KUTAY CEM ACAR

215251070

Exploratory Data Analysis Report

Section 1:

My Questions

1) Does age effects player's minutes per a game?
2) If a coach has coach of year award does it mean the coach's team wins a lot games?
3) Is there a relationship between minutes played and player's score?

I will choose question 3 for my hypothesis. By visualizing data and showing relationship between the 2 variables, I will be able to provide a yes/no

answer to the question. Ideally speaking, more minutes played makes more points. The correlation of the variables will show us the results.

Section 2:

I will use basketball_master.csv. It has minutes and points

- My variables will be player's minutes in and player's points
- **For reading csv file**(the file has to be same path with Jupyter notebook or we have to give full address)
- In dataset, the variable minutes has a lot 0's because minutes was not recorded in old years. To eliminate 0's I use filter with lambda filter with apply(filter).

Section 3:

- First I found mean, standard deviation, variance, minimum value and maximum value of minutes and points. Average of NBA player's minute played is 1242.81 points. Average of NBA player's points of a season is 534.96 points. Most minutes played in one season in the data is 3882 minutes, basically 64.7 hours in one season. Best points value in our data 4029 points for a season.

  All the statistics gathered:

  Mean Of Minutes= 1242.81797806

  Mean Of Points= 534.965045303

  Standard Deviation Of Minutes= 968.337482692

  Standard Deviation Of Points= 513.892497367

  Variance Of Minutes= 937677.480386

  Variance Of Points= 264085.49885

Maximum Value Of Minutes= 3882

Maximum Value Of Points= 4029

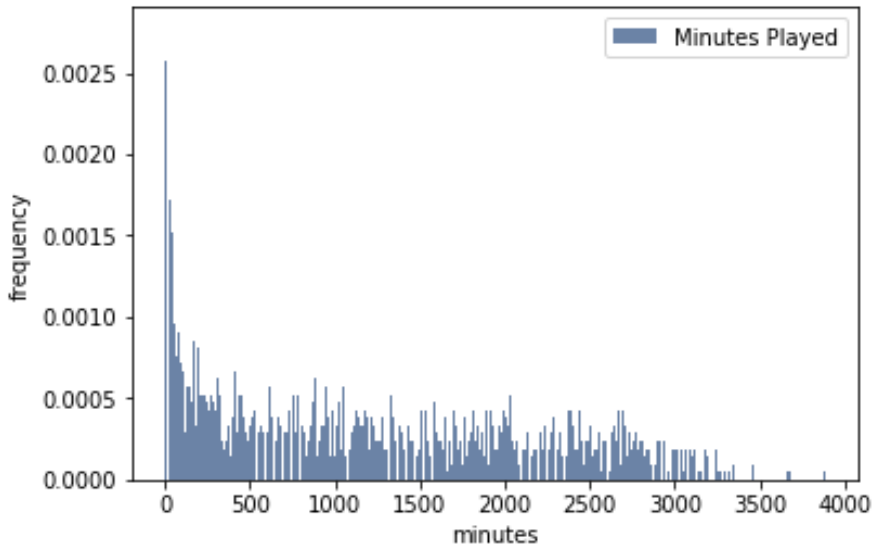
Minimum Value Of Minutes= 1

Minimum Value Of Points= 0

⇨ Then I show the histogram, PMF, and CDF of minutes and points:

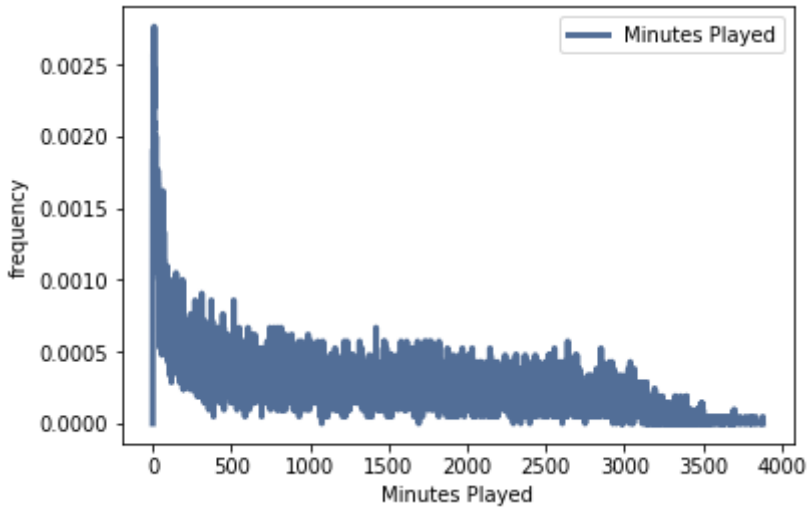
For minutes:

Histogram



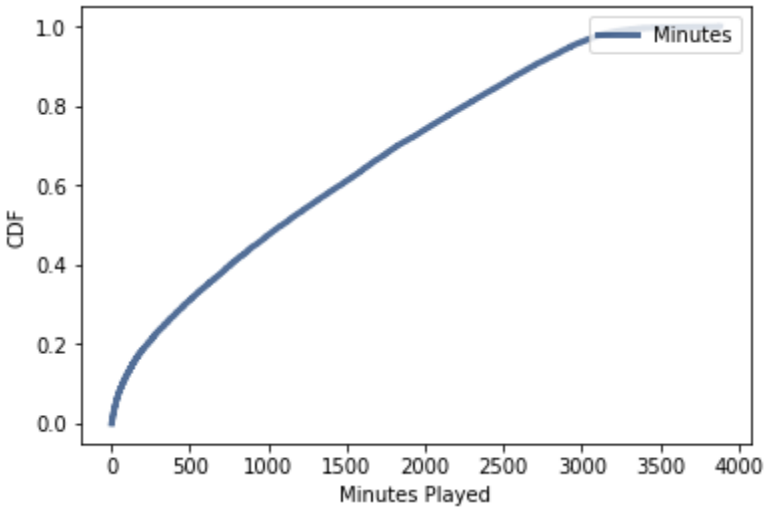
The histogram looks exponential

PMF



The PMF also looks exponential

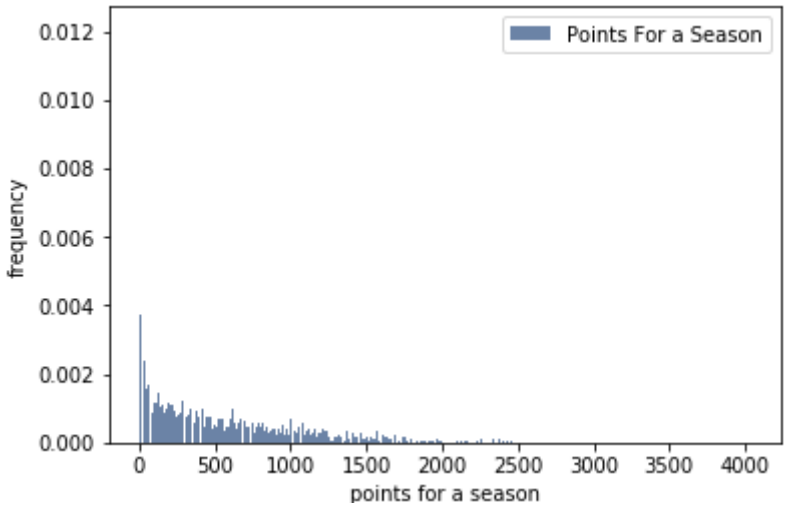
CDF



The CDF, finally, confirms that minutes is exponentially distributed.

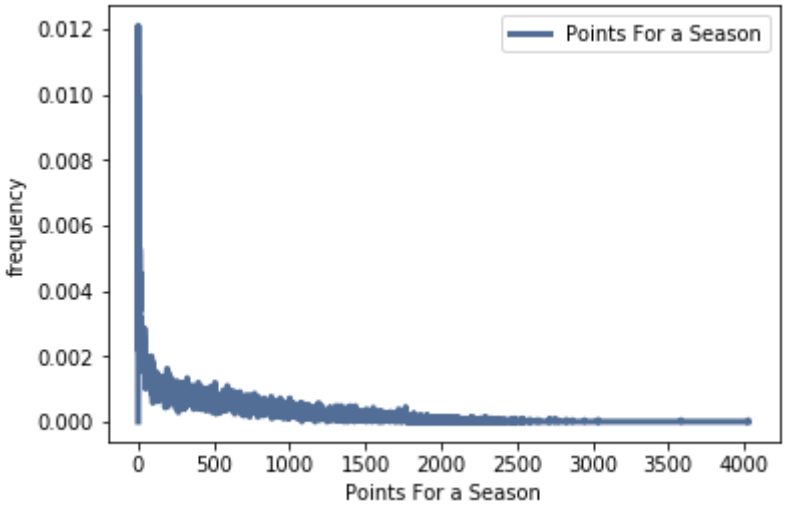
For Points:

Histogram



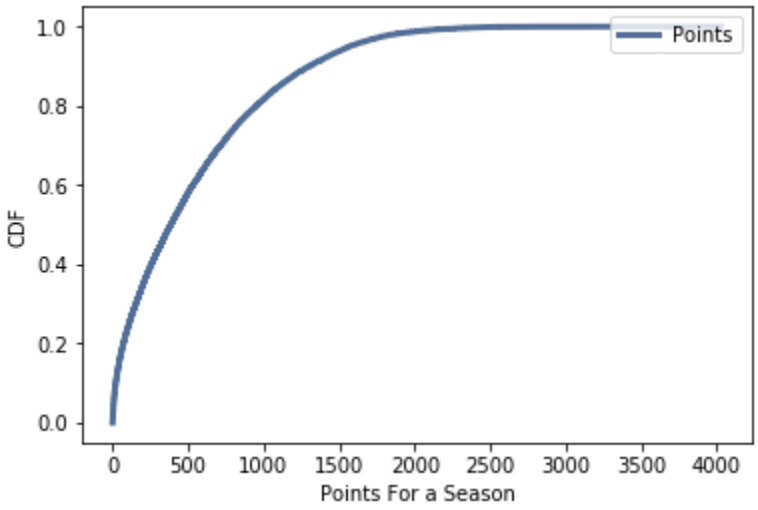
The histogram of points also seems to be exponential.

PMF:



The PMF of points look close to exponential.

CDF:

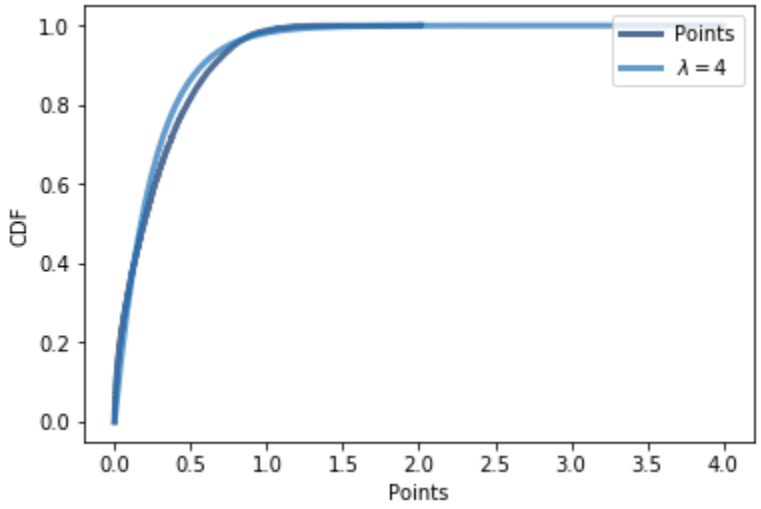


The CDF of points almost looks perfectly exponential. As shown by plots, both variables are exponentially distributed.

Section 4:

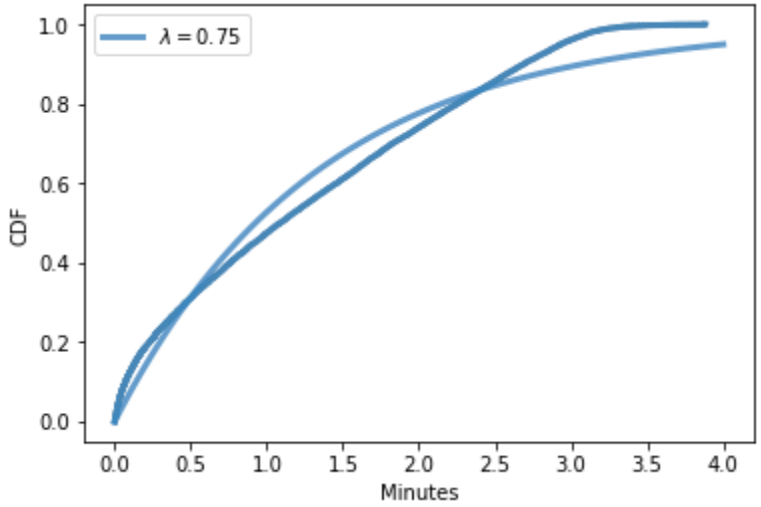
To model the data, I create 2 exponential CDF'S with lambda 4 and 3 for minutes and points and plot over original CDF's created by data:

For minutes:



The model fits data. Models summarize data because all we need now is just exponential function with lambda 4 and we have almost all data perfectly.

For points:

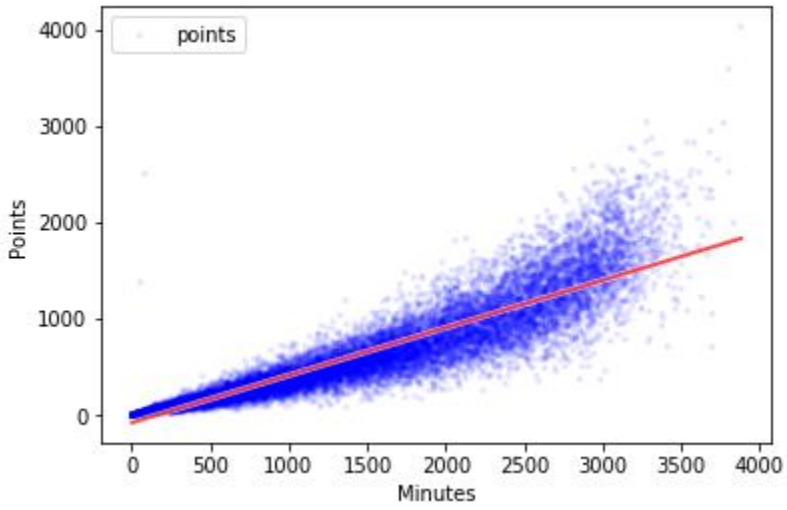


Exponential model with lambda 0.75 models the variable in a good way in many parts.

Section 5:

In this part I will show correlation between minutes and points according to the data. This will give a very important perspective to our results. To calculate Correlation I use the predefined functions in thinkstats2. The correlation calculated is 0.926, almost a perfect correlation. This means that if a player plays more minutes he is more likely to score more points. The reasons a player plays more in a season is known in basketball clubs: They are more valuable. Valuable players can change a team performance completely. This result confirms it.

I will also scatter the data points and draw linear least squares line to show the correlation:



Scatter plot shows the correlation in a good way. As player plays more minutes, points increase. Also good to notice the Right-Skewness of the points. Meaning that, for 0 to 1500 minutes, data points seem linearly increasing but when looking at the right tail of the plot the points start growing exponentially not linearly. Which means that if it was physically possible to play for example 10,000 minutes in a season, the points will grow exponentially, as shown by scatter plot.

Section 6:

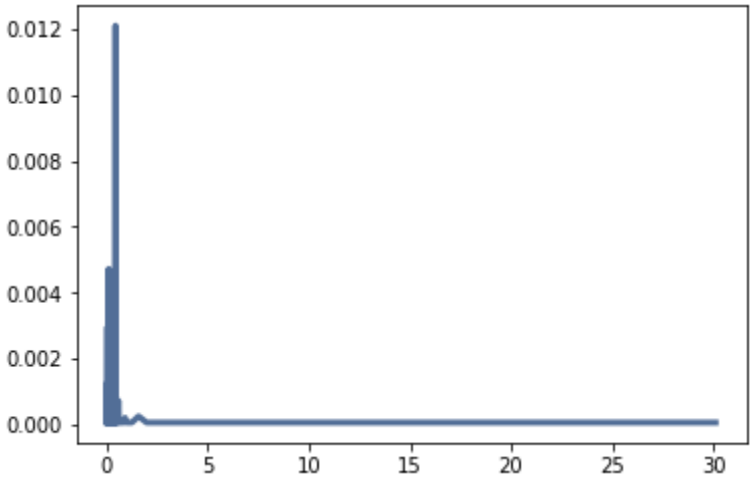
IN this part I chose

Null Hypothesis: No effect on score by minutes, and

Alternative Hypothesis: Minutes affect score.

My test statistic: absolute(points_mean*minutes_mean-x_point*x_minute)

Then I plot the test statistic result with probability. We get this plot.



Section 7:

In conclusion, as shown 2 variables minute per season and points per season are correlated. Average of NBA player's minute played is 1242.81 points. Average of NBA player's points of a season is 534.96 points. Most minutes played in one season in the data is 3882 minutes, basically 64.7 hours in one season. Best points value in our data 4029 points for a season. We also found the 2 variables to be exponentially distributed and can be modeled with lambdas 4 and 0.75. We found the correlation to be 0. 926 which mean the 2 variables are much correlated. More minutes played means more points, generally. Finally, there is Right-Skewness of the points. Meaning that, for 0 to 1500 minutes, data points seem linearly increasing but when looking at the

right tail of the plot the points start growing exponentially not linearly. Which means that if it was physically possible to play for example 10,000 minutes in a season, the points will grow exponentially, as shown by scatter plot.