Baseball Statistics Analysis

Which statistics lead to the best teams?

Brett Foster
DSC 530-T302 - Data Exploration & Analysis
Final Project

Questions

What are the most important statistics that lead to the most wins in baseball?

Do the claims in the book *Moneyball: The Art of Winning an Unfair Game*[(1)], about the 2002 Oakland Athletics, regarding on-base percentage's importance for success in baseball still hold true today, or is a new wave of statistical importance taking over the MLB?

There seems to be theories among the league that the National League produces lower Batting Averages and Earned Run Averages because of the lack of a Designated Hitter, forcing the active pitchers to hit. Alternatively, ERA and BA are higher in the American League. Is there any truth behind these claims?[(2)]

Many new changes were many in the 2020 season, what has its effects been on statistics comparatively, year over year?

Datasets

Multiple datasets consisting of the Major League Baseball (MLB) league standings from 2017 - 2020, and batting, pitching, fielding statistics from the same years, with a focus around the 2020 and 2019 seasons using earlier seasons for certain comparisons.

The datasets are 4 excel files for each category, standings, batting, pitching, and fielding, for a total of 16 excel files, all from baseball-reference.com. The similar datasets from each year are merged in my python file on the team name column, ‘Tm', and a few categories renamed because of functional issues.

**Potential issues with the datasets - 2020 kicked off a weird year for the MLB, changing many rules, some listed below, and their hypothesized effect on important statistics:
1) Universal Designated hitter - AL and NL: increased BA and ERA in NL
2) Pitchers must face 3 batters

Datasets Retrieved From [3]

Important Variables

The focus of my analysis surround numerous variables from the datasets, some based on experimental variable, and others based on assumed importance by the MLB and players, with a few variables used that are not listed below.

Winning Percentage (W_L_percent) - Experimental (ultimate variable that leads to the rank of a team)

Earned Run Average (ERA) - A leading pitching statistic measuring earned runs a team allows per 9 innings

Runs Batted In (RBI) - Statistic for batters, crediting a batter for a play that allows a run to score

Total Bases (TB) - Number of bases a runner has achieved, weighted by type:

1 - Single, 2 - Double, 3 - Triple, 4 - Homerun

Important Variables - Cont'd

Runs (R_x for per game average, R_y for total offense) - Number of runs scored by a team. Measure of successful offense.

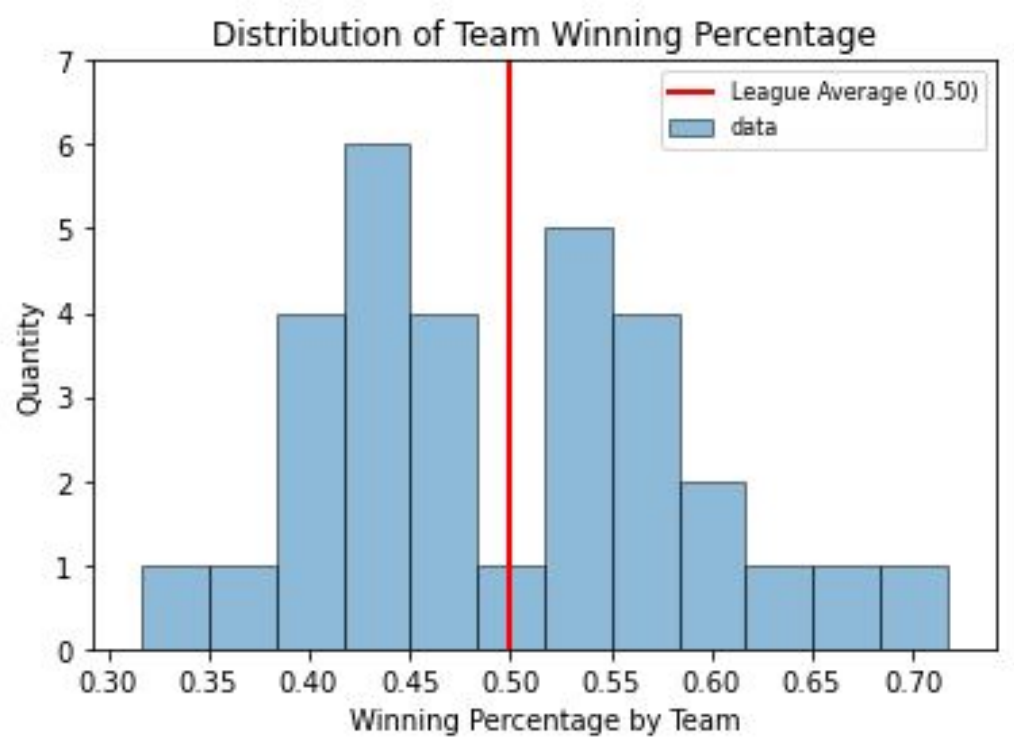
On Base Percentage (OBP) - Measures the frequency of a batter reaches base, either by hit, walk, or hit by pitch.

Batting Average (BA) - Frequency of successful hits by a batter by total plate appearances

Slugging Percentage (SLG) - Measures a batters productivity. Similar to Batting Average but adds more weight to extra-bases. Highly correlated to Total Bases.

Walks and Hits per Inning Pitched (WHIP) - Measure of a pitcher's performance. Pitchers ability to keep runners off bases.

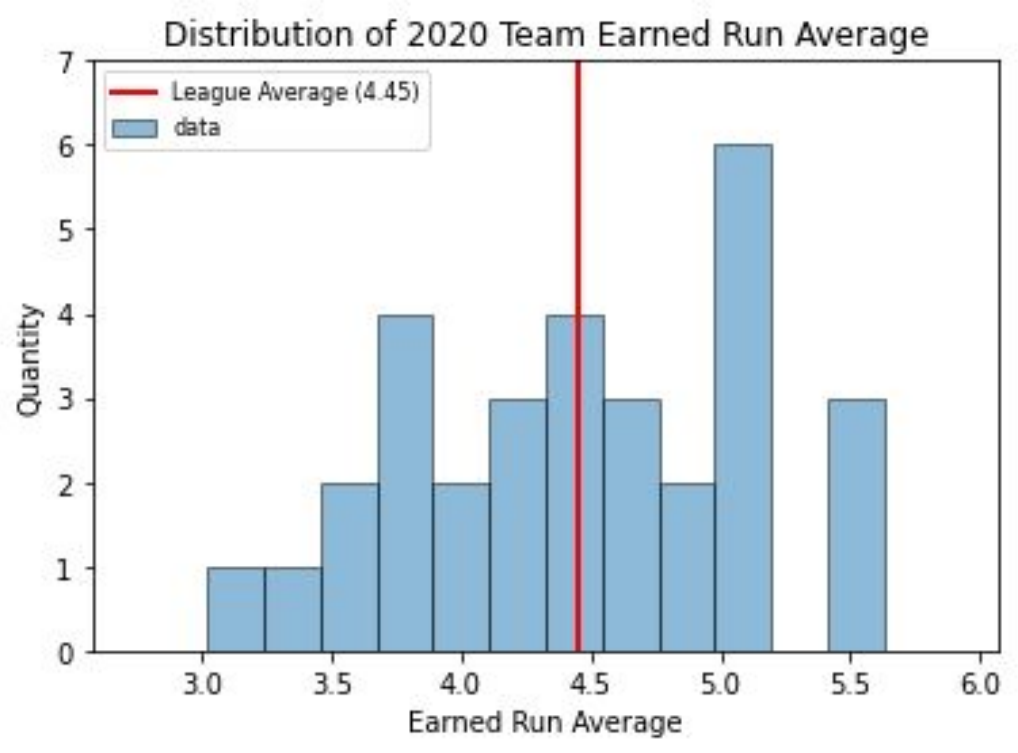
Winning Percentage (W_L_percent)



Mean:	.500
Median:	.483
Mode:	.433
Standard Deviation:	.090

It does appear to be slightly skewed right, implying more winning teams to offset the balance of a focus less than the mean.

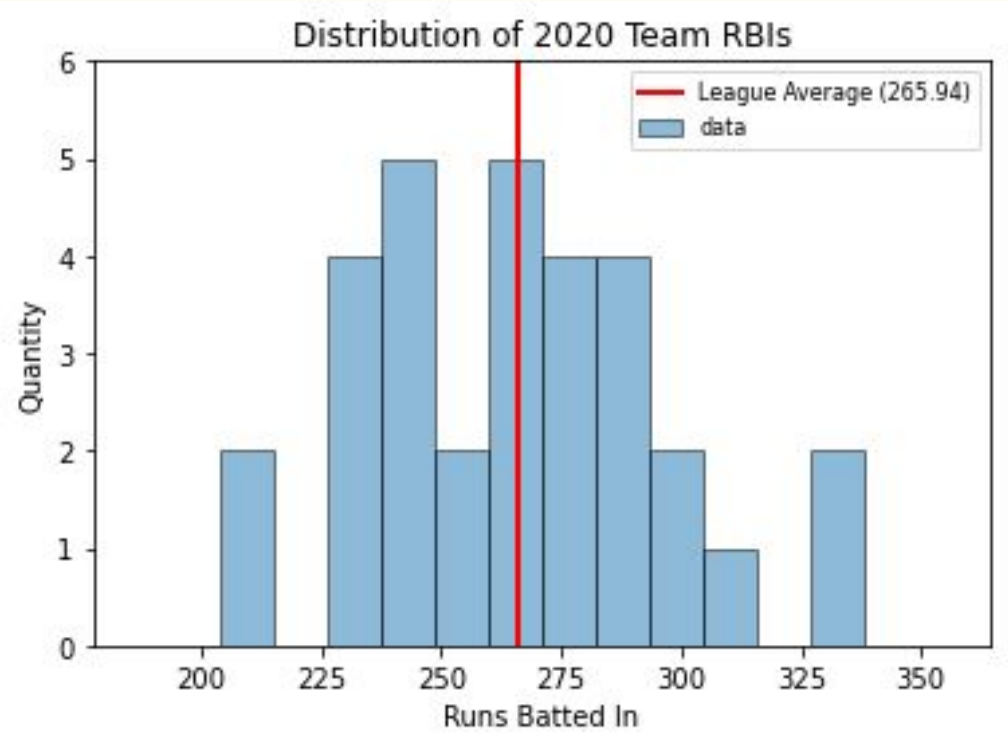
Earned Run Average (ERA)



Mean:	4.45
Median:	4.44
Mode:	5.09
Standard Deviation:	0.668

Based on the histogram, the data does appear to be just slightly skewed left. 3 data points separated from the group represent the 3 of the top 10 worst teams in the league.

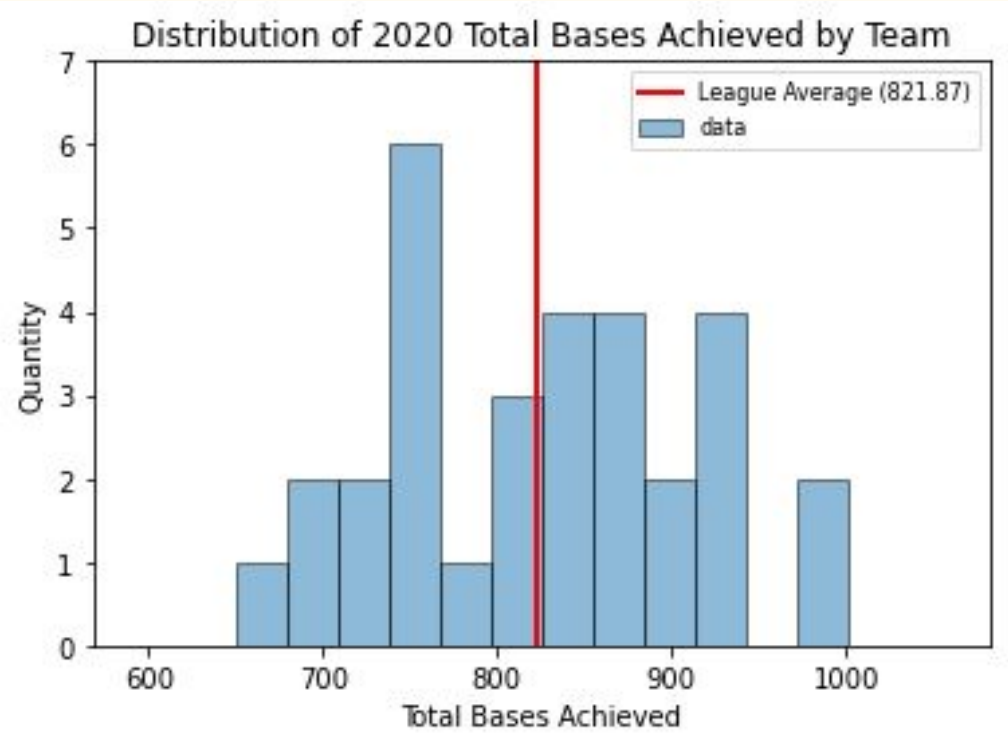
Runs Batted In (RBI)



Mean:	265.94
Median:	264.00
Mode:	264.00
Standard Deviation:	30.66

The plot looks reasonably uniform. The two smallest and largest values represent the 2 worst teams in the league (smallest) and 2 top ten teams, 1 being the World Series winner (largest).

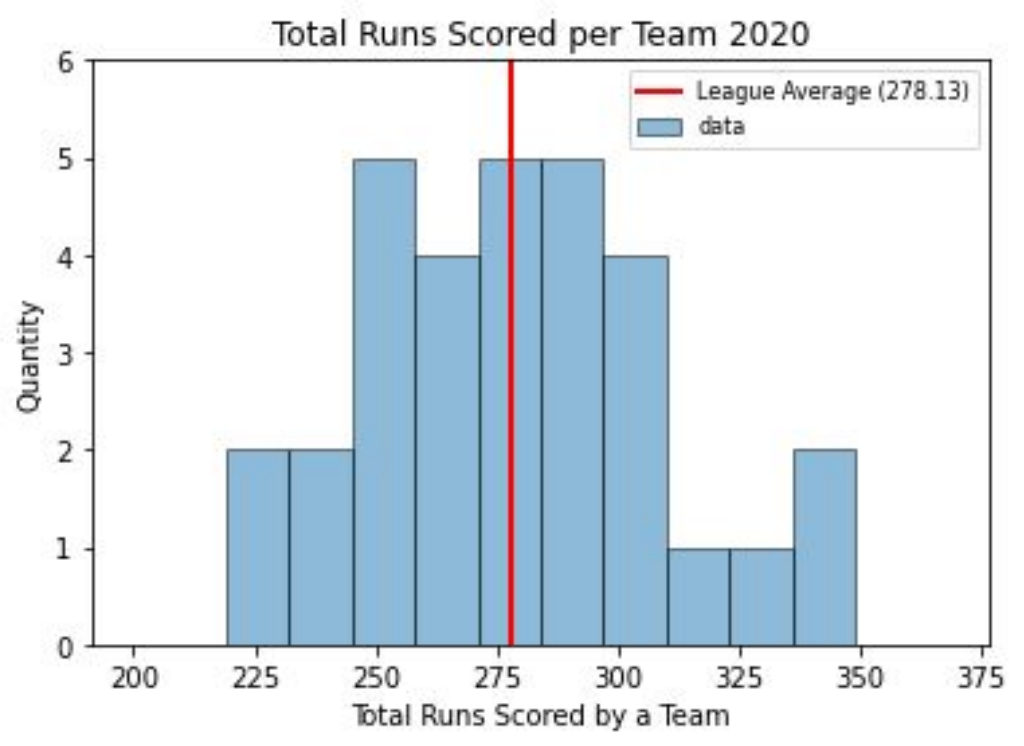
Total Bases (TB)



Mean:	821.87
Median:	828.00
Mode:	742.00
Standard Deviation:	88.06

The histogram does not seem perfectly uniform but there is a small gap right of the mean. The greatest values represent unusually great offensive batters and the teams were the World Series winners a the 6th ranked team.

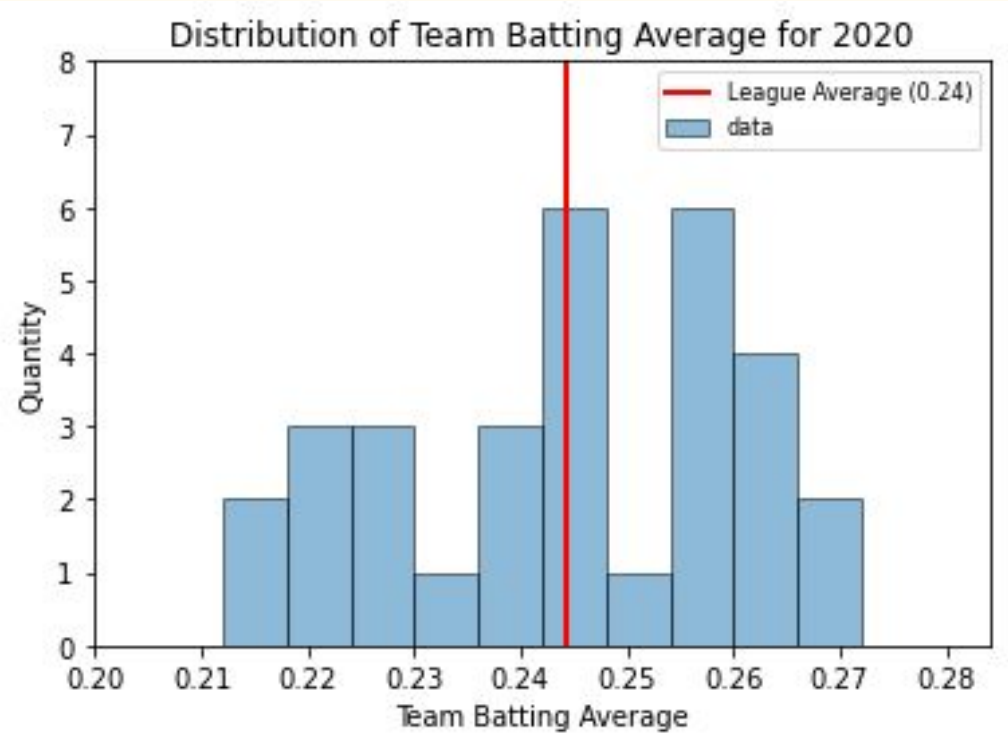
Runs (R_y)



Mean:	278.13
Median:	275.00
Mode:	248.00
Standard Deviation:	31.52

The chart looks moderately uniform without the presence of tails or skewness.

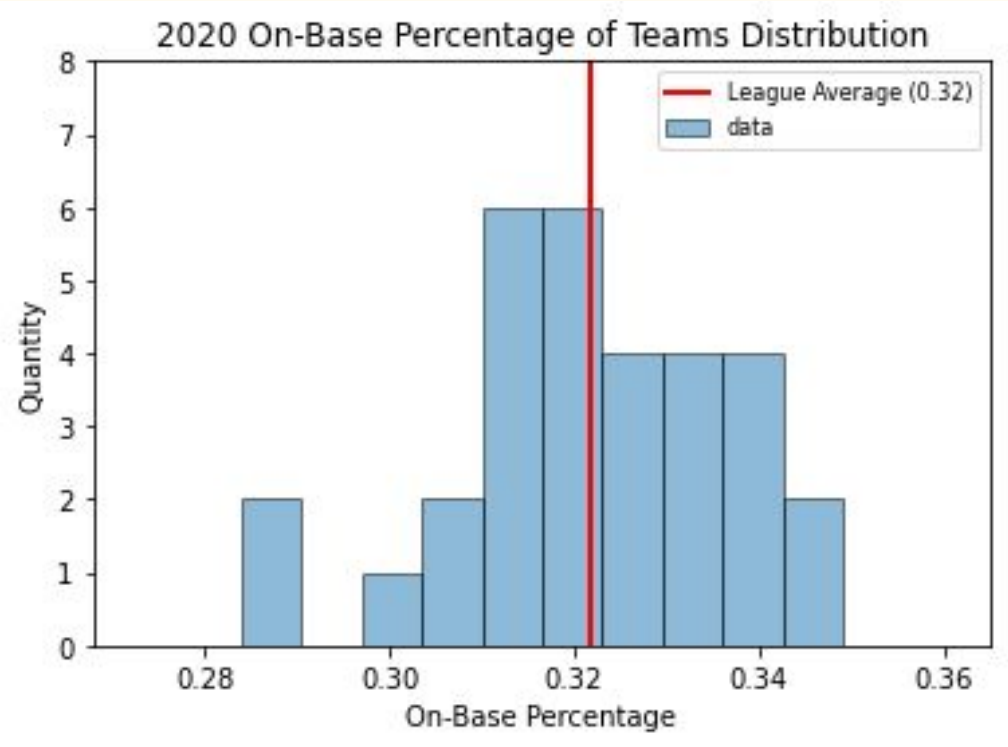
Batting Average (BA)



Mean:	.244
Median:	.245
Mode:	.257
Standard Deviation:	.016

The histogram is slightly skewed left but not to an overwhelming extent.

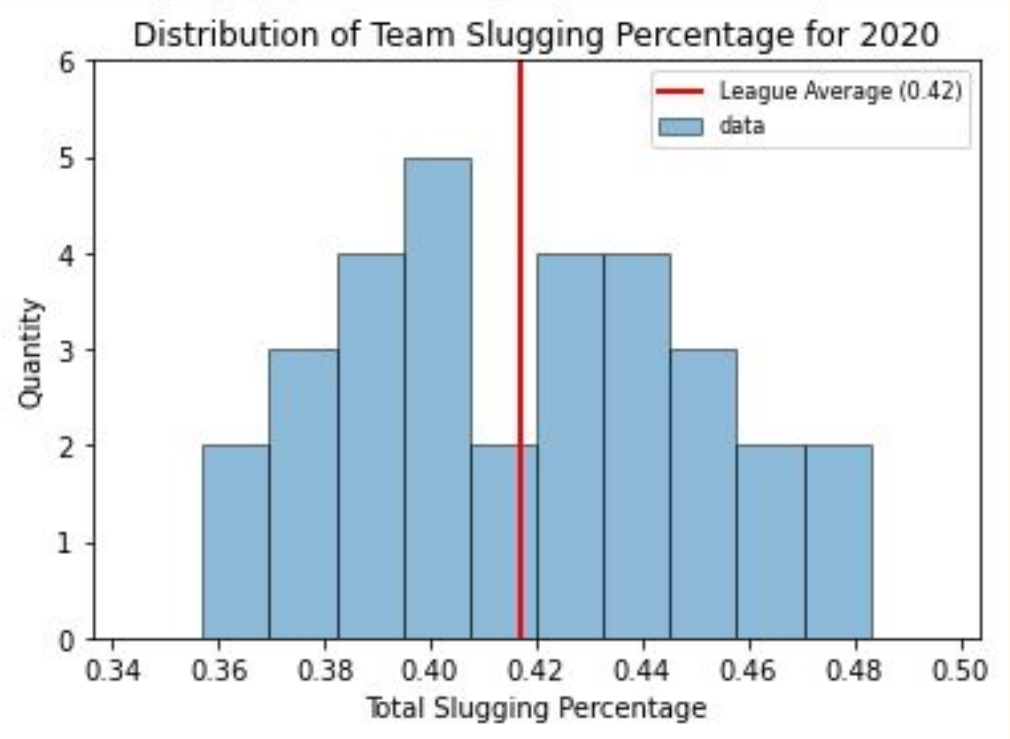
On-Base Percentage (OBP)



Mean:	.322
Median:	.322
Mode:	.312
Standard Deviation:	.015

The plot has a minimal right skew. It seems slightly offset by the terrible lowest valued OBPs. The represent terrible offensive ability and the 2 worst teams of 2020 season.

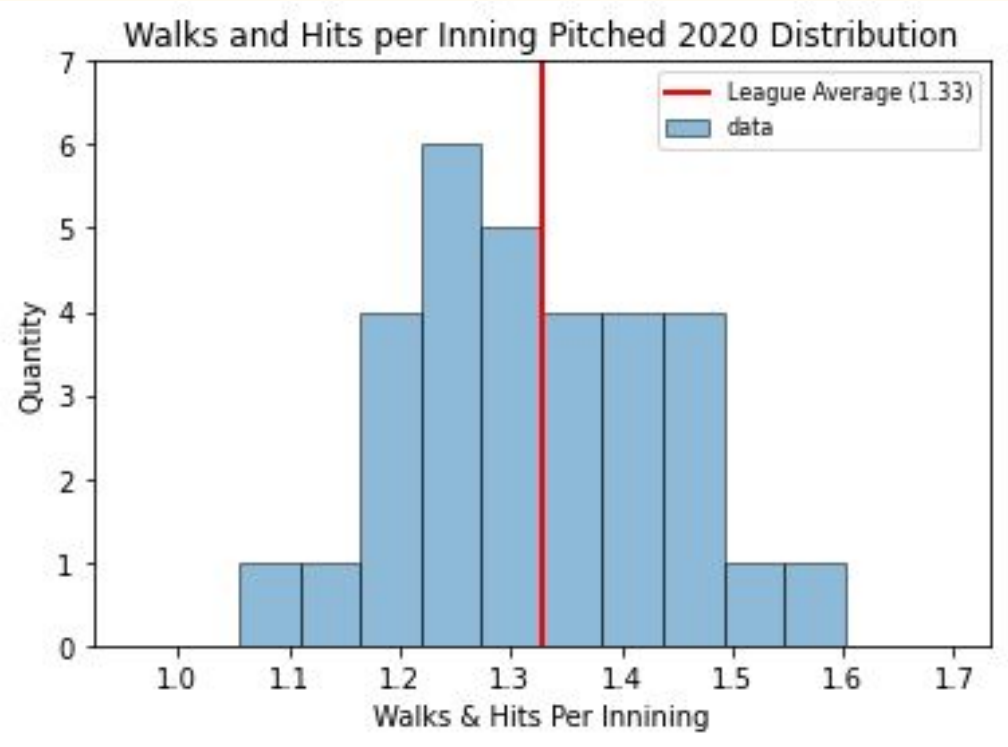
Slugging Percentage (SLG)



Mean:	.417
Median:	.418
Mode:	.483
Standard Deviation:	.034

Based on the Mean, Median, and Mode, the histogram is slightly skewed left, but the graph seems to have normality to it.

Walks & Hits per Inning Pitched (WHIP)



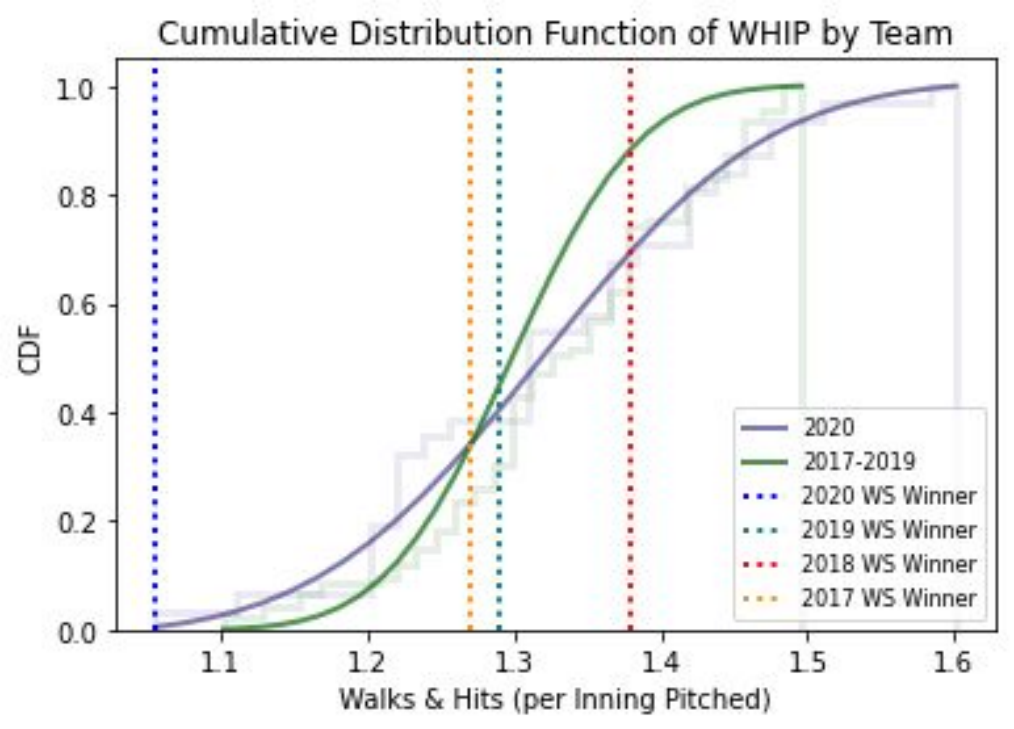
Mean:	1.328
Median:	1.321
Mode:	1.315
Standard Deviation:	.122

The WHIP distribution looks normal with a small variability. The WHIP is also a great representation of performance providing a great indicator of defensive capabilities.

Cumulative Distribution Functions - CDF

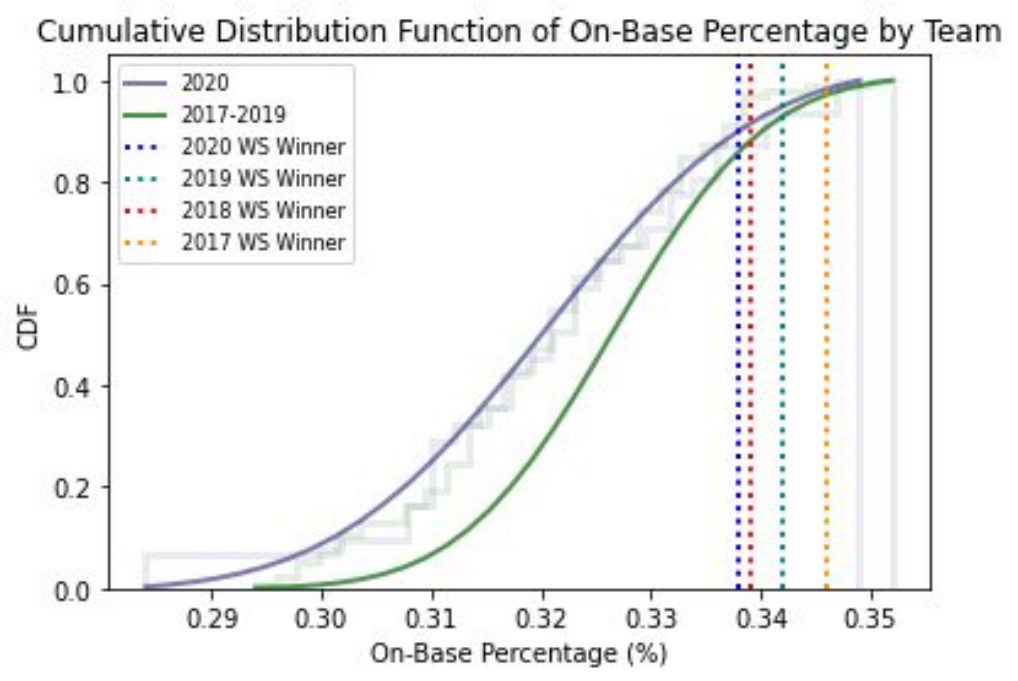
- Walks and Hits per Inning Pitched for Each Team. Specific Portrayal of a Pitcher's Prowess. Comparing the 2020 Seasons to the 2017 to 2019 Seasons.
- Team Earned Run Average (Earned Runs per 9 Innings Pitched). Representation of a Team's Pitching and Defensive abilities. (2020 vs 2017-2019 Seasons).
- Slugging Percentage Average per Team (Number of Bases per Possible Batting Attempts), comparing the 2020 season to the previous 3 seasons ('17 - '19). Similar to Batting Average but factors it base value per hit.
- Team Batting Averages, a team's offensive ability valued by reaching a base by hit, (Hits per Possible Batting Attempt) from 2020 compared to 2017 2019.

CDF - Walks & Hits Per Inning (WHIP)



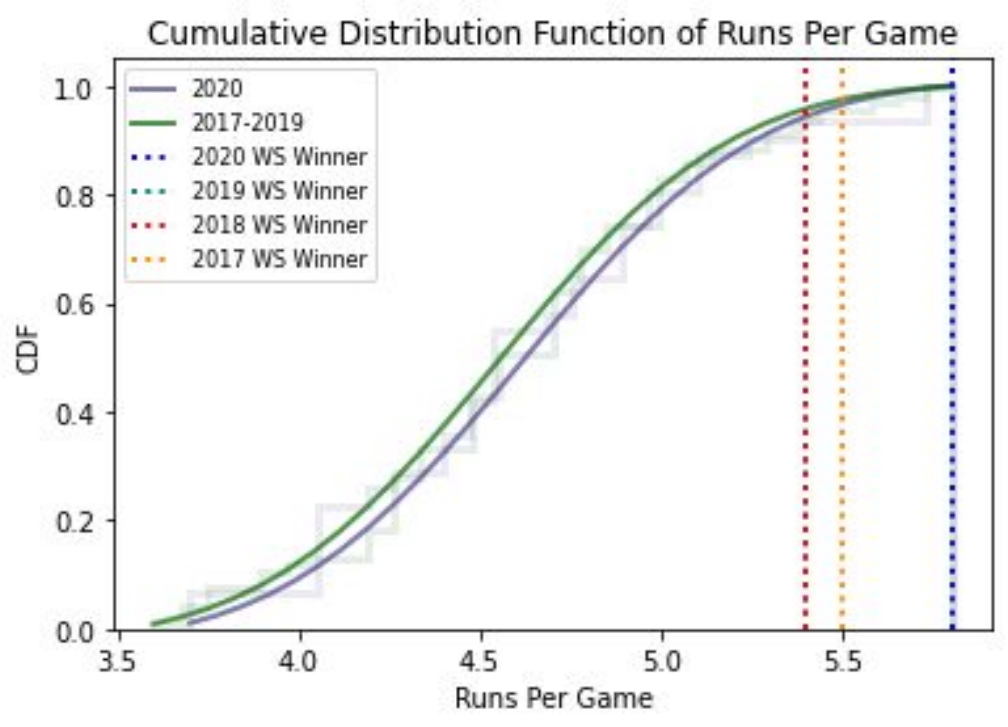
- 2020's Maximum value is ~6.66% greater and Minimum value ~4.5% less than 2017-2019 Seasons Maximum and Minimum
- Pitching in 2020 was better and worse compared to 2017-2019, implying volatile pitching staffs
- Does not produce great results that WHIP is a significant statistic for success in the MLB

CDF - On-Base Percentage (OBP)



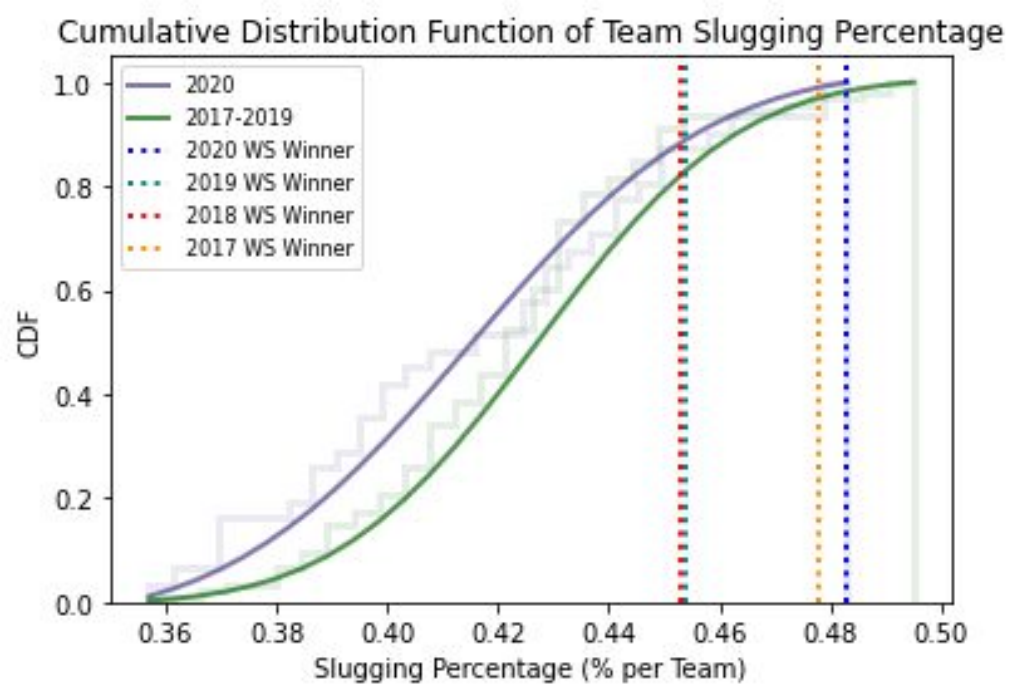
- 2020 produced worse OBP compared to 2017-2019
- The difference is not extreme but 2020s worst OBP is ~3.4% worse than the worst OBP team from the 2017-2019 seasons
- There is a great indication that on-base % is a leading statistic to success
- Past 4 World Series Winners are in the top 15% of best OBP teams

CDF - Runs Per Game (RunsPGame)



- The Distribution of Runs Per Game over the years are extremely similar
- The 2017-2019 seasons have a marginally smaller minimum which is understandable since its comparing 90 possible values vs 30 in the 2020 season
- Understandably, being a leader in scoring runs leads to a successful season with the last 4 World Series winners being in the top %5 of teams

CDF - Slugging Percentage (SLG)



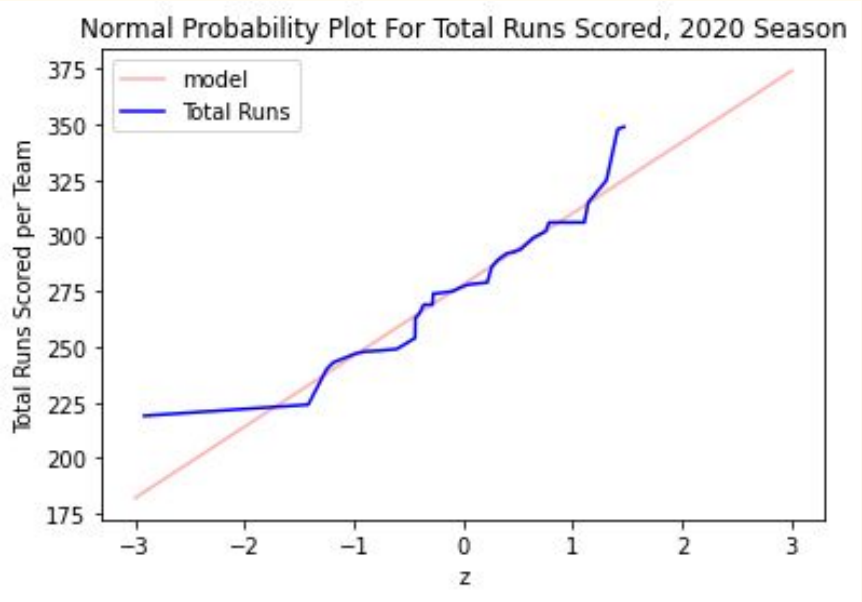
- SLG during the 2020 season is slightly worse compared to the 2017-2019 seasons, possibly from batters “going for more,” potentially leading to more strikeouts
- Specifically, the 2017-2019 seasons had higher maximum SLG values, at 0.495 vs 0.483
- Similar to On-Base %, SLG also appears to be an extremely important indicator of success during the season

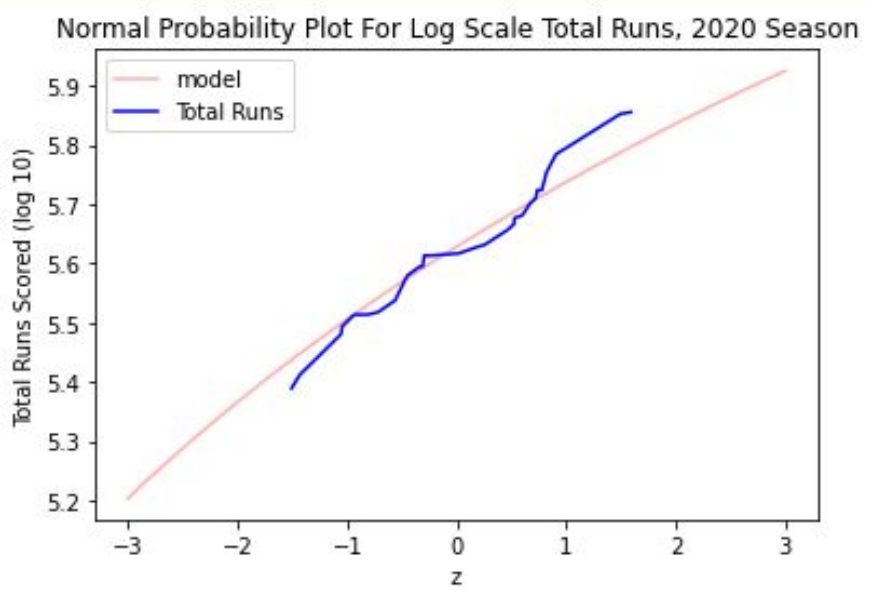
Results of CDFs - Main Takeaways

Based on the information provided through the Cumulative Distribution Functions, there are a few conclusions that can be made of the variables in the charts:

- Runs Scored per Game and On-Base Percentage are the most significant variables, that were plotted, with Slugging Percentage being the third most significant, that lead to a successful team, achieved by making the playoffs and, more specifically, winning the World Series
- Although the 2020 World Series winners, the Los Angeles Dodgers, had the lowest Walks and Hits per Inning Pitched (WHIP) in the past 4 years, there appears to be little importance on keeping batters off the bases compared to stopping runners from scoring, with the 3 previous World Series winners being close to the league average of WHIP

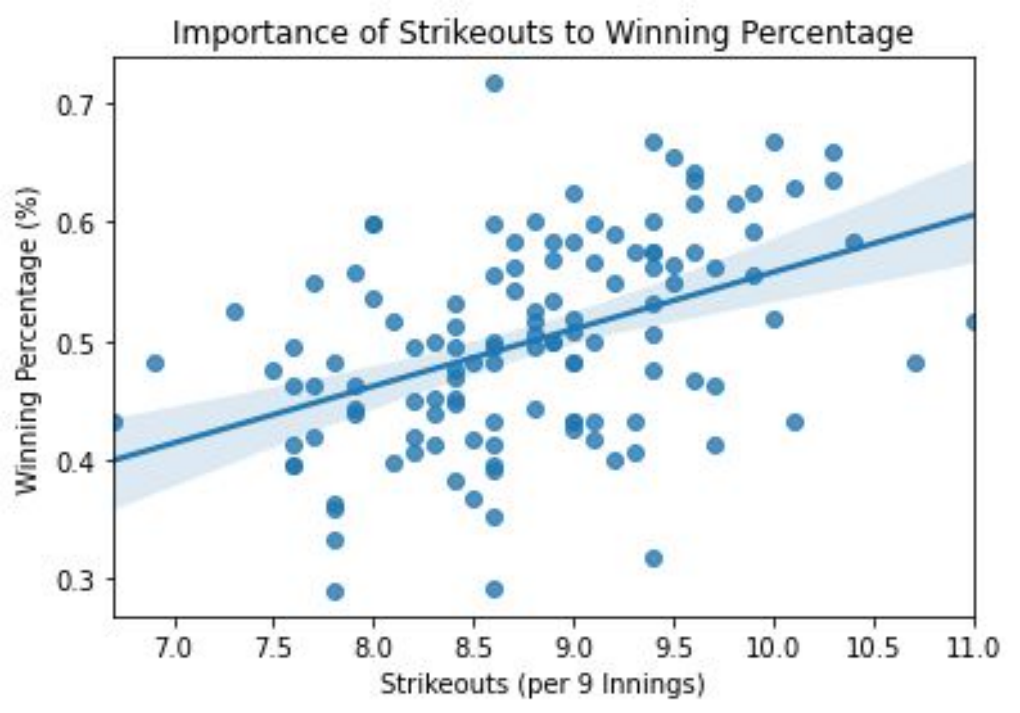
Normal Probability Plots - Total Runs





Based on the plots above, it does appear the log scale plot has a better relationship with the model compared to the normal plot. Unfortunately, neither represent an amazing match to the model comparatively.

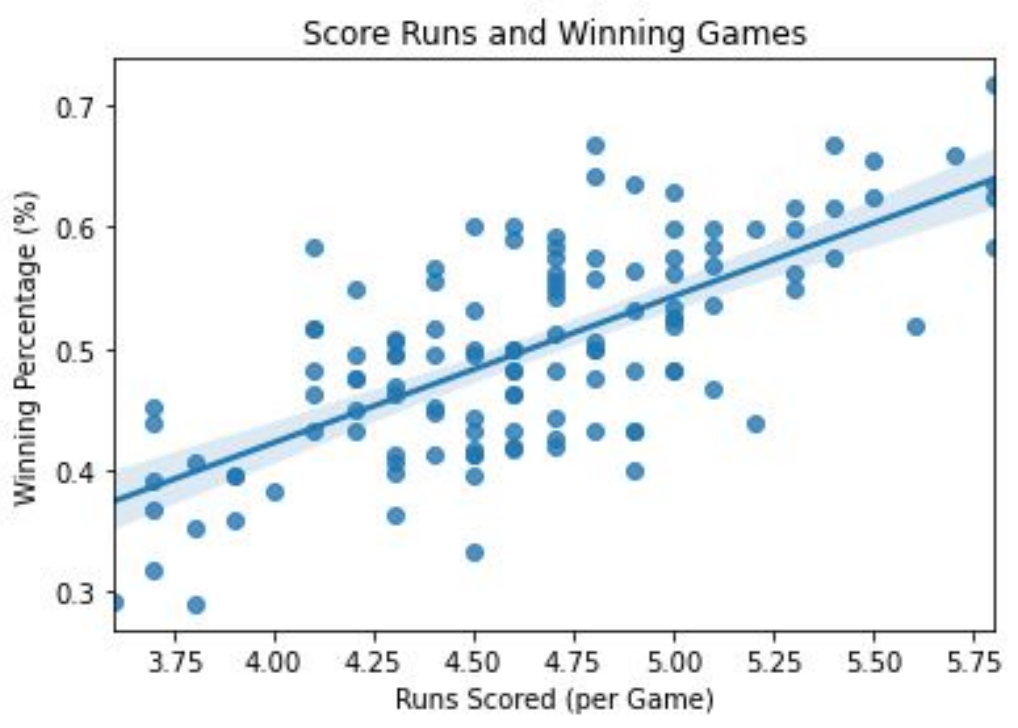
Comparing Win % and Strikeouts



Correlation:	0.441 (Pearson)
	0.474 (Spearman)
Covariance:	0.030
P-Value:	2.93e-07
	“Significant”

Strikeouts have not been a major statistic throughout my analysis but I am fascinated by this as a baseball fan. There is a positive correlation between winning and strikeouts but it is not as strong as I was anticipating prior to the analysis.

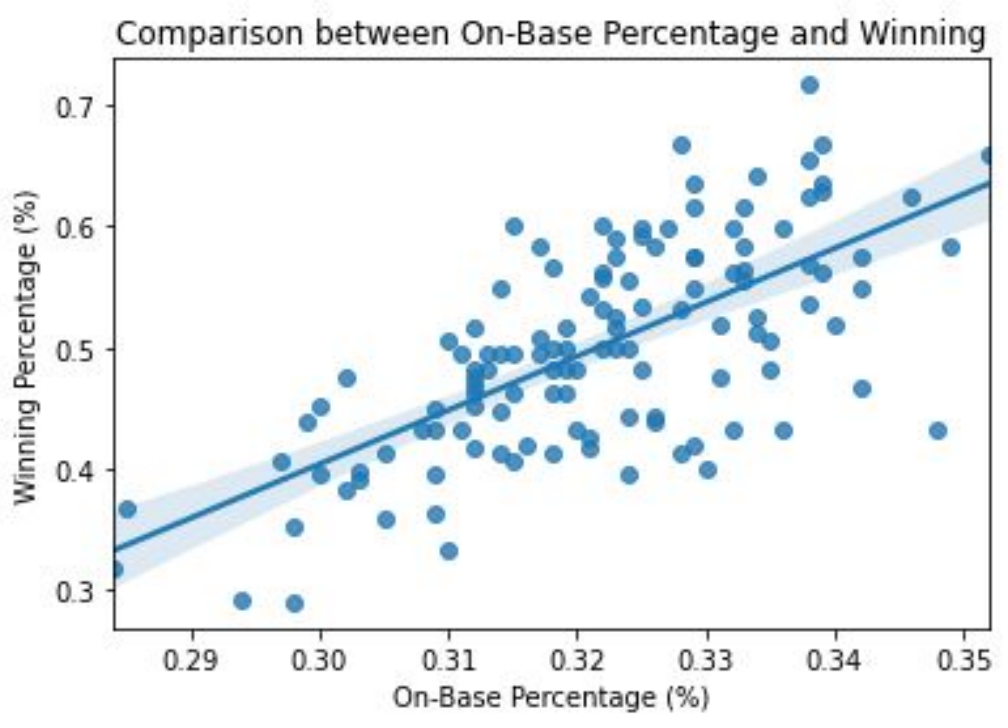
Scoring Runs and Winning Games



Correlation:	0.696 (Pearson)
	0.653 (Spearman)
Covariance:	0.030
P-Value:	3.06e-19
	“Significant”

It is understandable to think that scoring runs leads to winning baseball games. This plot makes that point obviously clear. Surprisingly, there are values more spread out than I was expecting but I attribute that to solid defense by the team.

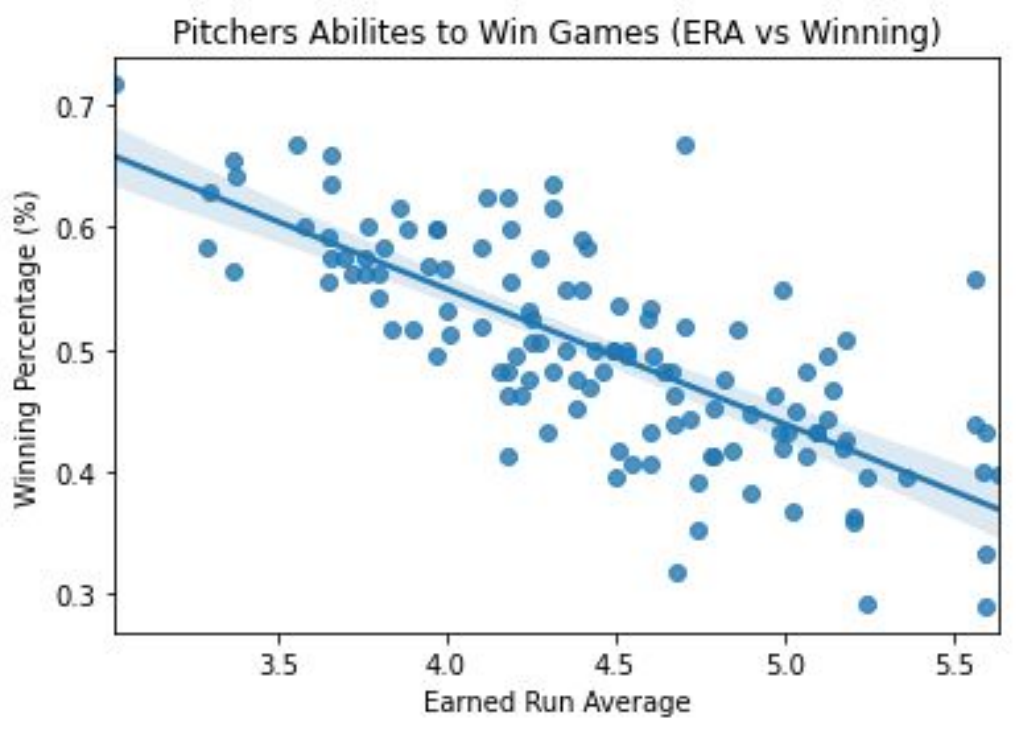
OBP's Influence on Winning



Correlation:	0.682 (Pearson)
	0.657 (Spearman)
Covariance:	0.0008
P-Value:	2.27e-18
	"Significant"

I was not shocked by this correlation. There was a stronger correlation between Runs scored and Winning but I felt it more obvious than OBP, SLG, or BA. OBP is the causation for the Winning.

ERA vs Winning



Correlation:	-0.747 (Pearson)
	-0.756 (Spearman)
Covariance:	-0.037
P-Value:	2.155e-23
	"Significant"

Earned Run Average and Winning have an extremely strong negative correlation between them. Meaning, the more runs a pitcher gives up, the higher the probability of losing are.

Outlook on the Variable Comparison

The comparison between strikeouts and winning percentage is the most fascinating as a baseball fan. Although the correlation is strong, the importance of WHIP and ERA to winning are far superior. With that said, there is a stronger relationship between strikeouts and ERA as well as strikeouts and WHIP, with correlations of -0.524 and -0.566, respectively, with strikeouts being a causation for ERA and WHIP.

There is also an extremely strong correlation between Runs Scored and OBP (0.871) so I felt OBP there are strong indicators for OBP being the most significant variable to winning baseball games as stated in “Moneyball: The Art of Winning an Unfair Game.”

The biggest takeaway from the correlation plots is the obvious importance of runs in winning baseball games, scoring many and giving up as few as possible, with numerous variables being influences on both offensive and defensive aptitude.

Multiple Regression Analysis - Winning %

OLS Regression Results

Dep. Variable:	W_L_percent	R-squared:	0.902
Model:	OLS	Adj. R-squared:	0.887
Method:	Least Squares	F-statistic:	59.95
Date:	Sun, 15 Nov 2020	Prob (F-statistic):	9.56e-13
Time:	02:11:28	Log-Likelihood:	66.713
No. Observations:	31	AIC:	-123.4
Df Residuals:	26	BIC:	-116.3
Df Model:	4		
Covariance Type:	nonrobust		

Multiple Regression Analysis - Winning %

OLS Regression Results

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.6900	0.098	7.030	0.000	0.488	0.892
TB	-0.0005	0.000	-3.152	0.004	-0.001	-0.000
TotalRuns	0.0024	0.000	5.702	0.000	0.002	0.003
ERA2	-0.0519	0.012	-4.466	0.000	-0.076	-0.028
ERA3	0.0061	0.002	3.595	0.001	0.003	0.010

Omnibus:	0.525	Durbin-Watson:	2.216
Prob(Omnibus):	0.769	Jarque-Bera (JB):	0.042
Skew:	0.041	Prob(JB):	0.979
Kurtosis:	3.159	Cond. No.	1.57e+04

Multiple Regression Analysis - Winning %

- Given the numerous variables available in the dataset, the modeling of the winning percentage was interesting to factor down.
- I found the overall best simplified model was [Winning % ~ ERA+ RBI] but the produced approximately an 83% R-Squared.
- This comprehensive model accounts for Total Bases, Total Runs, and two factors of Earned Run Average, ERA squared and ERA cubed.
- Depending on the audience of the presentation (MLB Mangers & Executives), I would model Winning % with only ERA and RBI to work with a simplified model and easily understandable coefficients anyone could quickly understand.
    - More RBIs => More Wins, Greater ERA => More Losses
- The coefficient I found interesting was the Total Bases which linearly decreases the model for more bases achieved. This is surprising given the positive correlation that exists between Total Bases and Winning %.

Hypothesis Testing: T-Test

The focus of the hypotheses testing will be on:

1. The difference between the AL and NL leagues, specifically prior to the changes made in 2020, where the National League do not utilized a designated hitter in place of the pitcher, forcing pitchers to hit, potentially leading to lower batting average, slugging percentage, and ERA for the National League
2. Whether or not the changes in the 2020 season, more games against opposite league teams (AL vs NL), three (3) batter minimum for relief pitchers, and full-time designated hitter in the National League, effected the statistical outcomes of teams in drastic ways

The main test that will be employed to analyze these hypotheses will be the t-test for significance.

T-Test Analysis: AL vs NL

American League vs National League Statistics (Statistics from 2017-2019):

- Batting Average
  - Test Statistic = 0.8865
  - P-Value = 0.3778, “Insignificant”
    - AL Average: 0.2526 vs NL Average: 0.2507
- Earned Run Average
  - Test Statistic = 1.3515
  - P-Value = 0.1801, “Insignificant - but close to consideration”
    - AL Average: 4.5282 vs NL Average: 4.3696
- Slugging Percentage
  - Test Statistic = 0.1195
  - P-Value = 0.9053, “ Insignificant”
    - AL Average: 0.4276 vs NL Average: 0.4190

T-Test Analysis: 2020 Changes Effects

2020 Season Changes:

- American League:
  - Batting Average
    - Test Statistic = -2.8841
    - P-Value = 0.0055, "**Significant**"
      - 2020: 0.2426 vs 2017-2019: 0.2526 (Decrease of ~4%)
  - Earned Run Average
    - Test Statistic = -0.5240
    - P-Value = 0.6022, "Insignificant"
  - Slugging Percentage
    - Test Statistic = -1.6839
    - P-Value = 0.0976, " *Reasonably Significant*"
      - 2020: 0.4137 vs 2017-2019: 0.4276 (Decrease of ~3.3%)

T-Test Analysis: 2020 Changes Effects

2020 Season Changes:

- National League:
  - Batting Average
    - Test Statistic = -1.2393
    - P-Value = 0.2201, “Insignificant”
  - Earned Run Average
    - Test Statistic = 0.5921
    - P-Value = 0.5561, “Insignificant”
  - Slugging Percentage
    - Test Statistic = 0.1195
    - P-Value = 0.9053, “Insignificant”

Hypothesis Testing: Pearson Correlation

The Pearson Correlation test variable's relationships with one another.

The majority of the hypothesis testing was done on the scatter plots:

- Winning vs Strikeouts
- Winning vs Runs per Game
- Winning vs On-Base Percentage
- Winning vs Earned Run Average

All of these variables share an extremely strong relationship with winning baseball games, which is partially the reason for choosing them. I was under the impression they were significant, with the exception of Strikeouts before further analyzing their importance. With that said, strikeouts also are reasonably correlated to ERA which is strongly correlated to Winning, which is the reason why I believe strikeouts become a significant factor for winning.

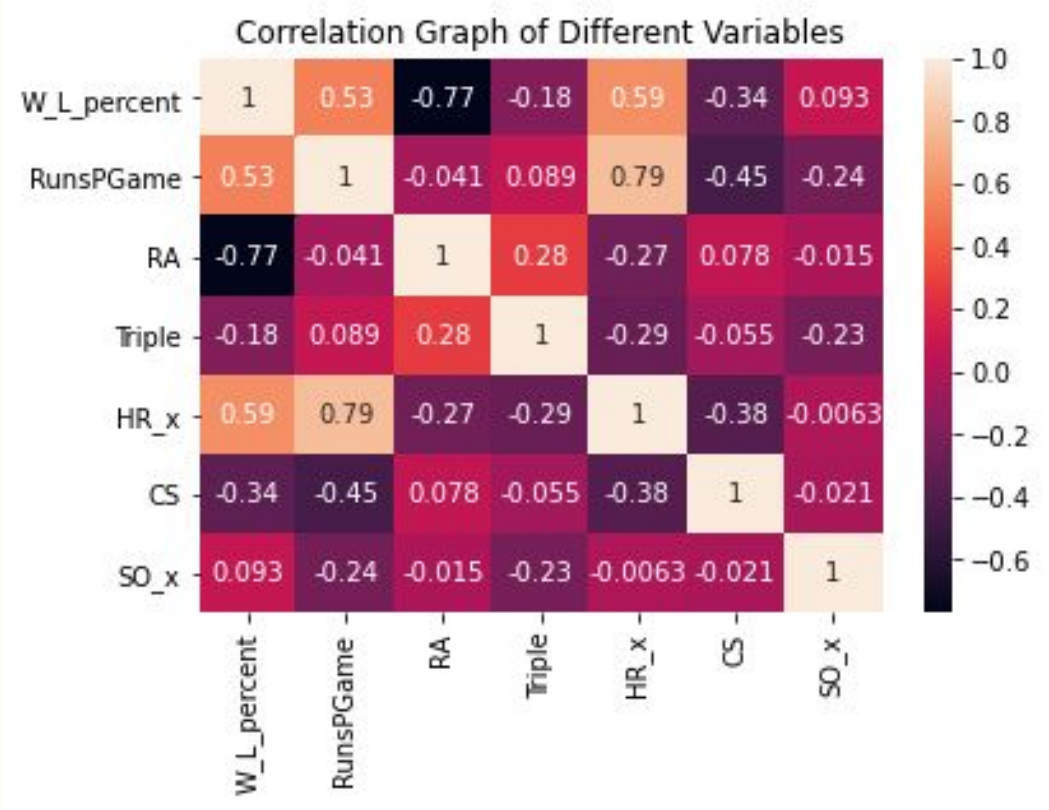
Hypothesis Testing - Conclusion

- The changes to the baseball rules in 2020 had a considerably greater effect on the American League, hitting considerably worse than previous years, than the National League, which was not anticipated.
  - A more complete analysis on the American League vs National League statistics supported my claims below regarding the Designated hitter in the National League, with the National League scoring better in 6 out of the 8 compared statistics
  - I mainly contribute this to the COVID-19 implementations of AL teams playing more games against NL teams than in a non-pandemic affected season
- There appears to be little significant difference between the American League and National League when it comes to the role of a Designated Hitter batting for the pitcher
  - The AL does have ~3.5% greater ERA as well as slightly better Batting Average and Slugging Percentage than the NL, which supports the difference in having a Designated Hitter but the values do not appear to be absolutely significant.

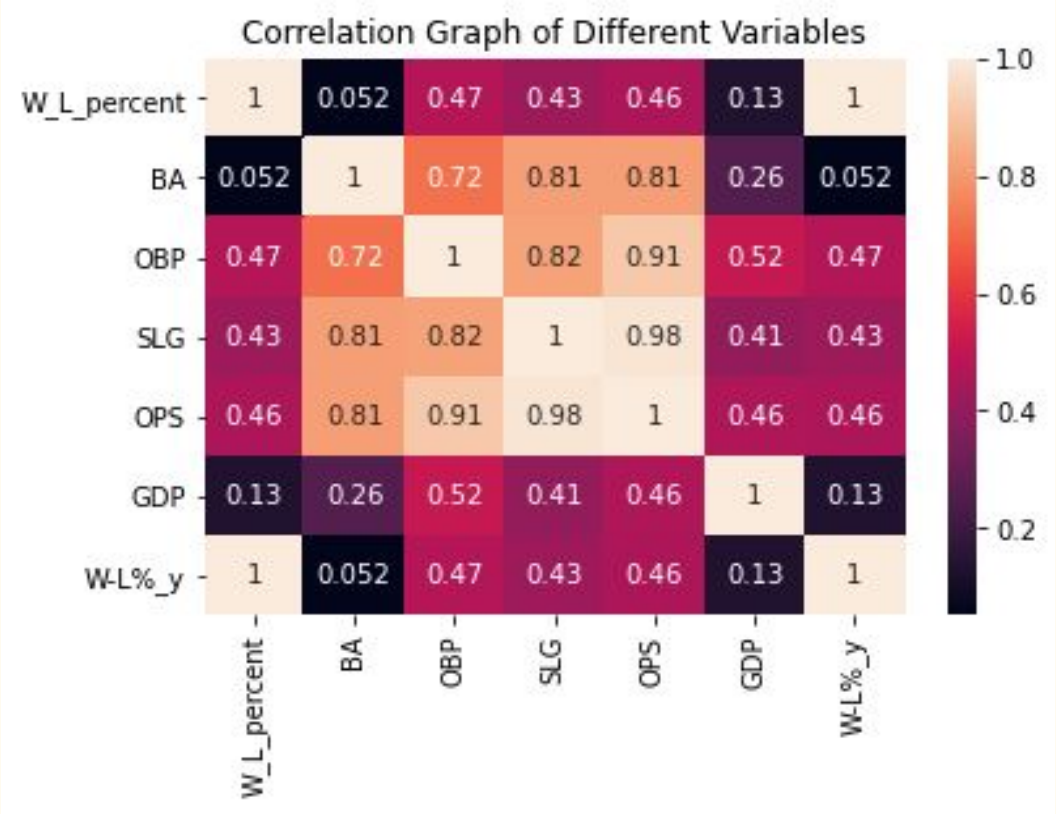
Additional Charts from the Analysis

- ★ The following slides are different distributions and heat maps that I had used throughout the analysis.
- ★ They either were not of utmost importance or were depicted another way
- ★ Many of them are packed with too much information at causing contrasting colors and lines or may not display the information the the most desired way
- ★ Hope they are helpful!

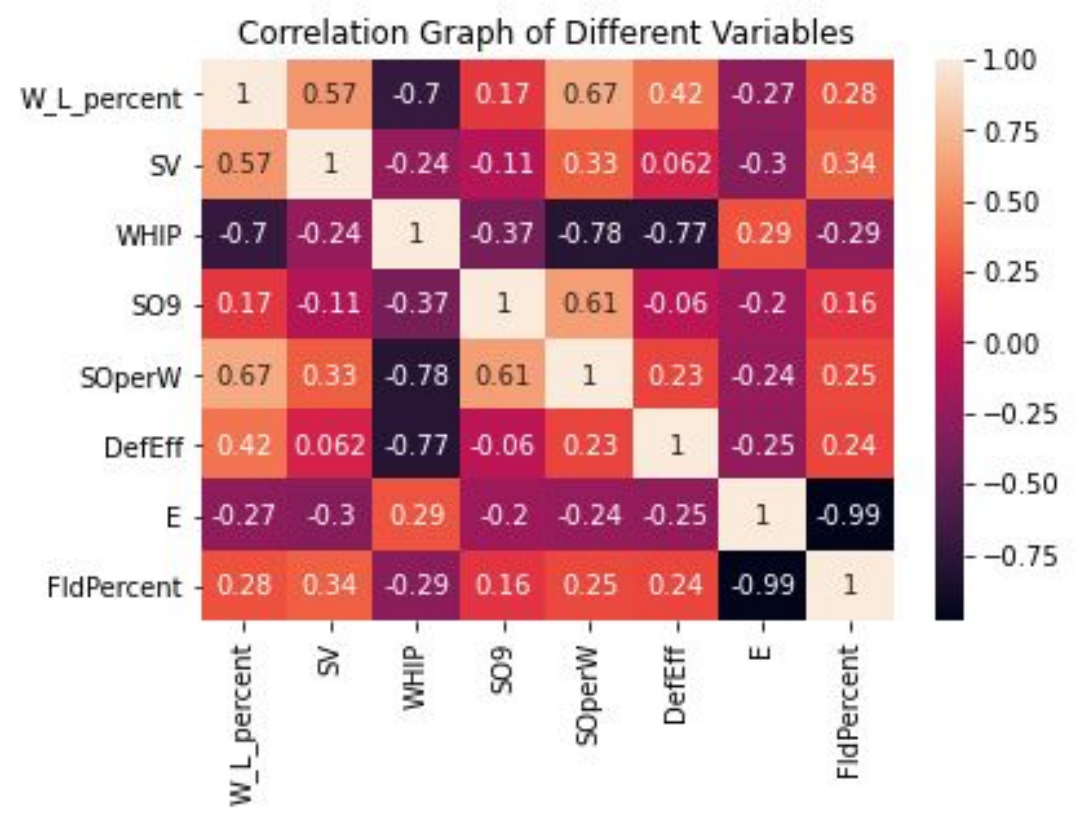
Correlation Heatmap of Some Variables



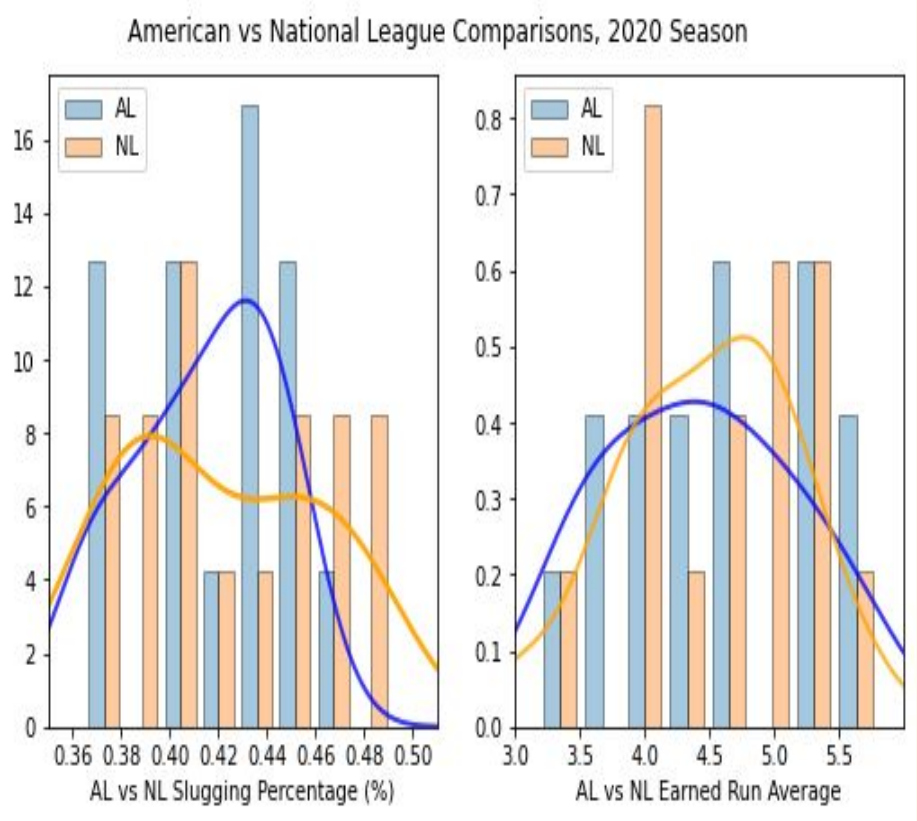
Correlation Heatmap of Some Variables

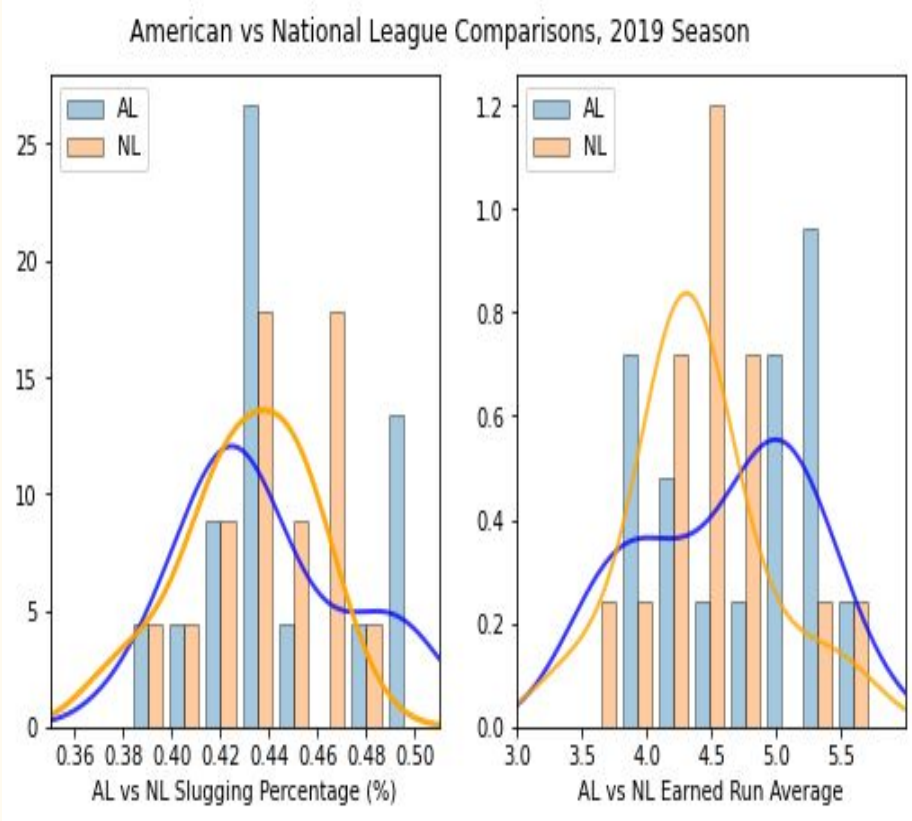


Correlation Heatmap of Some Variables

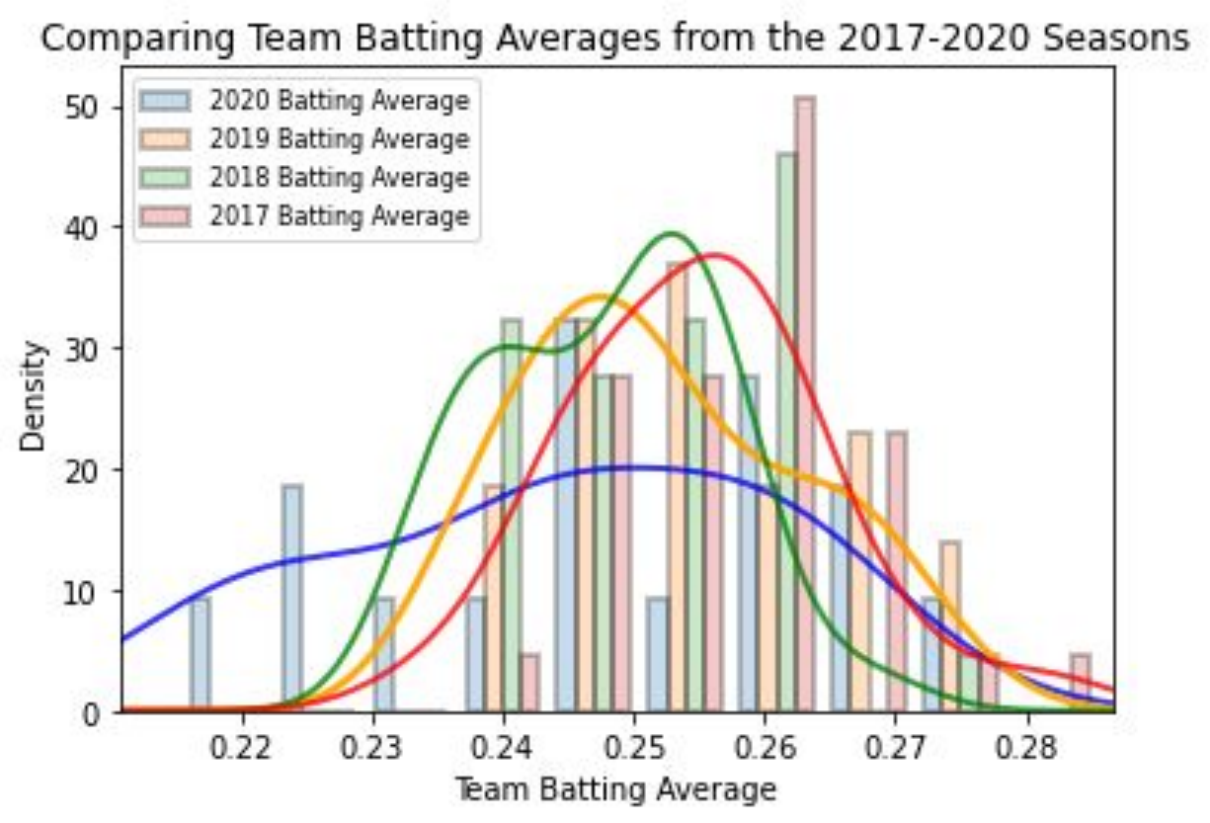


AL & NL Comparison: 2019 - 2020

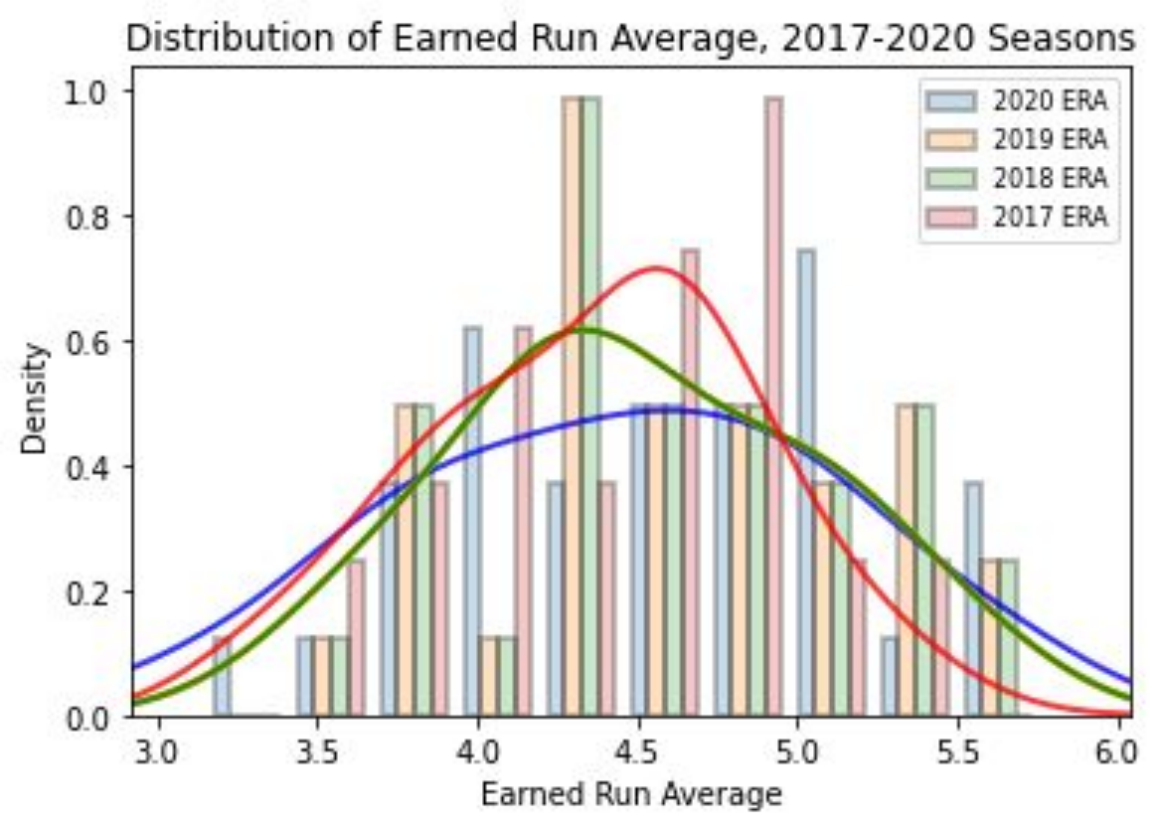




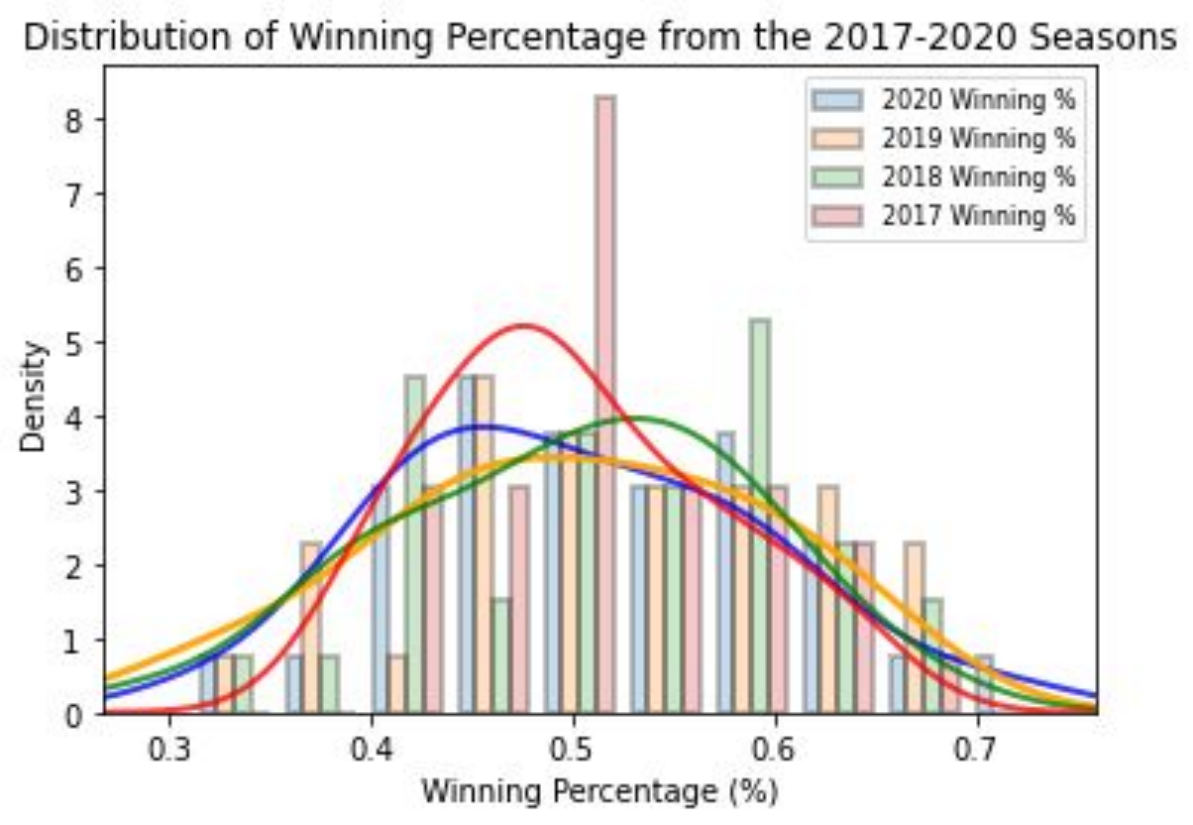
Distribution of Batting Average



Distribution of Earned Run Average



Distribution of Winning Percentage



References

1. Lewis, M. (2003). *Moneyball: The Art of Winning An Unfair Game*. NewYork: Norton.

2. Schoenfield, D. (2018, July 10). *The AL has more starts, but is the NL as better league?* ESPN. https://www.espn.com/mlb/story/_/id/24047323/the-al-more-stars-nl-better-league

3. *Major League Baseball Win Totals*. (n.d.). Sports Reference LLC. https://www.baseball-reference.com/leagues/MLB/

4. Castrovince, A. (2018, May 7). Mike Trout, Mookie Betts agree: This stat is best. MLB.com. https://www.mlb.com/news/mlb-players-vote-for-stats-they-value-most-c274986480

5. Rymer, Z.D. (2013, May 28). Which Is Baseball's Superior Overall League, the AL or the NL? BleacherReport.