CS 240 Project Report

Section 1

In this report I am going to try to find out something from the given data. Before getting into data, I will define a path of questions to see what will be my main goal.

-Do I need all the data for more granular analysis, or do I need a subset to ensure faster performance?

-Is there any data that is related or getting fed from another data?

-Is there any data which needs normalization to have a statistical approach?

For my project I am going to work on "Is there any data that is related or getting fed from another data?" question. This question is mainly focused on correlation between two different data. To measure this first I need choose which columns I will inspect.

I will check the correlation between Normal season wins and playoff wins. (W column and won column). I choose correlation because I want to see how they move/change together.

My hypothesis will be:

H_0: $\rho = 0$

H_A: $\rho \neq 0$ or H_A: $\rho < 0$ or H_A: $\rho > 0$

I wrote all of conditions in alternative hypothesis because my hypothesis must be mutually exclusive which means they must conquer the all of the possibilities.

Section 2

Data I will use are:

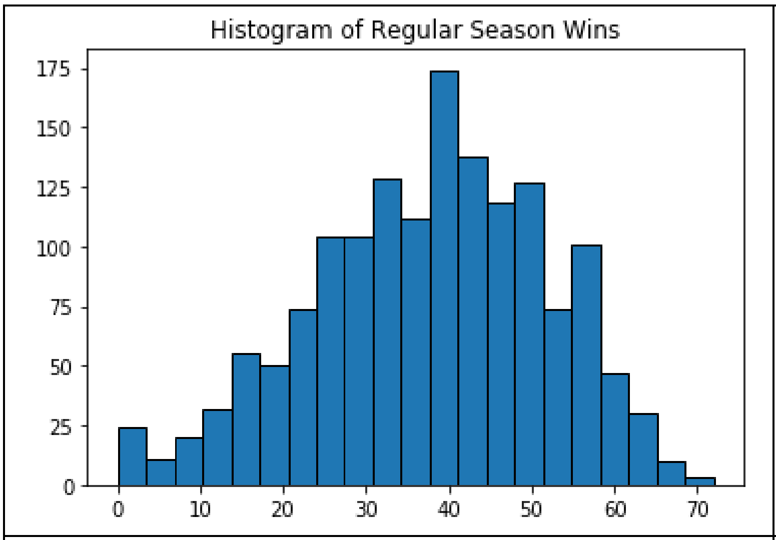
"W" column from "basketball_series_post" file &

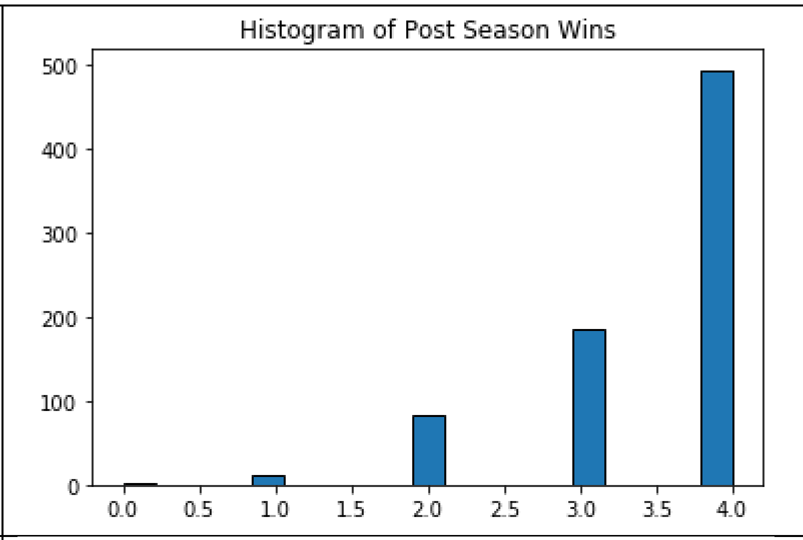
"won" column from "basketball_teams" file.

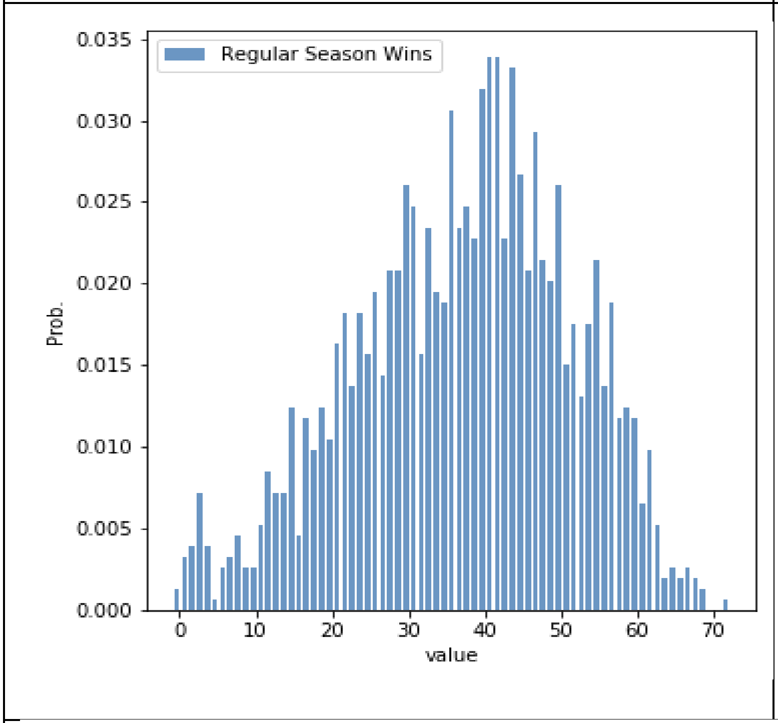
I don't need any data manipulations or cleaning on files because both of my data sets are available to use directly.To examine my data, won column is between 0-81 and w column is 0-4 however they show differ ranges proportion between them could be acceptable.

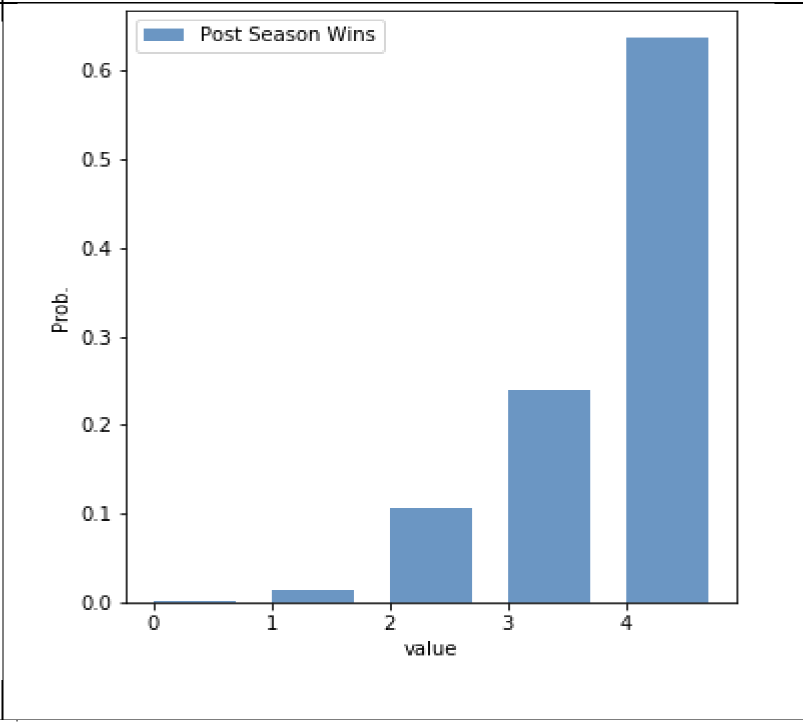
Section 3

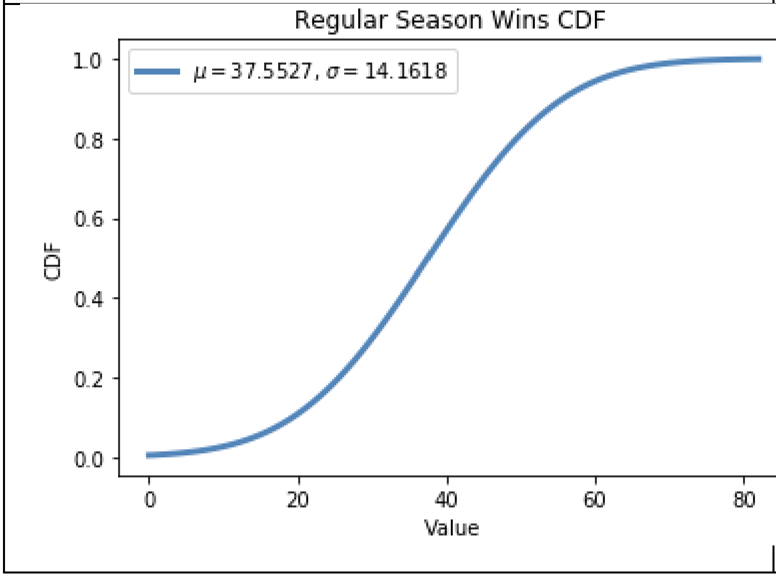
Descriptive Statistics	Win Count in Regular Season	Win Count in Playoffs
Row Count	1536	775
Mean	37.55	3.49
Standart Devaition	14.16	0.76
Minimum	0	0
First Quantile (%25)	28	3
Second Quantile (Median)	39	4
Third Quantile (%75)	48	4
Maximum	72	4

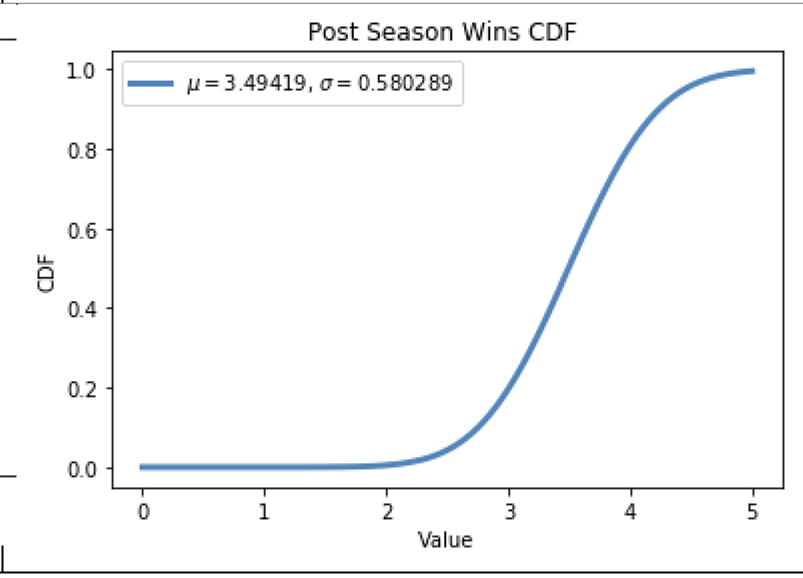

Histogram of Regular Season Wins


Histogram of Post Season Wins



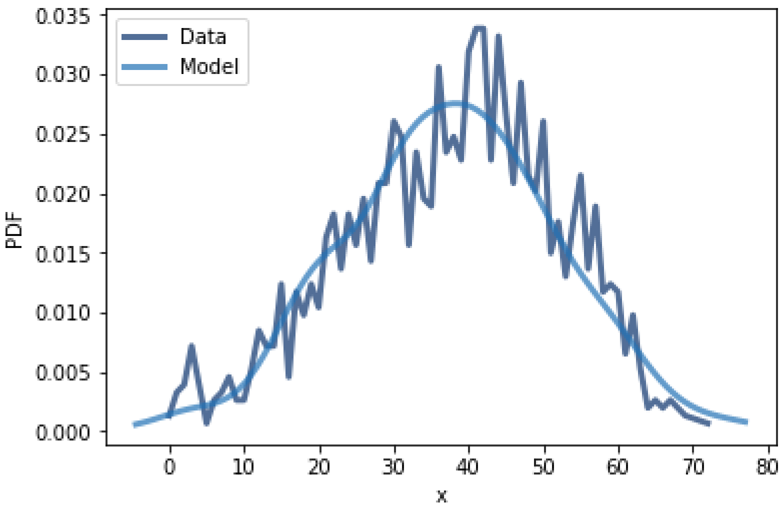



Regular Season Wins CDF


Post Season Wins CDF

Since data range of win count in playoff is smaller it has similar values in different quantiles and the same value in them shows that most of the values are separated in that value (4). But in other one we can see the variety in numbers and a normal distribution in quantile values.

Section 4



Section 5

Section 6

Section 7