

**BASEBALL RUN SCORE ANALYSIS**

MSBA 324 SF1: Web and Social Analytics

Instructor: Nabanita Talukdar

By

Punita Patel

Anuradha Patel

Veena Chintala

Golden Gate University

536, Mission Street

San Francisco, CA- 94105

|  |  |  |
| --- | --- | --- |
| Sr. No. | Content | Pg. No. |
| 1 | Executive Summary | 3 |
| 2 | Background | 4 |
| 3 | Business Problem | 4 |
| 4 | Data Collection | 5 |
| 5 | Descriptive analysis | 6 |
| 6 | Independent T-test | 7 |
| 7 | Regression analysis | 8 |
| 8 | Regression Diagnostic | 11 |
| 9 | Conclusion | 12 |
| 10 | References | 13 |
| 11 | Appendix | 14 |

**Executive Summary**

Baseball is one of the most popular games in the United States. It is admired and appreciated by both spectators and participants. Major League Baseball is a professional baseball organization. Thirty teams of MLB are divided into the National League and American League, each with three division- west, east and central (Baseball, 2018). The focus of this study is to analyze statistics of MLB’s all thirty team’s score.

According to Baseball (2018), during earlier days, Baseball team’s decisions such as player selection, run prediction, and player’s values were based on committee’s experiences and intuitions. Currently, more statistics are recorded and analyzed to make informed decisions (The Game Theory, 2018). The Oakland team was the first team to use analytics to analyze athletics data in the field of sport. Oakland Athletics evaluated the player’s performance and identified an undervalued performer to make their team competitive using player’s performance statistics with advanced analytics.

In this project, we performed a statistical analysis to identify statistically significant independent variables for predicting the total runs scored.

**Background**

Abner Doubleday founded a baseball in 1839 in New York. The first match was between the New York Nine and Knickerbockers; this was played in 1846 at the Elysian Fields, New Jersey. The following fifty years saw a steady development of baseball and an improvement in playing skills and equipment (Baseball, 2018).

The demand for baseball has grown beyond the live spectators to newspapers, radio, and television. The fanbase for baseball has grown over the years. The passion for baseball has developed many other industries including, dedicated sports television, sports magazines, dedicated radio stations, companies who take on sports trips, and active memorabilia industry (The Game Theory, 2018).

According to Baseball (2018), baseball is played between two opposite teams using ball and bat; these two teams take turns for fielding and batting. The game continues when a bowler from the fielding team called ‘pitcher’ throws the ball towards the batsman of the batting team; the batsman tries to hit the ball coming towards him with the bat. The goal of the batsman is to hit the ball hard and run around the bases counter-clockwise direction to score points which are called runs. The goal of the fielding team is to prevent the batsman from making runs and also preventing him from running around the bases. When a batsman hits the ball, and runs around the bases and touches the home base legally, the run is made. By the end of the game, whichever team scores the maximum runs wins the game.

**Business Problem**

Each team’s ultimate goal is to score the highest run so that they can win the game. The purpose of the paper is to analyze and identify the variables that have a significant impact on scoring runs.

**Data Collections**

The Baseball dataset was collected from the Kaggle website (History of Baseball, n.d.). The data set includes teams’ statistics of all baseball team from 1871 to 2015. There are 1926 records with 38 variables. Some of the variables are explained below:

*Run Scored*: Runs scored are recorded when a player successfully advances around first, second, and third base; and finally reaches home plate safely before the inning is over (Run or runs scored, n.d).

*At Bat*: At-bats are used for calculating batting average and slugging rate. An official at-bat comes when a batter reaches base via a fielder's choice (Glossary, n.d.).

*Base on Ball:* This statistic is the number of times a batter reaches base due to having received four balls during his at-bat, which both the batter and pitcher receive when it occurs*.* Base on balls is commonly known as a "walk"(Glossary, n.d.).

*Hits*: A hit is recorded when a batter strikes the baseball into ‘fair territory’ and reaches base (Glossary, n.d.).

*Doubles:* A batter is credited with a double when a player hits the ball into play and reaches second base. (Glossary, n.d.).

*Triples:* Triples are recorded when a batter hits the ball and reaches third base (Glossary, n.d.).

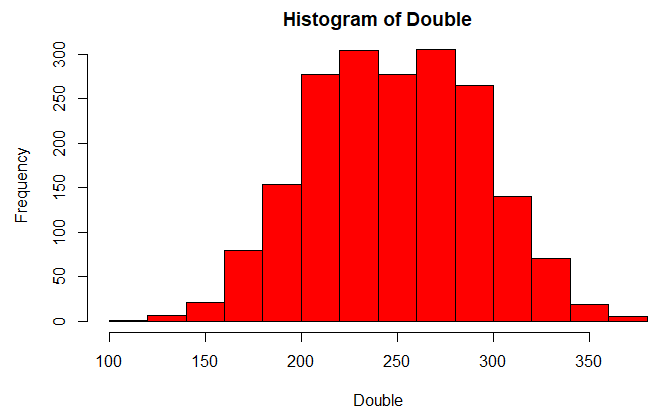
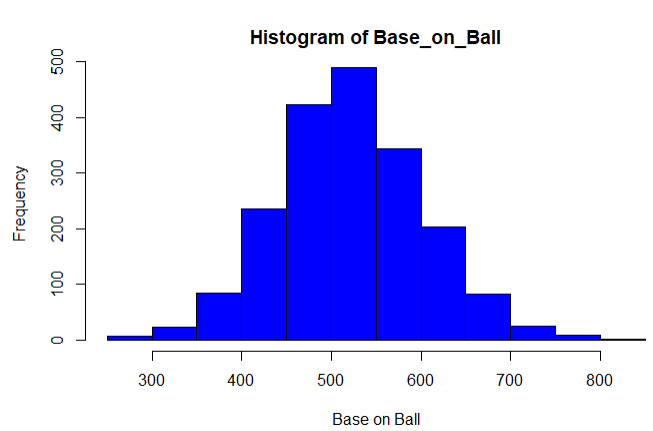
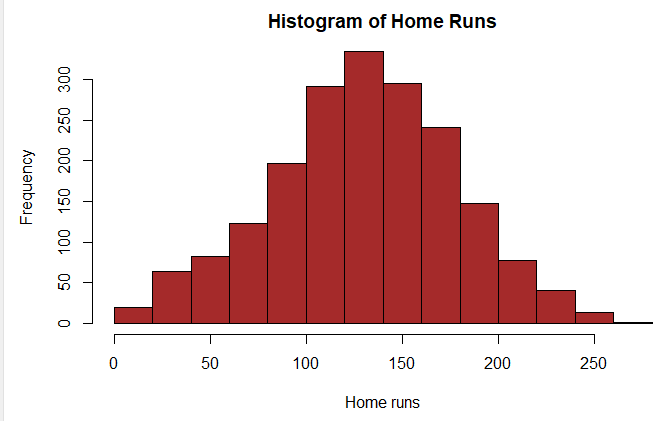
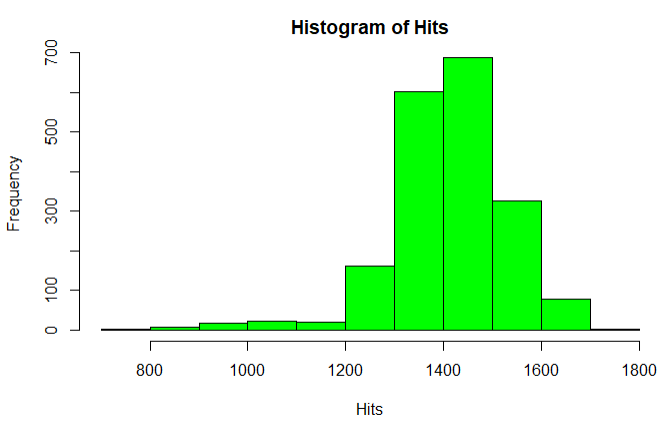
*Home runs*: Batter gets a home run point when he hits the ball to run around all the bases (Home run in Baseball topic, n.d.).

Stolen Bases: A stolen base occurs when a baserunner advances by taking a base to which he is not entitled (Glossary, n.d.).

**Descriptive analysis**

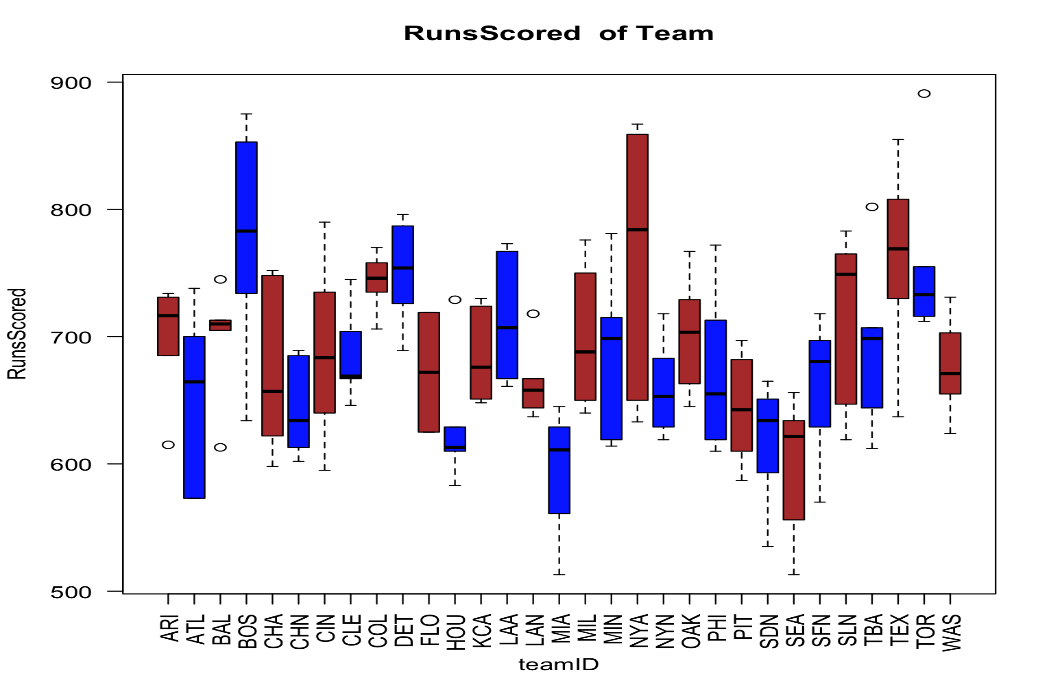
Descriptive analysis provides summaries and describes the sample and measures of the data. It helps to know the possible relationship between the variables. To carry out the descriptive statistics summary, histogram, box plot and were used to study further on data.

**Histograms:** A histogram gives the frequency distribution of the continuous data set. It provides a quick visualization by organizing a large amount of data.

Histogram of Home runs and base-on-ball has bell shape curve, whereas Double data is bi-modal as two data ranges occur most often. Hits data is left-skewed.

**Boxplot:** The box plot graph of Runs scored by each team gives a summary of their Runs scored over a span of 6 years by each team. From these box plots, we can easily compare median, max and min Run scored by each team and evaluate their performance. Higher the value of median (2nd quartile) better is the chances of team scoring run. Consistency in runs scored by the team can be related to the size of the interquartile range of box plot. Smaller the width of interquartile more consistent is the team with their performance in runs scored.

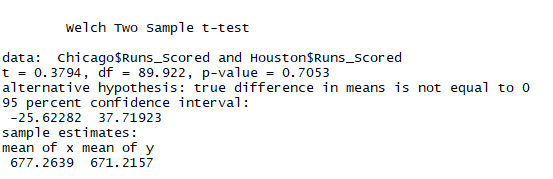
******

**Independent t-test**

An Independent t-test was calculated between two teams, Chicago Cubs and Huston Astros team for a run scored. The assumption for the t-test is that both groups are sampled from normal distributions with equal variances. MASS library used for running an independent test.

H0:There is no difference in the mean of run scored between the Chicago Cubs and Huston Astros.

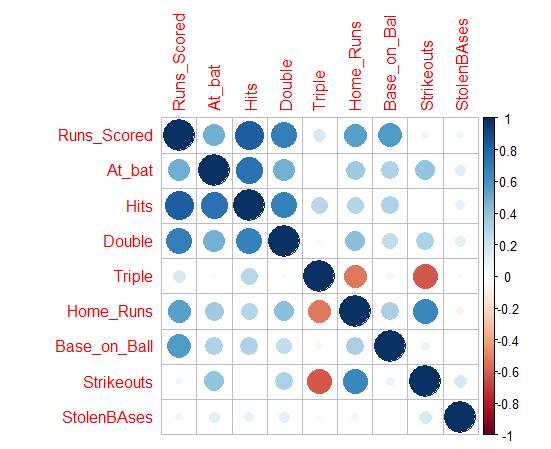
H1: There is a difference in the mean of run scored between the Chicago Cubs and Huston Astros.



P-value is higher than 0.05. Hence we failed to reject the null hypothesis. There is no significant difference in mean of runs scored between two teams

**Regression Analysis**

There is more than one predictor variable. Hence we performed multiple regression analyses. Before running the multiple regression, we used correlation analysis to understand the relationship between two variables.



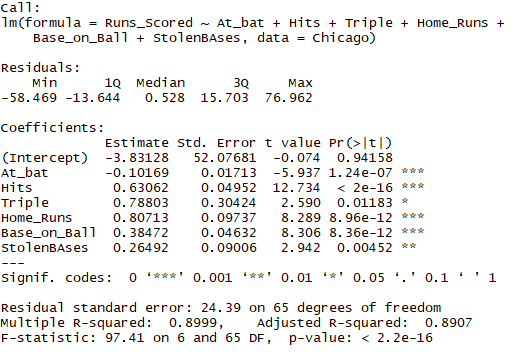
In the above correlation matrix, blue color shows the positive correlation whereas red shows the negative correlation. Runs scored has highest positive correlation hits, doubles, home runs and base on ball variables.

We built two regression models with two different teams; Chicago Cubs and Huston Astros. For each team, we used respective team data from 1871 to 2015. The Akaike Information Criterion (AIC) method has been used to compare two models. For both model we started building regression models using below mentioned independent and dependent variables:

|  |  |
| --- | --- |
| Dependent Variable | Runs scored |
| Independent Variables | At bat, Hits, Triple, Double, Home Runs, Base\_on\_Ball, stolen bases, Strikeouts |

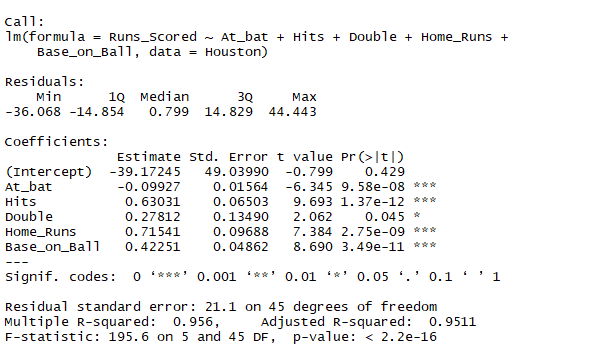
**A regression model with Chicago Cubs:**

Firstly, we ran the regression with all variables and examined the p-value of each variable. (Appendix-I). For the given predictor variables, p-values were used to evaluate whether or not there is a significant association between runs scored and the independent variables. It was observed that strikeouts have a p-value higher than 0.05, and we failed to reject the null hypothesis: coefficient associated with variables is 0. We removed strikeouts and reran the model; the second regression result got a p-value of ‘double’ variable less than 0.05. For the Chicago Cubs, we removed strikeouts and doubles from our final model. Final regression is shown below. The model has all the variables with a p-value of less than .0.05 and model Adjusted R-square is 0.89.



**A regression model with Huston Astros team:**

For Huston Actros team regression model, we followed similar steps-like Chicago clubs (Appendix-II). The initial result showed the triple, strikeouts and stolen bases have a p-value greater than 0.05. We removed all three variables from the analysis, and the final regression model is shown below. The model has all the variables with a p-value of less than .0.05 and the model Adjusted R-square is 0.95:



**Best model selection:**

AIC methodology was chosen to select the best model. The index considers a model’s statistical fit and the number of parameters needed to achieve this fit. Models with smaller AIC values—indicating adequate fit with fewer parameters—are preferred (Kabacoff, 2015, p.202).

Df AIC



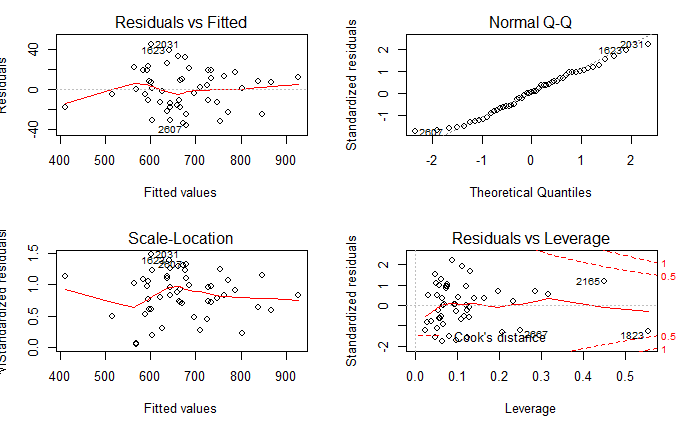
The Above table shows Huston Astros’s AIC is lower than the Chicago Cubs. Hence, Huston model was selected as the best fit.

Regression equation:

Run scored = -39.17 – (0.09) \* At bats + (0.63) \* hits + (0.27) \* Double + (0.715) \* Homeruns + (0.422) \* base on ball.

**Regression Diagnostic**

Even though we selected the best model through different analysis: p-value AIC and model summary, we still have to analyze whether this model has satisfied statistical assumptions underlying the model (Kabacoff, 2015, p.182). We can use the *plot (model)* function to verify these assumptions.



OLS regression assumptions are:

Normality: The Normal Q-Q plot shows that most of the points are on the straight 45- degree line.

Independence: Here we are assuming that data collection of dependent variables- Runs scored are independent.

Linearity: Residual vs. fitted graphs do not show any pattern. The is no symmetric relationship between residual and the predicted value.

Homoscedasticity: The assumption of constant variance is met as residual as numbers are randomly plotted around a horizontal line.

**Conclusion**

The analysis shows that each team should focus on variables that have a positive coefficient in the regression model to increase their total score. If a team selects a player with a higher base on ball, hits, double and home runs, then teams will be able to score higher runs. There are several statistics which are widely used in baseball analytics in determining the team winning- such as On-base percentage (OBP) and slugging rate. However, due to data inadequacy analysis of those variables were not part of overall team performance analysis. The variables analyzed in this paper can further be used to calculating the Batting average and Home runs per nine Innings.

**References**

Baseball Summary. (n.d). Retrieved from

<https://www.infoplease.com/people/baseball-summary>

Baseball. (2018, August 1). Retrieved from <https://en.wikipedia.org/wiki/Baseball>

Glossary. (n.d.). Retrieved from <http://m.mlb.com/glossary>

History of baseball (n.d.). Retrieved from <https://www.kaggle.com/seanlahman/the-history-of-baseball>

Haupert, M. J. (n.d). The Economic History of Major League Baseball. Retrieved from

<https://eh.net/encyclopedia/the-economic-history-of-major-league-baseball/>

Home run in Baseball topic. (n.d.). Retrieved from <https://www.ldoceonline.com/Baseball-topic/home-run>

Kabacoff, R. (2015). R in action: Data analysis and graphics with r. New York, NY: Manning Publications Co.

Run or runs scored- r definition-Sporting charts. (n.d.). Retrieved from <https://www.sportingcharts.com/dictionary/mlb/run-or-runs-scored-r.aspx>

The Game Theory of Baseball. (2018, July 19). Retrieved from

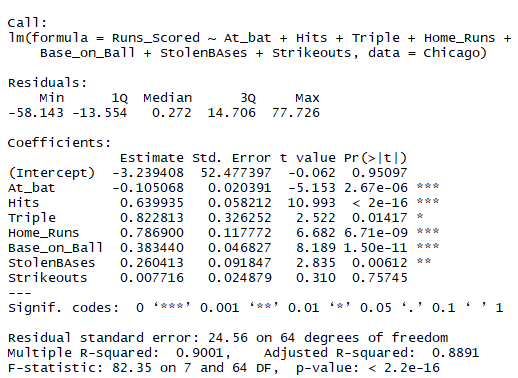
<https://www.bartleby.com/essay/The-Game-Theory-of-Baseball-F39JH6AWUDAX>

**Appendix**

**Appendix-1:**

**#Chicago Regression model 1st run:**

runs\_scored<-lm(Runs\_Scored~At\_bat+Hits+Triple+Double+Home\_Runs+Base\_on\_Ball+StolenBAses+Strikeouts,data=Chicago)

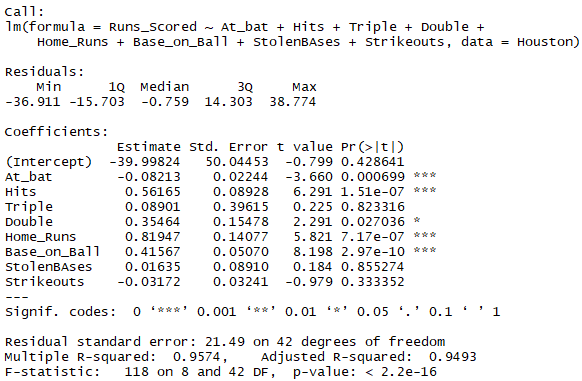
summary(runs\_scored) 

**Appendix-II**:

# **Huston** **Regression model 1st run**

runs\_scoredH<-lm(Runs\_Scored~At\_bat+Hits+Triple+Double+Home\_Runs+Base\_on\_Ball+StolenBAses+Strikeouts,data=Houston)

summary(runs\_scoredH)



**R Codes**

**#Installing libraries**

install.packages("readxl")

library("readxl")

install.packages("car")

library(car)

install.packages("corrplot")

library(corrplot)

library("Hmisc")

library("MASS")

**#Importing Dataset**

Team<- read.csv("Teams.csv")

**#Removing missing values**

Team<-na.omit(Team)

**#Subset of Chicago, Houston**

Chicago<-subset(Team, name=="Chicago Cubs")

Houston<-subset(Team, name=="Houston Astros")

**# Describe the data**

describe(Team[,9:17])

Team[, 9:17]

9 Variables 1926 Observations

-------------------------------------------------------------------------------------------

Runs\_Scored

n missing distinct Info Mean Gmd .05 .10 .25 .50

1926 0 468 1 704.5 114.3 540.0 578.0 640.0 703.0

.75 .90 .95

768.8 839.5 876.8

lowest : 329 348 370 378 382, highest: 1009 1027 1062 1065 1067

-------------------------------------------------------------------------------------------

At\_bat

n missing distinct Info Mean Gmd .05 .10 .25 .50

1926 0 630 1 5402 250.1 5082 5204 5346 5477

.75 .90 .95

5552 5616 5649

lowest : 3493 3507 3516 3521 3529, highest: 5756 5757 5767 5769 5781

-------------------------------------------------------------------------------------------

Hits

n missing distinct Info Mean Gmd .05 .10 .25 .50

1926 0 480 1 1413 127.4 1229 1286 1353 1415

.75 .90 .95

1490 1552 1588

lowest : 797 838 868 883 884, highest: 1684 1693 1698 1715 1724

-------------------------------------------------------------------------------------------

Double

n missing distinct Info Mean Gmd .05 .10 .25 .50

1926 0 215 1 249.5 49.44 179.0 195.0 216.0 250.0

.75 .90 .95

281.0 304.5 320.0

lowest : 119 133 134 135 136, highest: 358 363 371 373 376

-------------------------------------------------------------------------------------------

Triple

n missing distinct Info Mean Gmd .05 .10 .25 .50

1926 0 98 0.999 40.03 18.23 20.0 22.0 28.0 36.0

.75 .90 .95

47.0 62.5 76.0

lowest : 11 12 13 14 15, highest: 105 106 110 111 122

-------------------------------------------------------------------------------------------

Home\_Runs

n missing distinct Info Mean Gmd .05 .10 .25 .50

1926 0 233 1 130.5 53.91 43.25 64.00 100.00 132.00

.75 .90 .95

162.00 191.00 209.00

lowest : 10 12 14 15 16, highest: 246 249 257 260 264

-------------------------------------------------------------------------------------------

Base\_on\_Ball

n missing distinct Info Mean Gmd .05 .10 .25 .50

1926 0 380 1 522.8 90.78 392.2 425.0 470.0 520.0

.75 .90 .95

575.0 627.0 659.0

lowest : 275 278 283 284 294, highest: 770 775 783 823 835

-------------------------------------------------------------------------------------------

Strikeouts

n missing distinct Info Mean Gmd .05 .10 .25 .50

1926 0 821 1 864.9 268.3 445.2 518.0 709.2 890.0

.75 .90 .95

1032.0 1155.0 1232.8

lowest : 308 318 326 327 331, highest: 1442 1477 1518 1529 1535

-------------------------------------------------------------------------------------------

stolen bases

n missing distinct Info Mean Gmd .05 .10 .25 .50

1926 0 210 1 93.12 46.9 35.0 43.0 63.0 88.0

.75 .90 .95

119.0 148.5 167.8

lowest : 13 16 17 18 19, highest: 256 260 262 314 341

-------------------------------------------------------------------------------------------

**# Subset of Team data for last 6 years*:***

Team10\_15<-subset(Team, yearID < 2016 & yearID > 2009)

boxplot(Team10\_15$Home\_Runs~Team10\_15$teamID,las=2,xlab="teamID",ylab="Home\_Runs",col=(c("brown","blue")),main="Home\_Runs of Team")

**# Box Plot on Runs scored by team for last 6 years*:***

>boxplot(Team$Runs\_Scored~mydata$teamID,las = 2,xlab="teamID",ylab="RunsScored",col=(c("brown","blue")),main="RunsScored  of Team")

**# Histogram on Doubles**

hist(Team$Double,col='Red',main="Histogram of Double",xlab='Double')

**# Histogram on Home Runs**

hist(Team$Home\_Runs,col='brown',main="Histogram of Home Runs",xlab='Home runs')

**# Histogram on Hits**

hist(Team$Hits,col='green',main="Histogram of Hits", xlab='Hits')

**# Histogram on Base on Ball**

hist(Team$Base\_on\_Ball,col='blue',main="Histogram of Base\_on\_Ball",xlab='Base on Ball')

# **T-test between H0: There is no significant difference in mean run scored by team Chicago and Huston Astros**

t.test(Chicago$Runs\_Scored,Houston$Runs\_Scored)

**#Overall Team Correlation Analysis**

Teamcorr<-cor(Team[,9:17])

corrplot(Teamcorr)

**# Chicago Cubs Regression final fit**

Chicagofit<-lm(Runs\_Scored~At\_bat+Hits+Triple+Home\_Runs+Base\_on\_Ball+StolenBAses,data=Chicago)

**# Huston Regression final fit**

Hustonfit<-lm(Runs\_Scored~At\_bat+Hits+Double+Home\_Runs+Base\_on\_Ball,data=Houston)

**#AIC of two model**

AIC(Chicagofit,Hustonfit)

Df AIC



**##Huston Regression Diagnostic**

par(mfrow=c(2,2))

plot(Hustonfit)

