Databases Final Project: Final Report

Project Description:

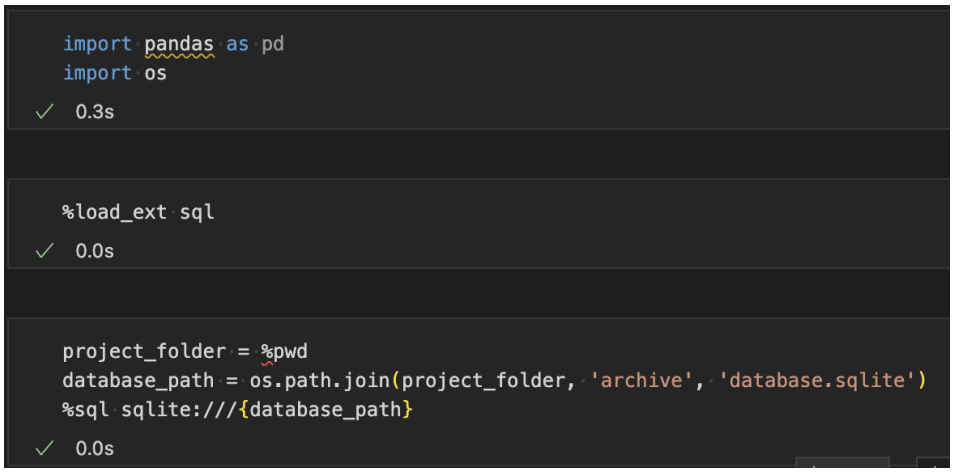
Our project was done in a Jupyter notebook. We imported data from kaggle and accessed it locally. Our project is an in depth analysis of a large dataset on tennis statistics at the ATP professional level. There are some user interactive portions in the notebook for our queries.

Changes from Phase I:

There are a lot of changes from the original plan. We have more queries that we wanted to display results for and analyze. We presented our data in various ways such as heatmaps, bar charts, and graphs. Our relational data model is much larger since we imported them from an ATP tennis database from kaggle. So instead of downloading data from an API, we used the csv files straight from kaggle. We found that there was a lot more we could work with and we weren't limited to data that we selected by hand. Fortunately, there was some country data in there already that we played around with. We still have some user input for our queries such as player names or country abbreviations that can be input to see data specific to those inputs. There is a slight emphasis on report generation for us since we are working locally and can utilize plots and data frames. We decided to do this instead of the interactive webpage because we believe this would generate more interesting results that we can learn from the giant amount of data. We were able to learn a lot more from the data using this approach compared to just querying results based on user inputs, so we prioritized our analysis of the data.

How we loaded the database with values:

We downloaded the csv files off of kaggle along with a .sqlite file. The sqlite file is a database file that uses the SQLite database management system. So this acted as our database that we ran queries from. The source is here: https://www.kaggle.com/datasets/guillemservera/tennis/data
After downloading the data, we import it into the Python notebook by:



Software platform:

We used a Jupyter notebook along with an sqlite file to query from our data files. Use the latest version of python if possible.

User's guide:

To run our code follow these steps:

1. Download the Jupyter notebook and ensure a python environment is running.

2. Download the data files and sqlite file. Visit: https://www.kaggle.com/datasets/guillemservera/tennis/data?select=database.sqlite and download the data. Unzip the file.
3. Put the archive file in the same directory as the Python notebook.
4. Hit “run all” in the Jupyter notebook. (you may have to install some modules via the command line)
5. Interact with the prompts to query specific country or player data.

Areas of specialization:

We have an interesting interface demonstrating significant accomplishment and expanded education. We are able to learn a lot of cool patterns and discoveries from our analysis. We kind of just asked some general interesting questions about the statistics and did a bunch of queries and plots to see if we could find any patterns. We would then go into specifics to further investigate our findings. Some of the patterns we saw proved to be unexpected. It’s pretty educational to learn that statistics in sports can help predict future results. We found it very interesting that patterns exist in something as complex as a sporting event and we got to participate in analyzing sports data to notice these patterns.

Selling Points:

- Size of our dataset: we are working with a lot of data (in the millions)
- The process it takes to find and analyze the data points we have: since we are analyzing our data, a lot of trial and error was required to find the queries that we wanted to display. Instead of choosing specific queries, we cycled through different queries and chose the most interesting data.
- Presentation of our findings: the visuals we came up with make it easy to see patterns
- Coverage: we cover a lot of aspects when it comes to tennis statistics (obviously not all), but we explore many different aspects that have potentially interesting findings such as age, country, handedness, and ranking
- How interactable our project is and how our project graphically displays the results of the queries from the user input: the data we have can be categorized in many ways, we chose to allow users to interact with specific player and specific country statistics; there is potential for others but our project allows this to be done easily

Limitations:

- Limited by the type of queries we have presented: obviously there are infinite amount of different queries we can come up with for our data, we can only present the ones that we came up with
- Interactability: we didn’t make all portions interactable, there is a lot of potential for interactability for users since we can prompt information
- With more time we can make it fully interactable. A user would be able to input and look up queries with a much wider variety of information.

Output:

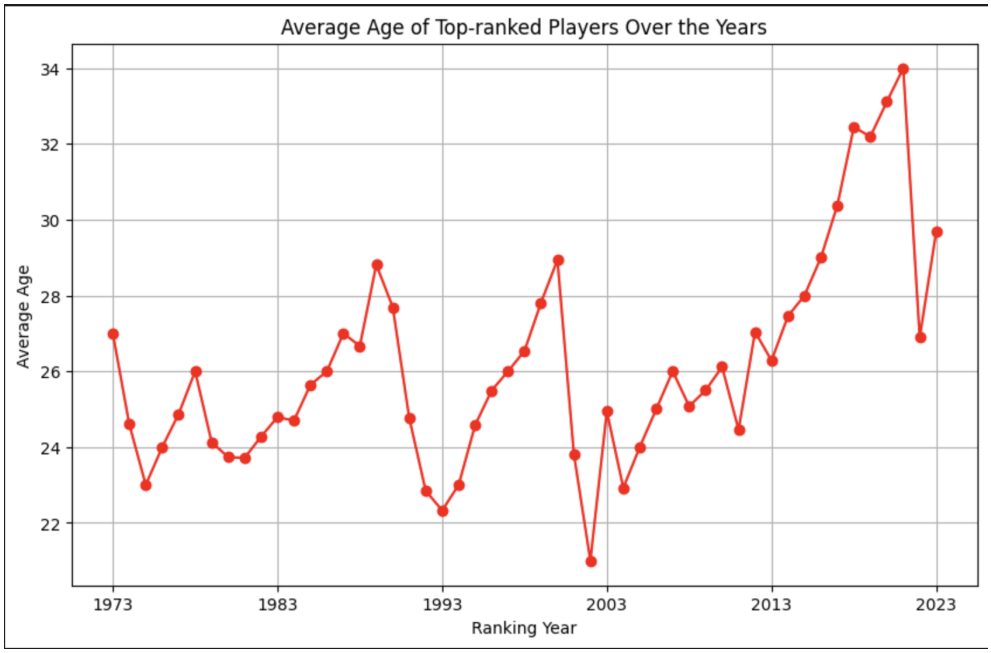
It is best to view the output in our Jupyter notebook. Here are some of the output that we had that is interesting to look at:

	ranking_year	number_of_players
0	1990	1602
1	1991	1603
2	1992	1697
3	1993	1693
4	1994	1729
5	1995	1831
6	1996	1901
7	1997	1990
8	1998	2062
9	1999	2042
10	2000	2033
11	2001	2091
12	2002	2135
13	2003	2188
14	2004	2296
15	2005	2395
16	2006	2528
17	2007	2447
18	2008	2435
19	2009	2365
20	2010	2270
21	2011	2362
22	2012	2489
23	2013	2691
24	2014	2766
25	2015	2843
26	2016	2762
27	2017	2536
28	2018	2523
29	2019	2265
30	2020	2034
31	2021	2384
32	2022	2776
33	2023	2488

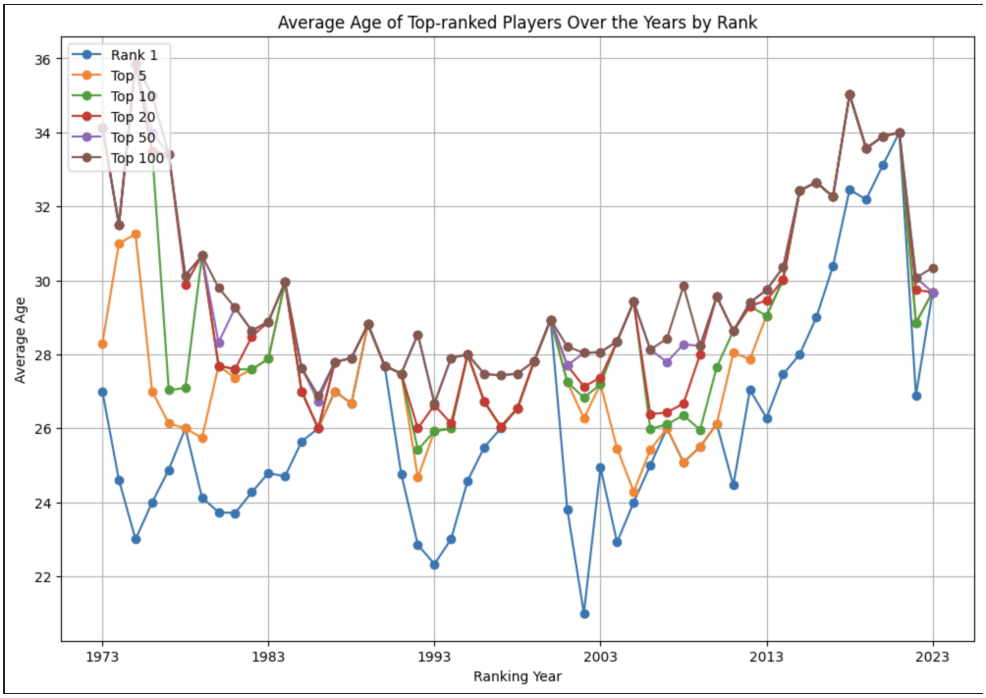
We created a heatmap to learn that the number of players over the years is increasing. We learned that the sport of tennis is likely growing.

player_name	ranked_first	years_being_first
Novak Djokovic	355	12
Roger Federer	307	9
Pete Sampras	285	8
Ivan Lendl	228	8
Rafael Nadal	193	10
John McEnroe	136	6
Jimmy Connors	104	8
Andre Agassi	100	5
Bjorn Borg	94	4
Lleyton Hewitt	80	3
Stefan Edberg	71	3
Jim Courier	58	2
Gustavo Kuerten	43	2
Andy Murray	37	2
Carlos Alcaraz	31	2
Mats Wilander	16	2
Daniil Medvedev	13	1
Andy Roddick	13	2
Boris Becker	12	1
Ilie Nastase	10	2
Marat Safin	9	2
Juan Carlos Ferrero	8	1
Yevgeny Kafelnikov	6	1
Thomas Muster	6	1
Marcelo Rios	6	1

We were able to find the players that have topped the rank the most weeks all time. Here we learned who the most successful players have been by the numbers.



This chart shows the age of the top ranked players each each. We learned that the pattern seems to indicate that a player dominates for a certain amount of time and then retires for a new player to take over. The chart is basically showing how our legendary players age.



We did this for other ranking positions and found similar trends. This might indicate that tennis comes in waves where each generation has a group of players that do well.

year	player_name	age
2022	Carlos Alcaraz	19
2000	Marat Safin	20
2001	Lleyton Hewitt	20
2023	Carlos Alcaraz	20
1977	Bjorn Borg	21
1980	John McEnroe	21
2001	Marat Safin	21
2002	Lleyton Hewitt	21
2003	Andy Roddick	21
1974	Jimmy Connors	22
1981	John McEnroe	22
1992	Jim Courier	22
1993	Pete Sampras	22
2003	Lleyton Hewitt	22
2004	Andy Roddick	22
2008	Rafael Nadal	22
1975	Jimmy Connors	23
1979	Bjorn Borg	23
1982	John McEnroe	23
1983	Ivan Lendl	23
1993	Jim Courier	23
1994	Pete Sampras	23
1998	Marcelo Rios	23
1999	Carlos Moya	23
2003	Juan Carlos Ferrero	23

We discovered the youngest players to have topped the rankings.

hand	cases	ranking_by_hand
L	278010	674.2971439876263
R	2234345	731.5845811636073

We found the average rankings for handedness (L-lefties, R-righties).

rank	lefties_percentage
1	20.0
2	31.41831238779174
3	19.53195319531953
4	15.01123595505618
5	15.985630893578806
6	15.92442645074224
7	12.398921832884097
8	9.712230215827338
9	13.689407540394974
10	12.455035971223023

```
%%sql
SELECT
    (SUM(CASE WHEN hand = 'L' THEN 1 ELSE 0 END) / CAST(COUNT(*) AS REAL)) * 100 AS lefties_percentage
FROM
    player_rankings
WHERE
    rank <= 100
```

```
* sqlite:////Users/thomasyu/Desktop/Project/archive/database.sqlite
Done.
```

lefties_percentage
14.816479737474994

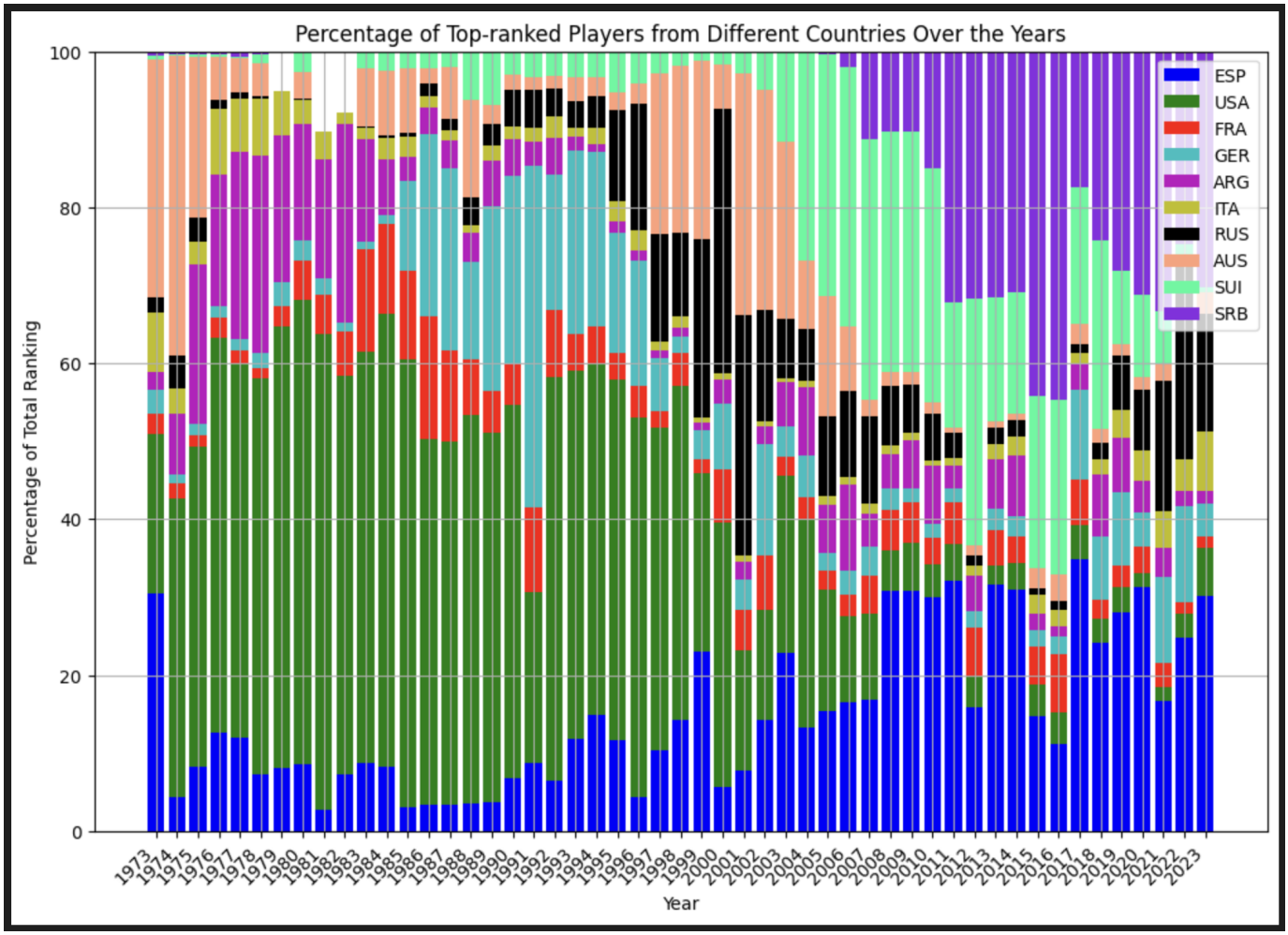
We learned the percentage of lefties for each top 10 position as well as what percentage of the top 100 are lefties.

hand	total_cases	ranking_by_hand	percentage_of_total
L	278010	674.2971439876263	11.065713245142506
R	2234345	731.5845811636073	88.9342867548575

This shows what percentage of players are lefties. The percentage is lower than any of the percentage of lefties in the high rankings. This must mean that being left-handed gives an advantage!

ioc	country_points
ESP	36384117.0
USA	32983420.0
FRA	26052278.0
GER	19536655.0
ARG	18709291.0
ITA	14498405.0
RUS	13312107.0
AUS	13006544.0
SUI	11621000.0
SRB	11267457.0
CZE	10570540.0
GBR	9477252.0
SWE	9420275.0
CRO	7190220.0
AUT	6691896.0
BRA	6433650.0
NED	6137481.0
BEL	5259736.0
JPN	4887741.0
CAN	4629628.0
SVK	4113360.0
CHI	3853831.0
RSA	3844179.0
ROU	2788844.0
UKR	2487952.0

We took a look at what countries have accumulated the most total points.



This chart shows the top 10 countries with the highest ranking players in the past 50 years. The larger the bar, the better the performance of that country in that year. In the current year, we can see that Serbia and Spain are dominating with Russia following in third.

player_name	rank	ioc	year
Jimmy Connors	1	USA	1974
Jimmy Connors	1	USA	1975
Jimmy Connors	1	USA	1976
Jimmy Connors	1	USA	1977
Jimmy Connors	1	USA	1978
Jimmy Connors	1	USA	1979
John McEnroe	1	USA	1980
John McEnroe	1	USA	1981
John McEnroe	1	USA	1982
Jimmy Connors	1	USA	1982
John McEnroe	1	USA	1983
Jimmy Connors	1	USA	1983
Ivan Lendl	1	USA	1983
John McEnroe	1	USA	1984
Ivan Lendl	1	USA	1984
John McEnroe	1	USA	1985
Ivan Lendl	1	USA	1985
Ivan Lendl	1	USA	1986
Ivan Lendl	1	USA	1987
Ivan Lendl	1	USA	1988
Ivan Lendl	1	USA	1989
Ivan Lendl	1	USA	1990
Jim Courier	1	USA	1992
Jim Courier	1	USA	1993
Pete Sampras	1	USA	1993

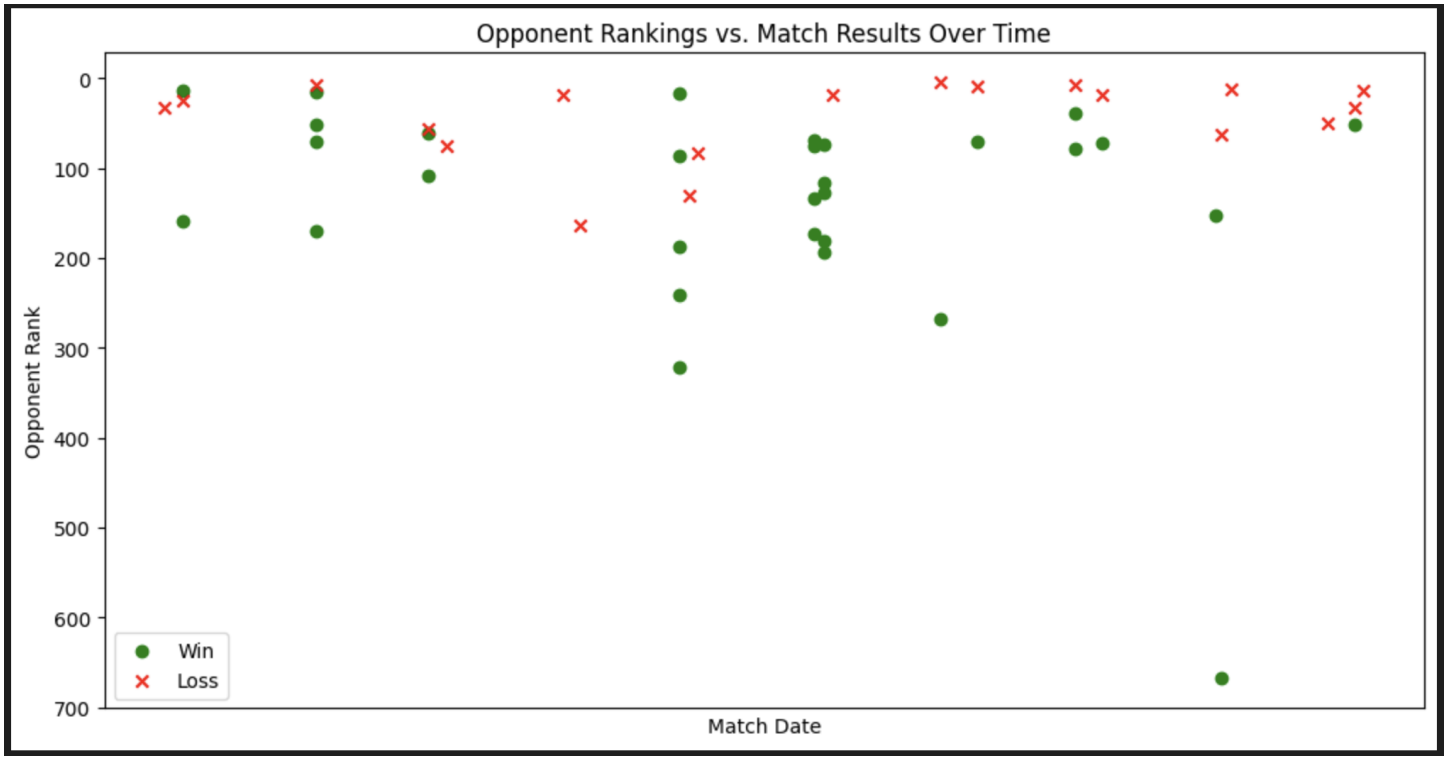
player_name	rank	ioc	year
Taylor Fritz	5	USA	2023
Taylor Fritz	7	USA	2023
Taylor Fritz	8	USA	2023
Taylor Fritz	9	USA	2023
Taylor Fritz	10	USA	2023
Frances Tiafoe	10	USA	2023
Frances Tiafoe	11	USA	2023
Frances Tiafoe	12	USA	2023
Tommy Paul	12	USA	2023
Tommy Paul	13	USA	2023
Frances Tiafoe	13	USA	2023
Frances Tiafoe	14	USA	2023
Tommy Paul	14	USA	2023
Frances Tiafoe	15	USA	2023
Tommy Paul	15	USA	2023
Ben Shelton	15	USA	2023
Frances Tiafoe	16	USA	2023
Tommy Paul	16	USA	2023
Ben Shelton	16	USA	2023
Frances Tiafoe	17	USA	2023
Tommy Paul	17	USA	2023
Ben Shelton	17	USA	2023
Tommy Paul	18	USA	2023
Frances Tiafoe	19	USA	2023
Tommy Paul	19	USA	2023

player_name	rank	points	ioc	year
Carlos Alcaraz	1	6820.0	ESP	2023
Carlos Alcaraz	1	6820.0	ESP	2023
Carlos Alcaraz	1	6820.0	ESP	2023
Carlos Alcaraz	1	7420.0	ESP	2023
Carlos Alcaraz	1	6815.0	ESP	2023
Carlos Alcaraz	1	6815.0	ESP	2023
Carlos Alcaraz	1	7675.0	ESP	2023
Carlos Alcaraz	1	7675.0	ESP	2023
Carlos Alcaraz	1	9675.0	ESP	2023
Carlos Alcaraz	1	9375.0	ESP	2023
Carlos Alcaraz	1	9225.0	ESP	2023
Carlos Alcaraz	1	9225.0	ESP	2023
Carlos Alcaraz	1	9395.0	ESP	2023
Carlos Alcaraz	1	9815.0	ESP	2023
Carlos Alcaraz	1	9815.0	ESP	2023
Rafael Nadal	2	6020.0	ESP	2023
Rafael Nadal	2	5770.0	ESP	2023
Rafael Nadal	2	5770.0	ESP	2023
Carlos Alcaraz	2	6730.0	ESP	2023
Carlos Alcaraz	2	6730.0	ESP	2023
Carlos Alcaraz	2	6730.0	ESP	2023
Carlos Alcaraz	2	6480.0	ESP	2023
Carlos Alcaraz	2	6780.0	ESP	2023
Carlos Alcaraz	2	6780.0	ESP	2023
Carlos Alcaraz	2	6780.0	ESP	2023

We took a look at top players from different countries. The first image depicts the best players from the USA ever (shortened list). The middle image shows the best current players from the USA. The right image shows the best current players from Spain. We allow the user to pick countries to check out and run the queries from their interactive input (these are just samples).

36	Brandon Nakashima	Win	70.0	20230731.0	7-6(5) 6-4
37	Jannik Sinner	Loss	8.0	20230807.0	W/O
38	Max Purcell	Win	78.0	20230807.0	7-6(2) 3-6 7-5
39	Lorenzo Sonego	Win	39.0	20230807.0	7-6(3) 6-0
40	Corentin Moutet	Win	72.0	20230828.0	6-2 7-5 6-3
41	Grigor Dimitrov	Loss	19.0	20230828.0	6-3 6-4 6-1
42	Aslan Karatsev	Loss	63.0	20230920.0	4-6 6-3 6-2
43	Ye Cong Mo	Win	668.0	20230920.0	7-5 6-3
44	Alex De Minaur	Loss	12.0	20230927.0	6-3 5-7 7-6(6)
45	Roman Safiullin	Loss	50.0	20231002.0	6-3 6-2
46	Tomas Martin Etcheverry	Loss	32.0	20231023.0	6-7(5) 6-3 6-2
47	Yannick Hanfmann	Win	51.0	20231023.0	7-5 6-4
48	Alex De Minaur	Loss	13.0	20231030.0	7-6(5) 4-6 7-5
49	Leandro Riedi	Win	152.0	20230915.0	6-7(7) 6-4 6-4
50	Total				50 Matches

Finally, we started looking at match data. This is a sample of a player's (Andy Murray) matches this year. We know that they played 50 matches this year and we know the results and ranking of their opponents.



We also charted the match data on winning or losing vs opponent rankings (for the selected player above, Andy Murray in this case) so that it is easier to visualize patterns in their performance. This player typically wins against lower ranked players, but struggles against the best.

Relational table specification:

We want to note that due to the magnitude of data in the database we worked with, displaying the relational tables here is pretty unrealistic. There are 81 columns for one of our data tables. It is best to view the tables through Kaggle, we used all data files here:
https://www.kaggle.com/datasets/guillemservera/tennis/data?select=database.sqlite
Note: We do have our old relational tables in our phase I submission. Some of the columns are in our final database, but we have a lot more data we are working with now.

SQL code:

Best viewed in our Jupyter notebook or attached is a PDF of our notebook output titled Databases_Project.pdf. The output for each query follows the query itself.