Final Project Report

University of New Hampshire

Shawn Bedard, Srija Gandhesree, Divyalata Godavarti,
Akshitha Ponnapalli, Sharmishta Tallapally, & Hannah Wirth

DS 801: Business Intelligence

Jing Wang

October 3, 2022

Executive Summary

The purpose of the final project is to investigate data about the ATP Tennis World Tours. The investigation into the analysis of tennis data comes from the interest surrounding the article "Why Tennis Is Still Not Ready to Play Moneyball". This final project aims to look more into what types of data about tennis are available, what kind of analyses can be done on the available data, and what types of insights can be generated from the available data. Identifying these three aspects of tennis data will contribute to the sparse landscape of tennis analytics. It also will help identify what significant gaps there are. The main findings of this project were that investigation into other datasets to either replace or supplement the existing dataset chosen is necessary. Additionally, the analyses conducted on the data in both SQL and external software revealed that while basic statistics and information can be procured from SQL analysis, more advanced analysis with Tableau or Python is more useful. The secondary findings of this project were the results of the data analysis performed on the dataset. It was found that left-handed players do not have an advantage over right-handed players, Spain, France, and the USA produce the highest percentage of winners, and age does not play a significant role in determining a winner, and surface type does appear to play a role in how high a winner is ranked.

Data Description

The dataset that was chosen for the final project is the "tennis_atp" dataset created by Jeff Sackmann on GitHub. The dataset is a master set containing data on players, historical rankings, results, and match stats. The creator continually updates the data with the last update being 10 days ago. The ATP match data files contained a variety of fields including tournament name, surface type, draw size, tournament date, winner and loser information, match length, number of serve points, number of breakpoints, etc. The primary benefits of using Jeff Sackmann’s dataset are the completeness of the data, the wide range of attributes, and the large amount of data available.

“tennis_atp” dataset attributes		
tourney_id	loser_entry	w_bpSaved
tourney_name	loser_name	w_bpFaced
surface	loser_hand	l_ace
draw_size	loser_ht	l_df
tourney_level	loser_ioc	l_svpt
tourney_date	loser_age	l_1stIn
match_num	score	l_1stWon
winner_id	best_of	l_2ndWon
winner_seed	round	l_svGms
winner_entry	minutes	l_bpSaved
winner_name	w_ace	l_bpFaced
winner_hand	w_df	winner_rank
winner_ht	w_svpt	winner_rank_points
winner_ioc	w_1stIn	loser_rank
winner_age	w_1stWon	loser_rank_points
loser_id	w_2ndWon	
loser_seed	w_SvGms	

The data that was ultimately migrated to the database for analysis is the ATP match data from 2010 to 2019. Although the master set has data from the early 1900s to the present, the decision to use match data only from 2010 to 2019 was made to keep the dataset at a reasonable size and to exclude missing data from 1973 to 1984.

Although the chosen dataset contains a large range of data to perform analysis on, there is more data on the ATP World Tour that is not included in the dataset but is still valuable to analyze. For example, data on a player's years of experience or the speed of a player's shot are not found in the dataset. A player's years of experience compared to their age or the speed of a player's shot compared to the hand they use (right or left) would be useful in determining what factors make a player better or worse on the court. Additionally, any data on betting agencies or outcomes is not included in the dataset. The "ATP Men's Tour" dataset on Kaggle contains data on betting agencies to determine a betting strategy predicting the outcome of a match. This type of data would be useful to predict match outcomes for betting purposes or analysis of factors that influence the outcome of a match.

There are a variety of analytical insights that can be generated from the fields within the chosen dataset. Some simple insights include determining whether left-handed or right-handed players have an advantage or tendency to win more matches, do certain players play better or worse on different types of surfaces (hard, clay, grass, etc.), and whether a player's ranking helps determine whether they win a match. Most of these analytical insights must be generated by importing the data into another software such as R, Python, or Tableau. However, basic data insights and simple analyses can be obtained through querying the database.

The analytical questions that this final project is interested in answering are the following:

1. Do left-handed players have the advantage over the right-handed players?
2. Which countries produce the best tennis players?

3. Does age play a significant role in determining a winner or loser?
4. Do winners play better or worse depending on the surface type of the court?

Database Design

To be able to implement the dataset into a functional database, the data from 2010 to 2019 was first combined into a single CSV file. The final dataset has 29,397 rows with 49 different fields total.

Next, the data was normalized into tables. The first table, “tournaments”, contains data on ATP tournaments from 2010 to 2019. There are 1,412 total tournaments identified by the “tourney_id” field.

tournaments	
tournament_id	Primary key for “tournaments” table
tourney_id	Identifier for each tournament
tourney_name	Tournament name
surface	Surface type of court
draw_size	Number of players in the draw
tourney_level	Type of tournament
tourney_date	Tournament date

The second table, “matches”, contains data on the individual matches within a tournament. For example, tournament 2010-339 in Brisbane 2010 had 31 different matches. The “winner_id” and “loser_id” fields are foreign keys to the “players” and “playerDetails” tables for information about the winners and losers of a match.

matches	
match_id	Primary key for “matches” table
tourney_id	Identifier for each tournament, foreign key
match_num	Match number within a tournament
score	Match final score
best_of	Number of rounds in a match
round	Round level
minutes	Match length in minutes
w_ace	Winner’s number of aces
w_df	Winner’s number of doubles faults
w_svpt	Winner’s number of serve points
w_1stIn	Winner’s number of first serves made
w_1stWon	Winner’s number of first-serve points won
w_2ndWon	Winner’s number of second-serve points won
w_SvGms	Winner’s number of serve games
w_bpSaved	Winner’s number of breakpoints saved
w_bpFaced	Winner’s number of breakpoints faced
l_ace	Loser’s number of aces
l_df	Loser’s number of doubles faults
l_svpt	Loser’s number of serve points
l_1stIn	Loser’s number of first serves made
l_1stWon	Loser’s number of first-serve points won
l_2ndWon	Loser’s number of second-serve points won
l_SvGms	Loser’s number of serve games
l_bpSaved	Loser’s number of breakpoints saved

l_bpFaced	Loser’s number of breakpoints faced
winner_rank_points	Winner’s number of ranking points
loser_rank_points	Loser’s number of ranking points

The third table, “players”, contains data about all of the players that have participated in an ATP tournament between 2010 and 2019. There are 1,272 unique players within the table.

players	
player_id	Unique player ID, primary key for “players” table
player_name	Name of player
player_ht	Height of player, in centimeters
player_hand	Which hand the tennis racket is held by player
player_ioc	Player’s three-character country code

The fourth table, “winners”, contains information about a match’s winner. It was necessary to normalize the winner’s information into two tables, “players” and “winners to minimize duplicate rows. Additionally, one winner can have multiple ages, seed, and rank based on the year of the tournament because a player can participate in multiple tournaments.

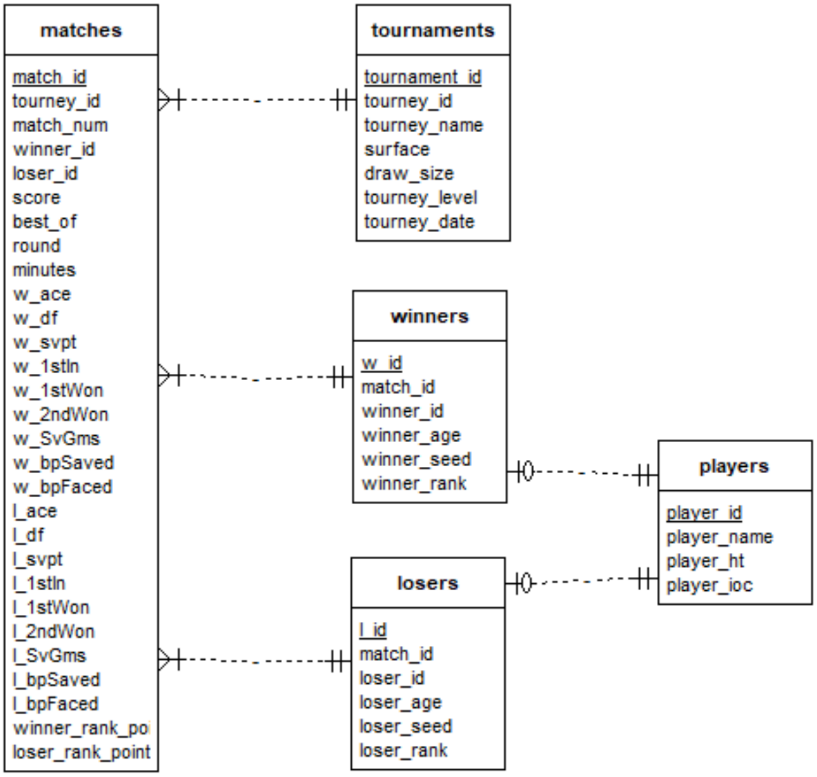
winners	
w_id	Winner identifier, primary key for “winners” table
match_id	Match identifier
winner_id	Player ID for the winner of the match
winner_age	Winner’s age in years as of the tournament date
winner_seed	Winning player’s ranking position

winner_rank	Winner’s ATP rank, as of the tournament date, or the most recent ranking date before the tournament date

The fifth table, “losers”, contains information about a match’s loser. Similar to the “winners” table, it was also necessary to normalize the winner’s information into two tables.

losers	
l_id	Winner identifier, primary key for “winners” table
match_id	Match identifier
loser_id	Player ID for the winner of the match
loser_age	Loser’s age in years as of the tournament date
loser_seed	Losing player’s ranking position
loser_rank	Loser’s ATP rank, as of the tournament date, or the most recent ranking date before the tournament date

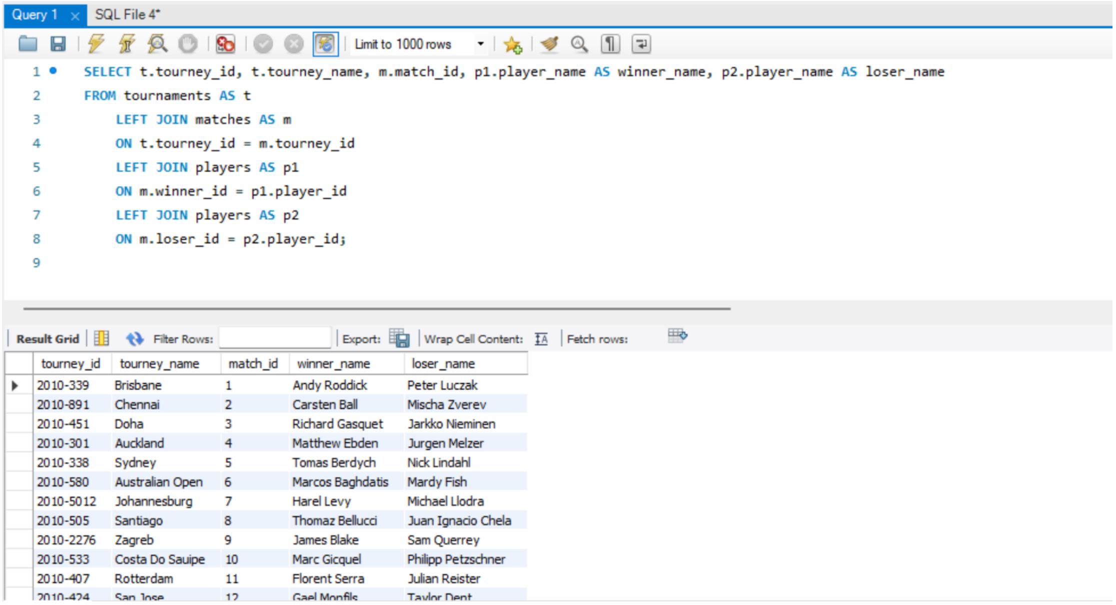
The relationships between the four tables can be visualized in the ERD diagram below. One tournament can have multiple matches take place. For instance, one tournament can have as few as 9 matches or as many as 30 matches. However, one unique match can only be held during one tournament. A match must have exactly one winner, but a winner can participate and win multiple matches. A match must also have exactly one loser, but a loser can participate and lose in multiple matches. A winner must be one player, but a player listed in the “players” table may not have played in a match, therefore not having ever won a match. Similarly, a loser must be one player, but a player listed in the “players” table may not have ever lost a match.



Database Potential

How the data is structured and how it was implemented into the database allows for easy querying to retrieve data and to run basic analytics on the data. Because the data was implemented into a MySQL database, the data tables can be joined on each other to determine the details of tournaments, matches, and players. Below are some example data queries and analyses that can be achieved through joins and filtering. A summary of the query and results can be found below.

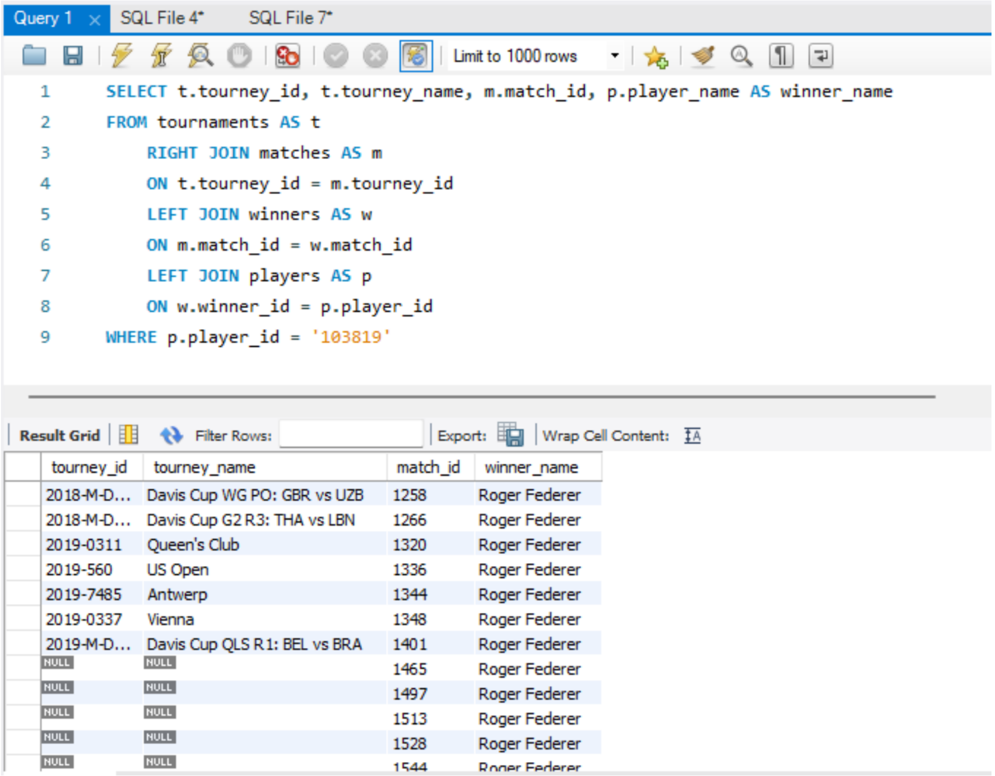
1. Find the names of the winners and losers of each match



```sql
SELECT t.tourney_id, t.tourney_name, m.match_id, p1.player_name AS winner_name, p2.player_name AS loser_name
FROM tournaments AS t
    LEFT JOIN matches AS m
    ON t.tourney_id = m.tourney_id
    LEFT JOIN players AS p1
    ON m.winner_id = p1.player_id
    LEFT JOIN players AS p2
    ON m.loser_id = p2.player_id;
```

tourney_id	tourney_name	match_id	winner_name	loser_name
2010-339	Brisbane	1	Andy Roddick	Peter Luczak
2010-891	Chennai	2	Carsten Ball	Mischa Zverev
2010-451	Doha	3	Richard Gasquet	Jarkko Nieminen
2010-301	Auckland	4	Matthew Ebden	Jurgen Melzer
2010-338	Sydney	5	Tomas Berdych	Nick Lindahl
2010-580	Australian Open	6	Marcos Baghdatis	Mardy Fish
2010-5012	Johannesburg	7	Harel Levy	Michael Llodra
2010-505	Santiago	8	Thomaz Bellucci	Juan Ignacio Chela
2010-2276	Zagreb	9	James Blake	Sam Querrey
2010-533	Costa Do Sauipe	10	Marc Gicquel	Philipp Petzschner
2010-407	Rotterdam	11	Florent Serra	Julian Reister
2010-424	San Jose	12	Gael Monfils	Taylor Dent

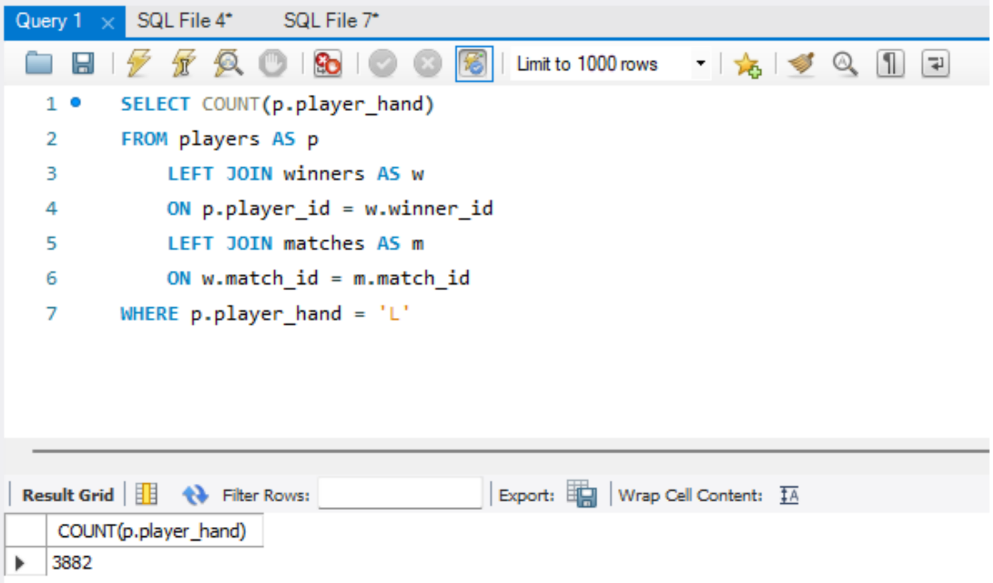
2. Find all of the matches won by Roger Federer (player ID 103819)



```sql
SELECT t.tourney_id, t.tourney_name, m.match_id, p.player_name AS winner_name
FROM tournaments AS t
    RIGHT JOIN matches AS m
    ON t.tourney_id = m.tourney_id
    LEFT JOIN winners AS w
    ON m.match_id = w.match_id
    LEFT JOIN players AS p
    ON w.winner_id = p.player_id
WHERE p.player_id = '103819'
```

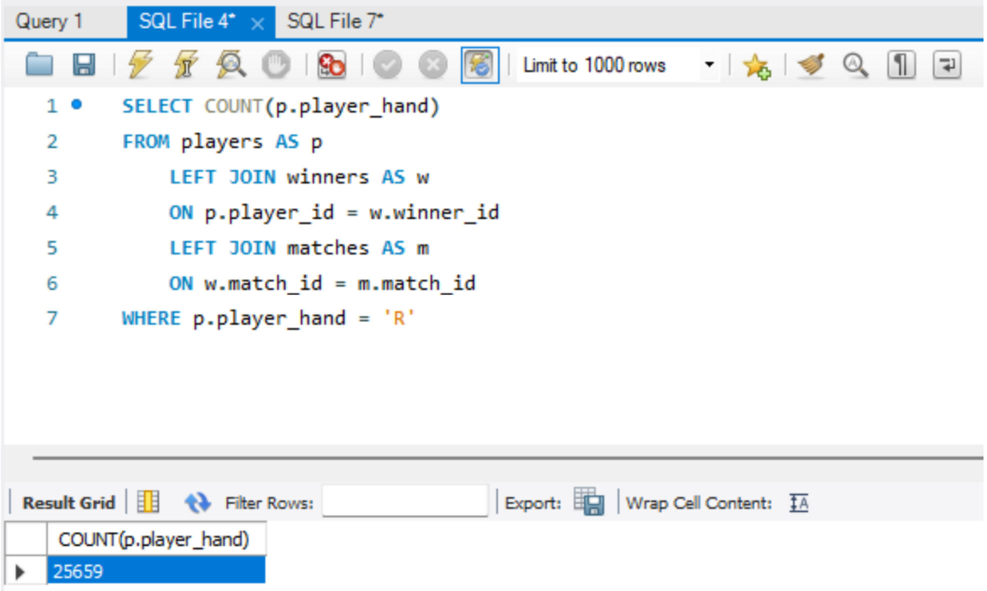
tourney_id	tourney_name	match_id	winner_name
2018-M-D...	Davis Cup WG PO: GBR vs UZB	1258	Roger Federer
2018-M-D...	Davis Cup G2 R3: THA vs LBN	1266	Roger Federer
2019-0311	Queen's Club	1320	Roger Federer
2019-560	US Open	1336	Roger Federer
2019-7485	Antwerp	1344	Roger Federer
2019-0337	Vienna	1348	Roger Federer
2019-M-D...	Davis Cup QLS R1: BEL vs BRA	1401	Roger Federer
NULL	NULL	1465	Roger Federer
NULL	NULL	1497	Roger Federer
NULL	NULL	1513	Roger Federer
NULL	NULL	1528	Roger Federer
NULL	NULL	1544	Roger Federer

3. Find the count of winners that were left-handed versus the count of winners that were right-handed



```sql
SELECT COUNT(p.player_hand)
FROM players AS p
    LEFT JOIN winners AS w
    ON p.player_id = w.winner_id
    LEFT JOIN matches AS m
    ON w.match_id = m.match_id
WHERE p.player_hand = 'L'
```

COUNT(p.player_hand)
3882



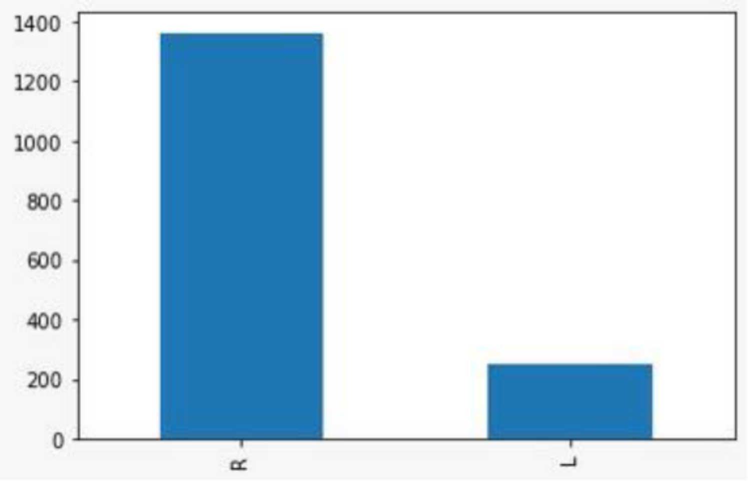
```sql
SELECT COUNT(p.player_hand)
FROM players AS p
    LEFT JOIN winners AS w
    ON p.player_id = w.winner_id
    LEFT JOIN matches AS m
    ON w.match_id = m.match_id
WHERE p.player_hand = 'R'
```

COUNT(p.player_hand)
25659

The database also allowed us to export the data into other software such as Tableau and Python to investigate more advanced analyses.

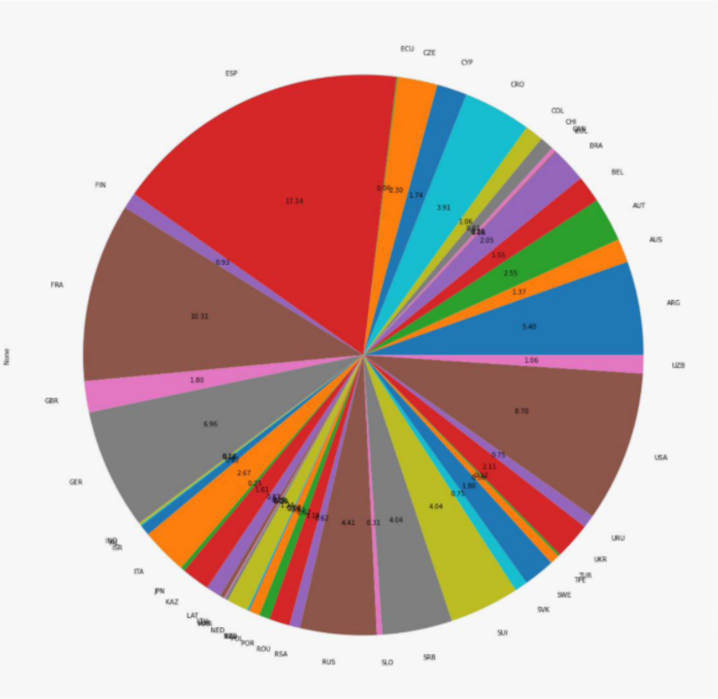
1. Do left-handed players have the advantage over the right-handed players?

Results: Left-handed players don't have an advantage over right-handed players as more right-handed players have won more matches.



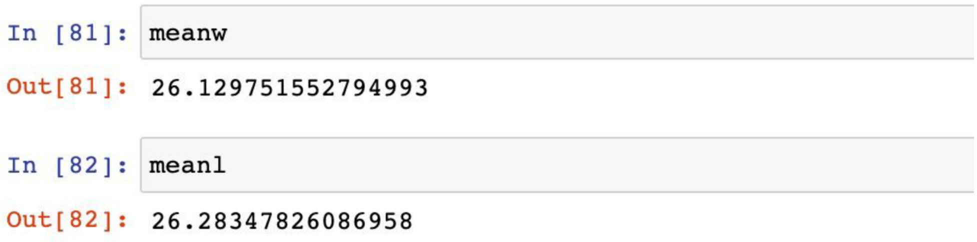
2. Which countries produce the best tennis players?

Results: From the created pie chart, it was found that Spain, USA, and France have produced the highest percentage of winning players.



3. Does age play a significant role in determining a winner or loser?

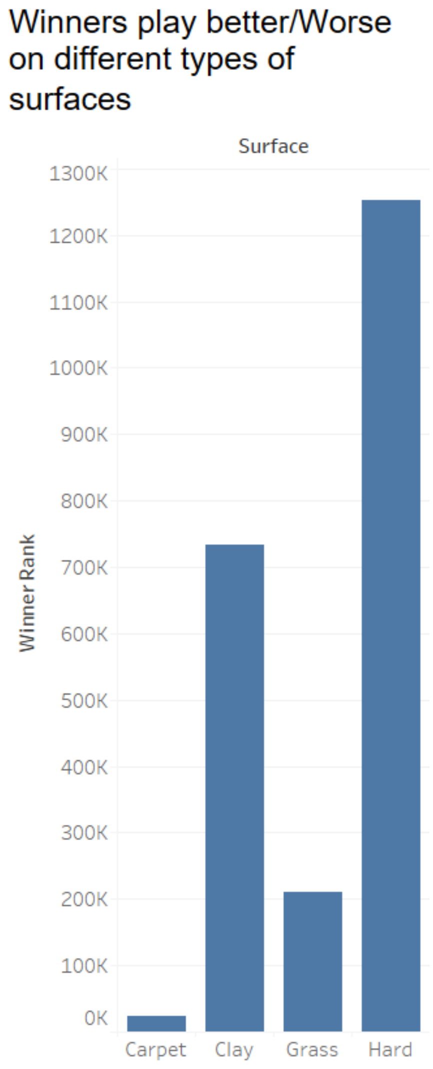
Results: Age does not play a significant role in determining a winner or loser given that the mean age of both winners and losers is almost equal.



```
In [81]: meanw
Out[81]: 26.129751552794993

In [82]: meanl
Out[82]: 26.28347826086958
```

4. Do winners play better or worse depending on the surface type of the court?



Winners play better/Worse on different types of surfaces

Sum of Winner Rank for each Surface.

Results: Winners had a higher rank when they played on a harder surface and the lowest when played on carpet.

Project Reflection

After choosing a dataset and completing our analyses, our group found that the most interesting thing we learned throughout the process was the ability to quickly move and manipulate the large amount of data that we sampled from an even larger dataset. The amount of information that was provided by our dataset left our group not knowing where to focus our analysis. Throughout the semester as our SQL knowledge increased our group's confidence increased, we became more interested in manipulating the database to answer our questions.

The primary tools that our team utilized for this project were an AWS relational database that allowed our team to work simultaneously on our dataset to allow multiple queries to be run from MySQL Workbench. This helped improve our team's overall productivity. Our group has also taken advantage of Tableau, a data visualization tool that enables us to better understand the thousands of rows of information into clear, consistent visualizations. The analysis performed within Tableau greatly helped understand how the player's match statistics influenced winning on an individual level and on the game of tennis in general.

This project could be improved in the future with the ability of utilizing more advanced analytical coding languages such as R or Python since they are more designed for data analysis and have the ability to answer more questions that SQL or Tableau could not answer. These other programming languages could allow us to create predictive models and programs that could provide more advanced insights to the dataset. The data analysis that we performed is only the surface level and can still be greatly improved upon. The large amount and diversity of match data could allow a complete dive into the inner workings of a tennis match and provide much more insight.

Throughout this project, our team encountered a variety of challenges that began with the normalization process of the dataset. The dataset contained such a large amount of data spread across different categories such as tournament data, match data, and player information. The dataset also included information such as age or player seed that constantly changed for the individual players from one tournament to another. Another difficulty that our group has faced for the entire team was to connect to the AWS database. The data import to the database was fairly simple after the normalization process but having it be accessible to the entire group was a challenge that we had to troubleshoot. After finally being able to access the dataset properly for SQL analysis, our final struggle was to narrow down what questions should be focused on for our visualizations. The vast amount of information allowed us to go as in depth as we possibly could but the group had to refrain in order to make the deadline of the assignment.