İSTANBUL
ŞEHİR
ÜNİVERSİTESİ

CS-240 FINAL PROJECT REPORT

Abdullah Furkan KAYA

211593808

Prof. Mehmet BAYSAN

College of Engineering

İstanbul Şehir University

03/06/2018

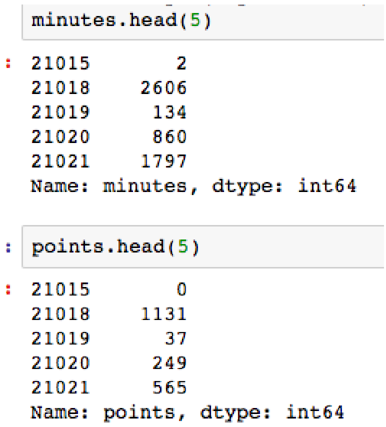
Part A

Possible questions:

- Is there a relationship between minutes and points?
- Are the players who has high number of rebounds have high number of assists?
- Are players who have high 2 point rates always have high 3 point rates?

1) From 2010, I filtered minutes and points. If minutes and points always goes hand in hand? We assume that more minutes are more opportunity for more points. We wonder whether our assumption can be proven statistically by the dataset.

2) My variables are “points” column, “minutes” column, and “year” column for filtering.

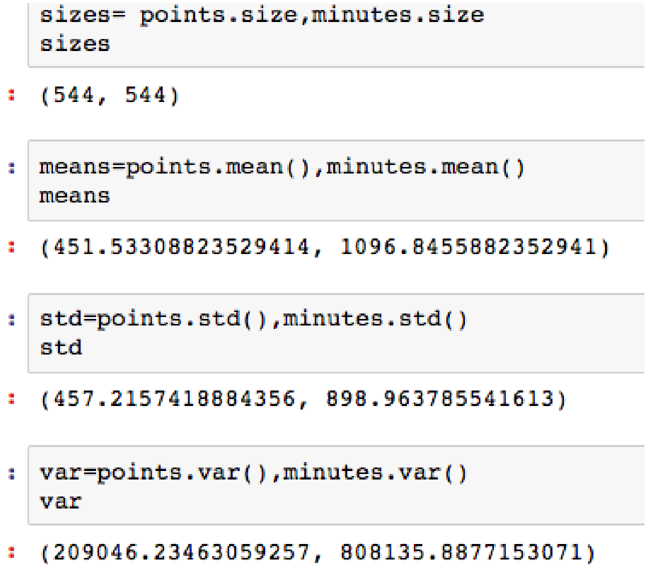


```
  minutes.head(5)

: 21015       2
  21018    2606
  21019     134
  21020     860
  21021    1797
  Name: minutes, dtype: int64

: points.head(5)

: 21015       0
  21018    1131
  21019      37
  21020     249
  21021     565
  Name: points, dtype: int64
```

I take only players that played in year 2010. So I filtered the dataset by column name “year”. Then I drop the NaN and Infinity values in order to clean my dataset. Then I was I able get better results from my correlation algorithm.

3) In order to further analyze my dataset, first I have looked at the sizes and means,

variance, standard deviation of my variables.



```
sizes= points.size,minutes.size
sizes
```

```
(544, 544)
```

```
means=points.mean(),minutes.mean()
means
```

```
(451.53308823529414, 1096.8455882352941)
```

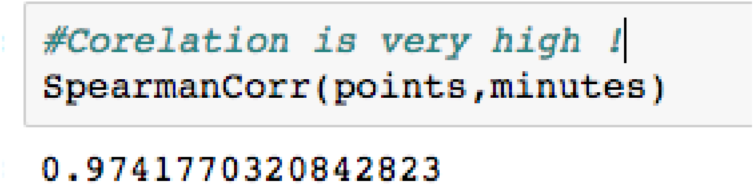
```
std=points.std(),minutes.std()
std
```

```
(457.2157418884356, 898.963785541613)
```

```
var=points.var(),minutes.var()
var
```

```
(209046.23463059257, 808135.8877153071)
```

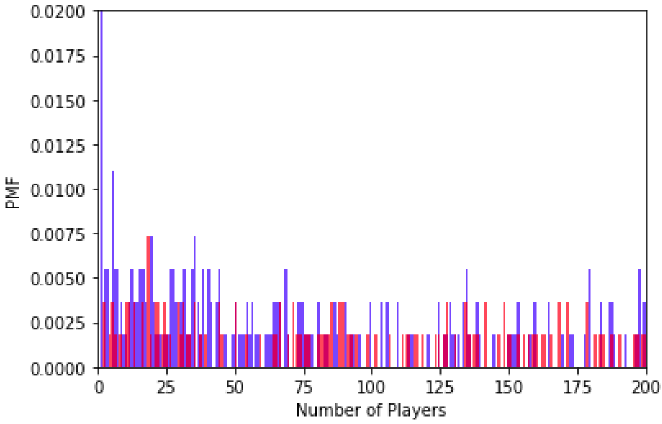
By using spearsman correlation I have calculated the correlation between them. According to my results shown below. There is a very strong correlation between point and minutes. But correlation does not mean there is a direct relationship between my variable. So I have to further analyse my data.



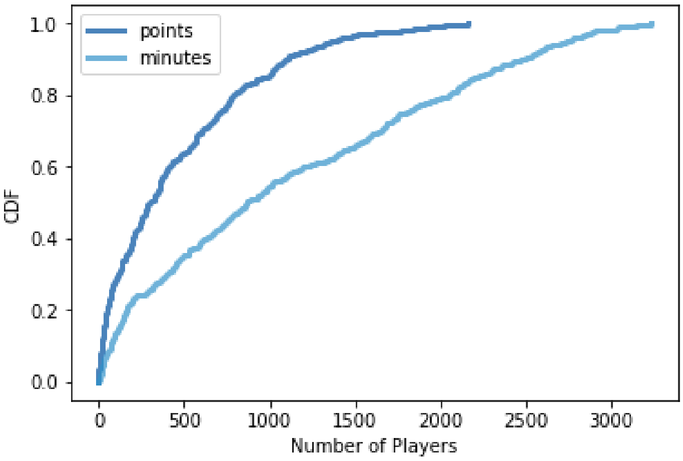
```
#Corelation is very high !
SpearmanCorr(points,minutes)
```

```
0.9741770320842823
```

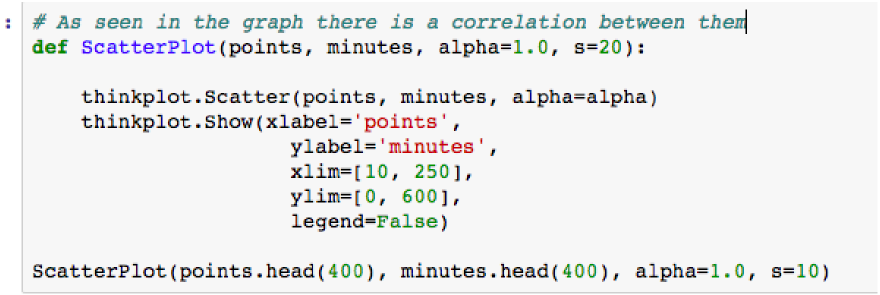
Another form of descriptive statistic is histogram. I plotted histogram to visually compare my values.



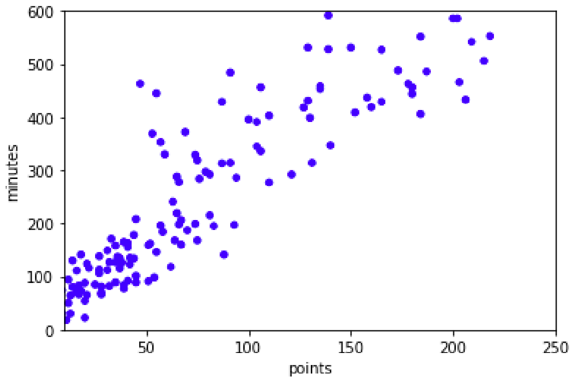
Last of my descriptive statistics is cumulative distribution. From my graph I can infer that there is similarity in the pattern.



5) My main assumption was that there is a relationship between the points that a players makes and how long does he stays in the game. To show that I have plotted a graph shown below.



```
: # As seen in the graph there is a correlation between them
  def ScatterPlot(points, minutes, alpha=1.0, s=20):

      thinkplot.Scatter(points, minutes, alpha=alpha)
      thinkplot.Show(xlabel='points',
                     ylabel='minutes',
                     xlim=[10, 250],
                     ylim=[0, 600],
                     legend=False)

  ScatterPlot(points.head(400), minutes.head(400), alpha=1.0, s=10)
```

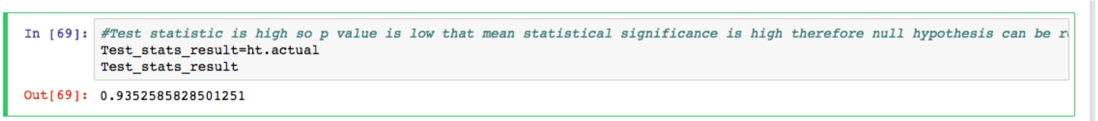


6) My null hypothesis: Points and minutes are not correlated.

Test statistics= Correlation result between points and minutes is used to determine the test statistics. My correlation value is really high so I can say that my test statistics is also has a high value.

P-value: has a negative relationship with test statistics. Because test statistics is high then p value is low.

Significance: Low p value mean that it is statistically significance and we can reject the null hypothesis



```
In [69]: #Test statistic is high so p value is low that mean statistical significance is high therefore null hypothesis can be r
         Test_stats_result=ht.actual
         Test_stats_result

Out[69]: 0.9352585828501251
```

7) As we have seen in test result that there is very strong relationship between the points that was made by a player and minutes he stayed in the game.