round	surface	count(winner_id)	sum(w_svpt)	sum(w_bpFaced)	sum(w_SvGms)	avg(w_ace)
▶	QF	Hard	20	1616	95	255	5.1500
	QF	Clay	22	1550	110	246	2.7727
	QF	Grass	3	170	7	31	2.3333

- Query and Results for Ferrer Statistics by Surface for Matches Lost:

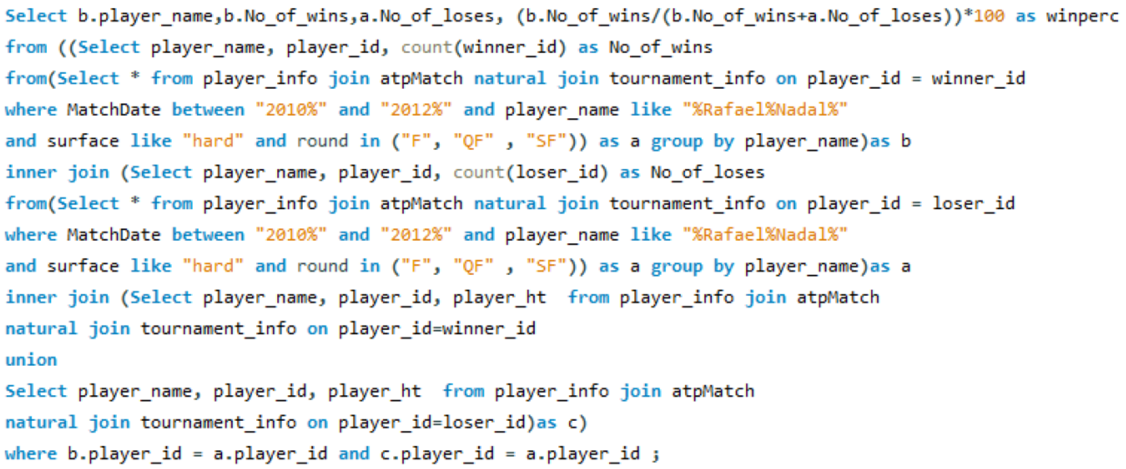


```
Select round,surface,count(loser_id), sum(l_svpt), sum(l_bpFaced), sum(l_SvGms), avg(l_ace)
from tournament_info
natural join atpMatch
join player_info on player_id = loser_id
where player_name like "%David%Ferrer%" and MatchDate between "2010%" and "2012%" and round in("F", "QF" , "SF")
group by surface;
```

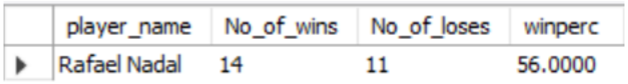


	round	surface	count(loser_id)	sum(l_svpt)	sum(l_bpFaced)	sum(l_SvGms)	avg(l_ace)
▶	QF	Clay	9	687	97	101	1.5556
	SF	Hard	12	917	93	145	2.0000
	QF	Grass	1	165	9	23	6.0000

- Query and Result for Nadal Matches won on Hard Surface:

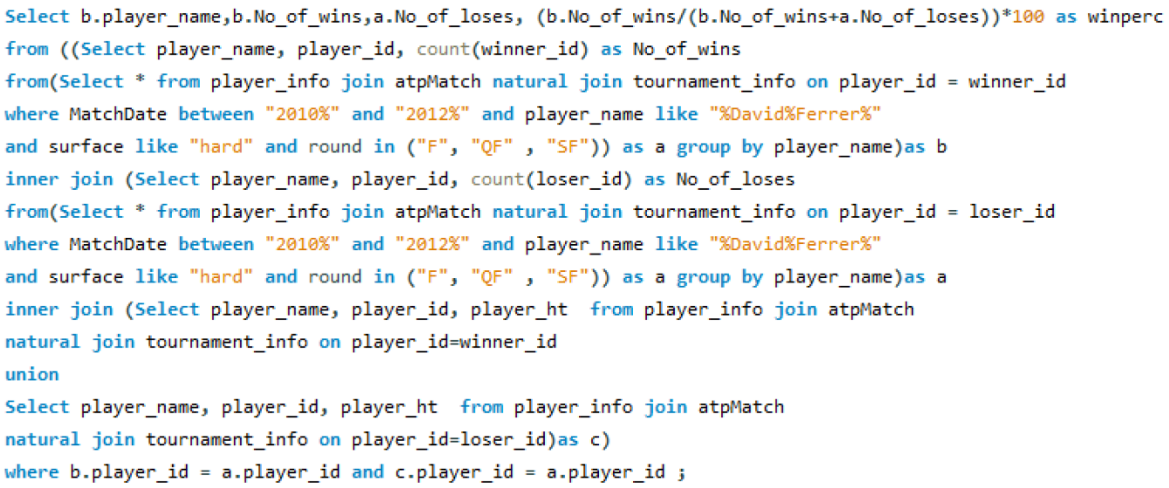


```
Select b.player_name,b.No_of_wins,a.No_of_loses, (b.No_of_wins/(b.No_of_wins+a.No_of_loses))*100 as winperc
from ((Select player_name, player_id, count(winner_id) as No_of_wins
from(Select * from player_info join atpMatch natural join tournament_info on player_id = winner_id
where MatchDate between "2010%" and "2012%" and player_name like "%Rafael%Nadal%"
and surface like "hard" and round in ("F", "QF" , "SF")) as a group by player_name)as b
inner join (Select player_name, player_id, count(loser_id) as No_of_loses
from(Select * from player_info join atpMatch natural join tournament_info on player_id = loser_id
where MatchDate between "2010%" and "2012%" and player_name like "%Rafael%Nadal%"
and surface like "hard" and round in ("F", "QF" , "SF")) as a group by player_name)as a
inner join (Select player_name, player_id, player_ht  from player_info join atpMatch
natural join tournament_info on player_id=winner_id
union
Select player_name, player_id, player_ht  from player_info join atpMatch
natural join tournament_info on player_id=loser_id)as c)
where b.player_id = a.player_id and c.player_id = a.player_id ;
```

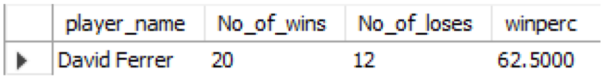


	player_name	No_of_wins	No_of_loses	winperc
▶	Rafael Nadal	14	11	56.0000

- Query and Result for Ferrer Matches won on Hard Surface:



```
Select b.player_name,b.No_of_wins,a.No_of_loses, (b.No_of_wins/(b.No_of_wins+a.No_of_loses))*100 as winperc
from ((Select player_name, player_id, count(winner_id) as No_of_wins
from(Select * from player_info join atpMatch natural join tournament_info on player_id = winner_id
where MatchDate between "2010%" and "2012%" and player_name like "%David%Ferrer%"
and surface like "hard" and round in ("F", "QF" , "SF")) as a group by player_name)as b
inner join (Select player_name, player_id, count(loser_id) as No_of_loses
from(Select * from player_info join atpMatch natural join tournament_info on player_id = loser_id
where MatchDate between "2010%" and "2012%" and player_name like "%David%Ferrer%"
and surface like "hard" and round in ("F", "QF" , "SF")) as a group by player_name)as a
inner join (Select player_name, player_id, player_ht  from player_info join atpMatch
natural join tournament_info on player_id=winner_id
union
Select player_name, player_id, player_ht  from player_info join atpMatch
natural join tournament_info on player_id=loser_id)as c)
where b.player_id = a.player_id and c.player_id = a.player_id ;
```



	player_name	No_of_wins	No_of_loses	winperc
▶	David Ferrer	20	12	62.5000

7 References

ATP and WTA Tennis Results and Betting odds Data | Kaggle. (n.d.). Retrieved October 7, 2022, from https://www.kaggle.com/datasets/hakeem/atp-and-wta-tennis-data?select=df_atp.csv

ATP World Tour tennis data - Dataset - DataHub - Frictionless Data. (n.d.). Retrieved October 7, 2022, from https://datahub.io/sports-data/atp-world-tour-tennis-data

Dargis, Manohla. “Throwing a Digital-Age Curveball.” *The New York Times*, The New York Times, 22 Sept. 2011, https://www.nytimes.com/2011/09/23/movies/brad-pitt-in-moneyball-by-bennett-miller.html.

GitHub - JeffSackmann/tennis_MatchChartingProject: Raw, user-submitted point-by-point data for pro tennis matches. (n.d.). Retrieved October 7, 2022, from https://github.com/JeffSackmann/tennis_MatchChartingProject

Newcomb, Tim. “The History of Tennis Umpiring, Hawk-Eye System.” *Sports Illustrated*, Sports Illustrated, 11 Nov. 2015, https://www.si.com/tennis/2015/11/11/history-of-hawk-eye-tennis-umpiring.