CS-240 FINAL PROJECT REPORT

Ahmet Semih ORHAN

Prof. Mehmet Baysan

College of Engineering

İstanbul Şehir University

03/06/2018

Part A)

Questions of Interest:

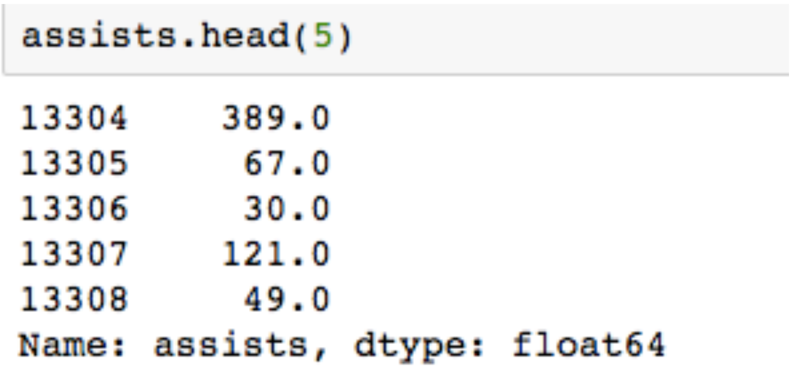
1. Is there a relationship between steals and assists?
2. Are the players who has high number of steals have high number of assists?
3. Players who stayed longer in the game scored more points.

1) If a basketball player has a high number of steals then the player has high number of assists.

We assume that the player that does a lot of steals is probably the point guard(which controls the game and rest of the team). He is more likely to find the player that is not defended by the opponent team therefor he is more likely to make assists. We wonder whether our assumption can be proven statistically by the dataset.

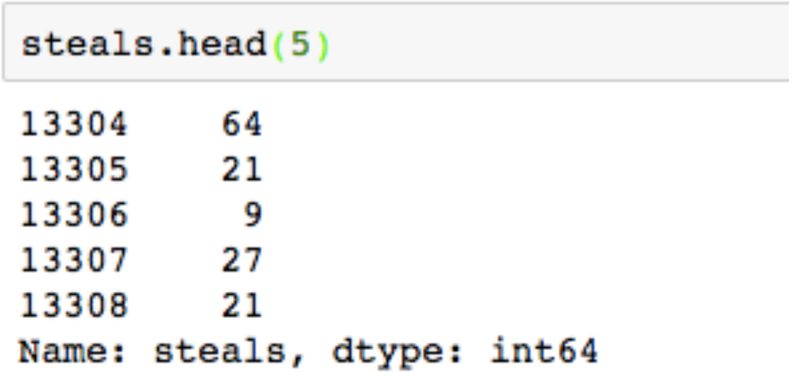
2) My variables are “assists” column, “steals” column, “lgID” column and “year” column.

I take only players that playing in the NBA league and also played in a certain year. So I filtered the dataset by column name “year” and “lgID”. Then I drop the NaN and Infinity values in order to clean my dataset. Results are more



```
assists.head(5)
```

```
13304    389.0
13305     67.0
13306     30.0
13307    121.0
13308     49.0
Name: assists, dtype: float64
```

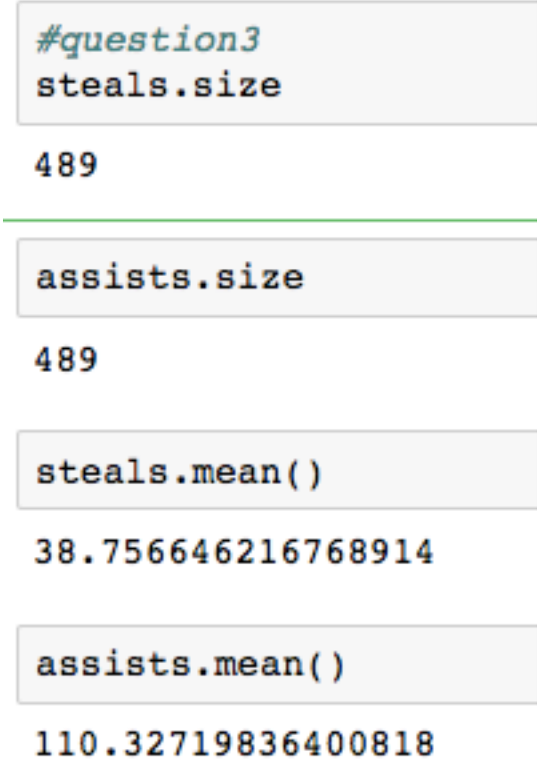


```
steals.head(5)
```

```
13304    64
13305    21
13306     9
13307    27
13308    21
Name: steals, dtype: int64
```

solid for the analysis after this step. Before I clean and filter my dataset, I faced that some of the functions like SpearsmanCorr, PMF or CDF are giving errors or NaN results.

3) In order to further analyze my dataset, firstly, I have looked at the sizes and means of my variables.



```
#question3
steals.size
```

```
489
```

```
assists.size
```

```
489
```

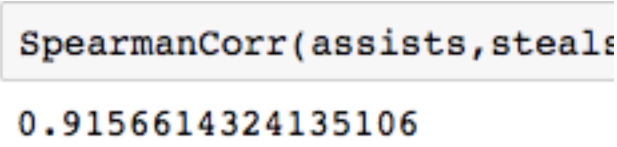
```
steals.mean()
```

```
38.756646216768914
```

```
assists.mean()
```

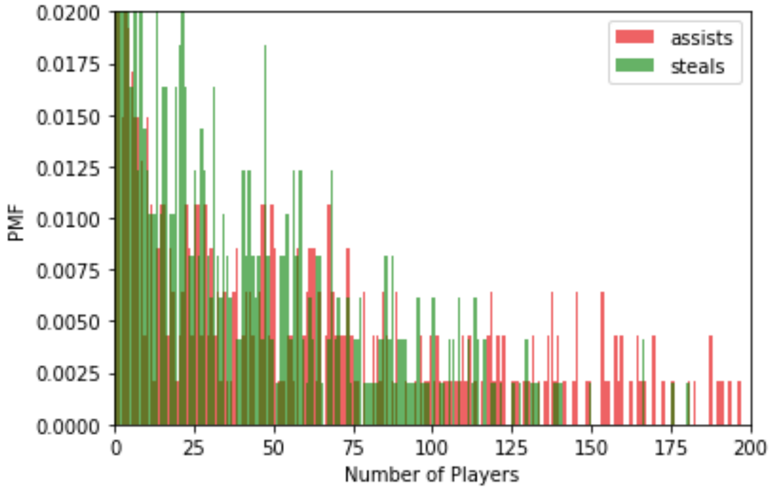
```
110.32719836400818
```

Then I have checked whether there is a correlation between steals and assists columns by using Spearsman Correlation. I have concluded that there is a correlation between. While steals increases, assists also increase. And my correlation result is very high So there may actually be a relationship between them.

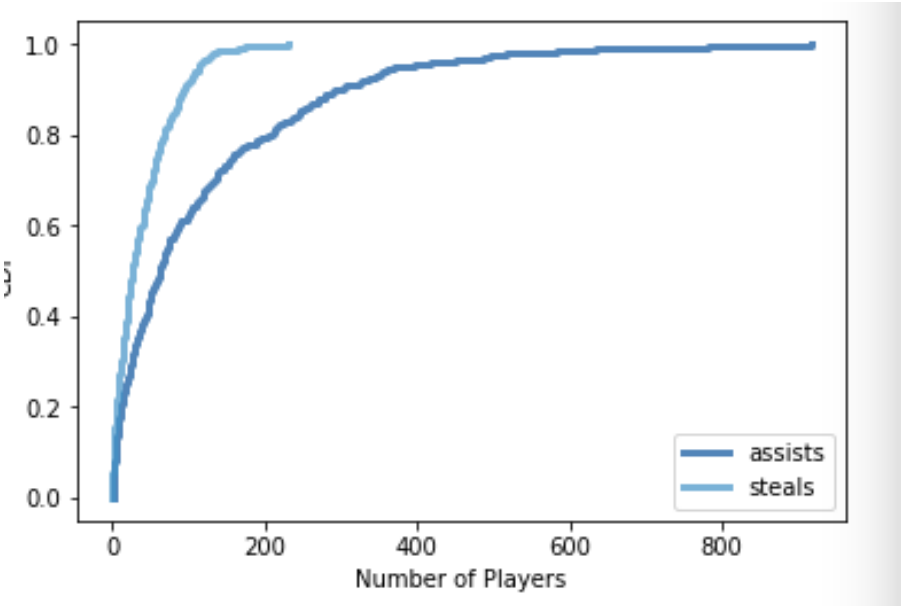


```
SpearmanCorr(assists,steals
```

```
0.9156614324135106
```

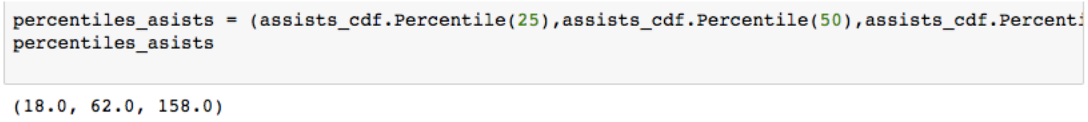


Then I have ploted a histogram to see my values in a visiual form. To analyse wheter there is a similar pattern between them. By looking at my histogram I can say that steals and assists values follow a similar pattern.



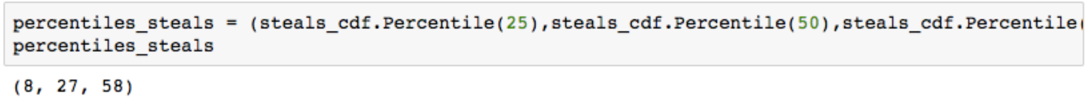
For this part lastly, I have drawn a CDF graph to compare the cumulative distribution of steals and assists and steals. By looking at my graph I can tell that steals occur less because it is harder to steal the ball from the other team than to make an assist. However they follow a similar pattern.

4) In this part, I have modeled my data using percentile ranks.



```
percentiles_asists = (assists_cdf.Percentile(25),assists_cdf.Percentile(50),assists_cdf.Percenti
percentiles_asists
```

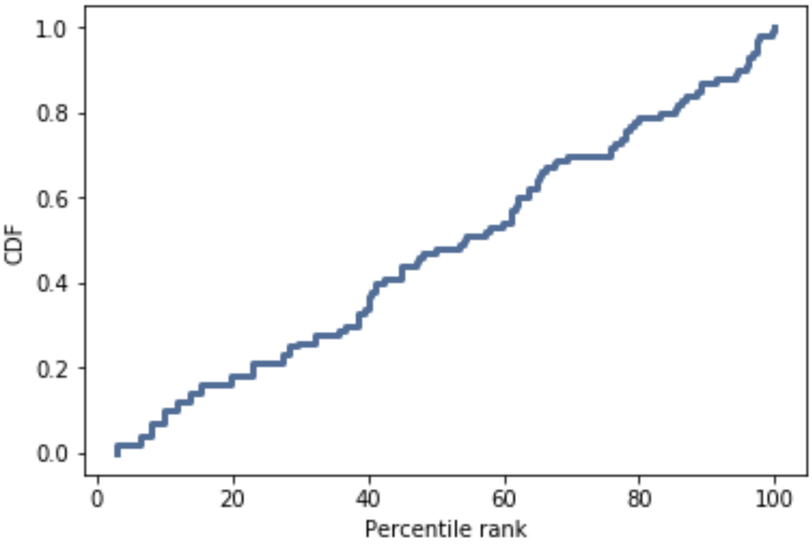
```
(18.0, 62.0, 158.0)
```



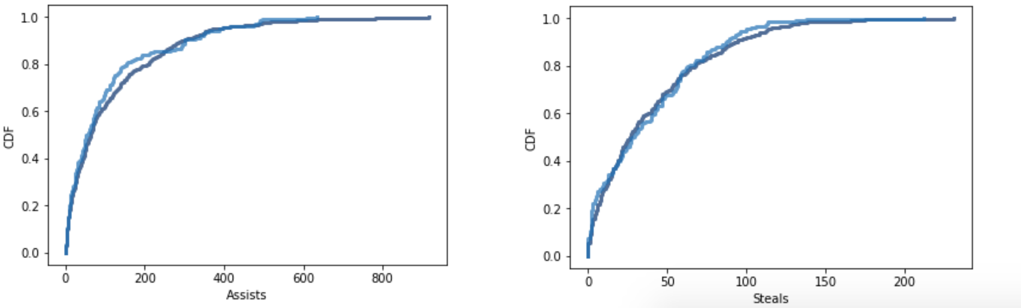
```
percentiles_steals = (steals_cdf.Percentile(25),steals_cdf.Percentile(50),steals_cdf.Percentile(
percentiles_steals
```

```
(8, 27, 58)
```

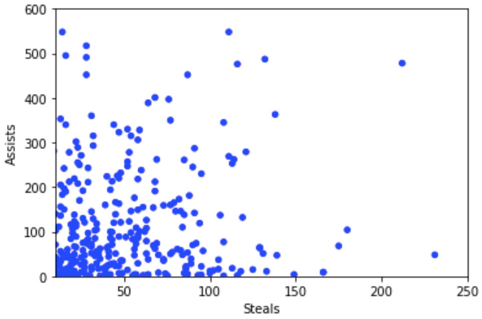
Then I draw a random sample from the observed assists and map each assist to its percentile rank.



Then I have generated random samles from the assitst and steals and compare the actual CDF from sample to actual data. I observed that samples do fit the actual model.



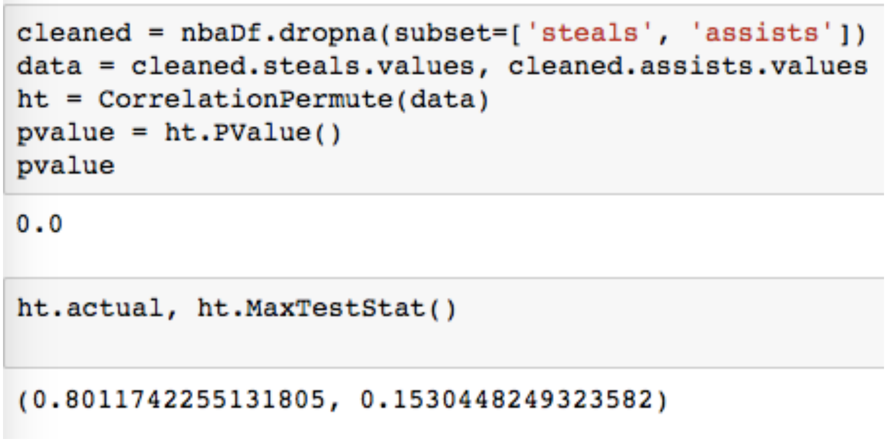
5) According to my question there is a relationship between steals and assist. Steals is assumed to be indipended variable wheras assists depends on the steals. Here is the visual of my scatered values from the assists and steals.



6) My null hypothesis:

H0= There is no relationship between the steals and assists.

Test statistics= I have used my correlation to calculate my test statistics. If the test statistics returns a high value then the my p-value is gets lower. In this case my test statistic results in a high value.



```
cleaned = nbaDf.dropna(subset=['steals', 'assists'])
data = cleaned.steals.values, cleaned.assists.values
ht = CorrelationPermute(data)
pvalue = ht.PValue()
pvalue
```

```
0.0
```

```
ht.actual, ht.MaxTestStat()
```

```
(0.8011742255131805, 0.1530448249323582)
```

P-value & significance= Since my p value is low then it has a statistical significance. So there is significant evidence to reject the null hypothesis.

7) I have learned that there is a relationship between the steals and assists as I can infer it from my correlation value being very high. Also it seems to be harder to make a steal than to make an assists.