# Predicting Tennis Match Outcomes Based on Playing Style and Court Surface

Supervised Learning Capstone

Martin Rothwell

# Research Question

**Can the outcome of a tennis match on different court surfaces be predicted by a player’s playing style?**

# The Data

https://archive.ics.uci.edu/ml/datasets/Tennis+Major+Tournament+Match+Statistics#

- Grand Slam tournaments from 2013
  - Australian Open (hard court)
  - US Open (hard court)
  - French Open (clay court)
  - Wimbledon (grass court)
- Men and Women events
- ~2000 data points across all eight events, about 250 each

| | Player | Round | Result | Sets | FSP | FSW | SSP | SSW | ACE | DBF | ... | BPC | BPW | NPA | NPW | TPW | ST1 | ST2 | ST3 | ST4 | ST5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Lukas Lacko | 1 | 0 | 0 | 61 | 35 | 39 | 18 | 5 | 1.0 | ... | 1 | 3 | 8.0 | 11.0 | 70 | 3 | 6.0 | 1.0 | NaN | NaN |
| **1** | Leonardo Mayer | 1 | 1 | 3 | 61 | 31 | 39 | 13 | 13 | 1.0 | ... | 7 | 14 | NaN | NaN | 80 | 6 | 6.0 | 6.0 | NaN | NaN |
| **2** | Marcos Baghdatis | 1 | 0 | 0 | 52 | 53 | 48 | 20 | 8 | 4.0 | ... | 1 | 9 | 16.0 | 23.0 | 106 | 4 | 5.0 | 4.0 | NaN | NaN |
| **3** | Dmitry Tursunov | 1 | 1 | 3 | 53 | 39 | 47 | 24 | 8 | 6.0 | ... | 6 | 9 | NaN | NaN | 104 | 6 | 6.0 | 6.0 | NaN | NaN |
| **4** | Juan Monaco | 1 | 0 | 1 | 76 | 63 | 24 | 12 | 0 | 4.0 | ... | 3 | 12 | 9.0 | 13.0 | 128 | 6 | 4.0 | 6.0 | 2.0 | NaN |

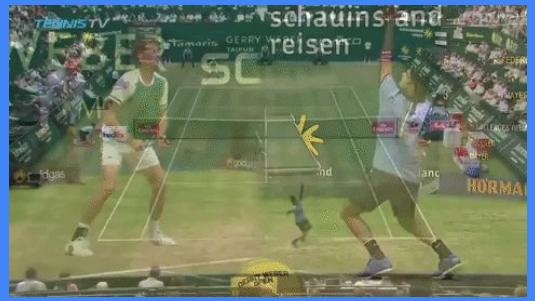
# Tennis Basics

- Courts
  - Grass courts are typically faster-playing with shorter rallies
  - Clay courts are typically slower-playing with longer rallies
  - Hard courts (various cement compositions) are somewhere in between



# Tennis Basics

- Shots
  - Ace
  - Double Fault
  - Winner
  - Unforced Error
- Three true outcomes
  - Winner
  - Unforced Error
  - Forced Error






Offensive/controlling playing style = taking more risks, resulting in more winners and unforced errors

# Data Cleaning

- French Open had no missing data!
- Wimbledon was missing **Total Points Won** for both men and women
- Australian Open had some **Aces** and **Double Faults** missing
    - I felt comfortable filling these with “0” since some players could be likely to go through a match without any of those
- US Open was missing **Total Points Won** for the women and **Winners** and **Unforced Errors** for the men
    - Since both the US Open and Australian Open are played on hard court, I looked at the distributions of other features to see and determined that it would make sense to use the median value for **Winners** and **Unforced Errors** from the Australian Open for the men data in the US Open.
- For **Total Points Won**, I had to get a little more creative...

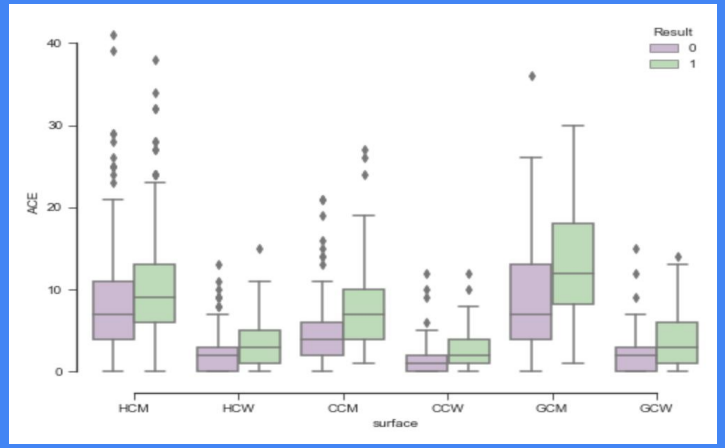
# Feature Engineering

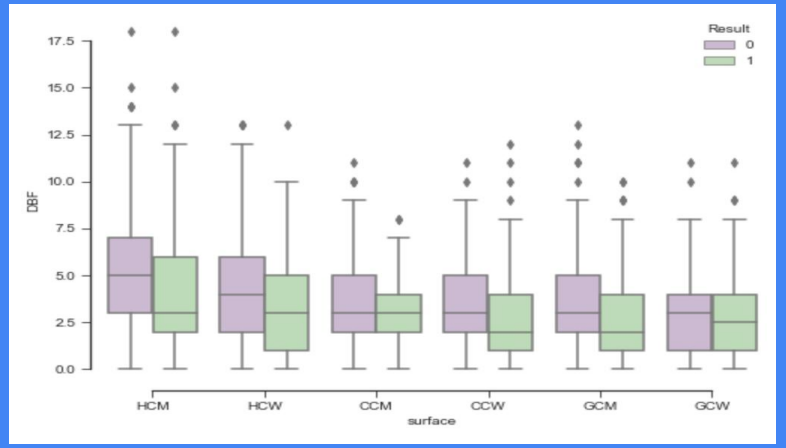
- **Total Points Won (TPW)** was missing in three of the data sets, so I created a new feature call **Total Games (tot_gms)** as a measure of match length. This feature was the sum of all games won by a player in a match, and made it possible to determine rate statistics from count statistics (i.e. **Ace, Double Fault, Winner,** etc.)

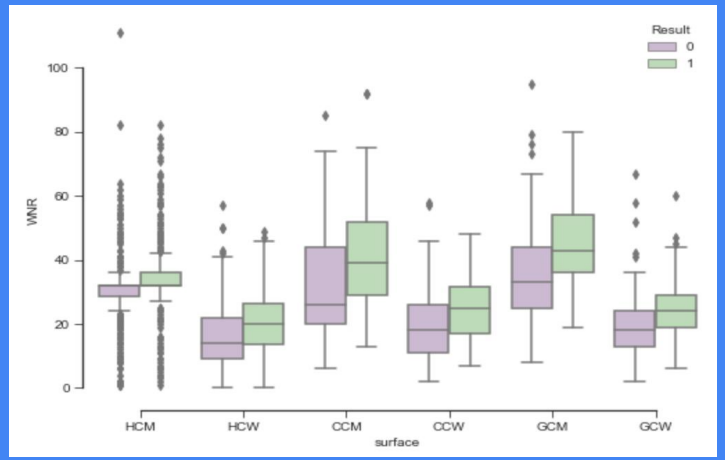
```
Australian Open Men 'tot_gms' to 'TPW' correlation:  0.8894388628
Australian Open Women 'tot_gms' to 'TPW' correlation:  0.855135485007
French Open Men 'tot_gms' to 'TPW' correlation:  0.904906629595
French Open Women 'tot_gms' to 'TPW' correlation:  0.883664889992
US Open Men 'tot_gms' to 'TPW' correlation:  0.913751151454
```

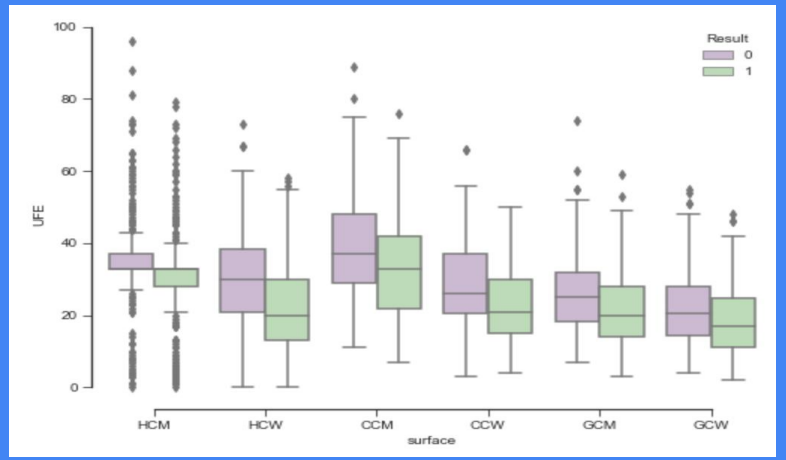
- Wanted a feature that would represent how much control a player had on a match.
  - This led to **Control Factor (ctrl_fact)**, which is a sum of the features that a player has control over: **Aces**, **Double Faults**, **Winners**, **Unforced Errors**.
- Also added dummy variables for court surface, tournament event, and player gender for better categorization

# Data Exploration -- Court Surface

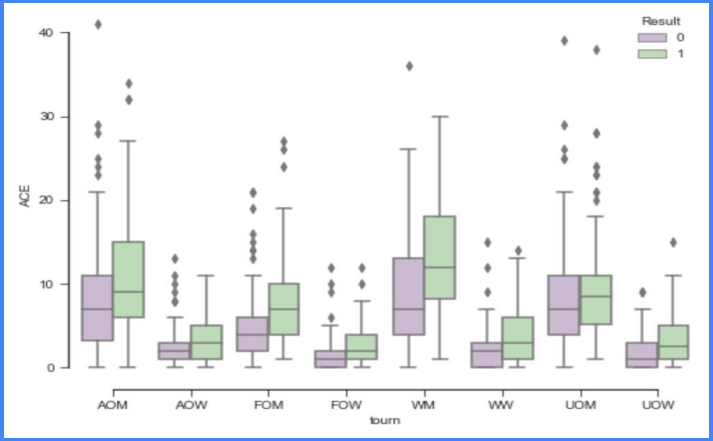


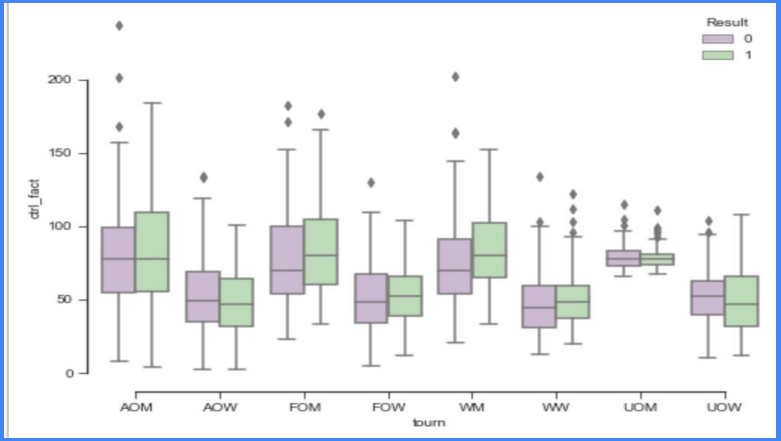



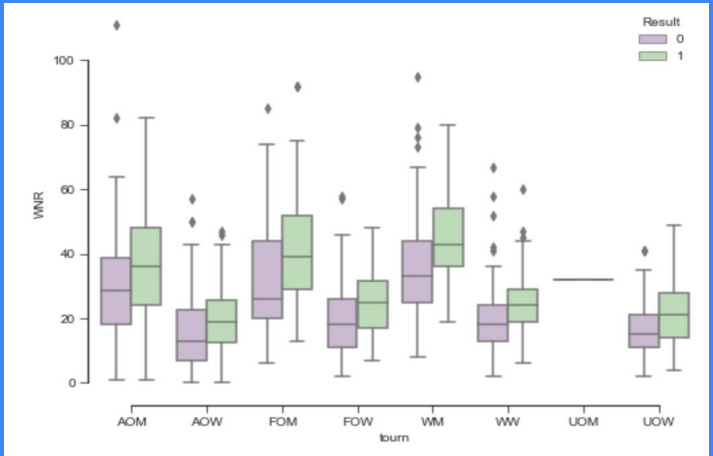



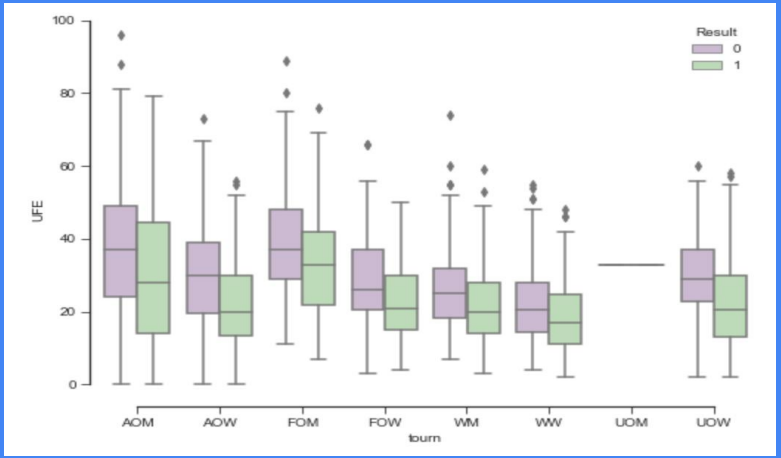



# Data Exploration -- Tournament



# Model Selection

Used the entire data set of approximately 2000 rows to test models. Tested K-Nearest Neighbors, Naive Bayes, Logistic Regression, and Random Forest.

| Features: 'Sets', 'FSP', 'FSW', 'SSP', 'SSW', 'ACE', 'DBF', 'WNR', 'UFE', 'tot_gms', 'ctrl_fact' | | | |
|---|---|---|---|
| KNN | Naive Bayes | Logistic Regression | Random Forest |
| 0.77 (weighted) | 1515/1886 (0.803) | 0.988 (10-fold) | 0.989 (10-fold) |

These models were quite overfit, so I thought about feature selection and which ones were really relevant to the “controlling” playing style.

| Features: 'FSW', 'SSW', 'ACE', 'DBF', 'WNR', 'UFE', 'ctrl_fact' | | | |
|---|---|---|---|
| KNN | Naive Bayes | Logistic Regression | Random Forest |
| 0.64 (weighted) | 1000/1886 (0.530) | 0.673 (10-fold) | 0.625 (10-fold) |

# Logistic Regression

- Used 10-fold cross validation to get more accurate results with the small amount of data

| | Australian Open (hard court) | US Open (hard court) | French Open (clay court) | Wimbledon (grass court) |
|---|---|---|---|---|
| Men | 0.677 | 0.630 | 0.674 | 0.660 |
| Women | 0.773 | 0.782 | 0.792 | 0.771 |

- From my intuition and knowledge of tennis, I expected there to be a significant spread between winning probability between clay and grass courts, given an offensive/controlling playing style, but there really was no measurable difference using this model and data.
- Interesting that the predictive accuracy for the women's matches is much higher, even though there is still minimal difference in playing surface.

# Conclusion

- Not a ton of predictive power in these results to differentiate between court surface.
  - This is a conglomeration of the top-100 or so players at the time of each tournament. There is a huge drop-off in skill level between the top-10 and the rest of the pack.
- I was expecting to see the control factor metric be much more important in predicting match winners on grass courts than on clay courts, but if anything, it was just slightly the opposite.
- With the “Control Factor” metric, Aces/double faults and winners/unforced errors may offset each other. It would be interesting to run a similar experiment but with a net positive factor instead of a sum.

# Future research

- Use more/better data
  - Attain 10+ years of data to see if that yields more decisive results
  - Use official ATP and WTA data -- more complete and accurate
- Look at data from different eras
  - Look at changes in playing styles/equipment and how that affects outcomes
- Use data from just top players
  - Create a metric such as “Positive Control Factor” where it looks at the net positive tallies instead of the basic sum
- Correlate rankings
- Remove outliers, such as the losing match with control factor of 200+
  - not enough rows in this data to be comfortable removing outliers, but it could improve the results if it there were more rows
- Look at shot velocity data on different surfaces and by different players