MONEYBALL ANALYSIS: ATP TOUR

Summary

Moneyball is a book, turned into a movie, that tells the true story of how the 2002 Oakland Athletics became the first professional baseball team to successfully use analytics to improve their team. The Oakland Athletics are a small market team, meaning that they do not have the big budget that some of the larger market teams do to spend money on the best players. Also, when they do have a player that becomes a superstar performer, they are not able to retain him for a long period of time because a larger market team can offer him a large amount of money that the Athletics can not match. Therefore, they focused on diving deep into the analytics and statistics of baseball to find players that were undervalued and who they believed could help them win more games while staying within their budget. Being the first organization to rely on analytics to guide their team to success provided a competitive advantage over the other teams in the league. They made the playoffs which was a huge success for the team and in the process, won 20 games in a row which was the longest streak ever at that time. This project is intended to apply the similar concepts to tennis that the Oakland Athletics applied to baseball and collect, analyze, and draw insights from professional tennis data.

The Oakland Athletics approach changed baseball to become a data centered and statistics driven sport. Now it is time for Tennis. Unlike baseball, tennis is not a team sport. The main focus of this project was not to assemble a competitive team, as in baseball, but rather to discover patterns and predictive variables in Tennis which could be used to pinpoint overlooked strengths and weaknesses. The project encompassed the full process from identifying the data, to visualizing and presenting the conclusions of the insights discovered. These insights could then be leveraged as either grounds for making better informed betting decisions by identifying hidden opportunities within a tennis match or for coaches and athletes to discover patterns and overlooked opportunities to improve their own performance. Considering the constraints of time, the project is only scratching the surface of what could be done with more resources, however we were successfully able to pinpoint areas of interest for further analysis, interesting patterns and findings valuable for both players, coaches and betters as well as clear next steps.

Description of the Data

Since the focus of this project was leaning towards identifying outliers and performance analytics, the approach quickly turned towards looking at betting data. When do the odds of the game get the outcome wrong? Why? What factors can we look at to attempt to find patterns within this question? The dataset we ultimately decided on using had data on every game in an ATP tournament dating back to 2008. For each game the dataset included characteristics of the players, such as years in the pro leagues, height, weight, nationality, rank at the time of the game, prefered hand as well as statistics from the game such as betting odds going into the game and the outcome of each set. A complete data dictionary can be found below:

Name	**Definition**	**Data type**
ATP	Tournament number, identifying each tournament	Integer
Location	Venue of tournament	String
Tournament	Name of tournament	String
Date	Date of match	date
Series	Name of ATP tennis series	Integer
Court	Type of court	String
Surface	Type of surface	String
Round	Round of match	Integer
Best of	Maximum number of sets playable in match	Integer
Winner	Match winner	String
Loser	Match loser	String
WRank	ATP Entry ranking of the match winner as of the start of the tournament	Integer
LRank	ATP Entry ranking of the match loser as of the start of the tournament	Integer
WPts	ATP Entry points of the match winner as of the start of the tournament	Integer
LPts	ATP Entry points of the match loser as of the start of the tournament	Integer
W1	Number of games won in 1st set by match winner	Integer
L1	Number of games won in 1st set by match loser	Integer
W2	Number of games won in 2nd set by match winner	Integer

L2	Number of games won in 2nd set by match loser	Integer
W3	Number of games won in 3rd set by match winner	Integer
L3	Number of games won in 3rd set by match loser	Integer
W4	Number of games won in 4th set by match winner	Integer
L4	Number of games won in 4th set by match loser	Integer
W5	Number of games won in 5th set by match winner	Integer
L5	Number of games won in 5th set by match loser	Integer
Wsets	Number of sets won by match winner	Integer
Lsets	Number of sets lost by match loser	Integer
Comment	Comment on the match (Completed, won through retirement of loser, or via Walkover)	String
B365W	Bet365 odds of match winner	decimal
B365L	Bet365 odds of match loser	decimal
PSW	Bet&Win odds of match winner	decimal
PSL	Bet&Win odds of match loser	decimal
MaxW	Maximum odds of match winner	decimal
MaxL	Maximum odds of match loser	decimal
AvgW	Average odds of match winner	decimal
AvgL	Average odds of match loser	decimal
EXW	Expekt odds of match winner	decimal
EXL	Expekt odds of match loser	decimal
LBW	Ladbrokes odds of match winner	decimal
LBL	Ladbrokes odds of match loser	decimal
SJW	Stan James odds of match winner	decimal
SJL	Stan James odds of match loser	decimal
UBW	Unibet odds of match winner	decimal
UBL	Unibet odds of match loser	decimal
pl1_flag	Winners Nationality	String
pl1_year_pro	Winners starting year as a pro	Integer
pl1_weight	Winners weight	Integer
pl1_height	Winners height	Integer
pl1_hand	Winners playing hand	String
pl2_flag	Losers Nationality	String
pl2_year_pro	Losers starting year as a pro	Integer
pl2_weight	Losers weight	Integer

pl2_height	Losers height	Integer
pl2_hand	Losers playing hand	String

Steps taken to alter the data

Our data provided us with the betting odds from a variety of sources like Bet365, Unibet. We filtered down to only the popular sites betting data to make our analysis more accurate and manageable. We also removed some variables like comment, court which were not useful for our analysis. We included weight and height although the data was not accurate so that we could get an overall picture of the player's physique.

Once our data was cleaned and all incomplete observations removed, we further altered the data to only look at a specific subset of tournaments. Our full data set consisted of eight different tournament series which comprise the complete ATP world tour and make up the year long, professional tennis season. However, we decided to only focus on one series called the Grand Slam. The Grand Slam of tennis consists of just four tournaments, however, these are the most popularized and viewed tournaments of the entire world tour. The tournaments that make up the Grand Slam are the Australian Open, French Open, Wimbledon, and US Open. Looking just at these four tournaments we were left with 6,800 matches to run analysis on.

We chose to focus our analysis on these four alone because they have the largest payouts for players so it is most commercially viable for a player to understand how to perform well at these four tournaments specifically. If a player performs well at one of the four tournaments they will not just win a larger amount of prize money, but their name becomes more recognized within the tennis community which could lead to increased likelihood of sponsorship deals. Also, from a betting perspective, these Grand Slam tournaments get the most world wide viewership which leads to larger betting volume and increased opportunity to profit from placing advantageous bets.

Excluding the tournaments that were not Grand Slam had implications for the value of some of our other columns. For example the original data contained a column for court which was either indicating an indoor or outdoor tournament. All Grand Slams are outdoor tournaments which makes this field consistent across all rows. Since we didn't see any true analytical value in this field when all games were outdoor ones we decided to remove the variable. Furthermore we excluded many of the site specific odds of the game and decided to

keep the average odds, max odds, bet 365 and Bet&Win. The objective was not to dive in depth about the difference between different betting sites but rather to create a general understanding of bets relation to player performance which the average odds variable does well.

Creating a primary key

A primary key is a uniquely defined attribute that identifies each row in an entity. It is useful for identifying each row in a table, establishing the relationships among the tables, searching records from a huge set of data, modifying the data to become simpler, and helping reduce the redundancy.

In our dataset, we do not have a unique key defined where we can identify a single row in the entity. Since it is a huge set of data, creating a composite from the existing attributes is also difficult. We have created a column for the winners of the game, giving a serial number to all the winners. The serial numbers given to the winners are unique and named as winnerID.

Since a winner can also be a loser, we cannot consider this as our primary key. In our data, we have observed that in each tournament there are a large number of games played. But each game in a given tournament has only one winner. Hence, the winnerID combined with the given game date gives us the unique ID which identifies each row in the table. We have converted the date column to text using an excel function. We have concatenated the date text column with the winnerID column and created a column UniqueID, which is our primary key in the entity.

The Design of the Database:

For the design of the database, we have followed four stages.

1. Requirement Analysis
2. Logical Design
3. Physical Design
4. Database implementation, monitoring, and modification

Requirement Analysis: We have examined our requirements to determine what type of data is required.For this project and analysis, we require ATP tennis data with specific attributes

that could tell us the potential of a player. This data should be helpful to analyze the different attributes so we can understand if a player could win or lose his next match. With the help of the requirements and observations defined by us, we have selected a dataset.

Logical Design: Before we even create a database, we need to have prior knowledge of what the database should look like, what the entities and attributes should be in the data, if the data is cleaned or not, and if the data contains primary keys. We also need to check whether the data is in normalized form. If at all, we have different tables, how is the relationship established between them? To comprehend all of the preceding, we must create an entity relationship diagram, which provides us with a summary of the logical and graphical representation of our dataset.

ERD

An Entity-relationship diagram, or ER diagram is a basic model or a graphical representation of data stored in a database. It is considered as a blueprint of the database. ER diagrams represent what kind of data is stored, entities, attributes and relationship between the entities which makes it easier to understand the data, maintaining and modifying the data.

Due to the restriction of time and the fact that we will not be updating the data we have not normalized the data before uploading it to our database. If one were to have done so however the normalization process would break up our data into three main tables. A Game table, a player table and a tournament table.

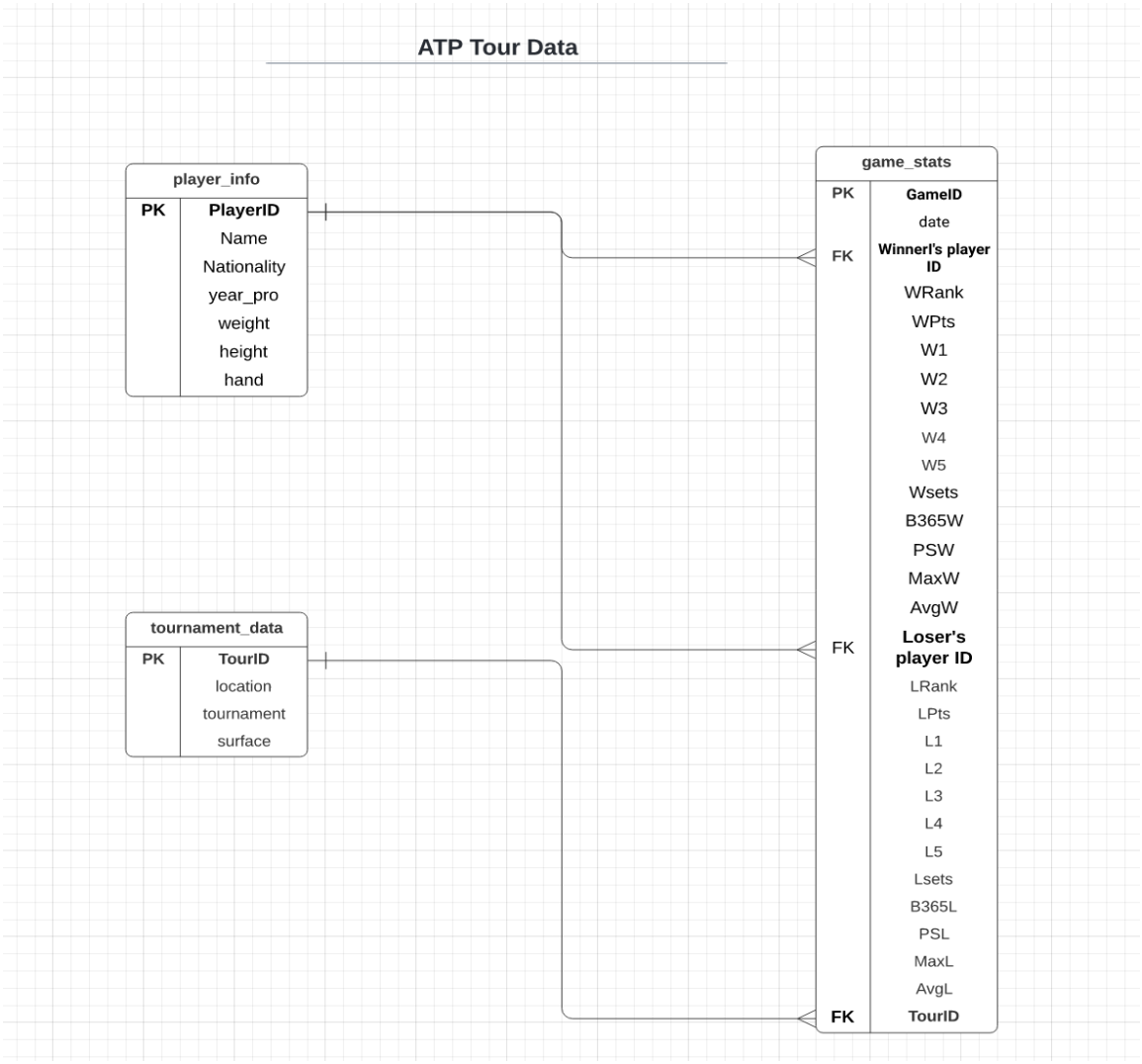
The player table contains all the information about the players. Their name, nationality, the year they went pro, their preferred hand and also their weight and height. At first we did not think the weight and height could go in the player table since it might be changing during the course of a player's career. After examining the dataset we had however we noticed that the height and weight was constant. Since our data structure is for analytical purposes and not updating we decided to store height and weight in the player table. Player ID is the primary key in this table.

The tournament table is storing all the tournament level information. This would be the location of the tournament, the tournament name and the surface it was played on. The tournament table has tournament id as its primary key. Each year of the same tournament has its own tournament id.

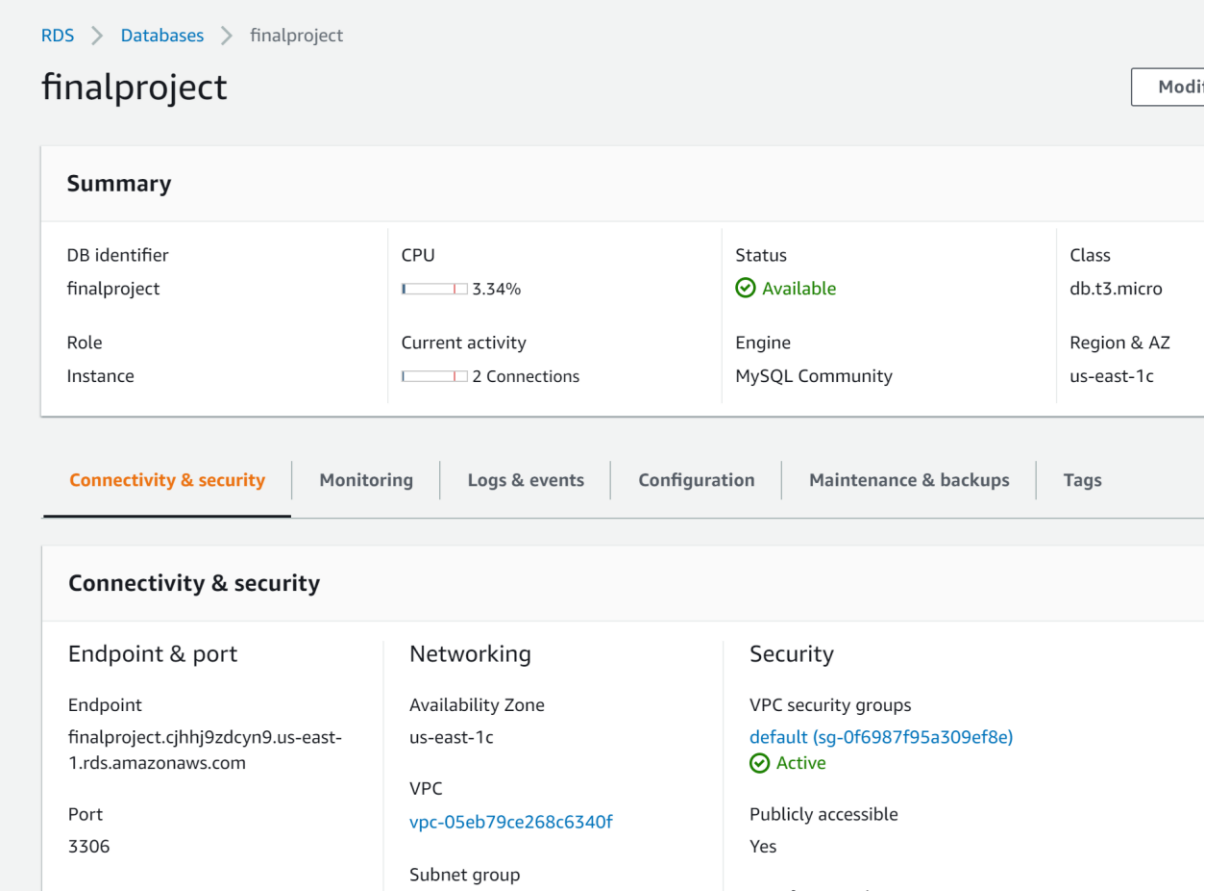
Lastly, the biggest table is the game table. The game table has all the game specific information. Each game ID describes only a single game and gives the winner and loser stats of that particular game. This would be the date of the game, the player id of the winner and

the loser of the game, which are both foreign keys from the player table, the game stats of points scored in each set, the rank of both players at the start of the game and all the game specific betting data on the odds of the winner and loser. The tournament id is also a foreign key in this table to keep track of which tournament the game was part of.

The relationships among all the three tables are established in such a way that each tournament can have a large number of games. Each player can play many games in a particular tournament. A player can be a winner or loser in the game. We can retrieve player information in a particular game using the winner's player ID or loser's player ID in the game stats table by joining the player info table. Because the tournament ID is a foreign key, we can also retrieve information from the tournament data, such as what kind of tournament it was and what kind of surface it was played on for a given GameID.



Physical Design: Now we have a clearer idea of what database is needed, we can physically create and implement databases. We have created a shared database as our final project. The reason for creating a shared database is that all the team members can access the same database and work on it instead of creating different databases and struggling to see what work was done by each individual. The database has been created on AWS.



Database implementation, monitoring, and modification: The created database is connected to MySQL using database credentials and authentication information. We have imported our dataset excel file into MySQL to create tables required for the analysis.

```
4 • USE FinalProject;
5 • SELECT * FROM Tables;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

UniqueID	ATP	Location	Tournament	Date	Surface	Round	Best of	Winner	Loser	WRank	LRank	WPts	LPts	W1	L1	W2	L2
20221171000	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Shapovalov D.	Djere L.	14	51	2593	1156	7	6	6	4
20221171001	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Kwon S.W.	Rune H.	54	99	1085	742	3	6	6	4
20221171002	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Lajovic D.	Fucsovics M.	39	35	1346	1457	6	3	4	6
20221171003	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Vukic A.	Harris L.	144	33	477	1473	4	6	6	3
20221171004	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Garin C.	Bagnis F.	19	72	2375	868	6	3	6	4
20221171005	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Martinez P.	Delbonis F.	61	38	1001	1347	7	6	3	6
20221171006	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Bonzi B.	Gojowczyk P.	63	82	979	802	6	3	6	3
20221171007	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Albot R.	Nishioka Y.	124	119	587	610	6	3	6	4
20221171008	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Alcaraz C.	Tabilo A.	31	135	1609	518	6	2	6	2
20221171009	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Kozlov S.	Vesely J.	169	78	401	824	7	5	6	3
20221171010	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Monfils G.	Coria F.	20	64	2373	976	6	1	6	1
20221171011	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Berrettini M.	Nakashima B.	7	68	4568	917	4	6	6	2
20221171012	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Khachanov K.	Kudla D.	30	105	1748	714	3	6	6	3
20221171013	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Mannarino A.	Duckworth J.	69	49	879	1166	6	4	2	6
20221171014	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Carreno Bust...	Etcheverry T.	21	131	2305	556	6	1	6	2
20221171015	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Bublik A.	Escobedo E.	37	141	1411	497	3	6	7	6
20221171016	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Nadal R.	Giron M.	5	66	4875	920	6	1	6	4
20221171017	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Korda S.	Norrie C.	43	12	1286	2900	6	3	6	0
20221171018	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Hurkacz H.	Gerasimov E.	11	106	3336	714	6	2	7	6
20221171019	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Koepfer D.	Taberner C.	53	108	1096	707	6	1	3	6
20221171020	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Moutet C.	Pouille L.	100	160	737	440	3	6	6	3
20221171021	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Griekspoor T.	Fognini F.	62	32	1001	1494	6	1	6	4
20221171022	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Mcdonald M.	Milojevic N.	55	139	1084	506	5	7	6	4
20221171023	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Kecmanovic M.	Caruso S.	77	146	836	458	6	4	6	2
20221171024	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Sonego L.	Querrey S.	26	110	1860	685	7	5	6	3
20221171025	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Hanfmann Y.	Kokkinakis T.	126	103	583	726	6	2	6	3
20221171026	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Opelka R.	Anderson K.	29	97	1776	744	6	3	6	4
20221171027	5	Melbourne	Australian Open	1/17/2022	Hard	1st Round	5	Millman J.	Lopez F.	89	109	788	696	6	1	6	3

The Potential of our Database:

Our dataset has many player stats data from the past 14 years. Data like betting odds, players world rank etc made our analysis easy. Though some of the data like weight were not accurate, we were able to understand the effect of those variables.

The database connection has been established to MySQL from the AWS database. We have imported our excel file into MySQL to create a table which can be used for the retrieval of different data in the dataset.

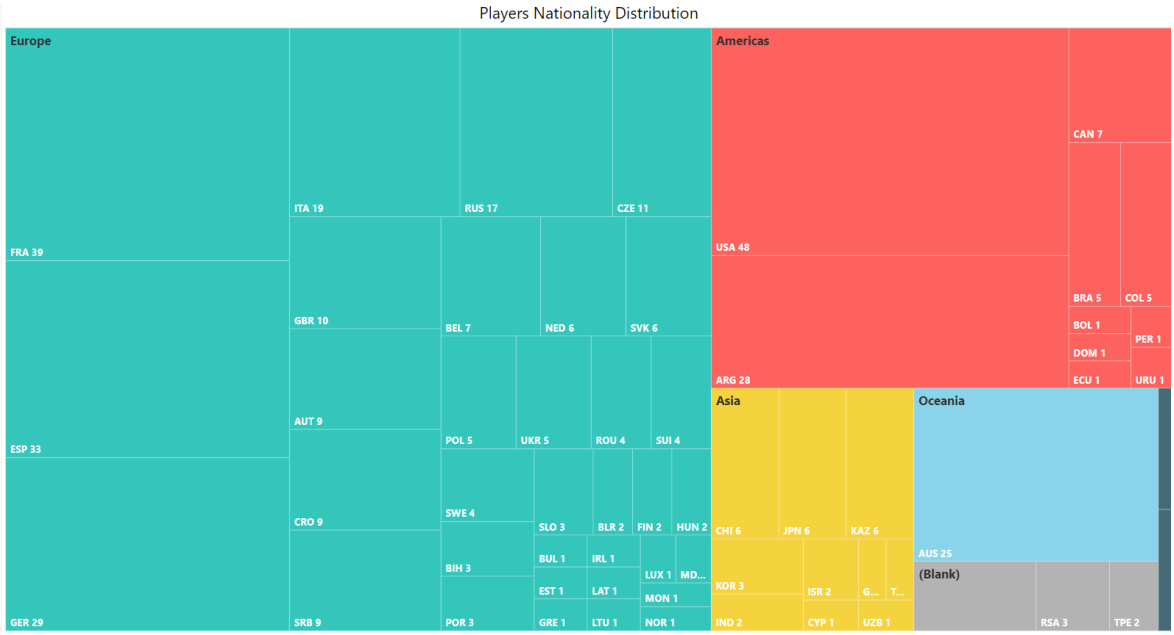
For the analysis, we used PowerBI and Tableau tools. For these analytical tools, the data set was imported from MySQL by establishing the connections in the tools. It is as easy as importing a file to connect to MySQL to retrieve the data for analysis.

This database could be used to from the perspective of both players and people betting on matches. They could use the database to query different past matches, see the results and betting odds, who won and lost, how many sets each player won, and the background information on the players involved in the match. This type of querying could be insightful for players to learn about their own performance or review the performance of a potential opponent. Bettors could also use this database to gather insights that could help them place more informed bets and increase their chances of profiting from those bets.

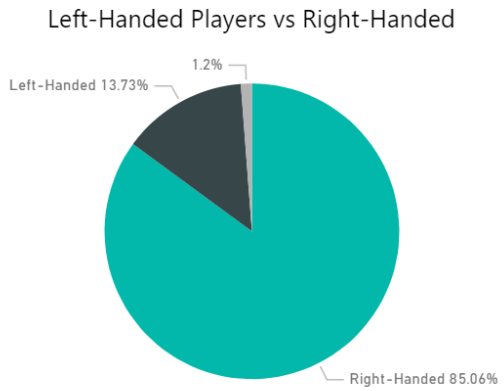
Analytical insights

The data analysis used a very exploratory approach. As most of the group was unfamiliar with Tennis we did not have many hypotheses of what we were expecting to find or which variables we expected to have most importance going into the process. As we were working with the data more and more research questions emerged. When we combined our own findings and analytics from the dataset with external research our understanding of the sport and influential variables deepened.

Initially we looked into some of the characteristics of pro tennis players in the grand slams. Most Tennis players are from Europe and Americas with the biggest national representation from the U.S, followed by France, Spain and Germany.



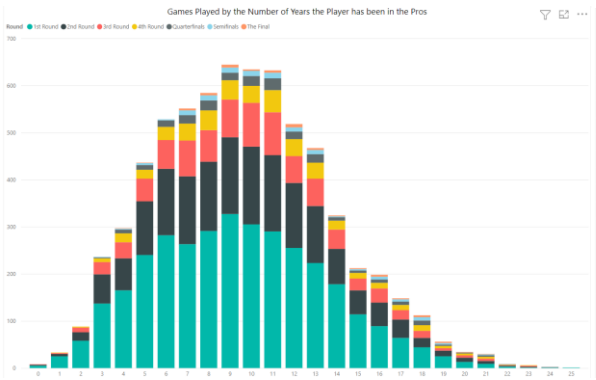
The distribution of right handed players vs left handed is representative of the normal majority of the population.



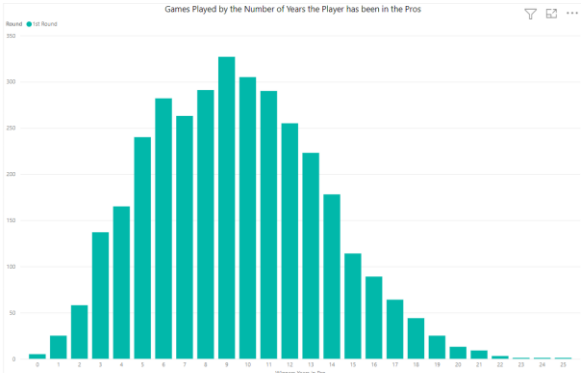
Plotting the distribution of games played by the years that the players have been in the pro leagues shows us that most games are played by players who have been pros for between 8-

12 years. The “years in pro” field was not originally in the dataset, we had to create it. This column is based on the current year of the game subtracted by the year that the player went pro. The curve appears to be following a normal distribution which intuitively is logical. As the sample size gets smaller when we are looking at specific rounds with fewer games, the shape of the distribution does however change. This can both be explained by the simple notion that the sample size is smaller however also from the idea that more factors are contributing to who makes it to the final rounds and that this distribution is not as normal as the one for the first round.

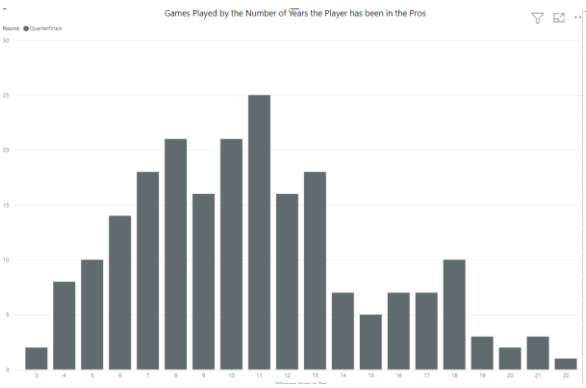
All rounds:



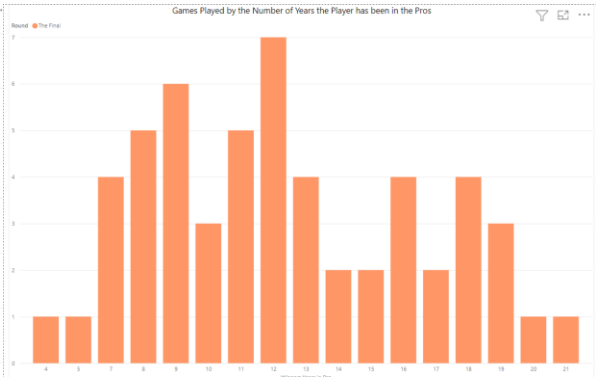
1st round:



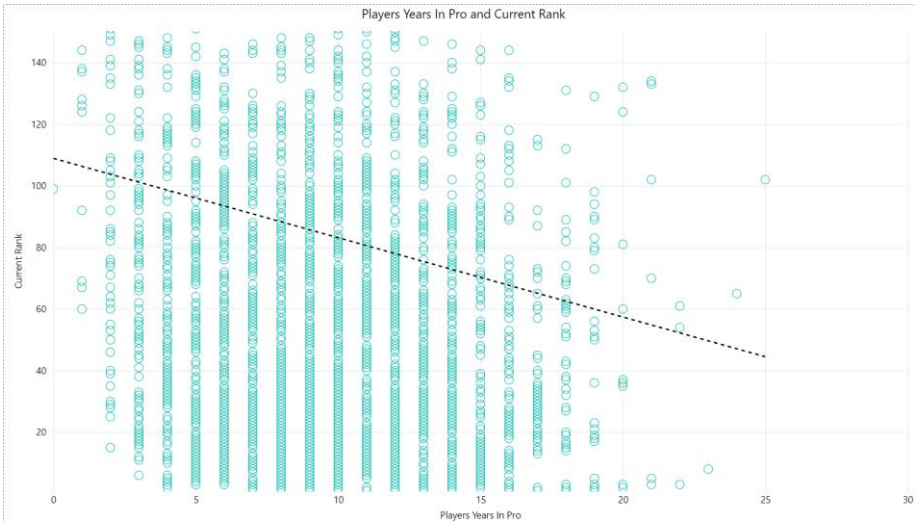
Quarterfinals:



Finals:



Furthermore we explored if the assumption that players' world rank would improve as they played more years in the pro league would hold true.The years in the pros field was used with the world rank to plot all the games. Each dot on the scatterplot is one player, their current world ranking by the time of the game and their total years in the pros at that time. By adding a trendline to this chart we can visualize that the trend is downward sloping indicating that our hypothesis is true. We are however curious to back this assumption up with more statistical analysis to see if the relationship is statistically significant.



Next we looked into the Grand slam final games and title winners. By analyzing winners over the years we can quickly note that three players stand out from the crowd. Nadal, Federed and Djokovic have all together won 45 out of the 55 available titles in the past 14 year. only six other players have been able to win any of the grand slam tournaments.

Grand Slam Title Winners over the years

Year	Cilic M.	Del Potro J.M.	Djokovic N.	Federer R.	Medvedev D.	Murray A.	Nadal R.	Thiem D.	Wawrinka S.	**Total**
2008			1	1			2			**4**
2009		1		2			1			**4**
2010				1			3			**4**
2011			3				1			**4**
2012			1	1		1	1			**4**
2013			1			1	2			**4**
2014	1		1				1		1	**4**
2015			3						1	**4**
2016			2			1			1	**4**
2017				2			2			**4**
2018			2	1			1			**4**
2019			2				1			**3**
2020			1				1	1		**3**
2021			3		1					**4**
2022							1			**1**
Total	**1**	**1**	**20**	**8**	**1**	**3**	**17**	**1**	**3**	**55**

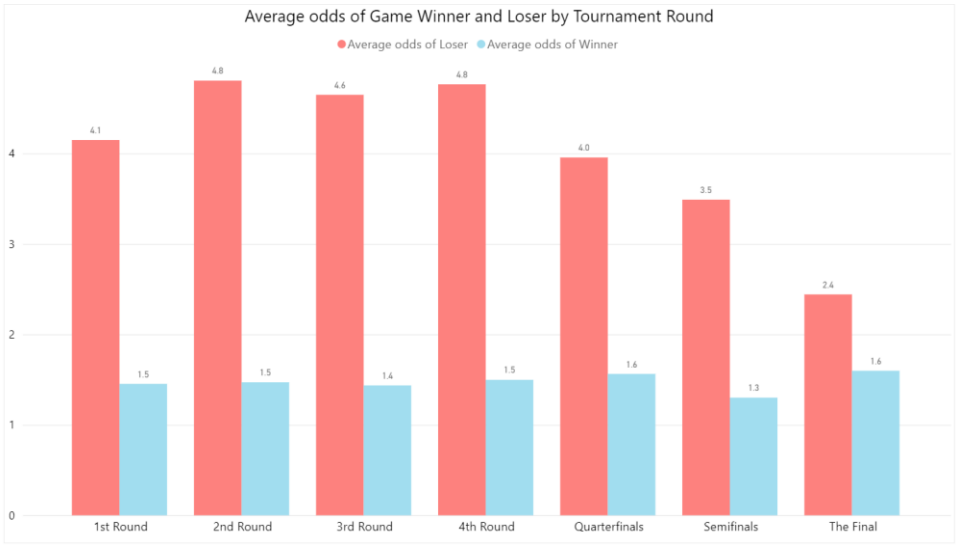
It is however noteworthy that many other players have made it to the final.. When looking at the final round losers there are 20 different players represented. Indicating that it takes something special, likely some element of mental toughness and resilience, to go all the way to winning. The fact that the title keeps going around with the same people even though many more are up playing for it in the final makes us curious as to what that special trait is that can

allow you to win a grand slam.

Grand Slam Final Losers

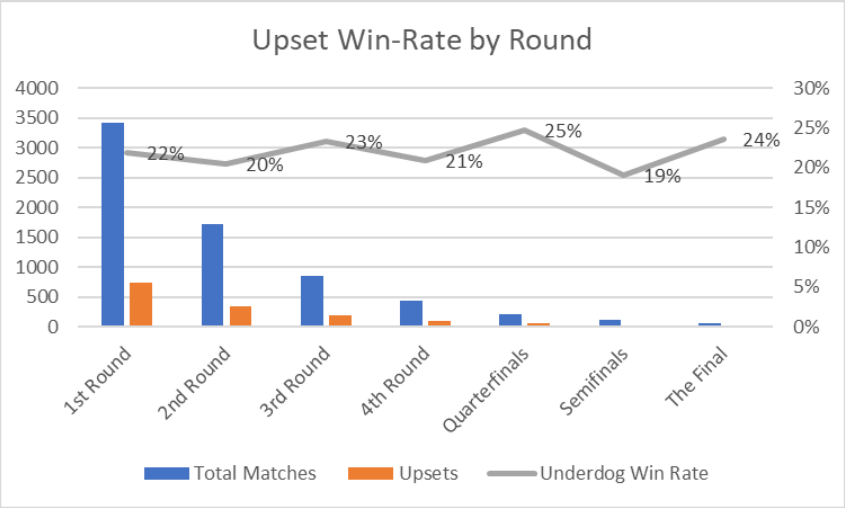
Year	Anderson K.	Berdych T.	Berrettini M.	Cilic M.	Del Potro J.M.	Djokovic N.	Federer R.	Ferrer D.	Medvedev D.	Murray A.	Nadal R.	Nishikori K.	Raonic M.	Roddick A.	Soderling R.	Thiem D.	Tsitsipas S.	Tsonga J.W.	Wawrinka S.	Zverev A.	T
2008							2			1								1			
2009							2							1	1						
2010		1				1				1					1						
2011							1			1	2										
2012						2				1	1										
2013						2		1		1											
2014						1	1				1	1									
2015						1	2			1											
2016						1				2			1								
2017	1			1							1								1		
2018	1			1	1											1					
2021			1			1			1								1				
2019							1				1					1					
2020						1										1					1
2022									1												
Total	**2**	**1**	**1**	**2**	**1**	**10**	**9**	**1**	**2**	**8**	**6**	**1**	**1**	**1**	**2**	**3**	**1**	**1**	**1**	**1**	

After an initial, broad, exploratory look at our analysis, we wanted to dive deeper into the area of our data that we found most interesting, the betting odds associated with the winner and loser of each match. In order to get an idea of when a profitable bet could be placed we looked at the breakdown of betting odds by tournament round. There are seven rounds in each tournament and we felt it would be beneficial to analyze the times in which the non-favored player (or underdog) won the match. The odds of the underdog winning the match are always lower than the favored player, so if someone bets on the underdog and they win, the payout will be higher than if you were to bet on the favorite. Therefore we wanted to pinpoint the best times to bet on the underdog.



The above chart shows the breakdown of the average odds of the winner and average odds of the loser of each month in every round of a Grand Slam tournament. The higher the odds number, the less likely that person is to win the match (according to the odds makers). Therefore, on average, the odds of the loser will always be higher than the odds of the winner because much more often than not, the person who is favored in the match wins. However, we noticed that the difference between average odds of the loser and average odds of the

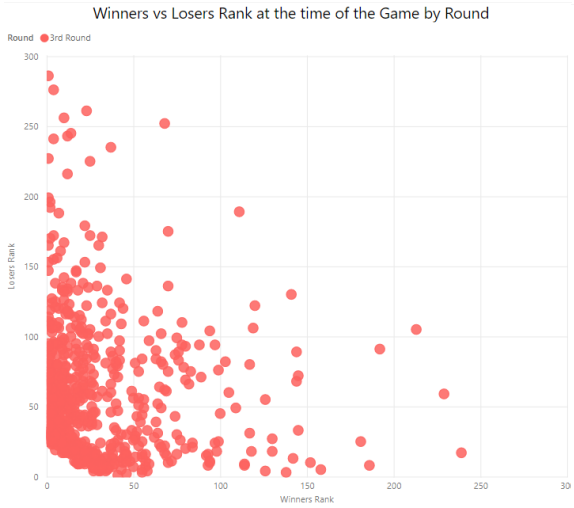
winner gets smaller as we move to the final 3 rounds of a tournament. Also, the average odds of the loser starts to decrease when we move into the final 3 rounds of a tournament. In the Final there is only one match and that match typically has 2 players of very equal ability, therefore it makes sense that the average odds of the loser and average odds of the winner are very close together and this does not provide us with much of an advantageous betting angle. This same logic can be applied to the Semifinal and in fact, although the average odds of the loser is decreasing in this round, the average odds of the winner is also increasing so this is also not an advantageous time to bet on the underdog because this suggests the chart suggests that the Semifinal is the round in which the favored played wins the most. However, the last of these rounds that stood out was the quarterfinals and after a bit of further analysis (Upset Win-Rate by Round) we did find that this is the best round to bet on the underdog player.

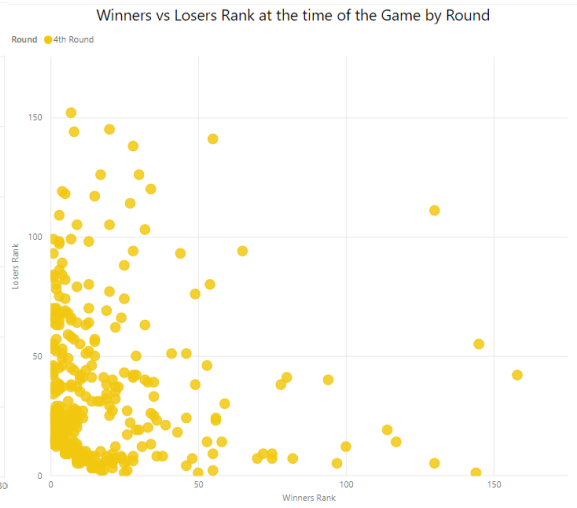


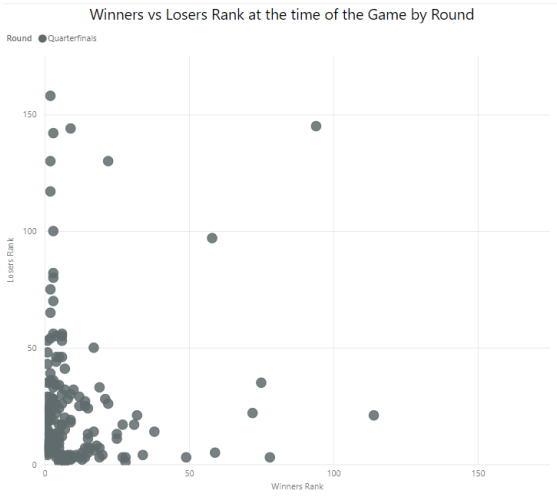
The chart above shows the total number of matches per round, total number of upsets per round (each match where the non-favored player wins), and the percentage of upsets by round. Consistent with our assumption from the previous analysis on upsets per round, the quarterfinals proved to be the best round to bet on a non-favored player to win, with an Upset Win-Rate of 25%. There are also just 4 matches in the QuarterFinals so if our analysis holds true, one match in every Grand Slam Quarterfinal will be won by an underdog. Choosing the correct underdog and placing a bet on them could lead to a much more profitable outcome then if one were to just bet on the favored players, and this analysis can certainly help narrow the search for the correct underdog.

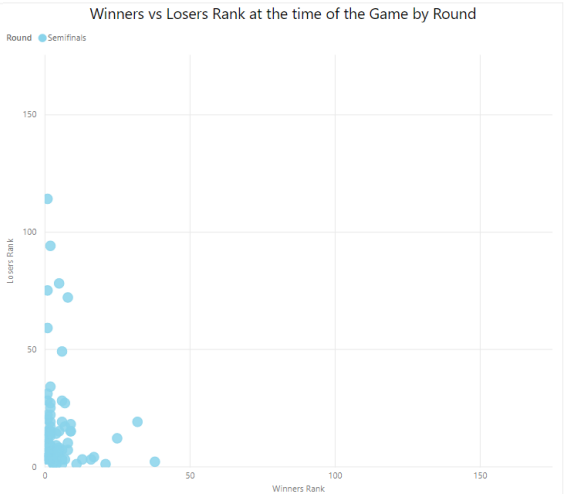
To visualize some of the more interesting games and trends we also plotted the current rank of the winners against the current rank of the losers to easier identify in which

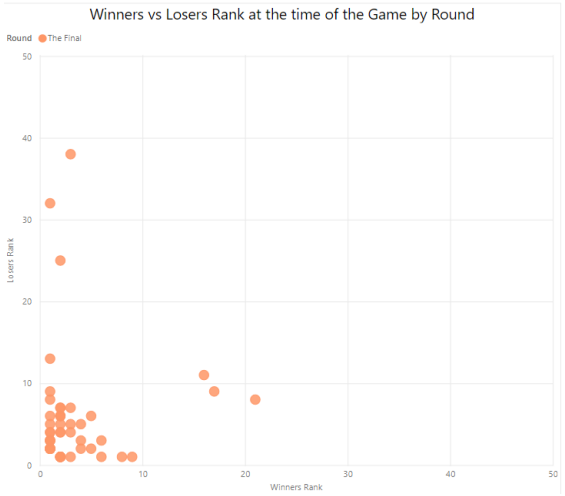
games the non-favored player had ended up being successful. This was done using a scatter plot. When diving deeper into the rounds of the tournament we could notice that the semifinals were usually won by the better ranked player however the other rounds did not show an overwhelming trend of advantage for the better ranked player. When combining this finding with our learnings from the previous chart we concluded that the quarterfinal is likely the most profitable roundn to be betting on and the round where you are most likely to see an underdog beat a better ranked player.







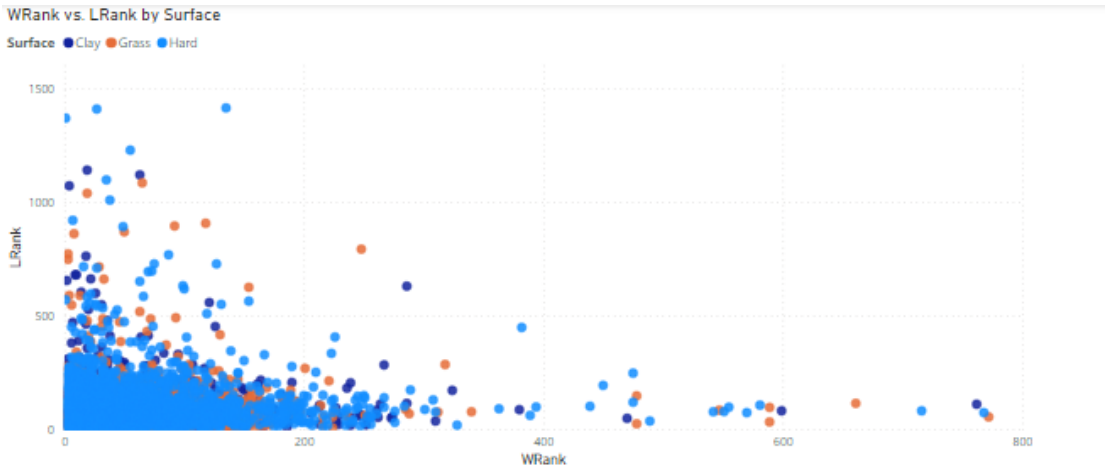




We paired these charts with an interactive table which would pull up the specific information of the game once you clicked on one of the instances.

Potential future explorations

In tennis, player strategy plays a major role than the physique of the player. Our analysis can further provide a player with the data in which rounds he has underperformed. The player can try to improve on those surfaces to level up his game. Players also can change their strategy based on the surface they are playing to increase their probability of winning the game.



The chart above shows a scatterplot of all the matches played in our dataset, the rank of the winner vs. rank of the loser, and segmented by color based on the surface that the match was played on. While analyzing the data throughout this project, we picked out the surface that the court is played on as a variable that we could focus in on and potentially draw insights from that could provide the type of advantage as the moneyball strategy used by the Oakland Athletics. We created this scatterplot to try and see which surfaces reveal matches where a lower ranking player beats a higher ranked player and if we could gather further insights from these outlier results. The four tournaments that make up the Grand Slam series are played on three different surfaces, and each surface is very different from each other so a player's strategy should change based on the surface they are playing on. The different surfaces cause balls to bounce differently, players movement looks and feels different, and fatigue can set in at different points during the match. We felt that a player that understands and utilizes this knowledge correctly can create a competitive advantage for themselves, similar to the advantages that the Oakland Athletics targeted.

The Australian Open and US Open are both played on hard surfaces. Hard courts are the toughest on the body and can cause accelerated fatigue. If we were to advise a player who is preparing for a match on hard court to potentially give them an advantage it would be a good strategy to conserve energy in the initial sets so your opponent is more tired at the end.

Of course it is hard to completely control the tempo of the match, so a more tangible strategy could be to focus more on increasing stamina in training prior to these tournaments rather than just strictly training their tennis skills.

The French Open is played on a clay surface which is considered the slowest surface because the ball bounces much higher on this surface compared to the others which causes it to lose a lot of its initial speed. Due to the fact that this surface plays slow, it is difficult for players to deliver an unreturnable shot and rallies tend to last longer on clay courts. This provides an advantage for players who don't rely on their power to win points, but instead rely on their technique. Our advice for a player training for the French Open would be to spend extra training time specifically on their technique as a player. If they can play shots with more spin and placement, as opposed to strictly power, it will be more beneficial to them on the clay court at the French Open to win matches.

Lastly, Wimbledon is played on a grass court which is the fastest surface a match can be played on. Grass courts provide an advantage to players that play aggressively, attack points early in a rally and play with more pace. Our advice for a player going into a match at Wimbledon would be to not play methodically or defensively because this could lead to an early defeat if the opponent is playing more offensively. The strategy on a grass court should be to play fast and try to win points early. Grass courts also favor players who have a powerful serve because it makes it much harder to return on such a fast surface. This point led us into thinking about some further analysis we could run to try and find matches where a lower ranked player could upset a higher ranked player in a grass court match.

From the scatter plot above we would only highlight the matches played on grass courts and add an additional variable which would be based on the average speed of each player's first serve. For simplicity sake we would make this a binary variable where the player is given a 1 if their serve is faster than 75% of all players in the tournament, and a 0 if it is not. We would then cross reference this new variable with the winner rank versus loser rank chart of all the matches on grass courts. We would then pinpoint the matches in which the winner of the match was a lower rank than the loser and see if that person qualified as having a fast serve and their opponent did not. We would hope to find evidence to support this theory that someone with a harder serve than their opponent will, more often than not, beat that opponent on a grass court. Then, if we are previewing a Wimbledon match where the lower ranked player has a harder serve than the higher ranked player, we would know that this would be a good time to bet on the lower ranked player to win.

Reflection on the project

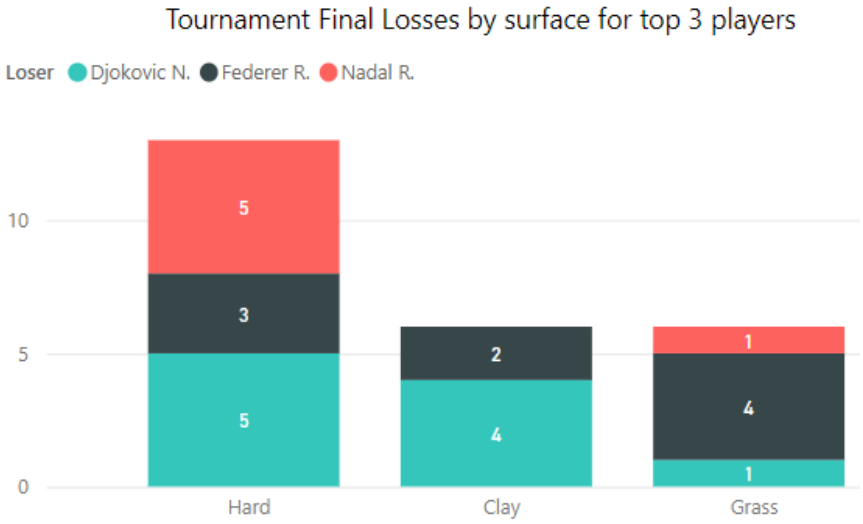
This project challenged us to push the boundaries in which we work with data. In past projects it may have been easier to overlook inconsistencies in the data or errors in our analysis, especially given our limited experience with this type of analysis. However, this project was a real world application, so it was difficult to apply skills we learned in class to practice making and correcting mistakes. At the end of this project we realized we had greatly increased our data literacy and were more likely to spot errors in our analysis. We learned to think about data in terms of how it was collected in order to conduct analysis without misinterpreting variables or assuming inaccurate results.

We found that the material we covered in the course helped us recognize errors we would not have known to look for before. For example, the data showed constant values for each individual's weight over the years, when in reality this is unlikely as players' weights tend to fluctuate, especially over a period of 15 years as was covered by this dataset. After taking this course we know to recognize this as an update error in the dataset. We recognize how this type of error is likely to occur, and can understand that such errors would have been less likely in a highly normalized dataset. In addition to recognizing pre-existing errors in the data, we made our fair share of mistakes in analysis, and were able to effectively peer review and help our group members spot inconsistencies. For example, one of us constructed a graph of player wins by court location (indoor or outdoor). The graph made it look like players are more likely to win when playing on an outdoor court, when really this result was skewed as there were far more outdoor games than indoor. Without peer review we may have included this and more erroneous observations in our analysis. Over the course of the project we learned how to work with each other, specifically how to review each other's work and understand each other's thought processes.

Our group members all came from varying data science backgrounds, and as a result we were able to teach each other different skills and software. For example, Justine had experience with Tableau but had not seen PowerBI before; she was able to become familiar with PowerBI from Lisa. We also learned what each other's strengths were, and where our teammates' knowledge could fill the gaps in our own understanding. Most of the group was not too familiar with the structure of tennis, but we were able to get insight and explanations from Ben. Looking back, our varying backgrounds proved to be a point of strength for our group's performance rather than a weakness.

One thing we found interesting about the project was the effect of the court type (clay, grass, or hard) on a player's performance. When thinking about factors that affect a match,

we initially thought the effect of court surface would be negligible. However, it was interesting to find that harder surfaces can be more taxing on a player's joints, and other surfaces may contribute to cardio fatigue during a match. Finding out how different players perform on different surfaces (pictured below) was interesting to all of us, regardless of our tennis background.



If we could continue this project in the future, we would likely create some predictive models to further analyze the tennis matches. It might be useful to have a model that could predict how well a top player (Nadal, for example) would perform on a day given parameters such as court type, temperature, humidity, etc. We would also like to broaden our analysis on player performance given different surfaces. As mentioned before, this was one of the more intriguing parts of our project, and continuing the study to analyze the nuances of player performance would be very interesting. This type of study could have practical applications as well; players could alter their training in preparation for a match on a certain surface.

Overall, the group exercised effective time management and communication skills. The guidelines provided by Jing were helpful, as having the data normalized and loaded into the database by the recommended date helped us to stay on track. The real world application aspect of the project helped us learn and apply our skills from class. We enjoyed conducting our analysis and are glad to leave the course with applicable skills.

Source for data set:

Edoardo Cantagallo. (2022, March). Atp Tennis Data with betting odds, Version 5. Retrieved september 10, 2022 from https://www.kaggle.com/datasets/edoardoba/atp-tennis-data