SI 601 Individual Project: **The Impacts of Weather on Tennis Player Performance**

Bing-Hsun Wu (bhwu)

Motivation

As a tennis fanatic, I have always wanted to analyze the numerous match data to see if I could find some key determinants behind victories. I myself have played tennis for a few years, and I sensed that there indeed are several pertaining factors that might impact my performance at that particular day. Players like me form inferences and some superstitious habits based on personal experiences without concrete and rational analysis, and wrong assumptions could actually pose threats to players' health status. Therefore, I would like to get the data of matches between professional players, and analyze it with weather data to determine whether humidity, temperature and such might make difference on the match result.

Dataset Sources

Tennis is a professional sport with dynamic bettors, and thus the data of all competitions is publicly accessible and sorted out well.

One dataset source is from Tennis-data.co.uk website, which I aim to get the following variables:

Winner, Loser, Wsets, Lsets, Location in CSV format

The other data set source is the OpenWeatherMap, which I could capture the following variables of interest in the current weather dataset for one location (current weather is used as the proxy variable for the historical weather trend data because free accounts can only access to current weather information. Despite API limitations, it is legitimate to presume the weather at a certain time point could be representative of the climate):

cityname, temperature, pressure, humidity, wind speed in JSON format

The key concept here is to use the location information in the tennis dataset as the GET request parameter in the OpenWeatherMap, so that further the current weather information could be mapped to the tennis dataset. Since this project aims at exploratory data analysis, the tennis dataset used is restrained on the most frequently appeared top 3 male players based on 2015 ATP (Association of Tennis Professional) records. Part of the reason is that one hypothesis to test is whether each player has one's own preferable weather condition, so if there are more game result data of the player could more representative of the player's overall performance. So the

original 2600+ game match data is reduced to 125+ matches concerning the players of interest. The data manipulation techniques in this project could be applicable to bigger dataset or female tennis professional data.

Data Manipulation Methods

Overview: I need to parse through these two datasets to collect the variables of interest, combine them based on the locations data and visualize them as boxplots. My expected output dataset would be the weather data of where the players of interest have won/lost.

The game result data for 2015 ATP in CSV format is from Exmerg, and its source is Tennis-data.co.uk. Since the game result is from their official records and a rather clean dataset with no missing values, it is used for further analysis.

The first goal is to know who are the most frequently appeared top 3 male players in this dataset. A list combining both the name of *Winner* variable and *Loser* variable was created, the imported Counter was used to count the frequency of each element in the list and then the result was outputted to players.csv. The result showed that "Djokovic N. appears 88 times, Nadal R. appears 80 times and Berdych T. appears 79 times". The *Location* variable with these players win or lose in 2,3,4,5 sets (*Wsets*, *Lsets* variables; the new rule in 2008 regulates all matches to be ended within 3 sets, the 4,5 sets were used to accommodate records before the new rule, or the Grand Slams records) were further captured in lists. A list of unique locations was used rather than repetitive locations because the API limits calls per minute to be within 60 for free accounts and an assumption of players attending tournaments across different locations evenly throughout the year.

The second goal is using the locations collected in prior process as request sent to the API. A dictionary was created with the location *name* as key and captured *temperature*, *pressure*, *humidity* and *wind speed* as corresponding values. And then all the required data for generating boxplot was outputted to boxplot.csv.

The final goal is generating visualization from the boxplot.csv using R. The data was first read in, the graph margins were then modified for better visualization and boxplots were eventually generated with overall and player-specific results.

Challenges with Python:

I discovered a quick and simple way to count the elements in a list by importing the Counter.

Python code blocks

```
from collections import Counter
# Output to a players file to make a frequency plot
    output_file = open('players.csv', 'w')
    output_file.write("Player,Frequency\n")
    for (k,v) in sorted(Counter(players).iteritems(), key=lambda(k,v):v,
reverse=True):
        output_file.write(str(k)+","+str(v)+"\n")
    output_file.close()
```

Challenges with R visualization:

At first the graph output was squeezed together and was not legible, so I looked up documentation and made the adjustments for the graph margin. However, the x-axis were still huddled together. I then decided to use “las=2” to make the axis vertical for readability. And then I discovered a way to draw the player-specific results by using the “Result+Player”.

R code blocks

```
# Change inner margin
par(mar=c(10, 6, 2, 1))
# Change outer margin
par(oma=c(0.5,0,1,.5))
# Draw the Plot
boxplot(Temp~Result, data=tennis, ylab="Temperature (°C)", main = "Result
Differences based on Temperature Variations", las=2)
boxplot(Temp~Result+Player, data=tennis, ylab="Temperature (°C)", main =
"Result Differences based on Temperature Variations", las=2)
```

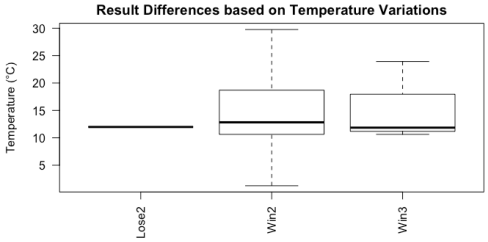
Analysis and Visualization

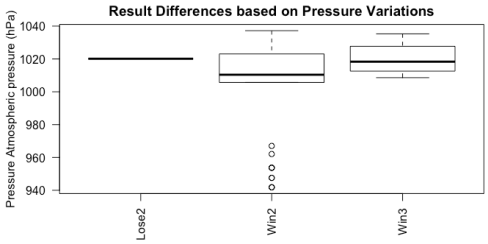
My goal is to combining the weather data and tennis matches results to see if anything insightful could be generated and may affect players’ future training or health management based on the data analysis results.

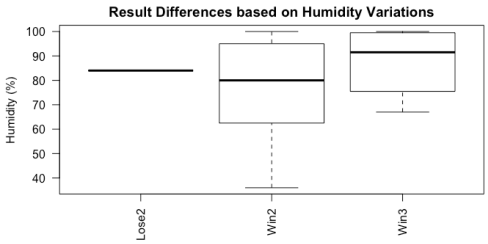
My exploratory data analysis attempts to use simple boxplots to pinpoint if the overall performance of three most frequent professional players in 2015 differs with their individual performance based on differences of temperature, pressure, humidity and wind speeds.

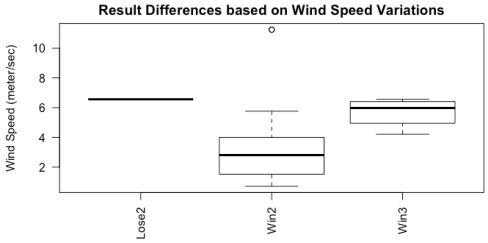
My results are as follows.

Overall Performance

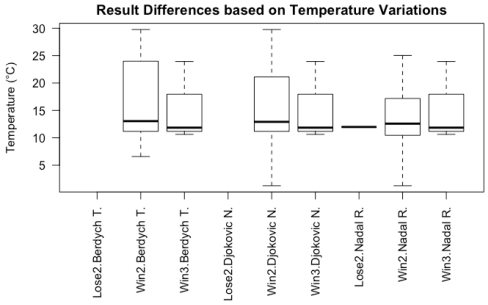


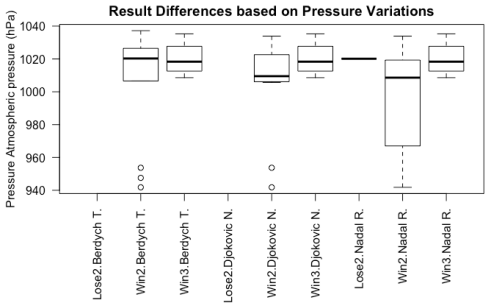


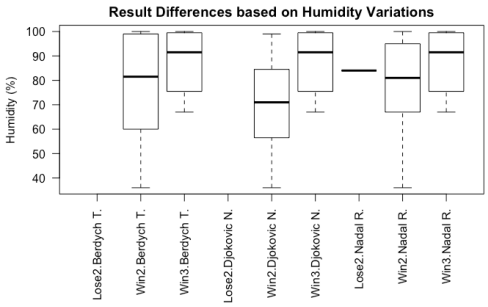


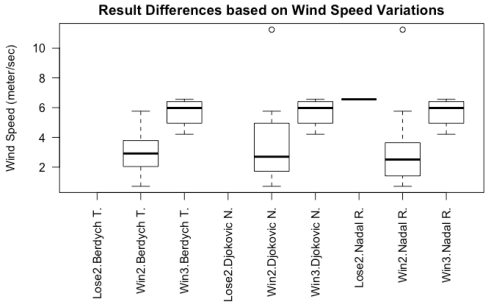


Player-specific Performance









First, they are apparently all top players who usually won by 2 or 3 sets, and losing in 2 sets seems like an outlier for them.

One interesting thing for the overall performance graph, winning in 3 sets seems to have a narrower range under every weather condition. There could be many interpretations, one might be that when players winning in 2 sets usually were in perfect conditions regardless of weather. For the wind speed part, it is apparent that slower wind speed helps players achieve quicker 2-set victory, and the reason might be with the serving accuracy.

The player-specific performance graph is significant in that it is like some kind of personal informatics but it is a weather-related and result-driven visualization. Such graphs could be beneficial for personal training and health management in knowing the suitable weather conditions for best performance. For example, Djokovic handles temperature really well, compared with its counterparts, while Nadal is good at a wide range of pressure.

Further analysis based on this project will be beneficial in providing more insights for enhancing tennis player performances, adjusting training activities and health management.

Reference

Exmerg http://www.exmerg.com/tennis-statistics/

Exmerg https://app.exmerg.com/

Tennis-data.co.uk http://www.tennis-data.co.uk/alldata.php

OpenWeatherMap http://openweathermap.org/current#parameter

Appendix: Player Frequency Distribution

Player	Frequency
Djokovic N.	88
Nadal R.	80
Berdych T.	79
Murray A.	78
Wawrinka S.	73
Federer R.	72
Bautista R.	70
Anderson K.	70
Ferrer D.	69
Nishikori K.	68
Isner J.	68
Tomic B.	64
Johnson S.	64
Sousa J.	64
Karlovic I.	63
Simon G.	63
Thiem D.	62
Gasquet R.	60
Troicki V.	60