# PROBABLILITY AND STATISTICS [TERM PROJECT]

Brief:

Abstract

[Author name]
[Email address]

# PROBABLILITY AND STATISTICS [TERM PROJECT]

## Brief:

Predicting the outcome of a Baseball Game in a complete regular season using Regression Model, and Normal Distributions, also analyzing the trends via resourceful graphs generated between resourceful elements from the selected Dataset.

# PROBABLILITY AND STATISTICS [TERM PROJECT]

**BRIEF:**

## THE DATA SET

The data used for this project came from *baseball-reference.com*. The initial dataset consisted of each team's seasonal hitting statistics from 1990–2018 (excluding 1994–1995 due to MLB strike). Data was split into 2 groups: 1990–2017 and 2018 for testing.

| | key_0 | Tm_x | W | L | W-L% | GB | Abv | Tm_y | #Bat | BatAge | R/G |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 25 | Tm_x | 65 | 97 | 0.401 | 26 | ATL | ATL | 46 | 27.4 | 4.21 |
| 1 | 4 | Baltimore | 76 | 85 | 0.472 | 11.5 | BAL | BAL | 47 | 27.3 | 4.16 |

| G | PA | AB | R | H | 2B | 3B | HR | RBI | SB | CS | BB |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 162 | 6084 | 5504 | 682 | 1376 | 263 | 26 | 162 | 636 | 92 | 55 | 473 |
| 161 | 6223 | 5410 | 669 | 1328 | 234 | 22 | 132 | 623 | 94 | 52 | 660 |

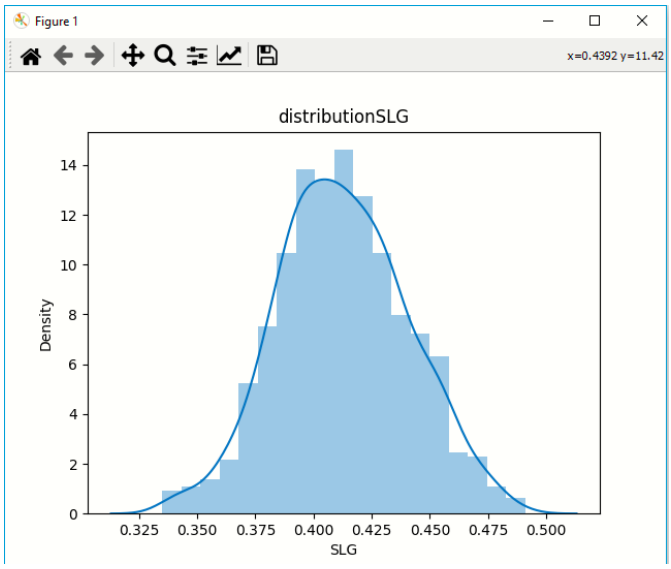
| BB | SO | BA | OBP | SLG | OPS | OPS+ | TB | GDP | HBP | SH | SF | IBB | LOB | Year |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 473 | 1010 | 0.25 | 0.311 | 0.396 | 0.706 | 90 | 2177 | 101 | 27 | 49 | 31 | 36 | 1074 | 1990 |
| 660 | 962 | 0.245 | 0.33 | 0.37 | 0.7 | 99 | 2002 | 131 | 40 | 72 | 41 | 50 | 1230 | 1990 |
| 598 | 795 | 0.272 | 0.344 | 0.395 | 0.739 | 104 | 2180 | 174 | 28 | 48 | 44 | 59 | 1233 | 1990 |

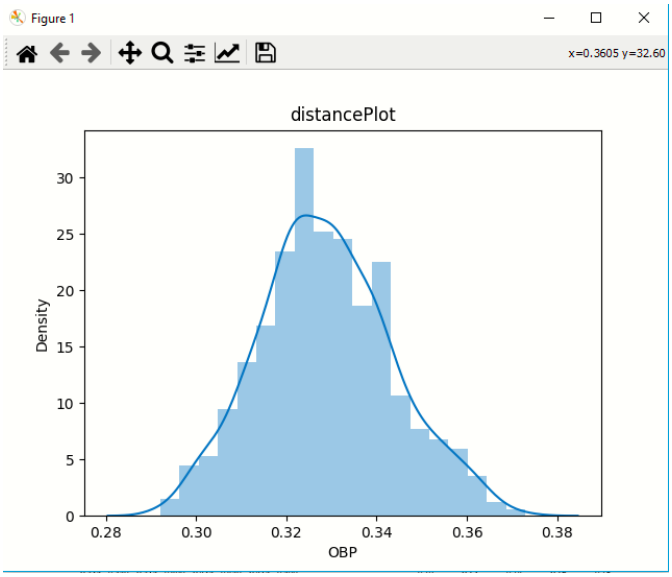
Using yearly team hitting statistics, the columns are

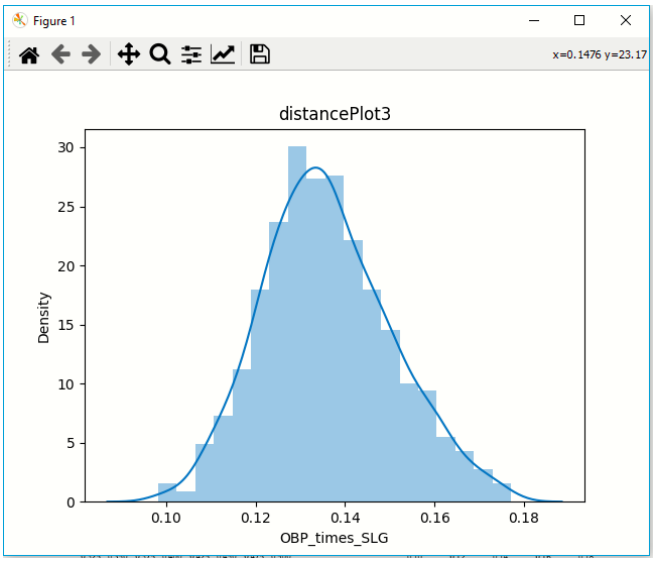
: 'Team', 'W', 'L', 'W/L', 'GB', 'Abv', 'Num_Hitters', 'BatAge', 'R/G', 'G', 'PA', 'AB', 'R', 'H', 'Doubles', 'Triples', 'HR', 'RBI', 'SB', 'CS', 'BB', 'SO', 'BA', 'OBP', 'SLG', 'OPS', 'OPS_Plus', 'TB', 'GDP', 'HBP', 'SH', 'SF', 'IBB', 'LOB', 'Year', 'H/G', 'Extra_Base_Hits', 'BABIP', 'OBP_times_SLG', and 'Age_of_Hitters'.

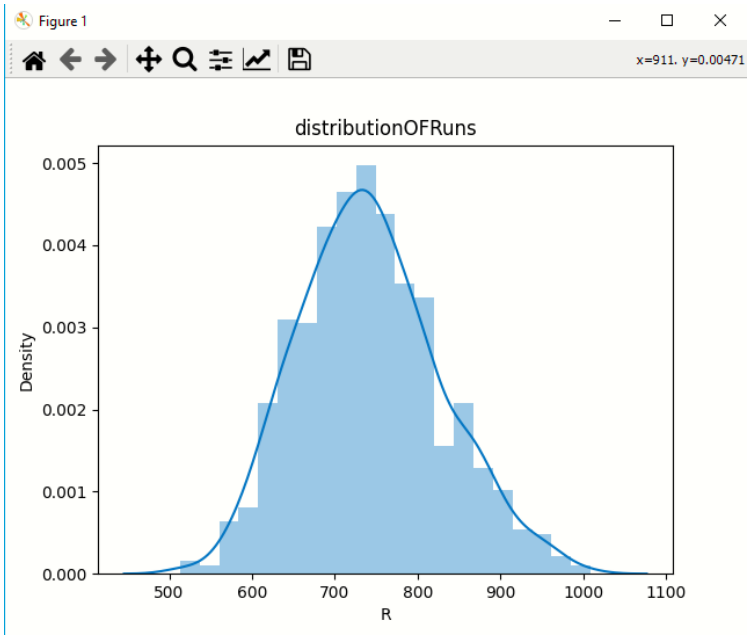
**TARGET VARIABLE: TOTAL RUNS SCORED IN THE SEASON**

## ANALYSIS ON RUNS SCORED

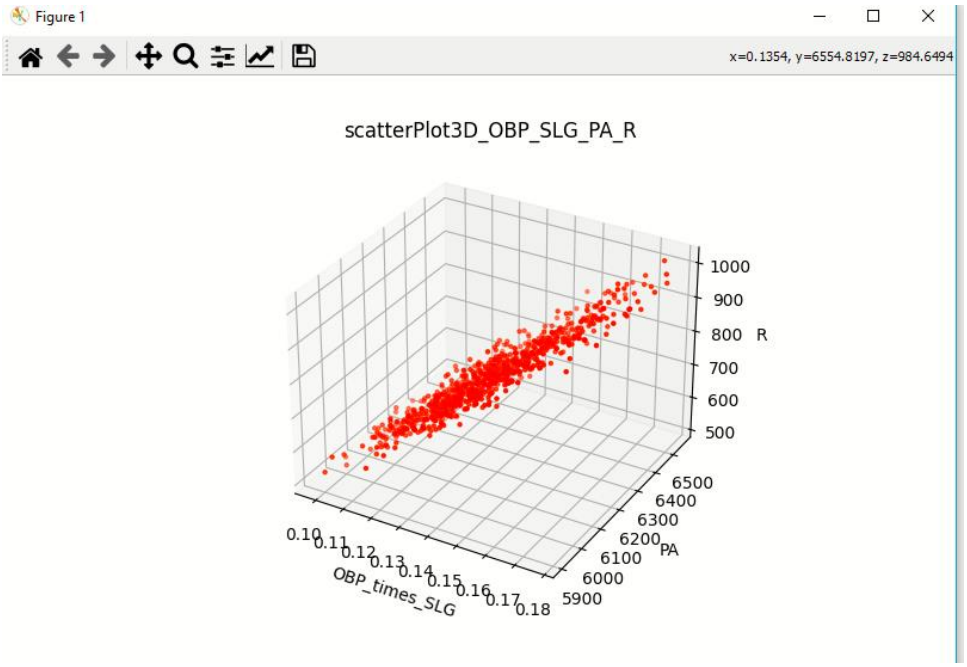


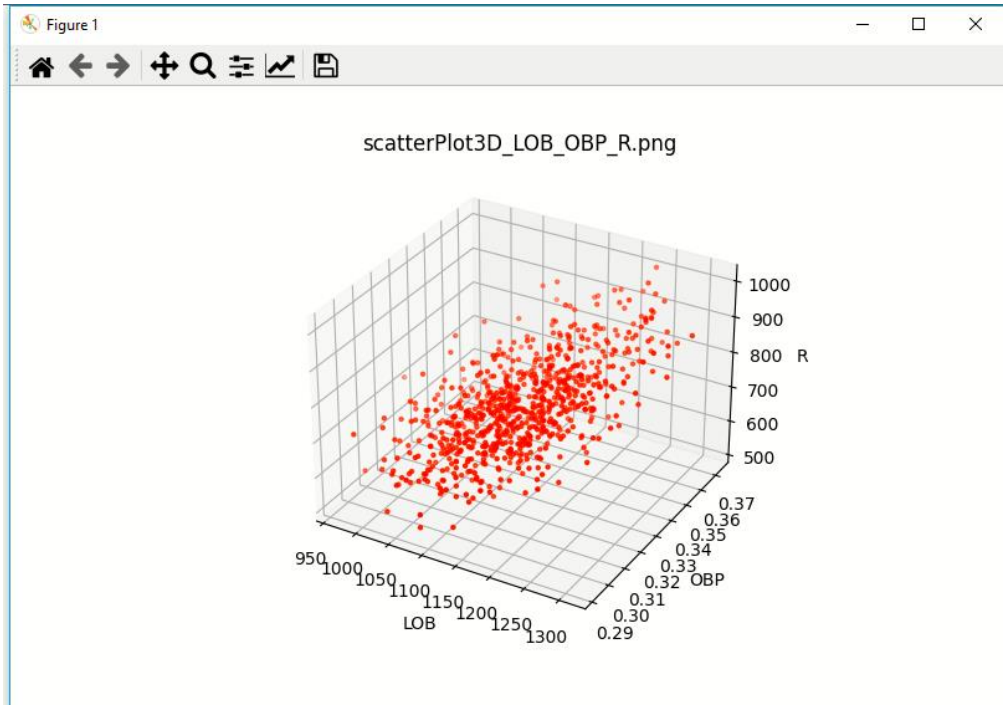



Looking at the distributions of some of the columns (runs, on base percentage, slugging, on base x slugging), they seems to be normally distributed which is an overall good sign.



On the left we see those incremental increases in OBP, and yearly plate appearances have a strong positive correlation with scoring more runs. This is intuitive because the more at bats a team has during a game, the higher chance they will have to score runs.

On the other hand, looking at left on base and on base percentage is a little bit different. Their correlation with runs per year is still positive but it is weaker. Obviously in baseball stranding runners on base is not a good thing. Those are missed opportunities for scoring runs. So, we would expect this to possibly have a slight negative correlation with runs per season. But actually, the opposite seems to be true. The graph on the right shows that incremental increases in left on base and on base percentage actually have an overall positive correlation with scoring more runs during the year. I believe this is true because although stranding runners is a bad thing, at the end of the day it still means that the team is putting itself in the position to score runs which is better than having a lower LOB and not scoring runs in the first place. The ironic thing about LOB is that you cannot tell how many runs the team scored based on it alone. A team with a low LOB could either have cashed in most of their base runners OR they could have never had runners on base in the first place.

# Linear Regression Model

After testing different sets of feature variables our final model was:

R ~ PA + LOB + OBP + OBP_times_SLG

## INTERPRETATION OF COEFFITIENTS:

Intercept: -1867.7977 (a team with 0 plate appearances will score -1868 runs in the season, funny I know)

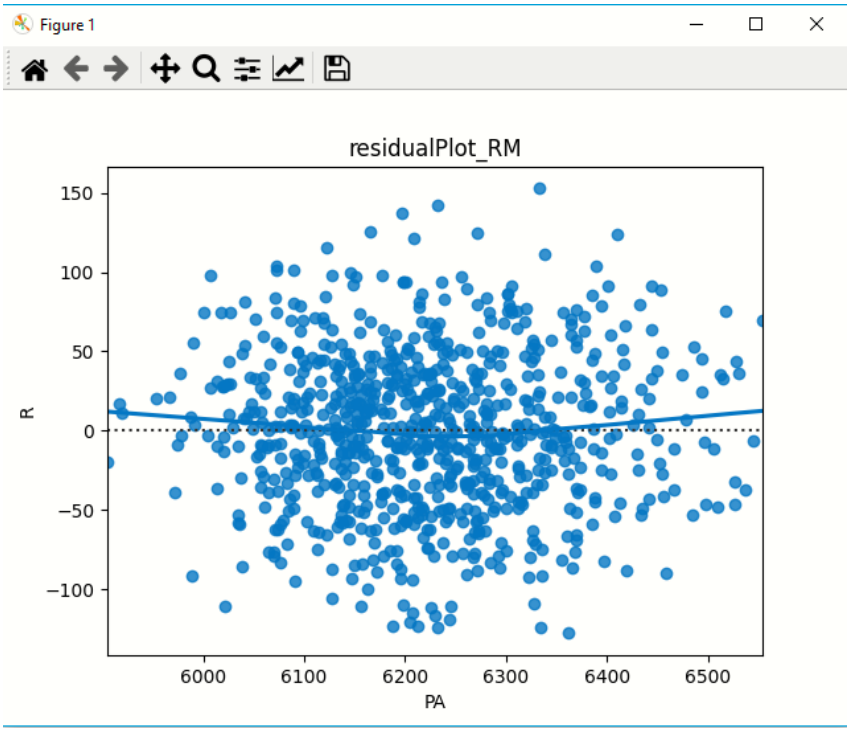
PA 0.3672: holding all else equal, a unit increase in PA will result in a 0.3672 increase in runs scored per season

LOB -0.6483: holding all else equal, a unit increase in LOB will result in a 0.6482 decrease in runs scored per season

OBP 2148.2308: holding all else equal, a unit increase in OBP will result in a 2148.2308 increase in runs scored per season. This number is so high due to the fact that OBP is a percentage between 0 and 1.

OBP_times_SLG 2633.9704: holding all else equal, a unit increase in OBP_times_SLG will result in a 2633.9704 increase in runs scored per season. This number is so high due to the fact that OBP_times_SLG is a percentage between 0 and 1.

R-Squared 0.960: this value represents that 96% of the variance shown in the model can be explained by our independent variables.



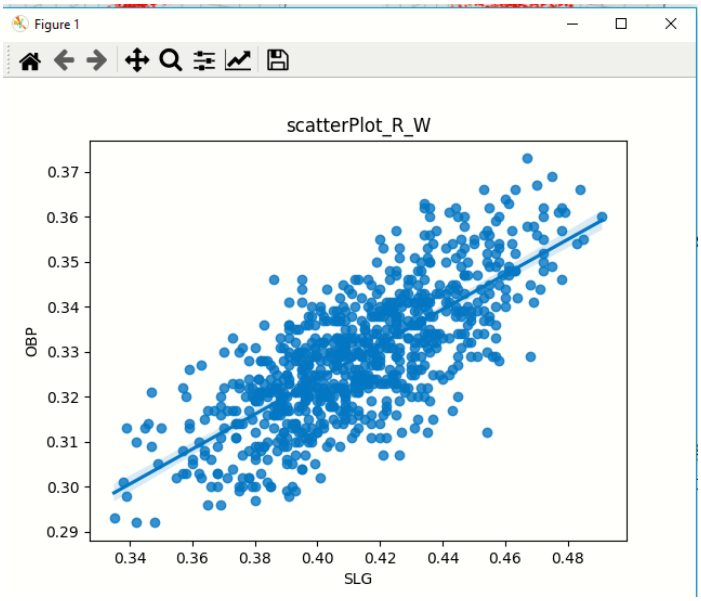

**The residuals seem to be normal distributed (great sign).**

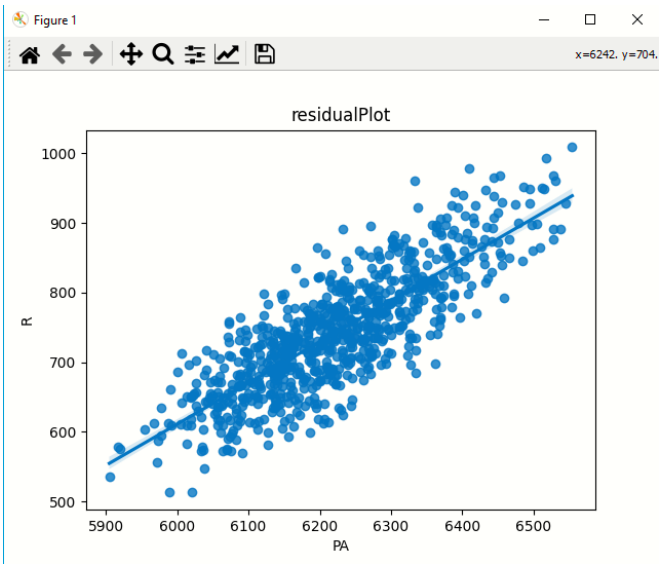
## TESTING THE MODEL ON UNSEEN 2018 DATA

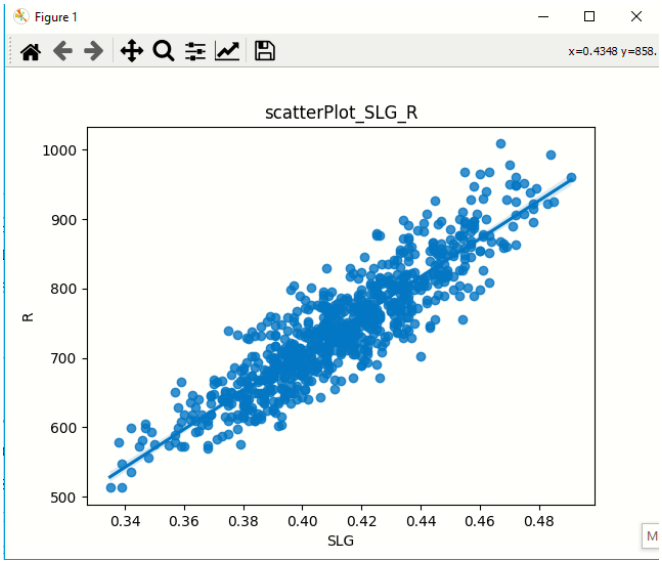
```
========================================
========================================

percentage of error: 2.299 %

========================================
========================================
========================================
========================================

Mean squared error: 303.6977184616725

========================================
========================================
========================================
========================================

RMSE: 17.426925100592833 runs
2018 mean runs scored: 721.0 runs
percentage of error: 2.417 %

========================================
========================================

========================================
========================================
```

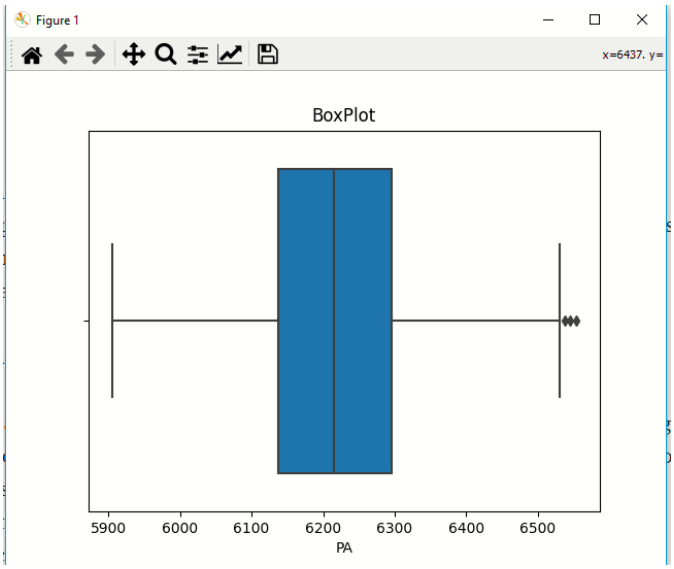
As seen above, the average error for each prediction is about 17 runs per season. Comparing this to the mean runs scored in 2018, we get an average error of about 2.45%

## ANALYSIS ON HIGHLY CORELATED ELEMENTS FROM DATA SET



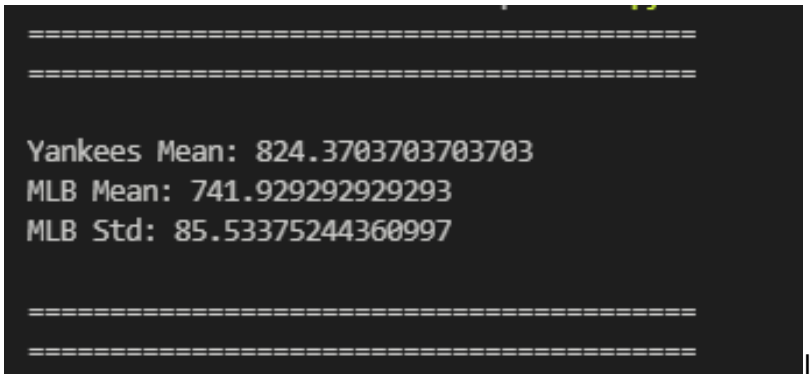
# TESTING ON A TEAM

**New York Yankees Hypothesis Test**

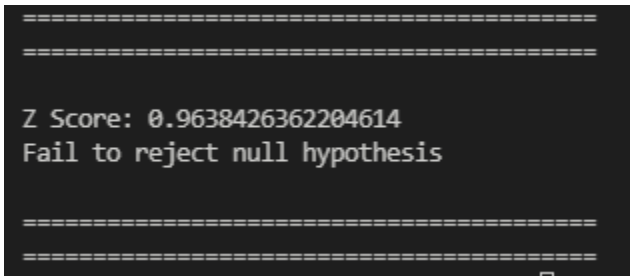
The New York Yankees have won the most World Series Championships within the span of 1990–2018 (5). Let's look into how they match up to the rest of the MLB in terms of runs scored.

H0: The mean runs for NYY = the mean runs for of all of MLB

Ha: The mean runs for NYY > the mean runs for of a



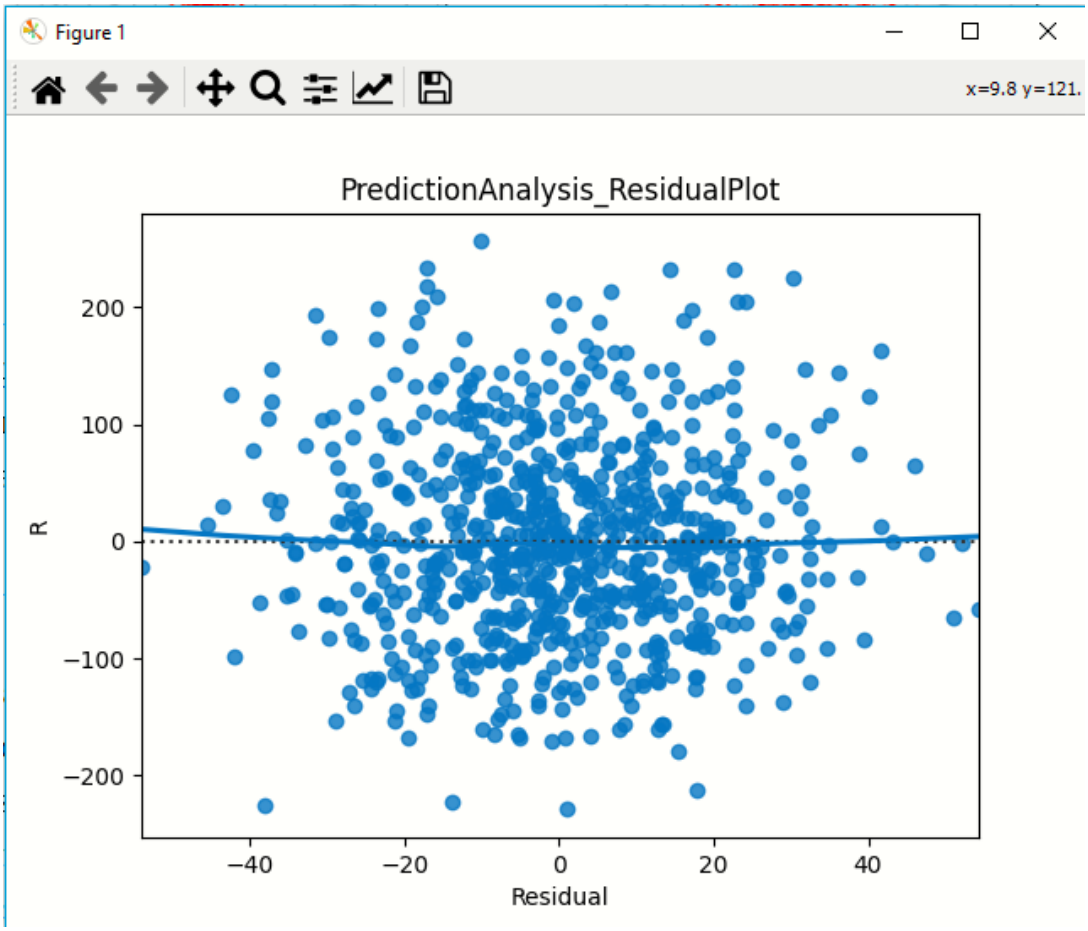

ll of MLB



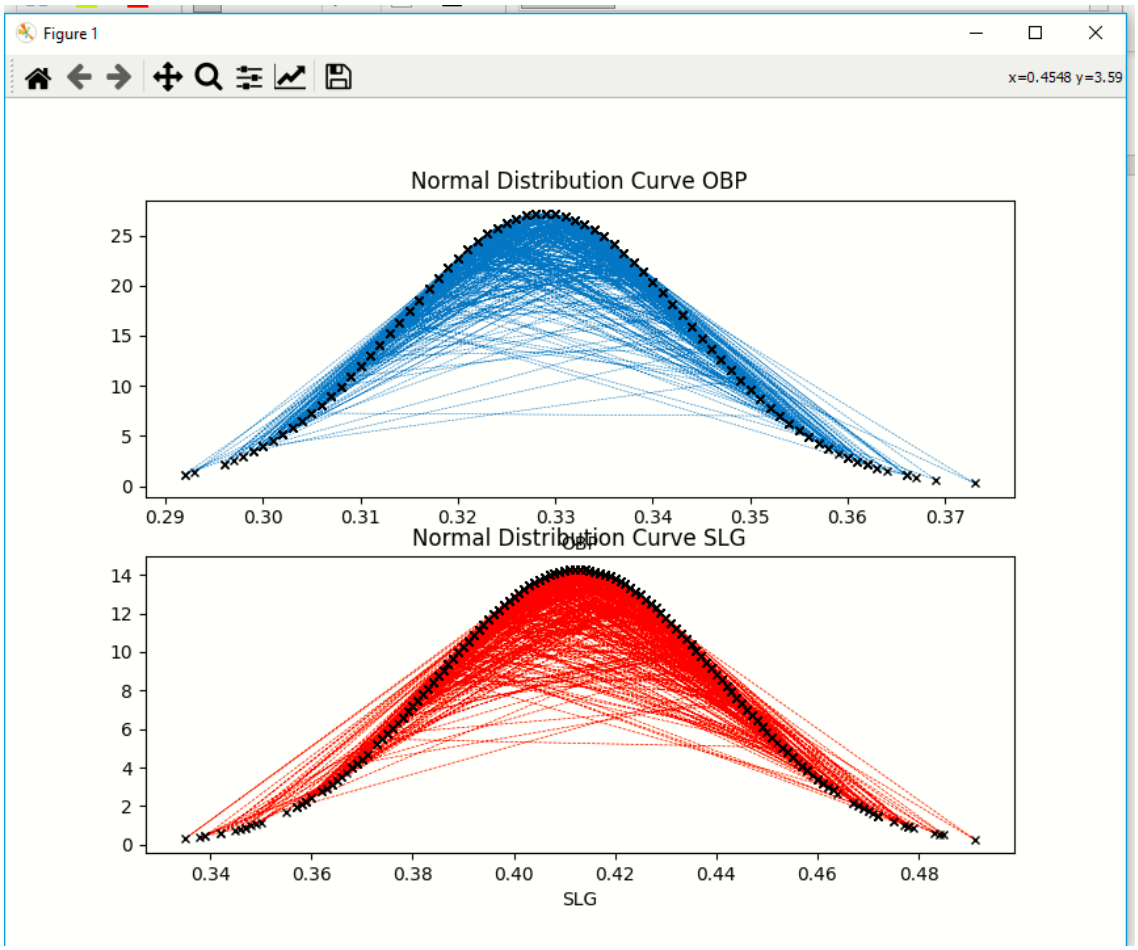

With a z score of .95 we fail to reject the null hypothesis that the New York Yankees significantly score more runs than the rest of the MLB

# Conclusion

After obtaining the data and doing my initial EDA, it seems that the most valuable hitting statistics for predicting runs scored per season are Plate Appearances, Left On Base, On Base Percentage, and On Base x Slugging Percentage. These variables together account for 96% of the variance in runs scored per season within the dataset. Going forward, I would like to test this model on the 2019 regular season as well as turn my focus toward the target variable of wins per season.




# Normal Distribution




To be Notes : Data is Highly Correlated and is Normal.

# QQ Plot Test

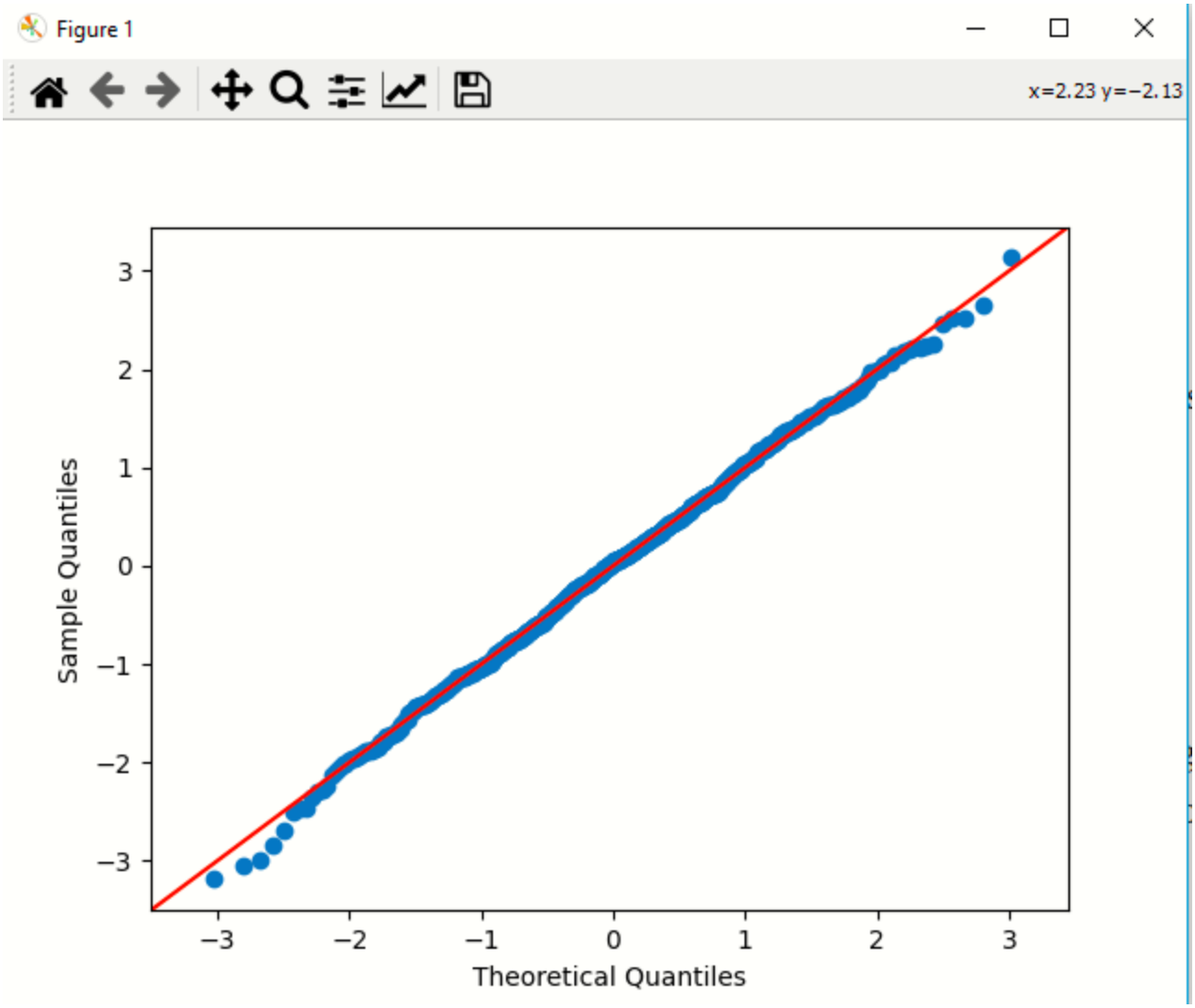


Figure 1
x=2.23 y=−2.13
Sample Quantiles
Theoretical Quantiles
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3