İSTANBUL ŞEHİR ÜNİVERSİTESİ

Exploratory Data Analysis

CS 240 Project Report

Ahmet Öztemiz - 213010595

03.06.2018

Abstract

In this report, I am planning to create 3 question to analyze based on basketball datasets and also, i will interpret outgoing results about these datasets.

Section 1-)

There are 11 different datasets about basketball and these datasets have informations about teams, players, coaches, awards and abbreviations. I decided to create 3 different question after that i analyzed these datasets.

- Do taller basketball players have more scoring rate than shorter basketball players? How does the height average affects teams winning?
- What is the importance of coaches 's age in terms of the their success? Is the age factor increase the likelihood of winning a award?
- Is there any crucial effect of basketball player's race on points and awards? How does race affects playing better in a game and winning awards?

For the first question, we can say that players length have crucial role on playing basketball. Also, we can easily observe that most of the basketball teams have very high height average based on their basketball players so we can say that height have positive relationship with scoring rate per player. There are some specific personal informations given in the basketball_master.csv dataset and also, we can find player informations in basketball_players.csv dataset.

In the second question, we are looking some correlation between won awards and certain age range. I can say that there will be positive relation between age and award because when coaches gain experience then they will win more game and winning award will be easier than inexperienced coaches. Most of the famous basketball teams and national basketball teams entrust older coaches for to win awards. Basketball_awards_coaches.csv and basketball_coaches.csv datasets could provide us required informations about coaches.

Lastly, i will analyze relationship between scoring rate and awards based on race of player. General aspect of race in basketball, people thinks that black people can play this game better than white people so mostly black people win awards about basketball because most of the black people have height and muscle opportunity than white people and also, in the NBA most of the famous basketball player's race is black. Therefore, we can say that black people dominate white people in playing basketball. Also, there should be positive relationship between number of awards and

scoring rate. In the basketball_master.csv, basketball_awards_players.csv and basketball_players.csv datasets player's race, awards and points of players given as a column so we can use these columns to create our datasets for analyzing our hypothesis.

I will work on third question which analyzes relationship between scoring rate and awards based on race and also, i will expect to see positive relationship our variables. Due to the fact that my null hypothesis claims that there is no significance between awards and scoring rate based on any race so being white or black does not have any importance on scoring or awards.

Section 2-)

I will use basketball_master.csv, basketball_players.csv and basketball_awards_players.csv datasets to analyze my hypothesis. In the basketball_master.csv file we can reach race of players and we can find total points of players in the basketball_players.csv file then we can find informations about awards from basketball_awards_players.csv file. But, we need to clear our dataset so we will find NaN values to remove in our data frame then we need to remove additional race type such as, '1' and 'O' because it has not any useful effect for our hypothesis. Also, there could be more than one row for each player in awards and points columns so we counted number of awards and points for each player. After that we found race of players than we merged all of them in same dataset and created separated two dataset based on their race which are named as white_df and black_df. After cleaning data, we reached 2460 black and 2431 white player. Following schema shows our new dataframe.I used bioID and playerID to match awards, points and race of players.

	playerID	award	points	race
0	abdulka01	35	38387	B
1	abdulma02	1	8553	B
2	adamsal01	1	13910	W
3	aingeda01	1	11964	W
4	aldrila01	1	7738	B

Section 3-)

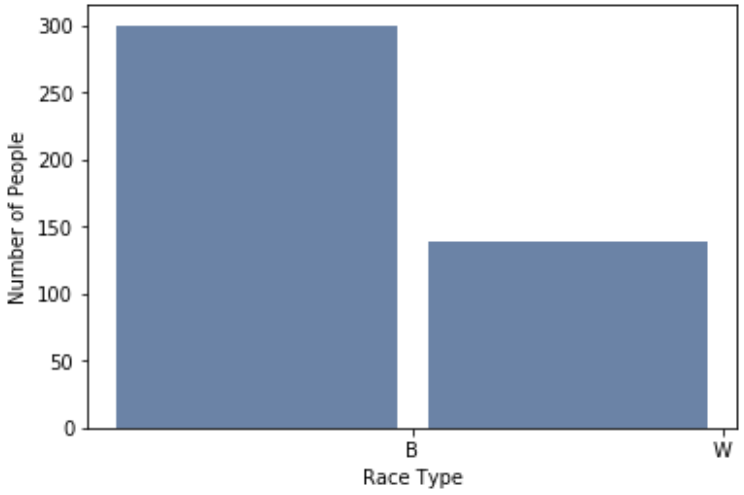
5 descriptive statistics for points column.

count	**439.000000**
mean	**10555.268793**
std	**7026.352579**
min	**125.000000**
max	**38387.000000**

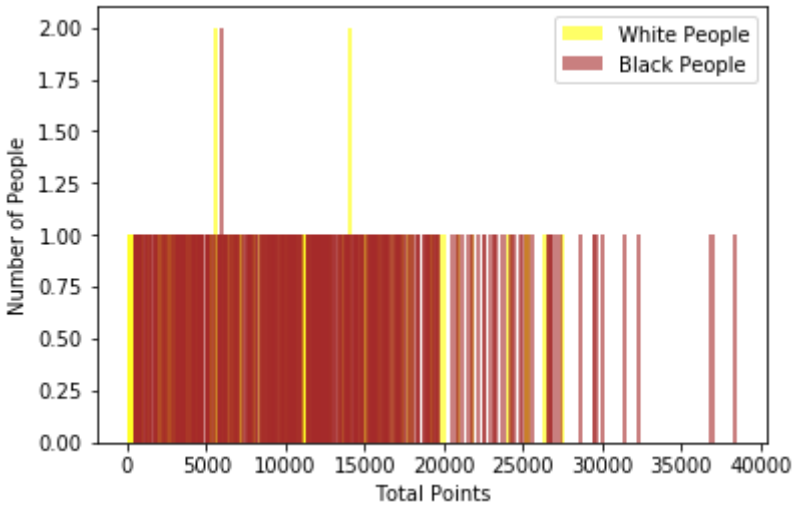
5 descriptive statistics for awards column.

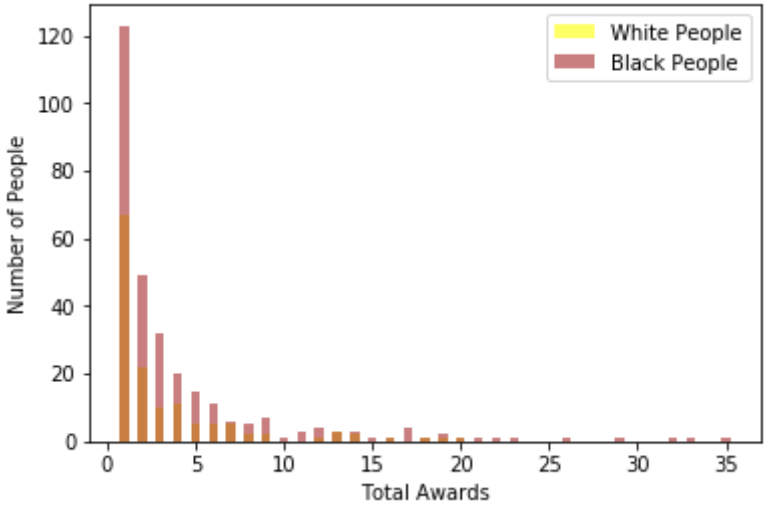
count	**439.000000**
mean	**3.899772**
std	**5.095792**
min	**1.000000**
max	**35.000000**

We can see the distribution of race in following graph. There are not too much player because the limited number of player won awards. After matching playerID with awards, we lose too much number of players in out dataframe. We divided our dataframe sd for that their race.

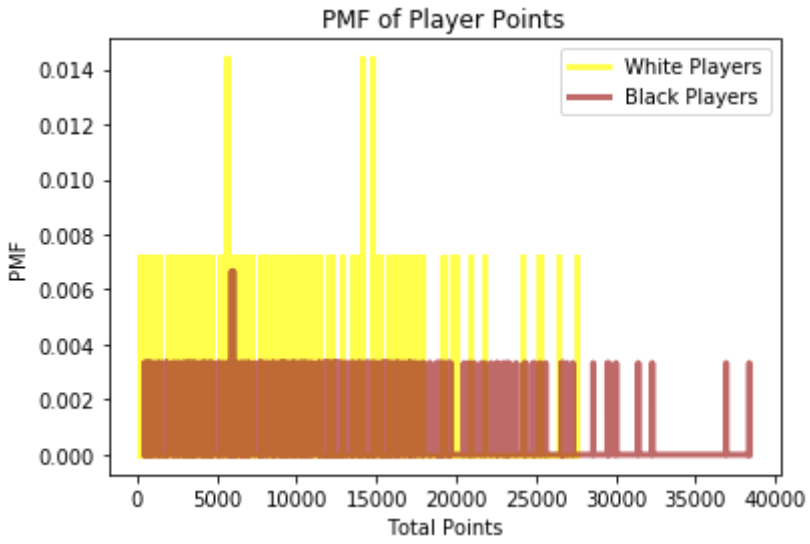


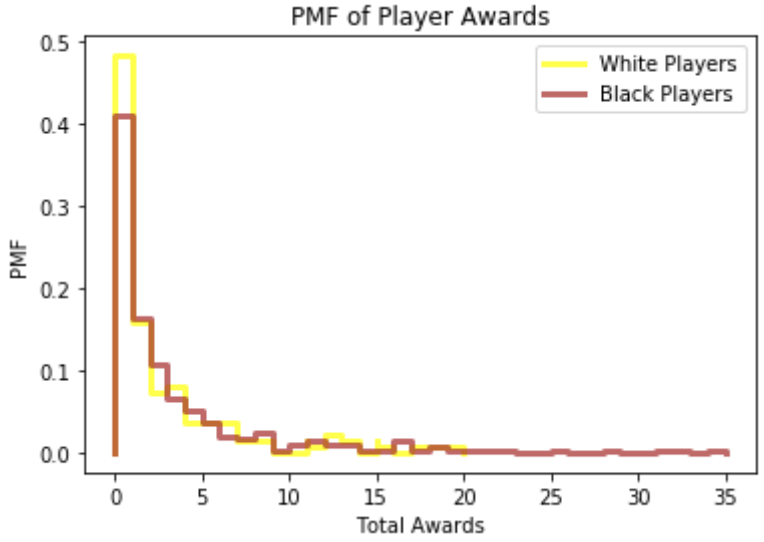
In the following histograms, we can analyze total points and awards for based on number of people for white and black race.



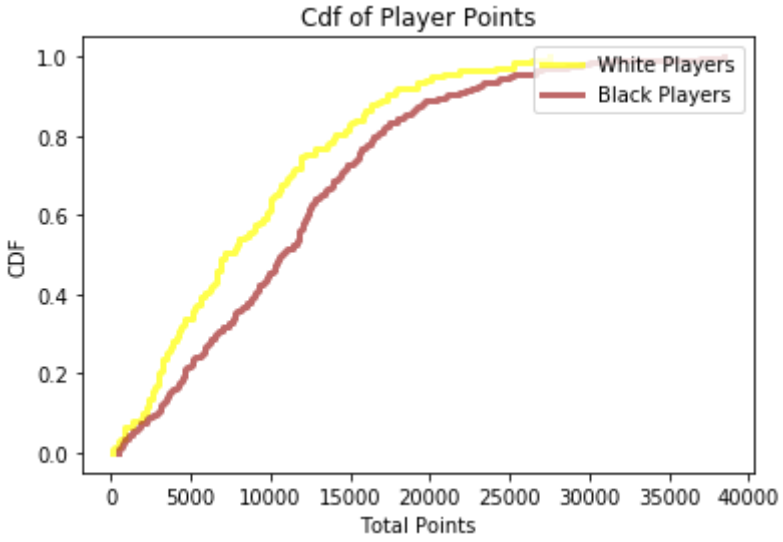


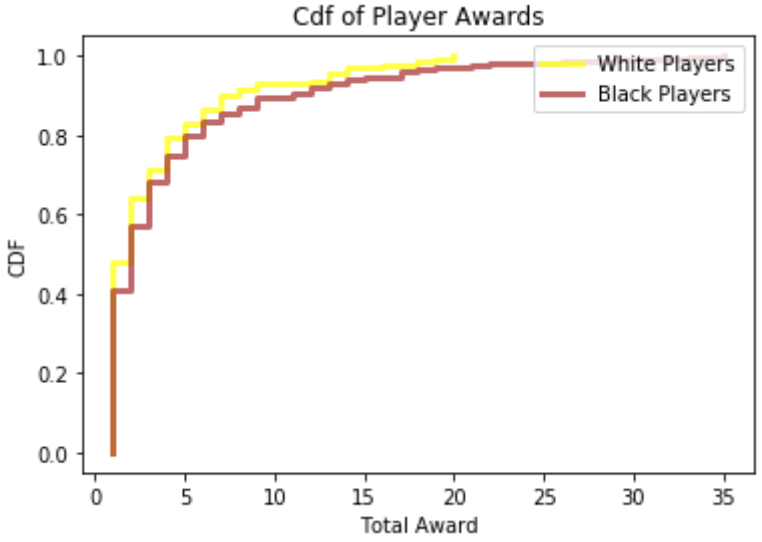
This Probability Mass Function (PMF) shows how many players could able to points and awards able to hit. The x axis shows the total points and total awards, and y axis shows probability of players are there. We can see that black players have more probability of points rate and awards than white players.





This Cumulative Distribution Function (CDF) shows how many players could able to points and awards able to hit. The x axis shows the total points and total awards, and y axis shows the probabilities in of players in percentiles. We can see that black players have more probability of points rate and awards than white players. Also, we can say that total points spread very wide range and awards spread very narrow range.

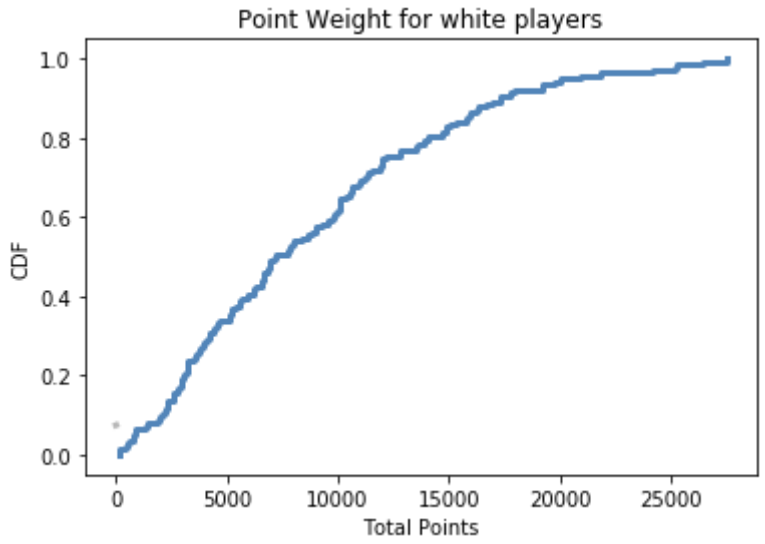


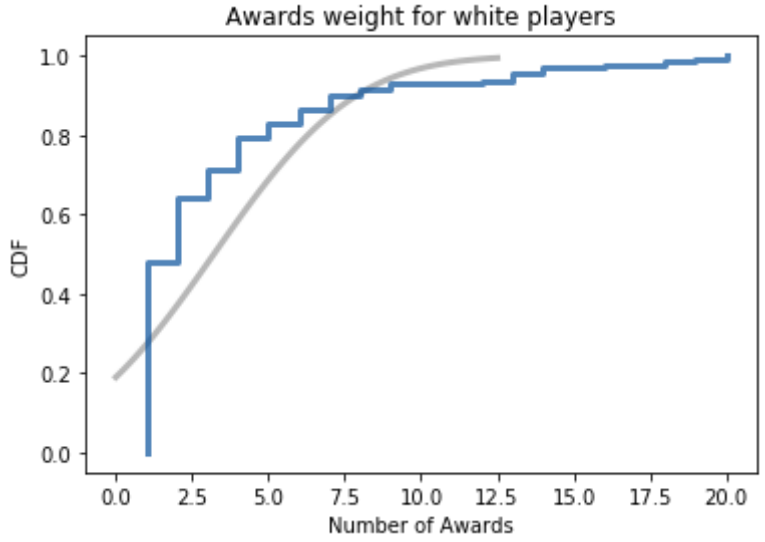


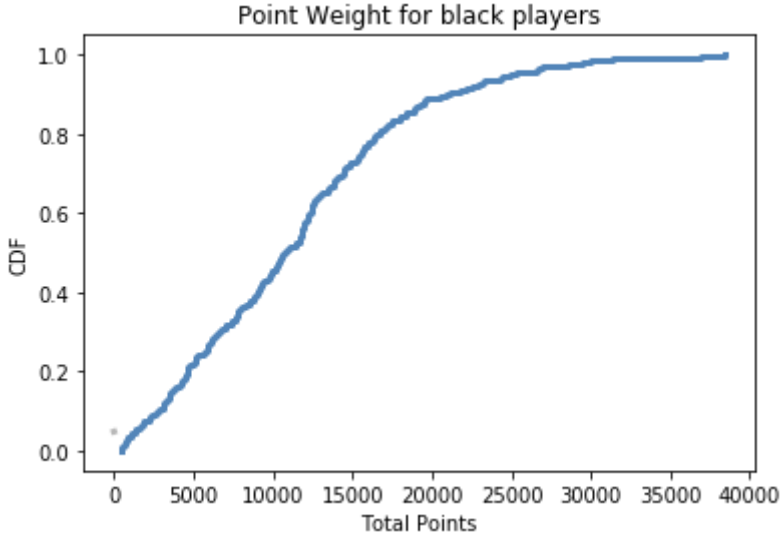
Consequently, we can observe from all the tables that number of players who have points and awards have very high dimensions on zero and slightly decreasing probability of players for each variable when we are increasing the dimensions of points and awards. Therefore, we can still say that there is a high relationship between white and black players from the tables and between total points and number of awards for each basketball players.

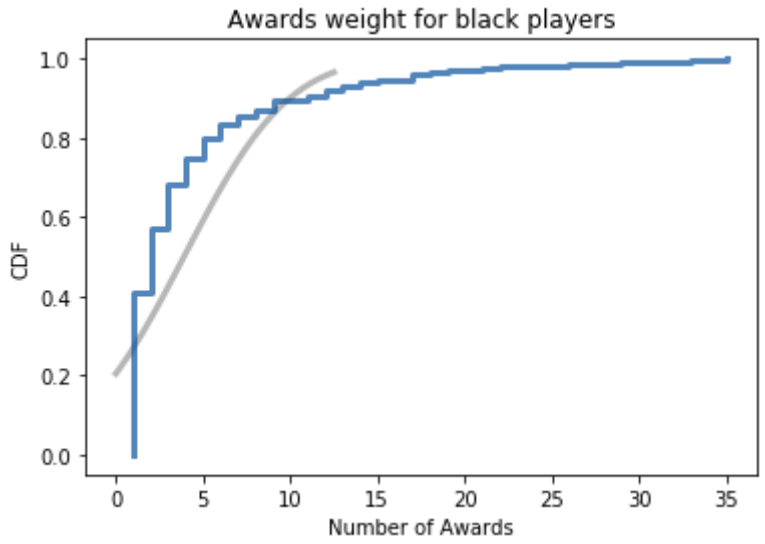
Section 4-)

In this section we used normal distribution for modelling and analyzing results of these modelling plots. The normal distribution is characterized by two parameters: the mean, μ, and standard deviation σ. The normal distribution with μ = 0 and σ = 1 is called the standard normal distribution. Its CDF is defined by an integral that does not have a closed form solution, but there are algorithms that evaluate it efficiently. My data fitted well for awards and race difference.

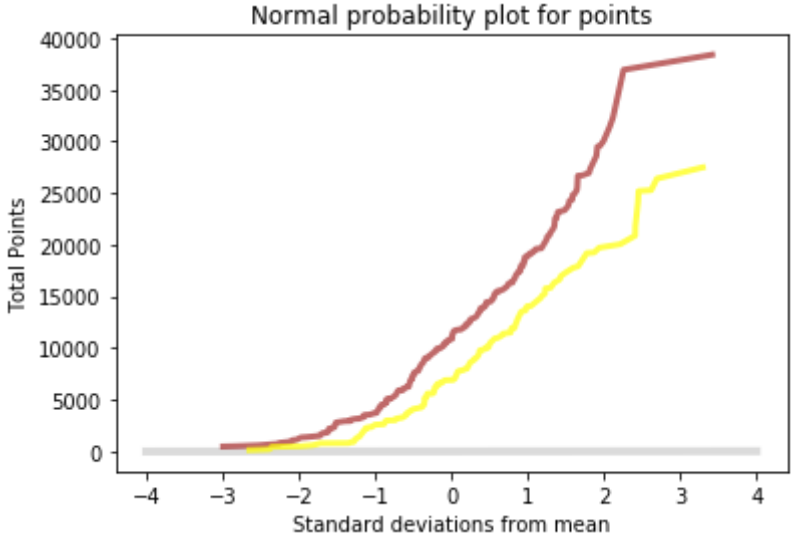

Point Weight for white players
CDF
1.0
0.8
0.6
0.4
0.2
0.0
0
5000
10000
15000
20000
25000
Total Points

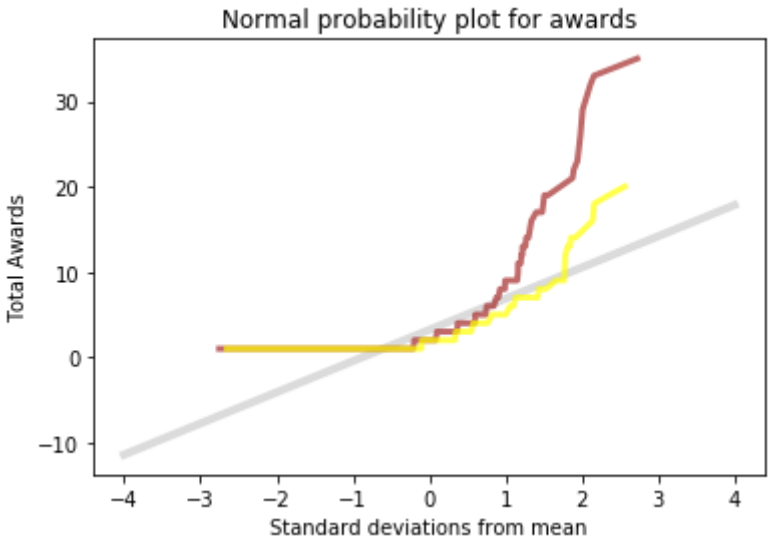

Awards weight for white players
CDF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Number of Awards


Point Weight for black players
CDF
1.0
0.8
0.6
0.4
0.2
0.0
0
5000
10000
15000
20000
25000
30000
35000
40000
Total Points



As we see from the plots that black and white players have similar slope for total points and Number of awards and our normal probability plot for total points concentrate between -3 to -2 and also, normal probability plot for total points concentrate on zero for awards plot.





Section 5-)

Our question is analyzing relationship between points and awards based on player's race. We need to find relationship between our variables. Is there positive relationship or negative relationship so we need to check dependency factor. Both variable increasing or decreasing together then we can say that there is positive relationship but there will be negative relationship between variables when one of the variable decreasing when, other one increasing. We calculated covariance and correlation because covariance test which will give us the tendency of two variables to vary together and correlation test which will give us the strength of the relationship between two relationships.

On the other hand, we can see from the covariance result, two variable have high tendency to each others. Also, our first correlation result shows correlation between total points and number of awards based on white player and second correlation result used for black players. We can compare two graph to have an idea about correlation between these variables on different race. Our results are positive so we can say that our both variable increasing or decreasing together. Also, our results show that there is high correlation between total points and awards variables. Our result claims that players who have high point in total, should have more awards than other players.

```
white_cov = Cov(white_df.points, white_df.award)
print "Covariance value between points and awards for white people:", white_cov
```

```
Covariance value between points and awards for white people: 16220.1589462
```

```
black_cov = Cov(black_df.points, black_df.award)
print "Covariance value between points and awards for black people:", black_cov
```

```
Covariance value between points and awards for black people: 24963.1941667
```

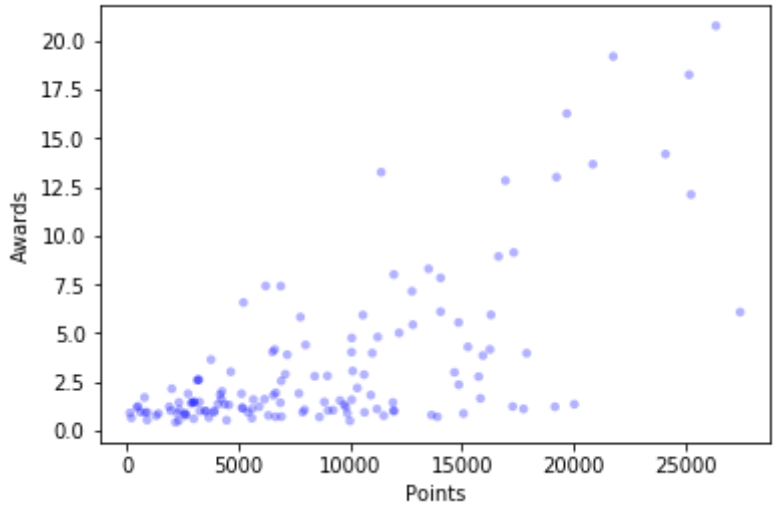
```
white_corr = Corr(white_df.points, white_df.award)
print "Correlation between points and awards for white people:", white_corr
```

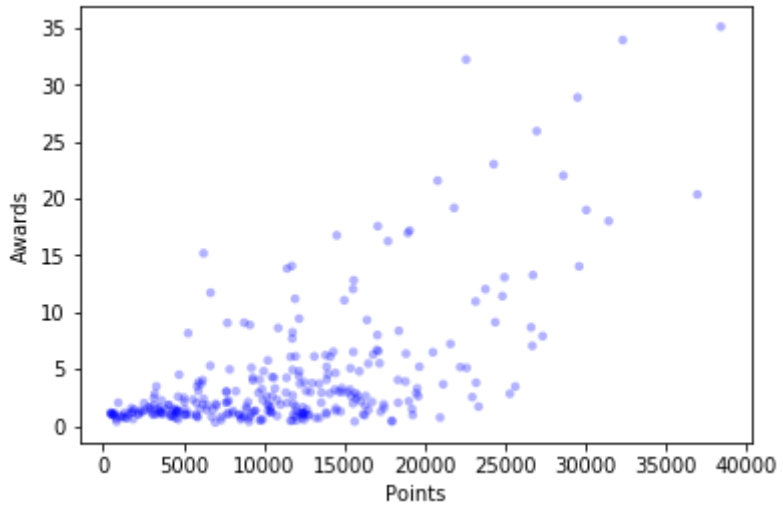
```
Correlation between points and awards for white people: 0.666696721028
```

```
black_corr = Corr(black_df.points, black_df.award)
print "Correlation between points and awards for black people:", black_corr
```

```
Correlation between points and awards for black people: 0.625987813598
```

Following plots show that if players could not have high points then they could not get too much number of awards. Also, increasing points increase probability of winning award and also, in the plot we can see that some of the players have high points but they do not have any award. There is not always positive correlation between points and awards but generally, they have positive correlation. However, when we compare two jittered plot, there is more density from black players plot. (There is not enough population because in the awards dataframe just we captured 1700 player.)





Section 6-)

Choosing the best test statistic depends on what question i am trying to address. In this section i created 4 plot. Two of them analyzing test statistics based on their race and the other two plot shows test statistics of points and awards.

I wanted to test it from two different perspective which are (points, awards) and (white players, black players). Also, i combined all of them from the plots shown in below.

Test Statistics 1: There is a high relationship between total points of players and winned number of awards. If the players have high points, most probably they have awards in the any basketball competition.

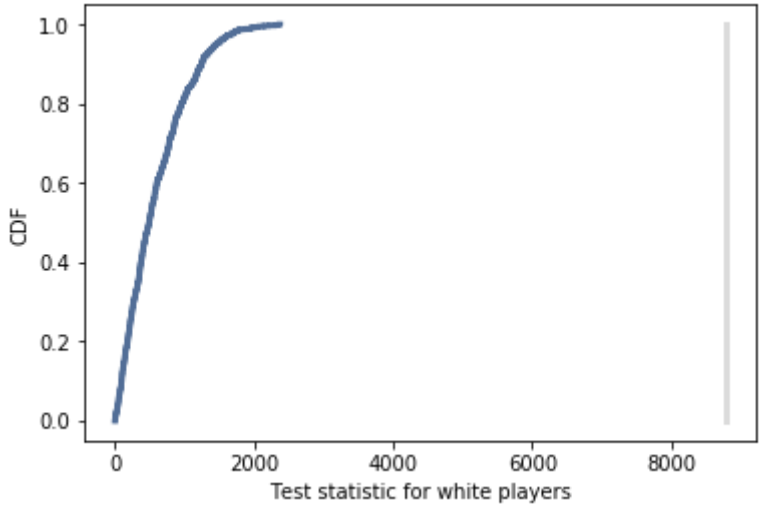
Null Hypothesis 1: There is no relationship between total points and number of awards.

Test Statistics 2: There is a high relationship between black players and white players. Players could have high points and awards as for that their race.

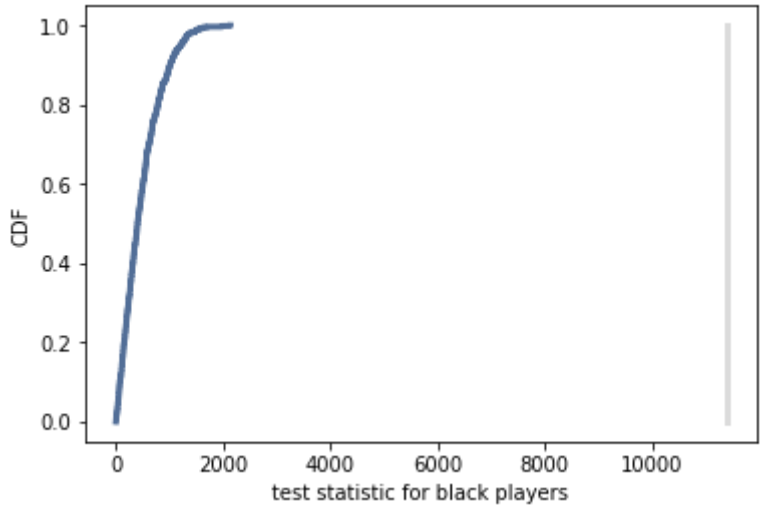
Null Hypothesis 2: There is no relationship between white players and black players.

P-values: I interpret p-values following passages with test statistic plots.

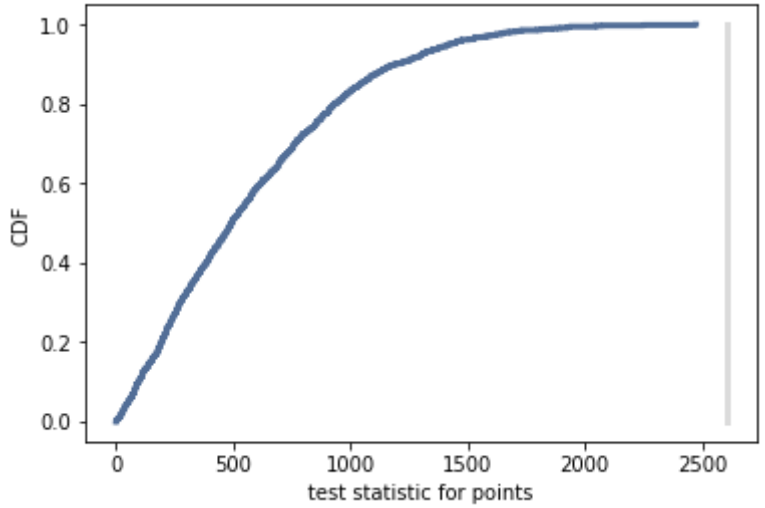
For the white players version of the test, the p-value is 0.00. In general the p-value for a one-sided test is about half the p-value for a two-sided test, depending on the shape of the distribution. In this hypothesis, we are measuring points and awards based on white players and the p-value is very small. The difference is statistically significant.



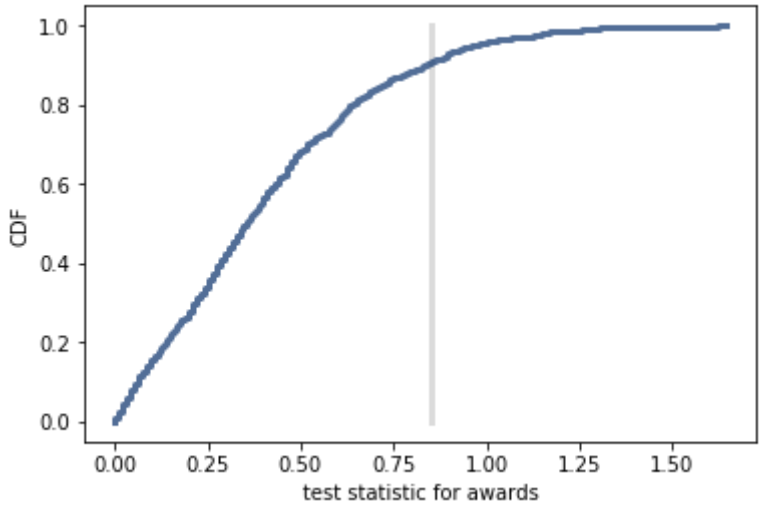
For the black players version of the test, the p-value is 0.00. In the second hypothesis, we are measuring points and awards based on black players and the p-value is very small as white players. The difference is statistically significant.



For the total points version of the test, the p-value is 0.002. In the second hypothesis, we are measuring points based on black and white players and the p-value is very small as white players. The difference is statistically significant.



For the total points version of the test, the p-value is 0.095. In the second hypothesis, we are measuring points based on black and white players and the p-value is very small as white players but the difference is not small as enough to be statistically significant.



Section 7-)

On the whole, I can say that there is a high correlation between total points of players and number of awards for each player. Also, we can say that black players could get more awards and mostly, they have more point than white players. We observe 5 descriptive statistics for our variables and the Histograms, Pmfs, and Cdfs the values are plotted little differences based on points, awards and race factors. Later when we checked the covariance and correlation between our variables. There is correlation between points and awards for both race. At the end of the project, i calculated Hypothesis Test, we saw that out test statistic is statically significant for all conditions except test statistic of awards based on different race.

To conclude that our variables have positive relationship and correlation between each other. If players have high points then they can get awards.