Homework 1

Josh Sherback

11/2/2019

##What is the goal of the project? It is a well known conspiracy that the MLB is juicing their baseballs to help players hit more home runs, and in turn drive up ratings for America’s Pastime. My goal is to support the idea that more home runs are being hit in baseball than ever before with statistics, as well as project the home run totals for a few years down the road.

# 

##1. Who will use this and how will they use it? I can use this for a supporting argument to prove that the MLB is using different baseballs.

#2. Obtaining the dataset for analysis. I found a website online that has a database with all the teams and all of their statistics in baseball history. There are a lot of other interesting datasets that I would like to examine in the future, however I think homerun totals will suffice for this specific project. It has league specific totals as well as the cumulative, showing the entire MLB.

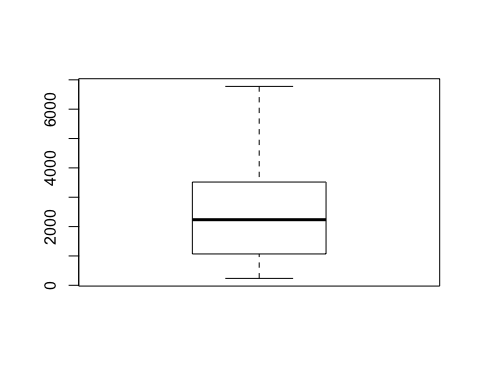
library(readxl)  
HR <- read\_xlsx("/Users/joshsherback/Documents/Data\ Mining\ Applications/Homeruns.xlsx"  
)  
str(HR)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 119 obs. of 4 variables:  
## $ Year: num 1901 1902 1903 1904 1905 ...  
## $ AL : num 228 258 184 156 156 137 104 116 109 147 ...  
## $ NL : num 227 98 151 175 182 126 141 151 150 214 ...  
## $ Both: num 455 356 335 331 338 263 245 267 259 361 ...

head(HR)

## # A tibble: 6 x 4  
## Year AL NL Both  
## <dbl> <dbl> <dbl> <dbl>  
## 1 1901 228 227 455  
## 2 1902 258 98 356  
## 3 1903 184 151 335  
## 4 1904 156 175 331  
## 5 1905 156 182 338  
## 6 1906 137 126 263

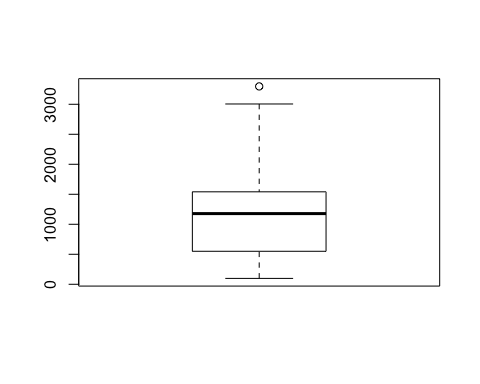
boxplot(HR$Both)



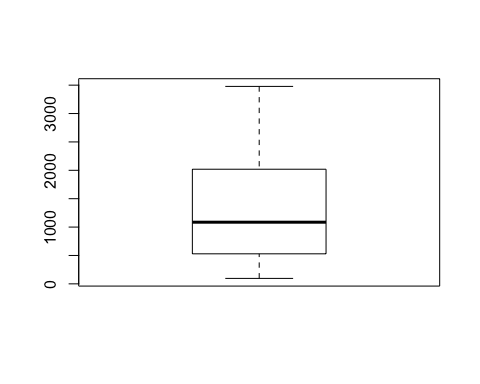
# 

#3. Is there missing data? If so, how should it be handled? Are there outliers? Luckily, this dataset is complete. It dates back to 1901 and is accurate, as I checked it with a few other sites to be sure it met integrity and accuracy bylaws. I created a boxplot to see if there are any outliers, and there do not appear to be any in the Both column or the AL column. However, there is an outlier in the NL. I created a value, OutVals, which will show me where the outlier is. The outlier is in 2019.

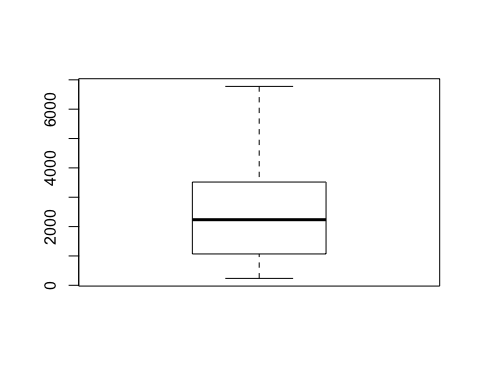
boxplot(HR$NL)



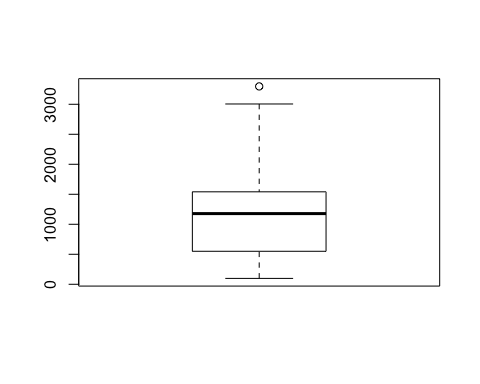
boxplot(HR$AL)



boxplot(HR$Both)



OutVals = boxplot(HR$NL)$out



OutVals

## [1] 3298

# 

#4. Reduce the data dimension, if necessary. Luckily, for this analysis there is no need to delete any unneeded data. This dataset includes only what I will need for conducting my experiment.

# 

#5. Determine the data mining task. I will be using the data for each years home runs for both leagues as well as for both leagues to predict what the future home run totals will be. I plan to run a linear regression and then a prediction on the cumulative total to compare how they size up to one another.

#6.Partitioning the Data

For this I will view a summary of each of the different variables to get some basic information. For all of the columns, the mean is greater than the median. This means that it is positively skewed and will violate the law of normality in the dataset. I will delete all years prior to 1960 to see if I can get the mean and median closer together. The subset seemed to do the trick. I will still analyze the old and the new dataset to see how switching it out will affect it.

summary(HR$AL)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 96 529 1086 1281 2019 3478

summary(HR$NL)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 98 550 1178 1223 1541 3298

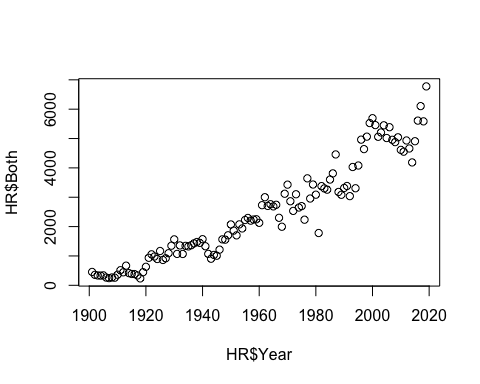
summary(HR$Both)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 235 1068 2235 2506 3518 6776

HRnew <- subset(HR, Year >= 1960)  
summary(HRnew$AL)

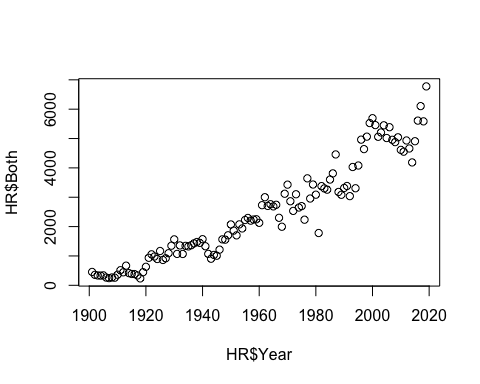
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1062 1552 2019 2030 2497 3478

plot(HR$Year, HR$Both)

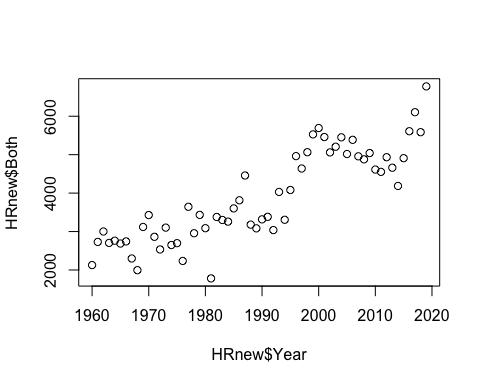


#7. Choosing the algorithm to perform the task. For this, I will use a regression. I will perform it on the old as well as the new to see how it works. First, I wanted to view the dataset in a plot to see if it is trending positive. The cumulative total seems to be showing a rapid increase in the HR dataset, almost to the point of being exponential. The HRnew dataset seems to be a little bit more linear, however you can see that the past ~10 years have shown a gigantic leap after a gradual decline in the previous 10 years. I used the ggplot function to break the data into AL and NL datapoints.

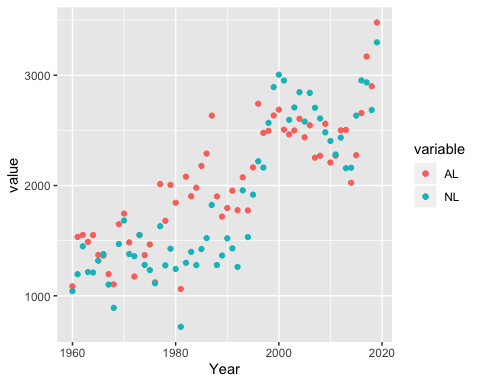
library(ggplot2)  
plot(HR$Year, HR$Both)



plot(HRnew$Year, HRnew$Both)



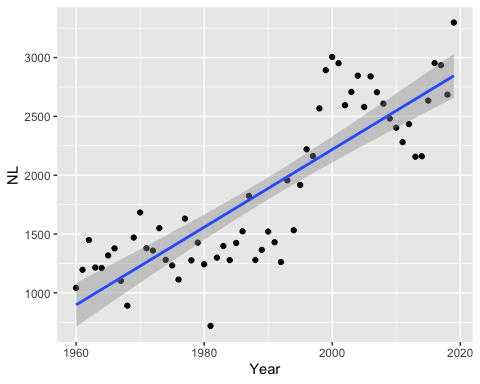
ggplot(HRnew, aes(Year, y = value, color = variable))+  
 geom\_point(aes(y = AL, col = "AL"))+  
 geom\_point(aes(y = NL, col = "NL"))



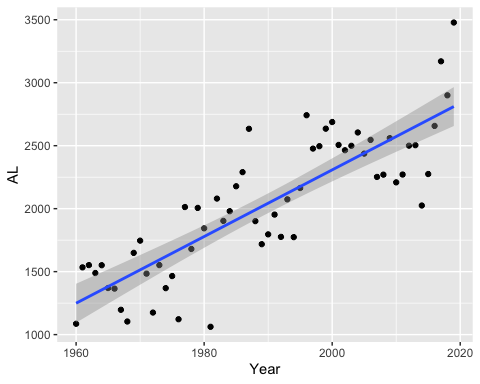
# 

#8.Using the algorithms to perform the task. I wanted to view the HRnew dataset with a slope line as well as the deviation along the x-axis. It is interesting to see how the slopes appear different, however the y-axis scale is different in each of the models so this will not prove it. I also performed the same plotting on the original HR dataset. The slope seemed to be more gradual. I will have to perform a linear regression and use the r squared value to find which regression is the best fit to project future home runs for both leagues.

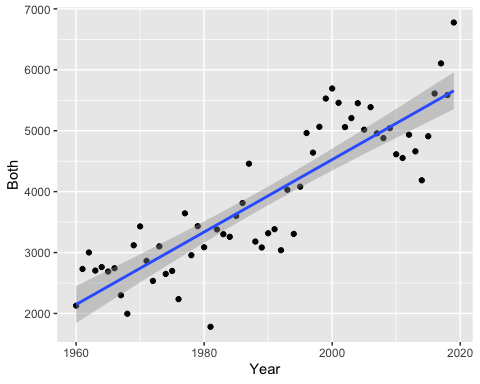
ggplot(HRnew, aes(Year, NL))+  
 geom\_point()+  
 geom\_smooth(method = "lm")



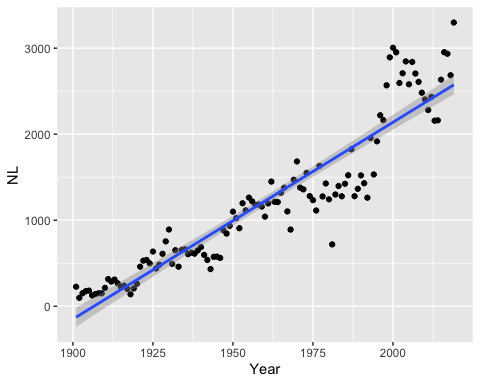
ggplot(HRnew, aes(Year, AL))+  
 geom\_point()+  
 geom\_smooth(method = "lm")



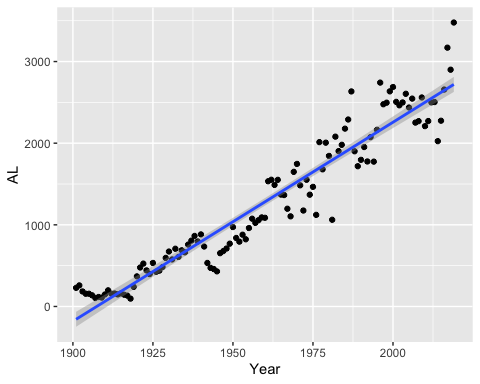
ggplot(HRnew, aes(Year, Both))+  
 geom\_point()+  
 geom\_smooth(method = "lm")



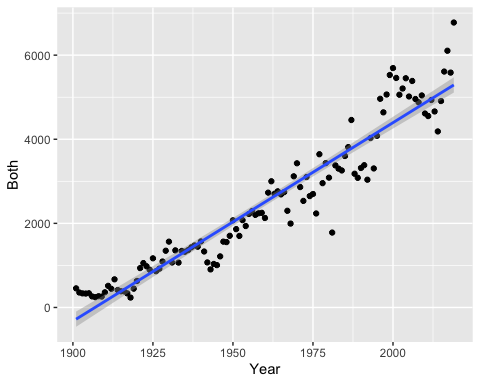
ggplot(HR, aes(Year, NL))+  
 geom\_point()+  
 geom\_smooth(method = "lm")



ggplot(HR, aes(Year, AL))+  
 geom\_point()+  
 geom\_smooth(method = "lm")



ggplot(HR, aes(Year, Both))+  
 geom\_point()+  
 geom\_smooth(method = "lm")



HRnewBoth <- lm(HRnew$Both ~ HRnew$Year)  
summary(HRnewBoth)

##   
## Call:  
## lm(formula = HRnew$Both ~ HRnew$Year)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1614.86 -333.17 -15.13 449.11 1167.18   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.144e+05 8.816e+03 -12.98 <2e-16 \*\*\*  
## HRnew$Year 5.947e+01 4.431e+00 13.42 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 594.4 on 58 degrees of freedom  
## Multiple R-squared: 0.7564, Adjusted R-squared: 0.7522   
## F-statistic: 180.1 on 1 and 58 DF, p-value: < 2.2e-16

HRBoth <- lm(HR$Both ~ HR$Year)  
summary(HRBoth)

##   
## Call:  
## lm(formula = HR$Both ~ HR$Year)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1716.72 -287.65 -9.65 305.58 1484.62   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -90008.458 2629.109 -34.23 <2e-16 \*\*\*  
## HR$Year 47.202 1.341 35.19 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 502.6 on 117 degrees of freedom  
## Multiple R-squared: 0.9137, Adjusted R-squared: 0.913   
## F-statistic: 1239 on 1 and 117 DF, p-value: < 2.2e-16

int <- -90008.458  
slope <- 47.202

# 

#9. Using the linear model for a prediction Though it seemed that partitioning the data to look at more recent years would be more affective, the old dataset actually had a higher R squared as well as t value and f-statistic making it the superior model. This may be because it includes more data, however I feel that partitioning it made it more of an exponential curve rather than a linear model. I will use the old model for the prediction.

lm2020 <- int + 2020\*slope  
lm2025 <- int + 2025\*slope  
lm2030 <- int + 2030\*slope  
  
c(lm2020,lm2025,lm2030)

## [1] 5339.582 5575.592 5811.602

# 

#10. Analysis I found that the model does not project future homerun totals well. I believe this is because in the past few years there has been a monumental uptick in homeruns to the point where it is no longer linear so doing a regression on the model just will not cut it. I think if I were to run an exponential prediction, accuracy would increase however we will save that for a later date. I tried using the predict() function, however the output I got had a prediction for all of the years rather than the data frame I created with the years I manually computed in the above chunk. This model cannot predict values accurately, however this still shows that there is an uptick in home runs. There is so much of an uptick that not even a linear model can quantify the change in slope!