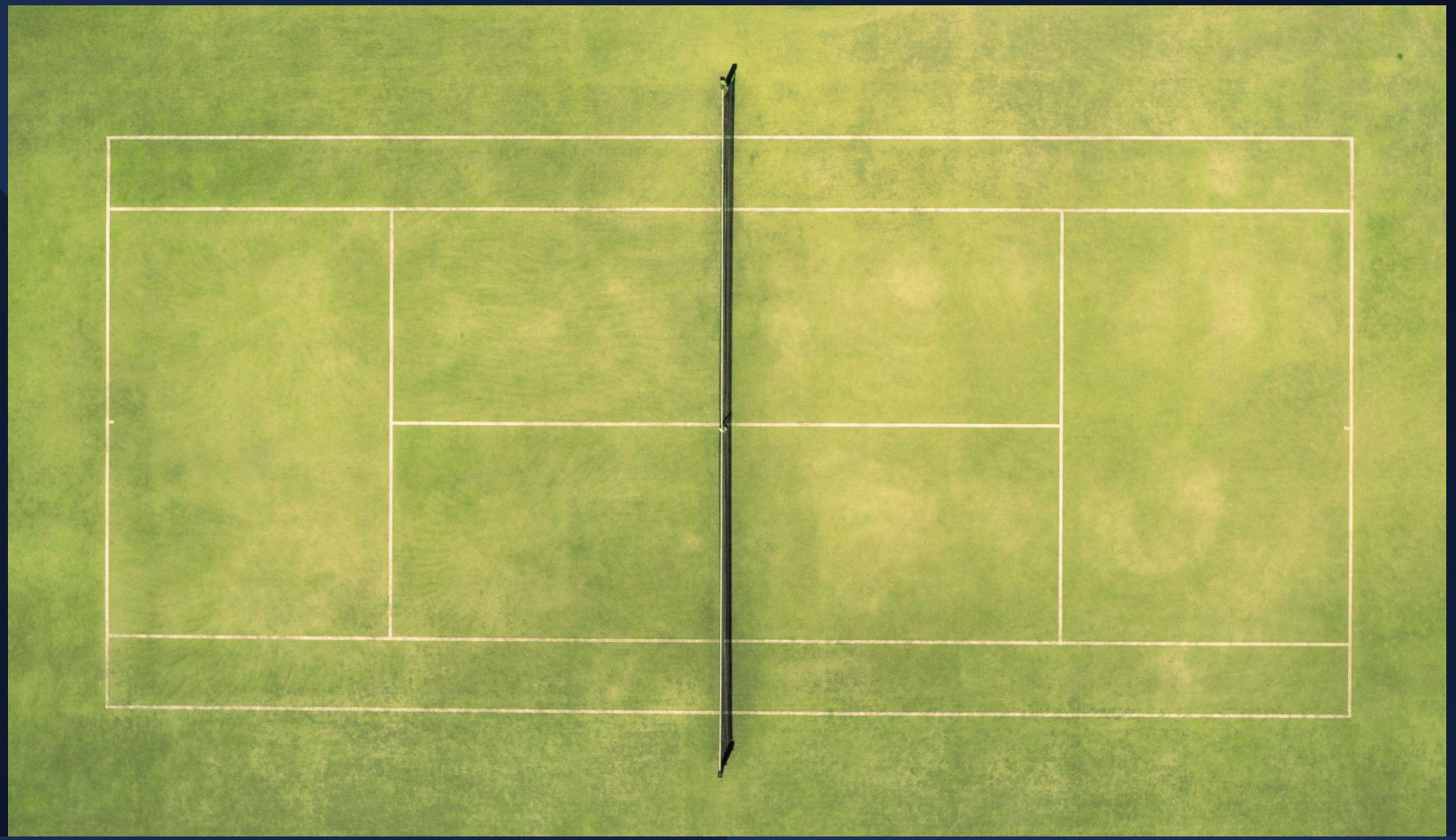
The Impact of Serve Stats in Tennis (ATP Tour)



Project Overview

Project Objectives

- For tennis fans who enjoy match analyses

Project Questions

- Which of the following three serve statistics have the greatest impact on the likelihood of winning a match on the ATP Tour?
    i. % First Serves In Play
    ii. % First Serve Points Won
    iii. % Second Serve Points won
- Did the most impactful stat (determined in question 1) change depending on court surface (clay, grass and hard courts)?

Project Overview

Skills & Tools Used

- Python (pandas, numpy, scikit-learn, seaborn, matplotlib)
- Project Design/Management
- Logistic Regression
- Tableau Visualizations and Storyboard
- Data Cleaning
- Data Wrangling
- Deriving New Variables

Links to Dataset, Data Dictionary and GitHub

- Dataset contains three data frames:
  - matches
  - rankings (unused)
  - players (unused)
- Download ATP Matches Dataset
- Link to GitHub repository

Analysis Steps

Step 1: Data Cleaning/Prep

Data contains match details (serve stats, court surface, etc.)

Removed data prior to year 2000 and with missing stats

Created subsets (by court surface)

Step 2: Deriving Variables

Derived three desired serve stats (in percentages) from available serve stats in data frame (# of points won)

Step 3: Exploring Relationships

Compared correlation between existing variables and each serve stat

Step 4: Data Wrangling

Wrangled winner/loser serve stats for logistic regression

Step 5: Answer Project Questions

Exploring Relationships

Average correlation coefficients between serve stat and # of bps faced (Match Winner Only)

%1st Serve Points Won: -0.542

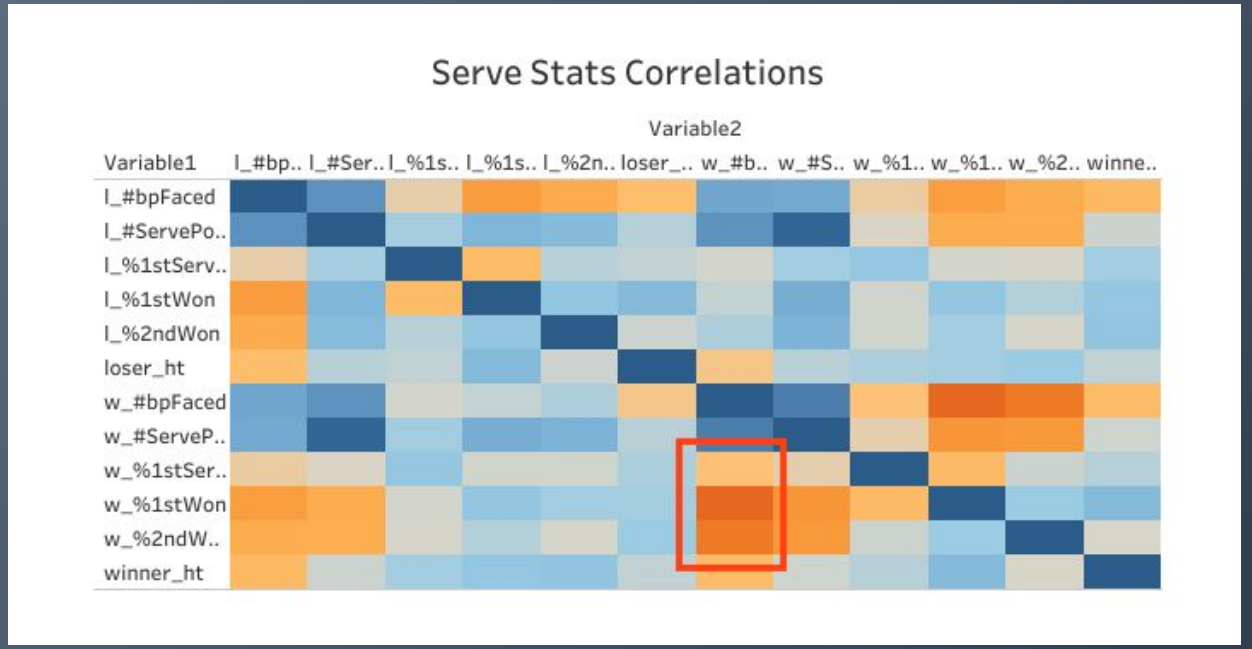
Strong Negative Correlation

%2nd Serve Points Won: -0.435

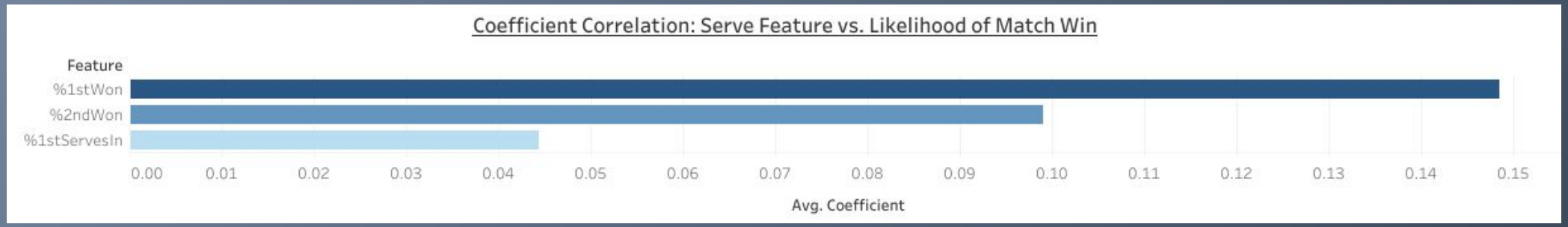
Moderate Negative Correlation

%1st Serves in Play: - 0.084

No Correlation



Q1. Which of the three serve statistics had the greatest impact on the likelihood of winning a match?



Created new dataframe for logistic regression (Step 4)

Serve stats by winner/loser of each match (Three Serve Stats)

New variable: 1 = match won; 0 = match lost)

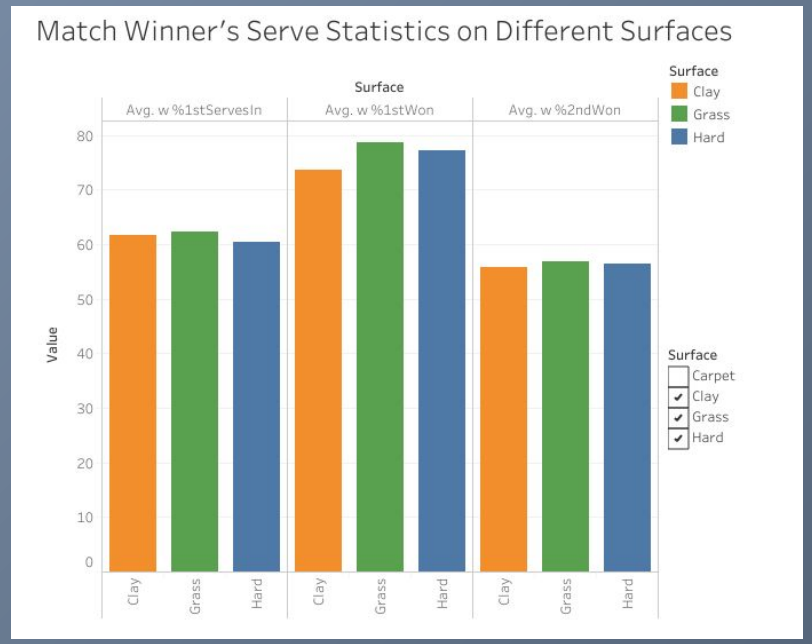
Insights

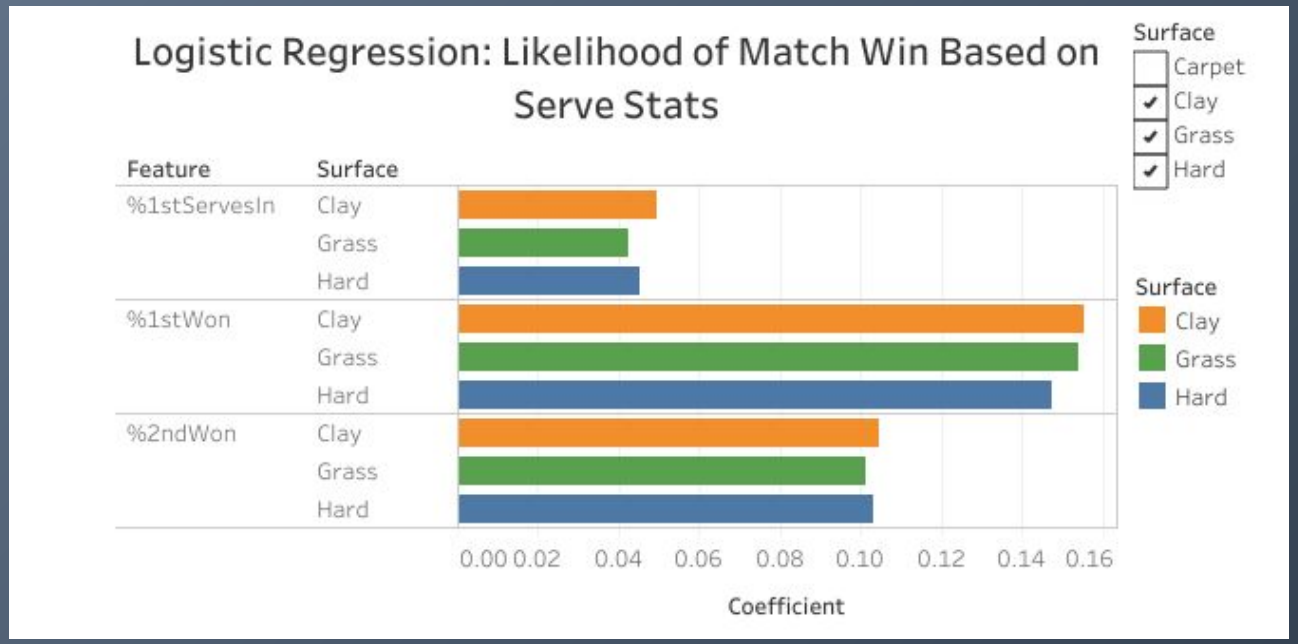
% points won on 1st serve had **weak correlation** with the likelihood of winning a match (0.15)

% points won on 2nd serve had **weaker correlation** with the likelihood of winning a match (0.10)

No correlation between **% of first serves in play** and match win

Q2. Did the correlation between points won behind first serve and match win vary by court surface?





<u>% of 1st serves points won by surface</u>

1. **Grass courts** (78.7%)
2. **Hard courts** (77.2%)
3. **Clay courts** (73.6%)

(Same order for % of 2nd points won)

<u>Correlation Coefficient (% 1st won and match win)</u>

- Consistent across all three surfaces
- Slightly highest on clay court (which had lowest average % of 1st serve points won)

Conclusions

What I Learned

- This was my first experience designing my own project from a self-chosen dataset.
- Through this project, I practiced with deriving new variables in python, as the three serve statistics I wanted to explore were not in the original dataset
- This was my first attempt at logistic regression and creating a new dataframe/variable in order to perform the regression

Next Steps

- Match wins cannot be determined by serve alone - it would be useful to research the impact of other statistics (such as winners, unforced errors, etc.) and how these variables differ across court surfaces
- Only one of the three available data frames were used for this project - there is potential for further research by using the other available data in this set (such as the impact of a player's height on serve, or the correlation between player rankings and serves, etc.)