

Data 621 Homework 1

Group 3: Amanda Arce, Austin Chan, Jithendra Seneviratne, Sheryl Piechocki

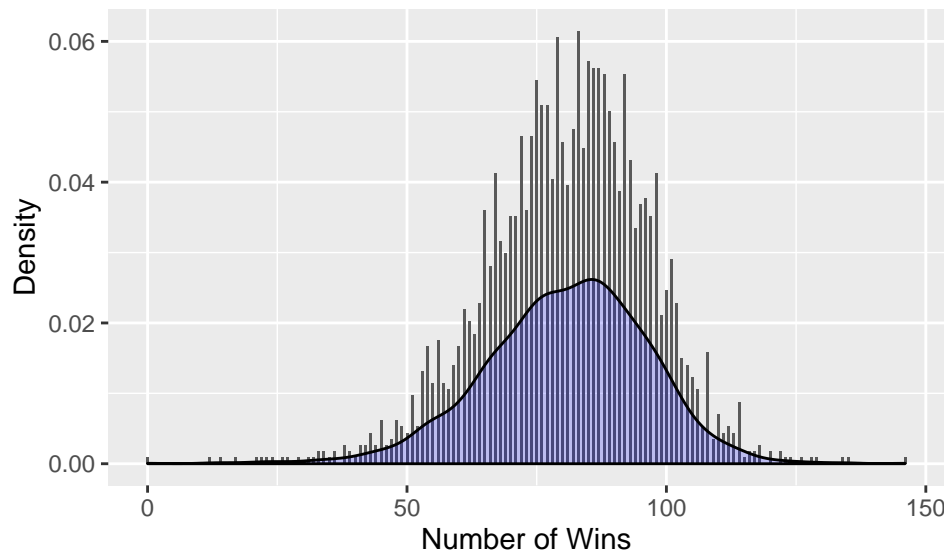
2/15/2020

Objective: Build a multiple linear regression model on the training data to predict the number of wins for baseball teams.

1. DATA EXPLORATION

The data used in this analysis, consists of performance statistics for baseball teams from the years 1871-2006. Each record represents the performance of one team for one year. There are 2,276 records and 17 baseball statistics, including the target variable wins. Statistics include batting information, such as hits, doubles, triples, homeruns, strikeouts, and walks. Also, given are pitching statistics of hits allowed, walks allowed, homeruns allowed, and strikeouts by pitchers. Other information regarding errors, stolen bases, caught stealing, hit by pitch, and double plays is also available.

The distribution of the target variable, Wins is below. It appears to be normally distributed, with a mean of 80.79 and standard deviation 15.75.



Summary statistics for each independent variable are provided below. The variables Caught Stealing and Hit by Pitch have a large number of missing values and therefore will be excluded from all subsequent analysis.

##	Hits	Doubles	Triples	Home Runs
##	Min. : 891	Min. : 69.0	Min. : 0.00	Min. : 0.00
##	1st Qu.:1383	1st Qu.:208.0	1st Qu.: 34.00	1st Qu.: 42.00
##	Median :1454	Median :238.0	Median : 47.00	Median :102.00
##	Mean :1469	Mean :241.2	Mean : 55.25	Mean : 99.61
##	3rd Qu.:1537	3rd Qu.:273.0	3rd Qu.: 72.00	3rd Qu.:147.00

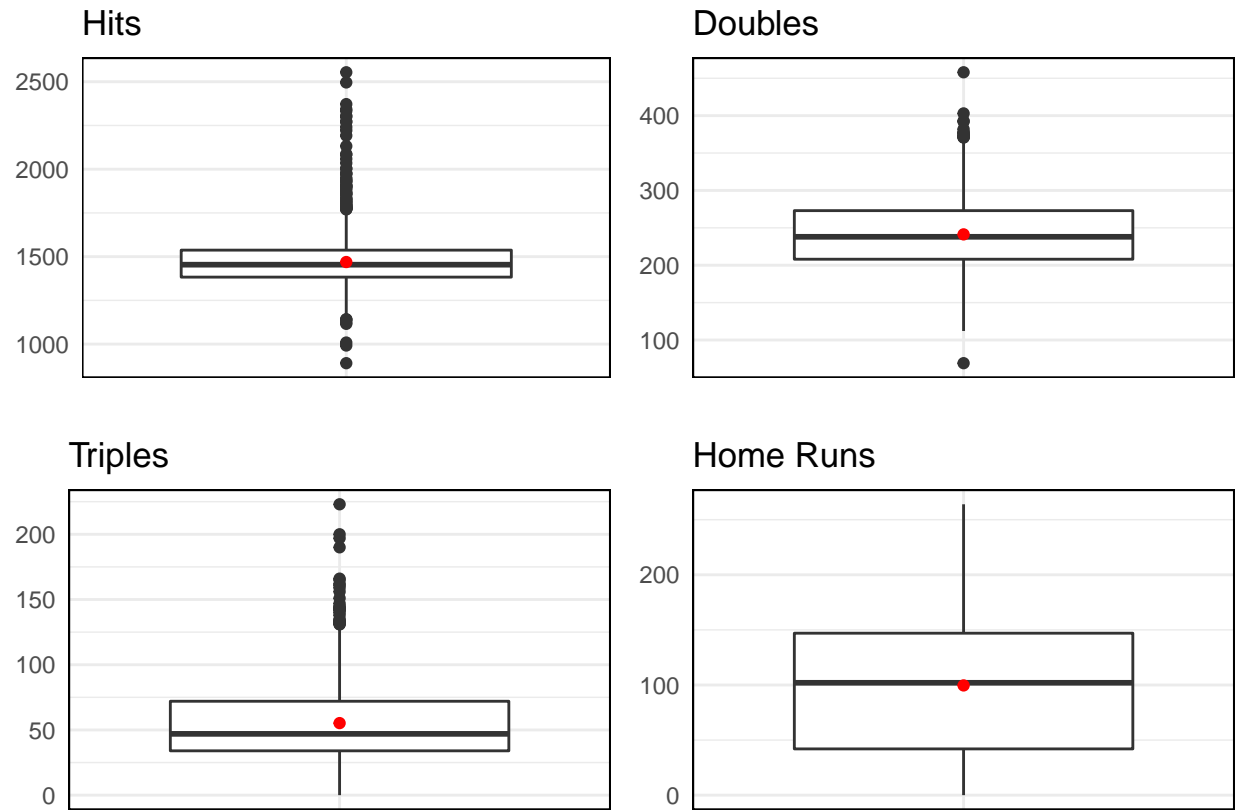
```

## Max. :2554 Max. :458.0 Max. :223.00 Max. :264.00
##
## Walks Batter SO Stolen Bases Caught Stealing
## Min. : 0.0 Min. : 0.0 Min. : 0.0 Min. : 0.0
## 1st Qu.:451.0 1st Qu.: 548.0 1st Qu.: 66.0 1st Qu.: 38.0
## Median :512.0 Median : 750.0 Median :101.0 Median : 49.0
## Mean :501.6 Mean : 735.6 Mean :124.8 Mean : 52.8
## 3rd Qu.:580.0 3rd Qu.: 930.0 3rd Qu.:156.0 3rd Qu.: 62.0
## Max. :878.0 Max. :1399.0 Max. :697.0 Max. :201.0
## NA's :102 NA's :131 NA's :772
## Hit by Pitch Hits Allow Home Runs Allow Walks Allow
## Min. :29.00 Min. : 1137 Min. : 0.0 Min. : 0.0
## 1st Qu.:50.50 1st Qu.: 1419 1st Qu.: 50.0 1st Qu.: 476.0
## Median :58.00 Median : 1518 Median :107.0 Median : 536.5
## Mean :59.36 Mean : 1779 Mean :105.7 Mean : 553.0
## 3rd Qu.:67.00 3rd Qu.: 1682 3rd Qu.:150.0 3rd Qu.: 611.0
## Max. :95.00 Max. :30132 Max. :343.0 Max. :3645.0
## NA's :2085
## Pitcher SO Errors Double Plays
## Min. : 0.0 Min. : 65.0 Min. : 52.0
## 1st Qu.: 615.0 1st Qu.: 127.0 1st Qu.:131.0
## Median : 813.5 Median : 159.0 Median :149.0
## Mean : 817.7 Mean : 246.5 Mean :146.4
## 3rd Qu.: 968.0 3rd Qu.: 249.2 3rd Qu.:164.0
## Max. :19278.0 Max. :1898.0 Max. :228.0
## NA's :102 NA's :286

```

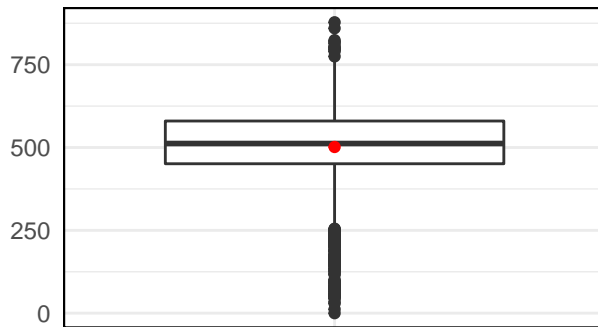
Box plots with the mean denoted as a red circle are provided.

The box plots of the batting variables show many outliers in Hits and Triples. The spread on Home Runs is large.

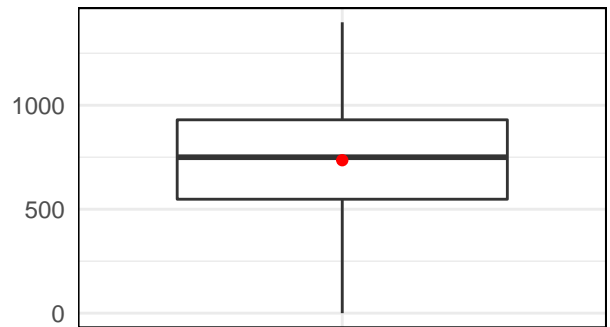


Further, Walks, Stolen Bases, and Errors also have quite a few outliers.

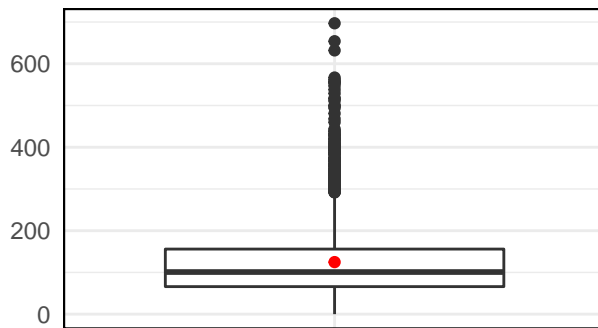
Walks



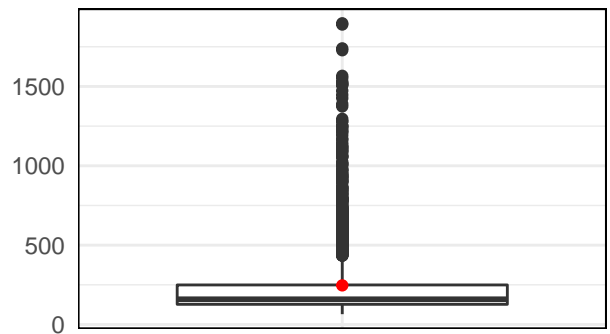
Batter Strike Outs



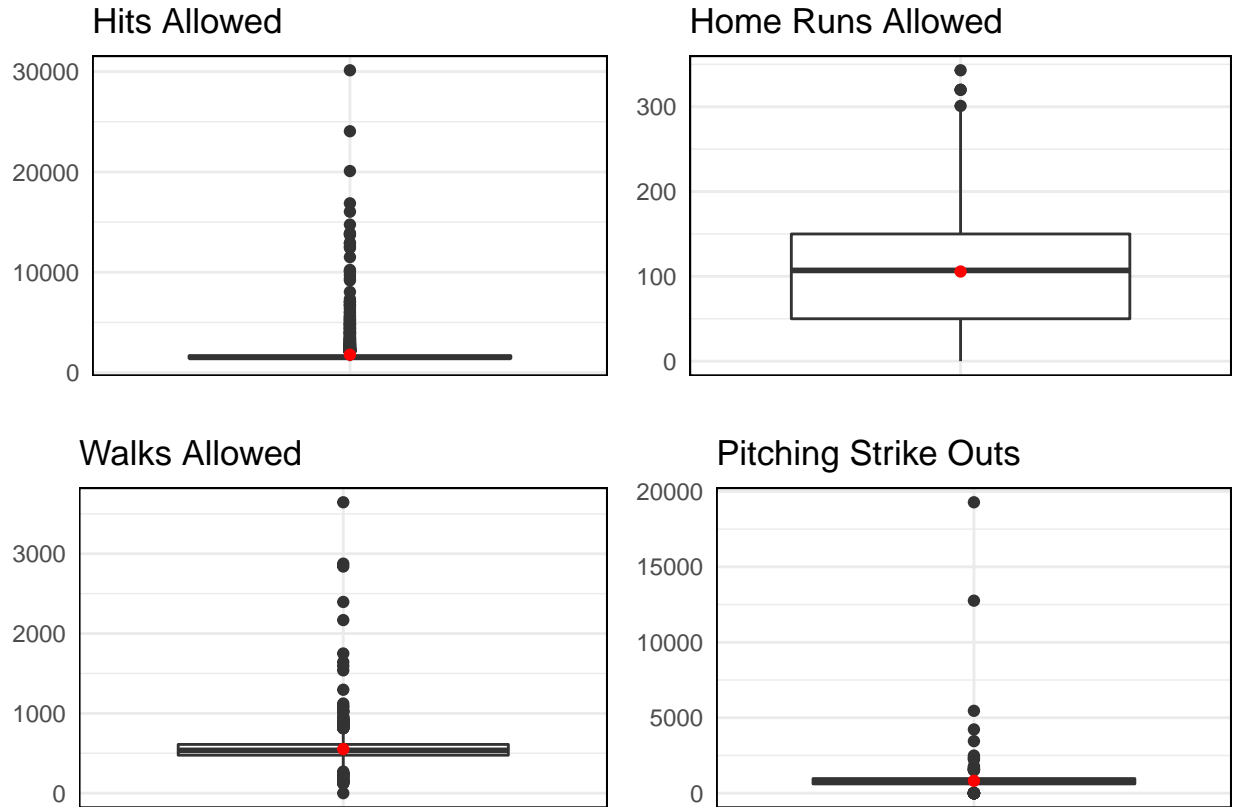
Stolen Bases



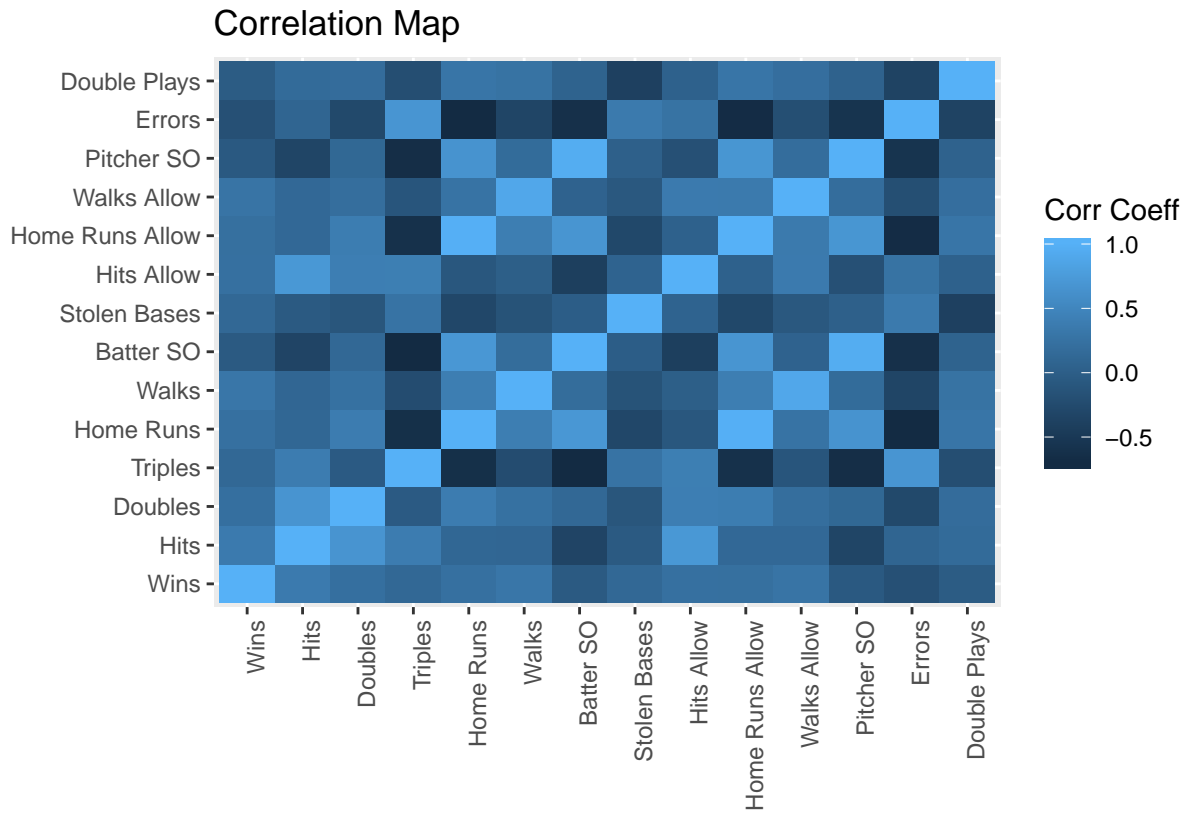
Errors



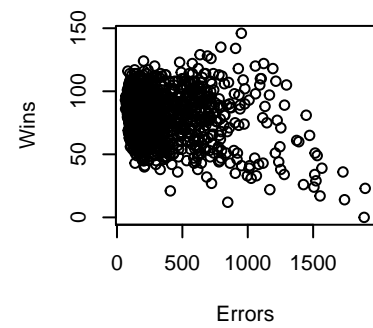
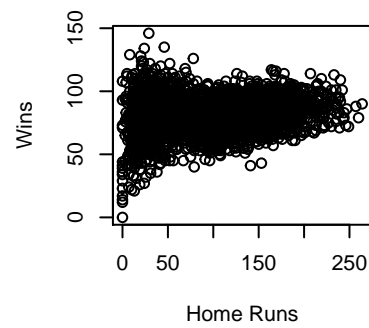
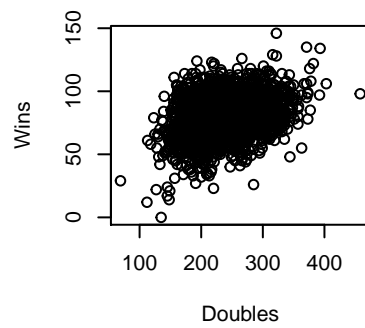
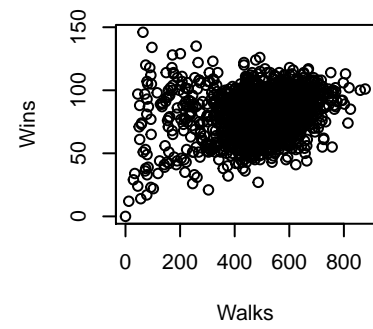
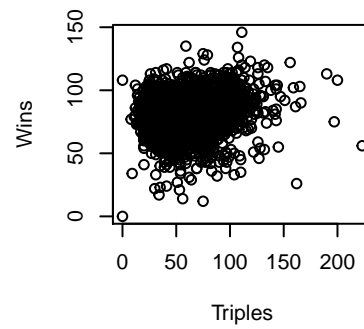
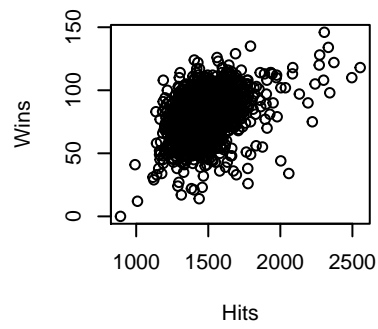
The pitching statistics box plots reveal many outliers in Hits Allowed, Walks Allowed, and Pitching Strike Outs. As was seen in Home Runs, the Home Runs Allowed has a large spread.



The correlation matrix below provides some insight into the data. Wins has the highest positive correlation with Hits and Walks, and negative correlation with Errors. In addition, the batting variables that have corresponding pitching variables are highly correlated, i.e. Walks is highly positively correlated with Walks Allowed, Strike Outs is highly positively correlated with Pitcher Strike Outs, etc. Other interesting correlations found are: Home Runs and Errors are negatively correlated, Triples and Batter Strike Outs are negatively correlated, and Home Runs and Batter Strike Outs are positively correlated.



Scatter plots of Wins versus a few of the independent variables are hard to decipher due to the large amount of records. The relationship between wins and hits, doubles, and walks looks to be positive linear, while the relationship between wins and errors appears to be negative linear.



2. DATA PREPARATION

3. BUILD MODELS

4. SELECT MODELS