Capstone Project 1 - Milestone Report

This capstone project's end goal is to conduct a frequentist inference analysis and draw a statistical conclusion on if ATP left hand players have advantages on serves when they play against right hand players. There were several steps taken to:

- Choosing data
- Data wrangling
- Data exploration

Choosing a data set was not easy. I had searched, downloaded and explored multiple data sets based on the initial ideas. Eventually, I have focused on tennis data. My audiences are individuals who are tennis fans and / or have interest to know tennis better. I like to watch grand slam tournaments and am a big fan of Rafael Nadal and Roger Federer. When we watched the playoffs between these two great players, I often heard people saying that Nadal has advantage because he is lefty. With 20 years ATP matches data, I would like to know if ATP lefties really have advantage against righties. You can read more on the Project Definition here.

Data is from Guithub (github.com/JeffSackmann/tennis_atp). I have manually downloaded and imported into dataframe for data wrangling.

Data wangling focused on below key areas:

1. Removing extraneous data – i.e.,
    - Removing rows with 100% missing metric values
    - Removing some categorical columns not needed
2. Feature engineering—i.e.,
    - computing and adding columns for data exploration and machine learning
3. Prepared data set of ATP matches between righties and lefties
4. Normalized the target data set

More details on cleaning/wrangling the data can be found in the Data Wrangling notebook.

How did lefties perform on serves against righties during ATP matches from 2000 to 2019? During data exploration, several observations of interest were found. Here are the high-level results but you can find the details here.

- Average height of lefties is taller than the righties played in the same year except year 2003, 2008 and 2009
- Compared mean of rank of each year, the average of rank of righties is higher than lefties
- Before 2004 and after 2012, the average age of lefties is older than righties
- Compared 20 year mean ace percentage for righties and lefties, the righties had higher ace percentage than lefties
- Lefties has a little higher double fault percentage than righties
- Compared 20 years matches, lefties do have higher first in server percentage
- The first serve won percentage seems no significant difference between righties and lefties
- Righties also have higher second serve in percentage than lefties
- The average second serve won percentage of righties through years are higher than lefties
- The average break points lefties faced are higher than righties
- Righties won more matches than lefties

Next, performed t-test to exam if significant difference between righties and lefties. Further, used bootstrap to generate sample mean distribution and compared calculated means, 95% confidence intervals. Also plot correlation for the data set and checked possible correlation between data fields. Details can be found here.

Again, to summarize, you can view the details for each step here:

- Choosing Data
- Data Wrangling
- Data Exploration (Storytelling and EDA)