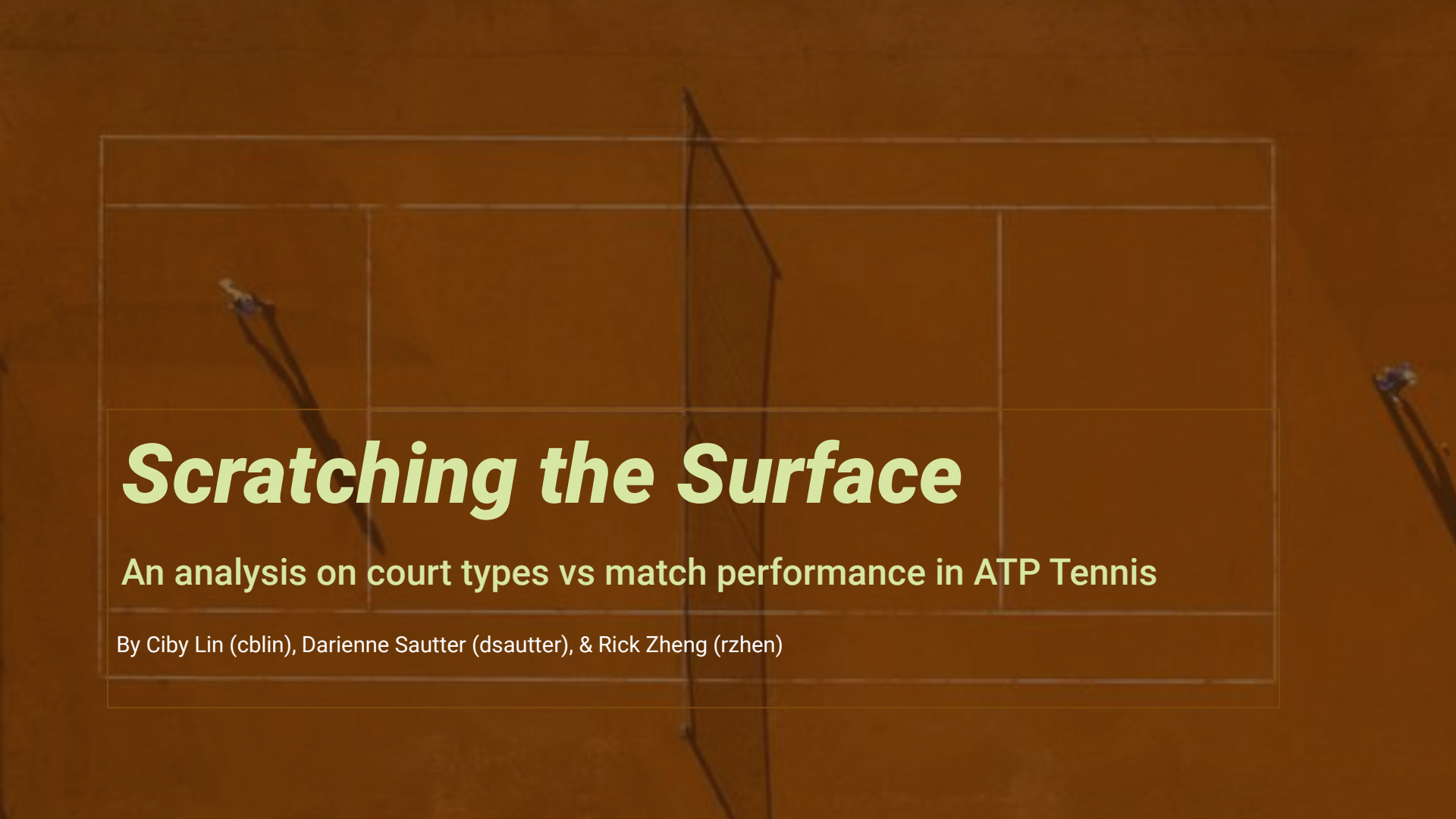


Scratching the Surface

An analysis on court types vs match performance in ATP Tennis

By Ciby Lin (cblin), Darienne Sautter (dsautter), & Rick Zheng (rzhen)

Project Motivation

Stakeholders: ATP

(The Association of Tennis Professionals)

Goal: Our goal is to determine whether or not court different surfaces affect a players ability to win matches

Data Sources *(data sets)*

3.1 Primary dataset description (tennis_atp):
11 csvs, each pertaining to a year of ATP sanctioned tournaments and matches, provides detailed information regarding match statistics. Saved a ‘m_matches’ in the notebook. Each row is identified by the features: tourney_id, match_num.

Key features

- winner_name: Name of match winner
- round: The current round the match is being played on
- tourney_name: Name of the tournament
- winner_rank: ATP ranking of the winner
- loser_rank: ATP ranking of the loser

3.2 Secondary dataset description (tennis_wta):
11 csvs, each pertaining to a year of WTA sanctioned tournaments and matches, provides detailed information regarding match statistics. Saved as ‘w_matches’ in the notebook. Each row is identified by the features: tourney_id, match_num.

Key features

- winner_name: Name of match winner
- round: The current round the match is being played on
- tourney_name: Name of the tournament
- winner_rank: ATP ranking of the winner
- loser_rank: ATP ranking of the loser

Size: 30859 rows, 50 columns

Location: https://github.com/JeffSackmann/tennis_atp

Data manipulation Methods

Handling missing or anomalous data (process stored in 'processing.ipynb')

There are no missing or anomalous data from the primary nor the secondary datasets from examination by column with the .unique() function.

However, as we merged the two dataset some ***'minutes'*** data were missing from the ***'atp_matches_2015'*** dataset. There are two approaches to imputation these missing data.

- First, we can use the mean or median calculated from the rest of the dataset.
- Second, we can look up the actual minutes played for the match on the internet then cite the source.

Joining the primary and secondary datasets

To merge the two datasets, we will use the ***'winner_name'*** feature (joined 'FIRST' and 'LAST' columns) in the primary dataset and ***'winner_name'*** feature in the secondary dataset as the level to do the merge.

Cleaning and Manipulation Challenges

- Ensuring ***'winner_name'*** string values are correctly matched in both datasets.
  - We thought about matching also on ***'COUNTRY'*** from primary dataset to ***'winner_ioc'*** from secondary, but multiple players have dual citizenship.
- Turn numerical value ***'Date'*** columns into datetime format and extract year, month, or date according to needs
- Concatenate ***'FIRST'*** and ***'LAST'*** columns to a new column to prepare for merging
- Query tournaments keeping ***'Australian Open'***, ***'US Open'***, ***'Roland Garros'***, ***'Wimbledon'***
- Created a new column in each dataset called "Association" to distinguish between WTA and ATP matches

Analysis

Comparative analysis of volume of matches played on each court type over 2010-2020
This analysis attempts to compare the distributions of court types used in each year of the ATP tour from 2010-2020. By using histograms, we expect to see if there are any trends or patterns in court type usage occurring year over year that may inform further research.

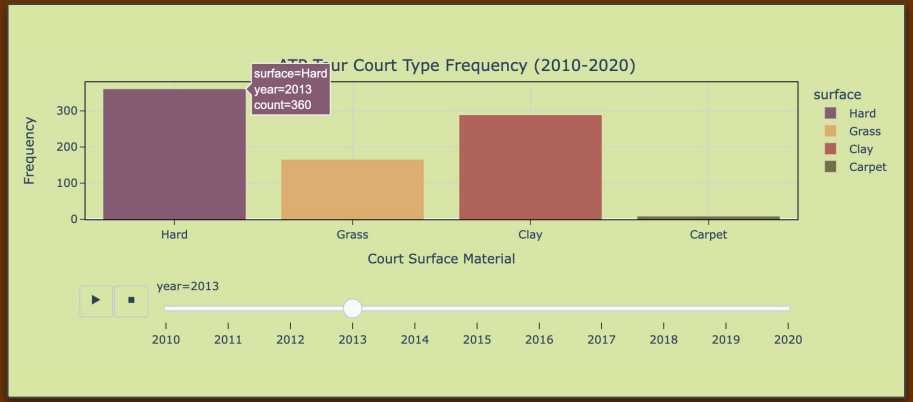
Comparative analysis of win rates by court surface type across all years
This analysis attempts to compare the distributions of win rates across each surface type during all ten years of ATP tour data. By using violin plots for each surface type, we expect to learn what, if any, distributions may be skewed or have less variation than others to determine the best court type for standardization.

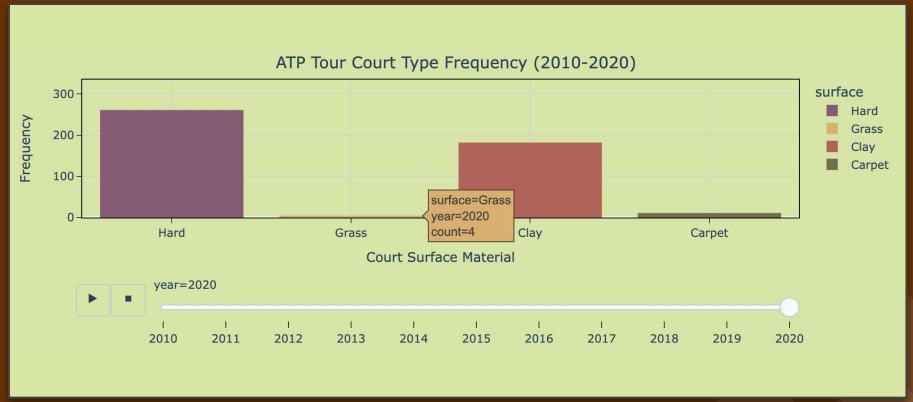
Comparative analysis of top ATP tennis players win rates by surface over 2010-2020
This analysis use OLS regression lines across a scatterplot to examine top players performance over time on each court surface to determine if they adapt to the ATP Tour's courts over time or struggle consistently on certain court types. If there is a significant disparity between performance on any court type, we will present this to the ATP for review.

Comparative analysis of match duration by surface and gender over 2010-2020
This analysis uses heatmaps to compare the difference in game duration and gender across surface to determine which surface leads to shorter games among male and female players. If there is a significant difference between the surfaces. We can determine the most optimal concentration time for the audience to determine which surface would be beneficial and preferred by tennis watchers by conducting additional surveys.

Visualization #1:

Interactive Time-series Histogram



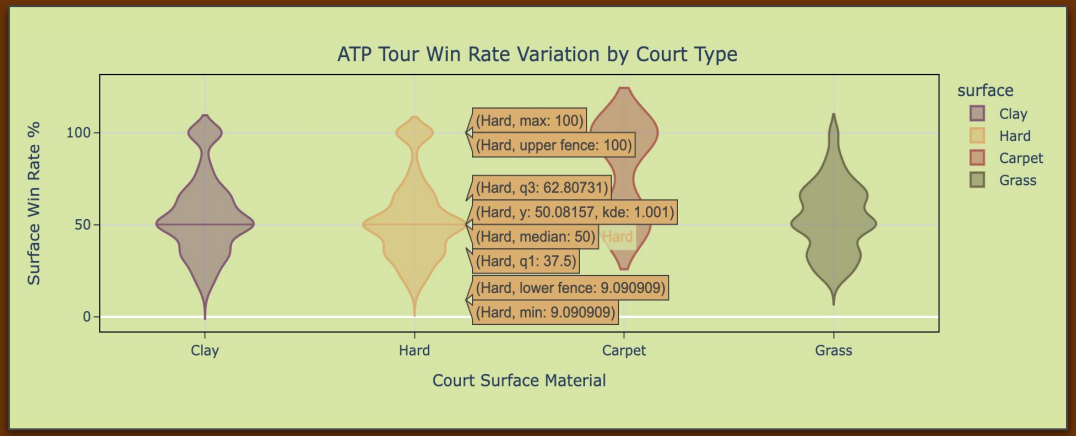


Using a histogram we are able to see hard courts are the most common surface type in the ATP tour, while carpet is the least. The ATP stopped using carpet courts in professional tennis in 2009, right before our data begins, which is likely why we see such a low frequency of carpet courts across the entire data set.

Grass courts are typically played on less than clay and hard courts due to maintenance; However, there was a large difference in the number of grass courts played on in 2020. The lowest amount of grass court matches over the past 10 years which likely occurred as a result of the 6-week shutdown due to the COVID-19 outbreak.

Visualization #2:

Violin Plot

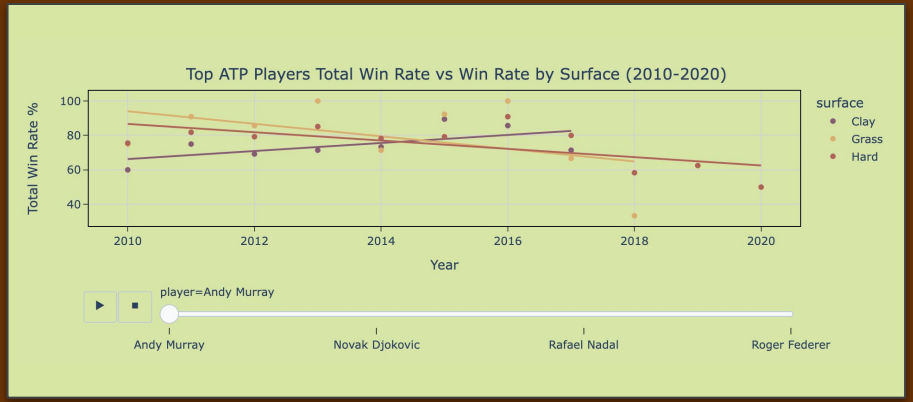


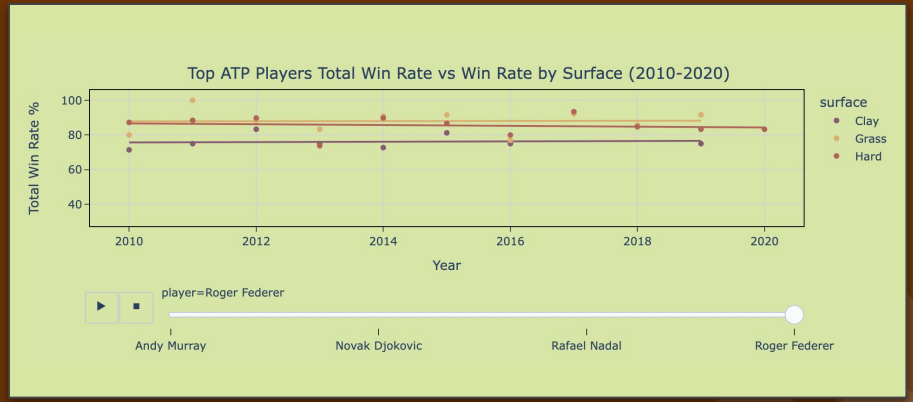
After removing zero percent from the win rates to mitigate noise, clay and hard court surfaces have similarly shaped distributions. The shape of the grass court distribution appears to be heavily influenced by the upper and lower quartiles, likely as a result of limited data on this court type. Clay and grass courts share the same upper and lower quartiles at Q3:66% and Q1:40% despite the disparity between data between them. Lastly, when compared against clay and grass courts, hard courts have a smaller variation in win rate percentage.

This visualization, along with our previous histogram, leads us to believe that we would need more data on hard and grass courts to determine if ATP needs to regulate them further, though our initial findings point towards hard courts enhancing players win rates.

Visualization #3:

Interactive Time-series Scatter Plot





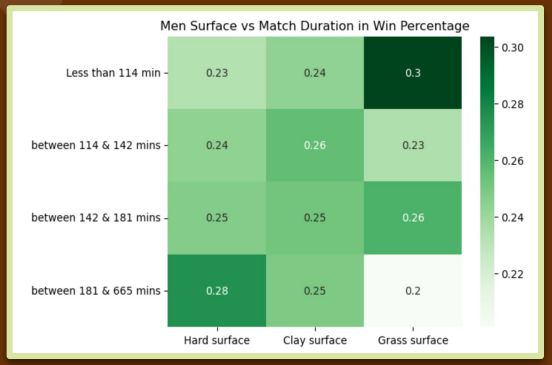
When looking at performance across time, Andy Murray was affected the most by court types. His win rate went down year over year when playing matches on grass and hard courts.

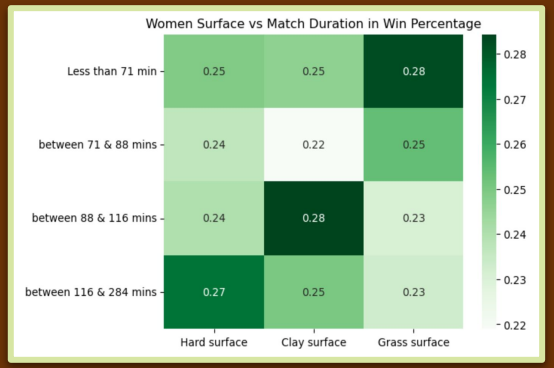
Of the top players in the ATP Tour from 2010-2018, Roger Federer's win rate was the least affected by court types. For instance, Federer's win rate on clay courts remains around 75%.

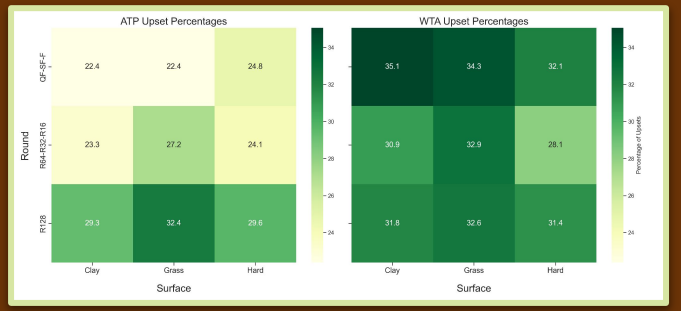
This visualization reveals most top players are unaffected by court type and show consistent performance on each court surface across time. However, it is inconclusive of whether or not win rate is affected by courts or increases/decreases over time in general.

Visualization #4:

Heat Maps







In men's best-of-five matches from year 2010 to 2020, grass courts have the highest win percentage (30%) of matches lasting less than 114 minutes. Hard courts have the second highest percentage (28%) in the longest category where game duration is between 181 and 665 minutes. Grass surface results in shorter games and hard surface results in longer games.

In women's best-of-three matches, which generally result in shorter game durations compared to men's single matches, 28% of the matches on grass lasted less than 71 minutes.

Clay courts were the second fastest, with 28% match durations falling between 88 and 116 minutes. Conversely, hard surface matches had the longest durations, with 27% of the matches lasting between 116 and 284 minutes.

Overall, through 2010-2020, there is a discrepancy between WTA upsets vs ATP Upset Grand Slam matches across the binned rounds. Additionally, the Grass Surface seems to hold the highest upset percentage in the first round of Wimbledon.

It is also seen that as the rounds progress for WTA Grand Slam matches, the upsets tend to increase. On the other hand, the upsets tend to decrease for ATP Grand Slam matches across all surfaces.

Ethical Considerations

Possible Misrepresentation

Since we are using data from ATP tennis tournaments for this analysis only, we do not have a full picture of player performance on court types. The sample is small for some surfaces such as carpet, even though we have over 10 years of match data.

Personal Information

The other issue is privacy. We have information on players performances based on each court type and their full names. If we were to publish this data publicly we would need to either swap names for unique identifiers or confirm with every player that their okay with the information being public. If we do not, many could be upset about information being available that could help competitors get insight into their match performances.

Statement of Work

Group:

- Held weekly Zoom meetings to discuss and set benchmarks for the project
- Creating the proposal, finding reliable and complex enough primary and secondary datasets for the overarching problem
- Creating the GitHub repository for code documentation and collection of datasets

Rick:

- Created the ATP/WTA upset percentage heatmap
- Supported in cleaning/merging of datasets
- Did the Data Sources (slide) of the final report
- Did the Visualization #5 (slide) of the final report

Darienne:

- Created Visualizations #1 (slide) of the final report
- Created Visualizations #2 (slide) of the final report
- Created Visualizations #3 (slide) of the final report
- Did the Project Motivation (slide) of the final report
- Did the Ethical Considerations section of the final report

Ciby:

- Loading/cleaning/merging of datasets
- Version control on Github repository
- Did the data cleaning and manipulation (Slide)section of the report
- Created ATP/WTP minutes vs win percentage heatmap
- Created visualization #4 (slide) of the final report