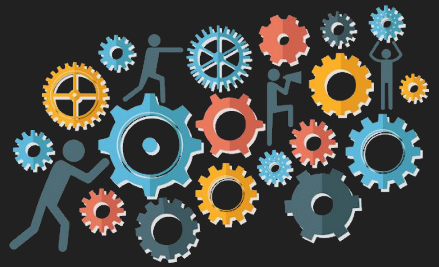
NBA Points

J Daniel Kim

Background

Question: Based on the other statistics can we predict how much points a player will score a season?

- Data visualizations
- Statistical questions
- Modeling
- Final Conclusions Preview: “I was able to predict the number of points scored per season within an RMSE of 33.34 and an R squared of 0.995.”



About the data

Original:

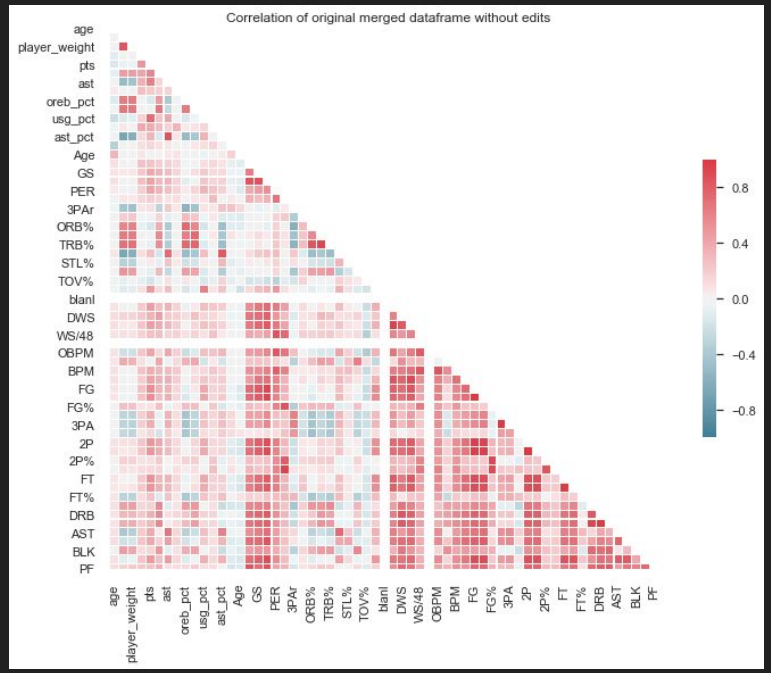
- Merged statistics from different Kaggle sources
- Individual player statistics of season
- Date ranges from 1950 to 2019

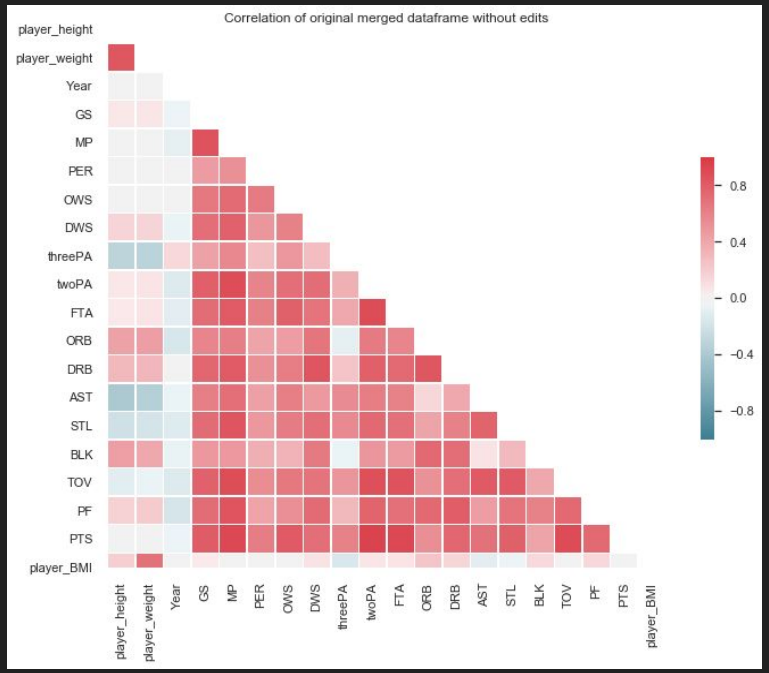
After:

- Three point line created 1979: Date range from 1980-2019
- Dropped any points related metric
- BMI feature created
- Dummy variable: Draft round & Team

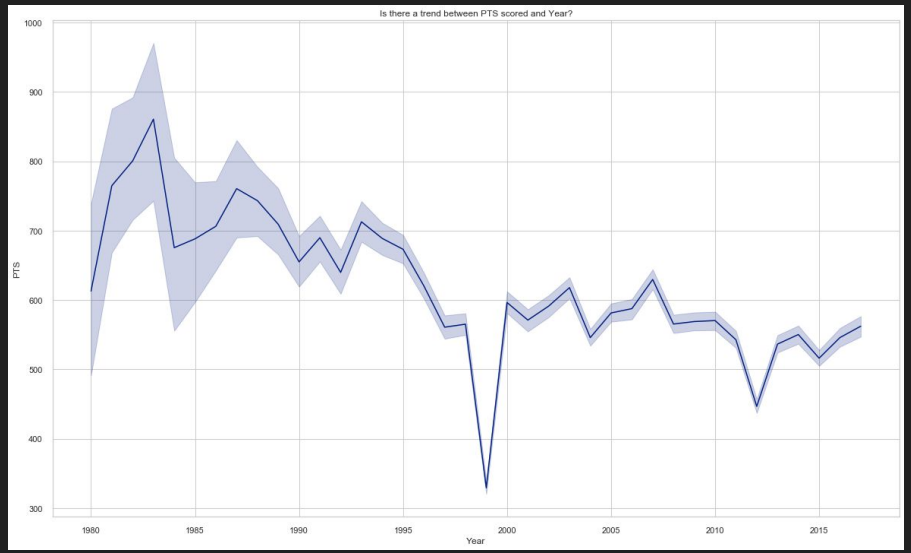
Final Features: Year, GS, MP, PER, OWS, DWS, ORB, DRB, AST, STL, BLK, TOV, PF, player_BMI

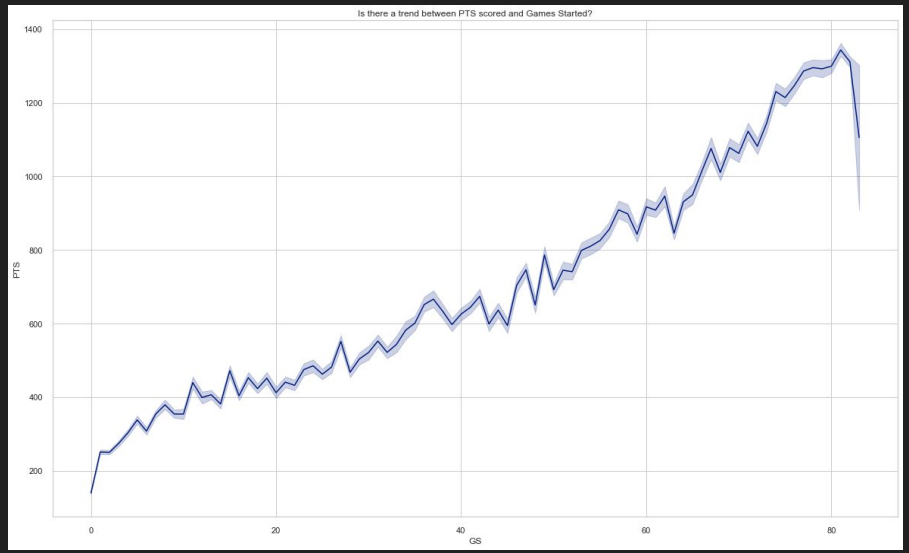
Correlation before and after cleaning data





About the data





- Graph #1: Relationship of the amount of points scored and the year. In 1998 there was a major dip, this was due to the lockout that shortened the season to 50 games.
- Graph #2: Relationship of games started and scoring the more points. As you can see, the more games you start, generally the more points you score.

Statistical Question #1

ANOVA Test: Test whether there is a statistically significant difference in mean of PTS scored among teams

Ho: The mean of points scored per season of all the teams is the same

Ha : The mean of points scored per season of all the teams is not

P-value: is 6.42 * 10^-205

Alpha: 0.05

Result: Reject the null

Statistical Question #2

ANOVA Test: Test whether there is a statistically significant difference in the mean of points scored per season based on the country of origin

Ho: The mean of points scored per season between all the country is the same

Ha : The mean of points scored per season between all the country is not

P-value: 0.922

Alpha: 0.05

Result: Fail to Reject the Null Hypothesis

Statistical Question #3

Two sample t-test: Do players that played in the 1980's and 1990's era score more points on average than one's from the 2000's?

Ho : mu 1900's points less than or equal to mu 2000's points

Ha: mu 1900's points more than mu 2000's points

P-Value <0.001

Alpha: 0.05

Result: Reject the null hypothesis

Before Model Adjustment Results

Model Type: Second degree polynomial Regression

R Squared: 0.9991

Testing RMSE: 14.92

Issues:

- Features for three point, free throw, and two point attempts hindered the prediction.
- Dummy variables did not allow a third degree polynomial due to creation of too many features that caused a loading error.

Changes: dropped three point attempt, two point attempt, free throw attempt, and all dummy variables.

Final Model Results

Model Type: Third degree polynomial Regression with final features from ‘About the data slide

R Squared: 0.996

Testing RMSE: 32.797

Does BMI help predictions?

Model Type without BMI: Third degree polynomial Regression

R Squared: 0.995

Testing RMSE: 34.01

THANK YOU