Data Narrative Assignment - 3

Course : Probability, Statistics and Data Visualisation (ES-114)

Nimitt
Computer Science and Engineering
IIT Gandhinagar
Gandhinagar, India
nimitt.nimitt@iitgn.ac.in

***Abstract*—This document is a data narrative based on collection of datasets containing statistics of matches played in major tennis tournaments played in the year of 2013. It analyses the dataset to answer questions posed.**

***Index Terms*—Data Narrative**

I. Overview of the Dataset

The eight datasets consist of match statistics of matches played in four tennis tournament held in the year of 2013 namely :

- French Open 2013
- US Open 2013
- Australian Open 2013
- Wimbledon 2013

Each dataset has the information of following attributes for each match played in the respective tournament :

- Round in which match was played
- Match Results
- Set-wise Scores
- Service Point Statistics
- Break Point Statistics

II. Questions to be addressed

The dataset is multi-dimensional and consists varied, though concentrated, ideas. The following questions cover all the aspects of this dataset:

- What is impact of progression of tournament into further rounds on statistical performance of players ?
- What are the most common final score lines (number of sets won by each player in the match) and set score lines (number of games won by each player in the set) and why ?
- What is the correlation between statistical attributes given in the dataset and result of the matches and which attributes are more correlated ?
- What are the chances of comeback after loosing first few sets in a match ?
- What impact does service strike rate (probability of winning point on serve) has on the overall performance of the player ?
- How good is conversion rate (of converting break point to win) in determining the performance of the player ?
- How are performance of a player and double service faults the player makes related ?
- How are the chances of a player wining a match dependent on number of aces he scores ?

III. Details of Libraries and Functions

For this assignment, Python was used for the computations and analysis. Various libraries used in this project have been listed below:

- Pandas
- Numpy
- Matplotlib.pyplot
- Plotly

```
import pandas as pd
import numpy as np
import seaborn
import matplotlib.pyplot as plt
import plotly.graph_objects as go
```

Fig. 1. Libraries used

All the functions used are part of the Python's built-in functions package and libraries mentioned above. The below list enlists the functions used in this assignment:

- pandas.DataFrame()
- pandas.Series()
- pandas.read-csv()
- len()
- pandas.Series.describe()
- pandas.Series.value-counts()
- pandas.DataFrame.plot()
- pandas.DataFrame.where()
- matplotlib.pyplot.scatter()
- matplotlib.pyplot.bar()
- pandas.rename()
- pandas.head()
- pandas.merge()
- np.svd()
- plotly.express.scatter-3d()
- np.argsort()
- np.argpartition()

IV. Information about Tournaments

The tournaments covered by dataset include :

- French Open 2013
- US Open 2013
- Australian Open 2013
- Wimbledon 2013

All these tournaments have knockout format where each match loosing player gets eliminated. The following details more :

- Common name : Grand Slams
- Total Numbers of Players : 128
- Total Number of Rounds : 7
- Total Number of Matches : 127

Technical Terms involved:

- Ace for player 1 : Player 1 serves and player 2 can't make contact with the ball
- Winner for player 1 : Player 1 serves and ball bounces twice before reaching player 2
- Service Strike Rate : Probability of player wining point on service

V. Answers to Questions

A. Effect of Progression of tournament on statistical performance of players

Dataset : French Open Men 2013

Tennis is a very competitive sport. It is quite difficult to make into and win grand slams (the four major tennis tournaments). The fact that these tournaments are in knockout format increases the standards of playing higher. With progression of tournament into further rounds, the competence among players increases. This can studied by analyzing the French Open Men 2013, by trying cluster the matches played, based on statistics given for each match, according to the round in which they were played.// Approach :

- First, create data matrix by selecting data given in dataset but dropping attributes : Names of players, Result of the match, Round (Round is the target attribute for clustering)
- Using, principal component analysis extract important dimensions in data matrix
- Plot data points according to its class i.e. round which it was played in

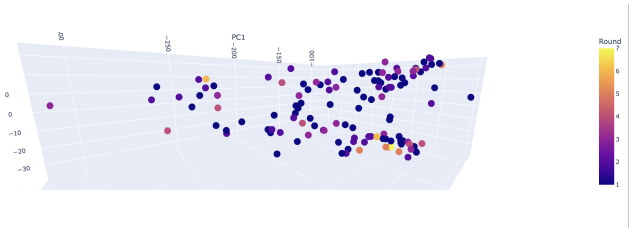

Fig. 2. Clustering according to Round

More visual analysis of the fig2, shows that matches played in rounds 5, 6 and 7 (yellowish tone) seem to cluster together. This indicates that certainly there is some variation in match conditions and statistics based on the round. This is probably due to the fact that knockout tournament imposes stiffer completion in later rounds as only stronger players are able to make in the final rounds. This can studied further by analyzing the distribution of statistical values after classifying data into initial and final rounds. Approach :

- First, divide the previously created data into initial (1st , 2nd and 3rd rounds) stage and final (5th, 6th and 7th)
- Take average values of data points for each attribute for both the initial and final stages
- Create a difference array storing the difference between attributes of initial and and final stage
- Create a ratio array storing factor by which mean values of attributes decrease for final stage
- Plot the difference array

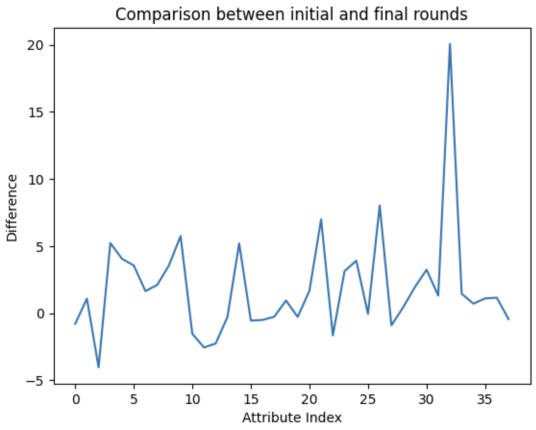

Fig. 3. Comparison of mean values for different attributes

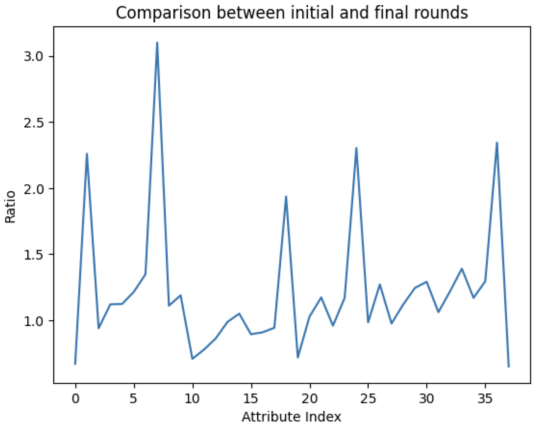

Fig. 4. Comparison of mean values for different attributes

Clearly, fig3 and fig4 shows that Final and Initial stages of the tournament differ greatly. The general trend shows that Final Rounds' matches have lower mean values for almost all attributes than the initial stage matches. The mean value of attributes for initial stage matches is actually 2-3 times than final stage matches. Further the significance of this can

be emphasised by looking at what attributes differ most. Approach :

- Get indices of attributes of maximum difference from the difference array
- Get the names of these attributes from the data matrix

```
frenchopen_men_features.keys()[indices]
✓ 0.2s
Index(['WNR.1', 'SSW.1', 'ACE.2', 'SSP.1', 'TPW.1', 'FSW.1', 'UFE.1', 'FSW.2',
       'WNR.2', 'TPW.2'],
      dtype='object')
```

Fig. 5. Attributes of maximum difference

Fig 5 shows that attributes which are most important in determining the result are:

- Number of Winners Scored by players
- Number of Aces scored by players
- Service Winning strike rates

From all the above discussion, it comes to conclusion that these above attributes have lower values in later rounds of the tournament i.e. the probabilities of a player scoring a winner, ace or making a winning serve in the later rounds are quite lower than the initial rounds. Above data shows that, it is actually 2 to 3 times tougher to score a ace or winner in final rounds than initial rounds. This thing goes for most of the attributes in the data matrix. This lays off quite interesting facts about the tournament :

- The difficulty to perform better in terms of statistical measures, show a high level of progressing competition as the tournament goes into final rounds.
- Final rounds are more intense and competitive than the initial rounds
- This can be explained by the format of the tournament : Knockout because it leads to swifter elimination of weaker players giving rise to immense competition in later rounds

In conclusion, the competition progressively increases as the tournament turns into final rounds.

B. Most Common Score Lines

Score line is actually the most important stat for a match. It becomes important to understand which score lines are common in a tournament and what are the possible reasons for this. Dataset used : Australian Open Men 2013
In each match, the player who wins most sets out of 5 wins match and player first winning 6 games with the lead of 2 games wins the set. So, there are two kinds of score lines :

- Final Score Line : Number of sets won each player in match
- Set Score Line : Number of games won each player in set

Beginning with set score line, the analysis of scorelines can be done by first extracting scorelines' data from the dataset and then plotting correspondingly. Approach :

- As the dataset contains only raw data of each player winning games in a set, first set score lines need to be constructed
- This can be done by concatenating set scores of each player together
- Then, distribution can be visualized by counting the frequency of each scoreline throughout the tournament and then plotting the frequency distribution

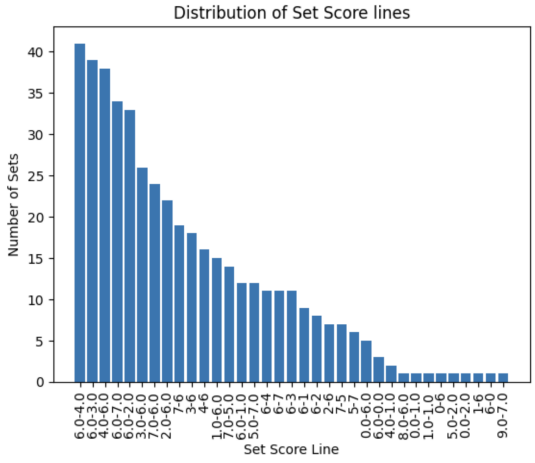

Fig. 6. Set score lines

Fig 6 has some interesting observations :

- The most common score line comes out to be "6-4"
- The ties "6-7" are not very common but happen frequently

So, it shows the fact that comebacks from being down some games in a set are frequent but in most of the cases these comebacks don't happen and the player with the lead wins. Now, it is a perfect time to analyze the match score lines. Approach :

- First, Final Score lines need to be extracted from the dataset
- Then, plot scorelines according to their frequencies

Fig 7 has some remarkable results. On analyzing the plot, it can be found that final scoreline "3-0" (considering 3-0 and 0-3 as identical as same for others) is most common scoreline in the tournament, followed by "3-1". But it can be observed that very few matches go to the fifth set for getting a result. This thing is a bit tricky to explain but analysis of distribution of score lines across different rounds is the key to this problem. The thing is that the tournament being a knockout tournament has more matches in the initial rounds and fewer matches in the later rounds. But we already saw that initial rounds have very less competition and generally have one sided matches. Both of these facts lead to the above observations. This can be studied more deeply by plotting of distribution of scorelines across rounds. Approach :

- Amending the data to consider scoreline "A-B" identical with "B-A"

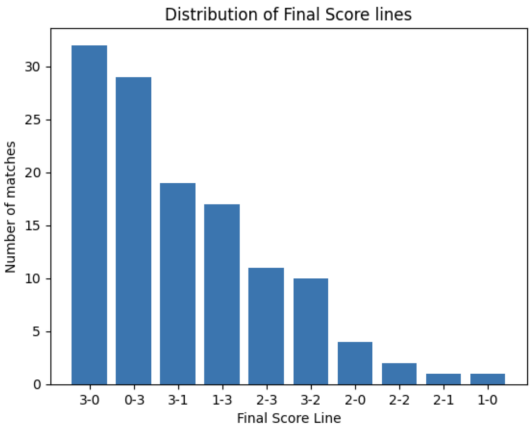

Fig. 7. Final Score lines

- First grouping final score lines according to the round
- Calculate the fraction of matches with each scoreline in each round
- Plot the data

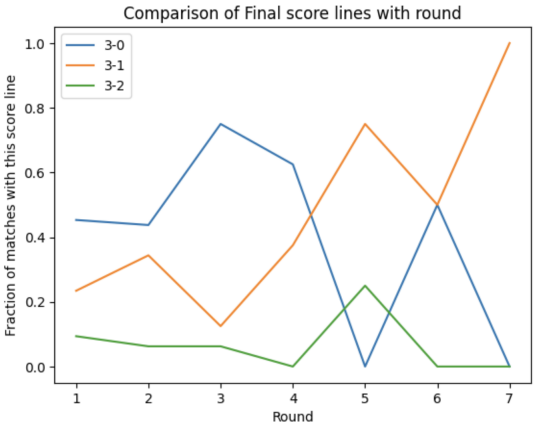

Fig. 8. Final Score lines across rounds

Now, Fig **??** shows that :

- Initial rounds have greater fraction of matches with score line "3-0" than the later matches
- But we already know that, tournament being a knockout has considerably more matches in the initial stages
- Due to this, the initial rounds have dominance in determining the frequency distribution of final score lines
- Hence "3-0" is the most common scoreline

C. Comeback rates after going down a few sets in a match

These tennis tournaments are quite intense and margin of errors is very low. It is easy to go down a few sets in a match but coming back in match again can be really tough. This section studies the comeback rate after going down a set in a match. Dataset : Australian Open 2013 The comeback rate can be calculated by checking in how many matches a player wins after loosing first set. Approach :

- First, creating an array to store which player won first set in each match
- Then, other create other array to store which player won the first match
- To calculate the number of matches, in which, the player winning first set wins the game see every element of these arrays and calculate common entries
- Plot this data

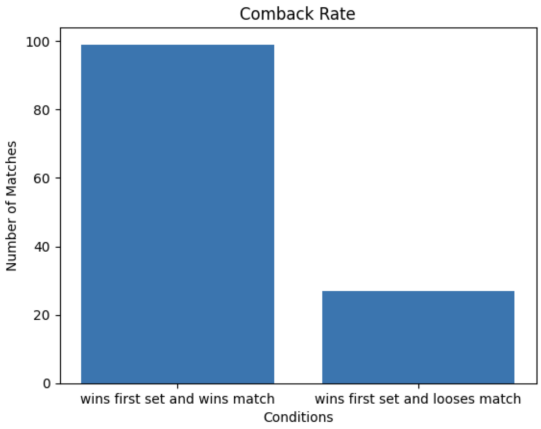

Fig. 9. Comeback rate

Clearly, Out of 127 matches 99 times the player winning first set wins the match. This shows that this many times the player losing first set loses the match. Fig 10 shows that the

```
print('Probability of Winnig a match after winning first set : ',count/(count_+count))
✓ 0.2s
Probability of Winnig a match after winning first set :  0.7857142857142857

print('Probability of Losing a match after loosing first set : ',1-count/(count_+count))
✓ 0.1s
Probability of Losing a match after loosing first set :  0.2142857142857143
```

Fig. 10. Comeback rate

probability of wining a game after losing first match is small approx. 0.2. In conclusion, in these tournaments the margin of errors is very minimal and chances of being able to comeback after loosing first few sets are minimal.

D. Correlation between match stats and results

The match stats given for each match in the dataset are quite important and seem to have very close influence on the results of the match. This section analyses this correlation. Dataset : French Open Women 2013 Visualizing multi-dimensional data can be bit tricky. This can be simplified with dimensionality reduction using principal components analysis and then visualizing with important components. Approach :

- First, create data matrix by dropping score statistical attributes as well as result(this is target) attributes.
- Using principal components analysis, extract major components from the dataset.

- Plotting each datapoint based on these components and target, i.e , result of the datapoint

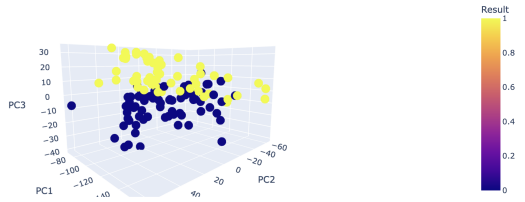

Fig. 11. Correlation between attributes and result

Clearly, matches with player 1 wining clustered together and matches with player 2 wining clustered together. This is important because even on removing all the stats related to scores of players were removed both classes separated from each other. This shows that there is certainly effect of attributes on result of the match. So, the result of the match can be easily predicted solely on the basis of these attributes. Thus, looking at the most influential attributes becomes important. This again can be done by singular value decomposition by extracting the influencing factor for each of the attributes. Approach :

- Use singular value decomposition, to train a linear regression model based X on data matrix and target
- Now, X represents the relative significance of each attribute in determining the result
- Use X to plot significance of each attribute

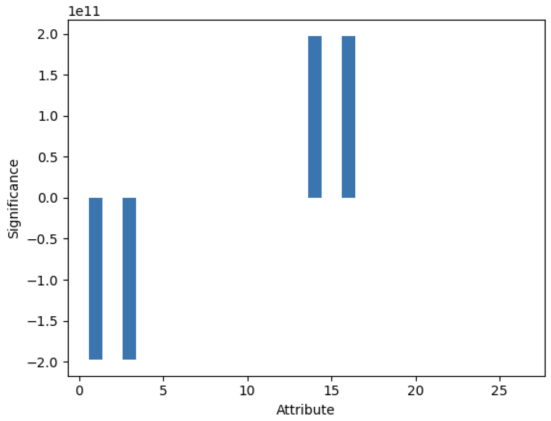

Fig. 12. Correlation between attributes and result

```
print(frenchopen_women_features.keys()[x_tilde.argsort()[-6:-1]])
print(frenchopen_women_features.keys()[x_tilde.argsort()[0:6]])
```

✓ 0.5s

```
Index(['NPW.1', 'ACE.1', 'TPW.1', 'BPW.1', 'SSP.2'], dtype='object')
Index(['SSP.1', 'FSP.1', 'BPC.2', 'BPC.1', 'ACE.2', 'SSW.2'], dtype='object')
```

Fig. 13. Correlation between attributes and result

The major influencing attributes come to be :

- Aces scored by players
- Service strike rates of the players

E. Effect of Service striker rates on performance

Previous sections showed that service strike rates are an important factor in determining the results of a match. This section explores effects of better striker rate on performance of a player

Dataset : US Open Men 2013 First, the study of raw first and second service rates is important:

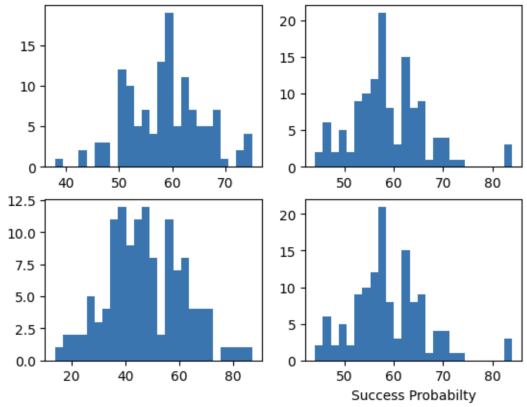

Fig. 14. First serve stats

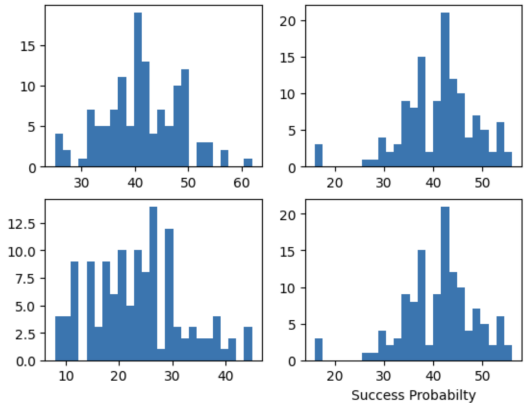

Fig. 15. second serve stats

- The second serve wining rates are smaller than the first serve wining rates
- This is due to the fact second serve fault results into a loss so player tend to serve more defensively in the second chance

The dataset has raw data of first and second serve percentages, but calculation of net service strike rate is necessary for further analysis. The following equation is used for calculating service strike rate.

$$StrikeRate = P(first) * P(win) + P(second) * P(win)$$

Approach :

- First, calculate strike rates for each player based on their aggregate performance in the tournament.

- Then, calculate the number of matches won by each of these players in the tournament which reflects their overall performance.
- Then, plot strike rates and number of matches won by each player

Fig. 16. Effect of strike rate on performance

The visualization in fig 16 as some interesting observations :

- Clearly, the players who won had higher strike rates than who lost in initial rounds but there seems anomalous response in the plot
- Interestingly, in the initial rounds players with higher strike rate also lost
- This is probably due to the fact that they have played against stronger players in the initial rounds.
- It also shows that performance depends on multiple factors rather than a single factor only

In conclusion, service strike rates play an important role in determining the match result but there are other factors also involved in determining the result.

F. Effect of double service faults

Double serve faults is a very good measure of a player's abilities. This section analyses how number of double serve faults influence the scoring statistics of a player. Dataset : Wimbledon 2013 The analysis can be done by aggregating number of matches each player won and average double service fault they made. Approach :

- For each player, calculate number of matches won and their respective average double service faults each match
- Then plot the corresponding data

It can be seen in fig 17 that players who won more matches have moderately values of average double service faults. Interestingly, this is very similar to fig 16. This thing can be explained by the fact that a player who has won more games has played more rounds in the tournament. Also, as the competition increases in later rounds the difficulty level rises and thus the players tend to make more mistakes which detiorates the player's performance stats. In conclusion, players wining more matches have moderately values of error making.

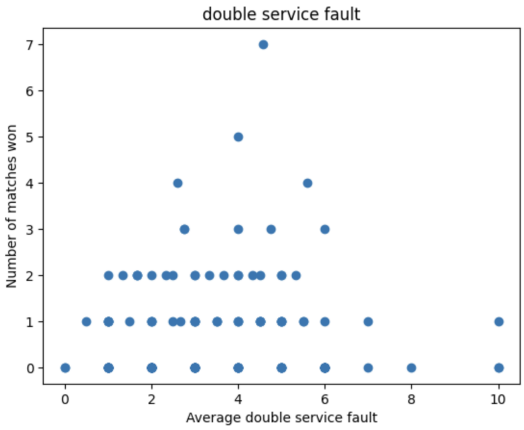

Fig. 17. Effect of strike rate on performance

G. Effectiveness of conversion rate in determining the performance of a player

In tennis break point is the most important moment in game. Converting break point in wining point is an important aspect. This section details the correlation between conversion rate and overall performance of the player. Approach :

- First, calculate the conversion rates for each of the players.
- Then, plot frequency distribution of

	Name of Player	Break points created	Break points won	Conversion rate
83	Na Li	64.0	35.0	0.546875
32	Dominika Cibulkova	60.0	34.0	0.566667
37	Eugenie Bouchard	60.0	31.0	0.516667
0	Agnieszka Radwanska	62.0	29.0	0.467742
7	Ana Ivanovic	51.0	27.0	0.529412
103	Simona Halep	46.0	24.0	0.521739
38	Flavia Pennetta	60.0	23.0	0.383333
119	Victoria Azarenka	49.0	22.0	0.448980
75	Maria Sharapova	52.0	22.0	0.423077
48	Jelena Jankovic	31.0	22.0	0.709677
12	Angelique Kerber	39.0	20.0	0.512821
41	Garbine Muguruza	36.0	18.0	0.500000
81	Monica Niculescu	28.0	18.0	0.642857
63	Kurumi Nara	27.0	17.0	0.629630
27	Casey Dellacqua	41.0	16.0	0.390244
26	Caroline Wozniacki	40.0	16.0	0.400000
72	Madison Keys	26.0	16.0	0.615385
34	Ekaterina Makarova	38.0	16.0	0.421053
21	Bojana Jovanovski	22.0	15.0	0.681818

Fig. 18. conversion rate

Most of the players have moderate conversion rate. Now, lets look at the players who made into the the latter rounds of the tournament. It can be seen that top players have comparable conversion rates of about 0.5. So, conversion rate is an important measure of a player's ability.

H. Effect of Number of aces on results

Previous sections that number of aces a player makes is an important attribute in determining his match result. Dataset : Wimbledon Women 2013 Approach :

- First, report the winners of each match

	Name of Player	Break points created	Break points won	Conversion rate
83	Na Li	64.0	35.0	0.546875
32	Dominika Cibulkova	60.0	34.0	0.566667
37	Eugenie Bouchard	60.0	31.0	0.516667
0	Agnieszka Radwanska	62.0	29.0	0.467742
7	Ana Ivanovic	51.0	27.0	0.529412
103	Simona Halep	46.0	24.0	0.521739
38	Flavia Pennetta	60.0	23.0	0.383333
119	Victoria Azarenka	49.0	22.0	0.448980
75	Maria Sharapova	52.0	22.0	0.423077
48	Jelena Jankovic	31.0	22.0	0.709677
12	Angelique Kerber	39.0	20.0	0.512821
41	Garbine Muguruza	36.0	18.0	0.500000
81	Monica Niculescu	28.0	18.0	0.642857
63	Kurumi Nara	27.0	17.0	0.629630
27	Casey Dellacqua	41.0	16.0	0.390244
26	Caroline Wozniacki	40.0	16.0	0.400000
72	Madison Keys	26.0	16.0	0.615385
34	Ekaterina Makarova	38.0	16.0	0.421053
21	Bojana Jovanovski	22.0	15.0	0.681818

Fig. 19. conversion rate

```
ausopen_women.where(ausopen_women['Round']==4).dropna(how='all')
```
✓ 0.3s

	Player1	Player2	Round	Result	FNL1	FNL2	FSP.1	FSW.1	S
112	Serena Williams	Ana Ivanovic	4.0	0.0	1.0	2.0	66.0	43.0	3
113	Casey Dellacqua	Eugenie Bouchard	4.0	0.0	1.0	2.0	61.0	33.0	3
114	Na Li	Ekaterina Makarova	4.0	1.0	2.0	0.0	80.0	26.0	2
115	Angelique Kerber	Flavia Pennetta	4.0	0.0	1.0	2.0	64.0	35.0	3
116	Jelena Jankovic	Simona Halep	4.0	0.0	1.0	2.0	67.0	32.0	3
117	Dominika Cibulkova	Maria Sharapova	4.0	1.0	2.0	1.0	69.0	27.0	1
118	Agnieszka Radwanska	Garbine Muguruza	4.0	1.0	2.0	0.0	73.0	34.0	1
119	Sloane Stephens	Victoria Azarenka	4.0	0.0	0.0	2.0	72.0	24.0	2

Fig. 20. Better players

- Then, collect for each match who scored more number of aces
- Now, iteratively see for each match is the player scoring more aces wins or not.
- Plot

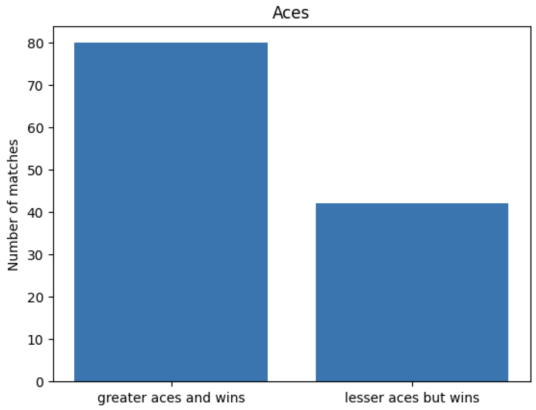

Fig. 21. Effect of number of aces in match result

Clearly, a player scoring more number of aces has greater probability of wining the match. But still the probability of wining after not scoring greater number of aces is also considerable. The match results are effected by overall performance of the player which is altered by multiple factors one of them being number of aces he scores in the match

VI. Summary of Observations

The analysis of this dataset concludes some very striking observations. Some of them are listed below:

- The progression of tournament into final rounds increases competition.
- Some of the score lines are more common in these tournaments. This is particularly due to format of the tournament i.e. knockout which promotes one-siders in initial rounds.
- Comeback rates after loosing firs few sets is very low in a match. This explains the stiff competition among players at this stage of play.
- The match stats given in the dataset effect the result of the match to a very great extent as the result of the match can be predicted solely on the basis of these stats.
- Some of the attributes especially were more correlated with the result of the match like service strike rate, number of aces made, number of winners, number of break points created.
- Top players tend to have moderately values of service strike rates, mean number of aces etc. This can be explained by the fact that progress of tournaments into final rounds increases competition which halts the performance stats of the top players.
-

VII. Acknowledgment

I would like to thank Prof. Shanmuga R for helping me through this assignment. He taught the concepts used for this assignment. His teachings have a great contribution in this assignment. My friends and the teaching assistants also shaped my ideas and thoughts to help me complete this assignment. I would like to thanks all of them.

VIII. References

References

[1] Pandas Documentation, Available: Pandas Documentation
[2] Numpy Documentation, Available: Numpy Documentation
[3] W3 Schools,W3 Schools website
[4] Dataset : UC Irvine ML Repository, Available: Dataset Used
[5] Stack Exschange, Stack Excahnge Site