biological\_data\_set; Nick Kendrick

## Discovering Biological Datasets

The first dataset of interest involves events that have been recently brought up in the news about the Houston Astros cheating scandal. The Astros won the World Series in 2017 and have been caught cheating by the MLB. The cheating scandal involved stealing signs from the opposing team at home.

I would like to look at the statistical differences between home vs away for the Hoston Astros to see if the aligations are true or is it a witch hunt.

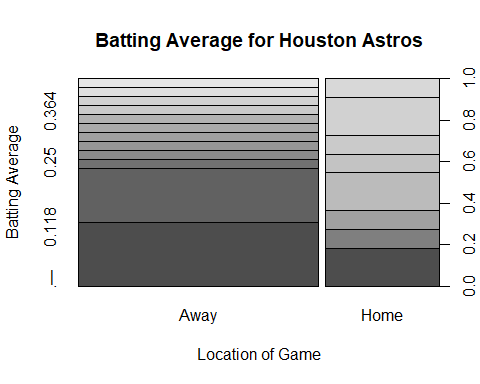
The issues that I expect to see are the even tho the players “had the signs” it still is very diffiult to hit a pitch. Will I be able to find a correaltion between home vs away and the increase in batting average.

setwd("C:/Users/kendrick.nicholas/Desktop/Bioinformatics books/BIFX551")  
cheating\_Astros <- read.csv("Astros\_world\_series\_2017.csv")  
summary(cheating\_Astros)

## RK Home.Field Player Team Pos   
## Min. : 1.000 Away:23  Altuve, J : 2 HOU:34 P :11   
## 1st Qu.: 5.000 Home:11  Bregman, A : 2 LF : 6   
## Median : 9.000  Correa, C : 2 C : 4   
## Mean : 8.912  Fisher, D : 2 1B : 2   
## 3rd Qu.:11.000  Gattis, E : 2 2B : 2   
## Max. :17.000  Gonzalez, M: 2 3B : 2   
## (Other) :22 (Other): 7   
## G AB R H   
## Min. :1.000 Min. : 0.000 Min. :0.00 Min. :0.000   
## 1st Qu.:1.250 1st Qu.: 0.250 1st Qu.:0.00 1st Qu.:0.000   
## Median :3.000 Median : 5.500 Median :1.00 Median :1.000   
## Mean :2.559 Mean : 7.176 Mean :1.00 Mean :1.647   
## 3rd Qu.:3.000 3rd Qu.:12.750 3rd Qu.:1.75 3rd Qu.:3.750   
## Max. :4.000 Max. :18.000 Max. :4.00 Max. :7.000   
##   
## X2B X3B HR RBI   
## Min. :0.0000 Min. :0 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.0000 Median :0 Median :0.0000 Median :0.0000   
## Mean :0.3824 Mean :0 Mean :0.4412 Mean :0.9412   
## 3rd Qu.:0.7500 3rd Qu.:0 3rd Qu.:1.0000 3rd Qu.:2.0000   
## Max. :2.0000 Max. :0 Max. :3.0000 Max. :5.0000   
##   
## BB SO SB CS AVG.   
## Min. :0.0000 Min. :0.000 Min. :0.00000 Min. :0 .--- : 9   
## 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:0.00000 1st Qu.:0 0 : 6   
## Median :0.0000 Median :1.000 Median :0.00000 Median :0 0.333 : 3   
## Mean :0.6471 Mean :1.588 Mean :0.08824 Mean :0 0.167 : 2   
## 3rd Qu.:1.0000 3rd Qu.:3.000 3rd Qu.:0.00000 3rd Qu.:0 0.25 : 2   
## Max. :4.0000 Max. :6.000 Max. :1.00000 Max. :0 0.308 : 2   
## (Other):10   
## OBP SLG OPS   
## .--- : 9 .--- : 9 Min. :0.0000   
## 0 : 6 0 : 6 1st Qu.:0.0000   
## 0.167 : 3 0.25 : 2 Median :0.3235   
## 0.188 : 2 0.5 : 2 Mean :0.4493   
## 0.286 : 2 0.071 : 1 3rd Qu.:0.8123   
## 0.333 : 2 0.091 : 1 Max. :1.5330   
## (Other):10 (Other):13

## Exploritory Plots

Box Plot for Astros Cheating Scandal for Home vs away for Jose Altuve:



The second dataset involves the number of concussions that occur in the NFL during the kick return. It has multiple interesting data points that will focus on type of hit, velocity of impact, and number of G’s received during the impact.

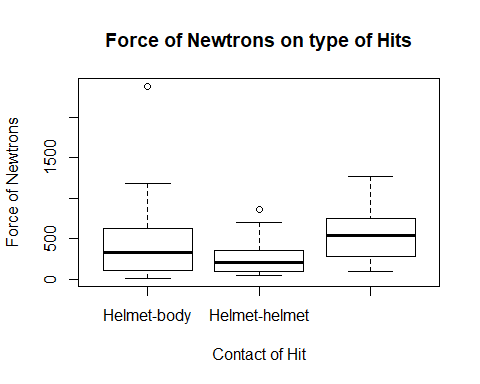
I am interested in this topic because of work and practical experience. My main projects at work target concussion biomarkers and how these impacts triggers the upregulation or supression of concussion biomarkers.

setwd("C:/Users/kendrick.nicholas/Desktop/Bioinformatics books/BIFX551")  
concussion <- read.csv("Combined\_Punt\_Plays\_Excel.csv")  
summary(concussion)

## Year Player GSISID Play Role   
## Min. :2016 Partner:33 Min. :23259 Min. : 183 PR :12   
## 1st Qu.:2016 Primary:33 1st Qu.:29526 1st Qu.:1407 PRG : 6   
## Median :2016 Median :31887 Median :2342 GR : 5   
## Mean :2016 Mean :30813 Mean :2225 GL : 4   
## 3rd Qu.:2017 3rd Qu.:32713 3rd Qu.:2918 PDR1 : 4   
## Max. :2017 Max. :33841 Max. :3746 PLG : 4   
## (Other):31   
## Punting.Team Number Position Weight Name   
## No :27 42 : 3 OLB :13 Min. : 78.90 Antonio Andrews: 1   
## Yes:39 48 : 3 WR :13 1st Qu.: 91.25 Arrelious Benn : 1   
## 49 : 3 CB : 8 Median :101.50 BJ Goodson : 1   
## 51 : 3 ILB : 7 Mean : 99.85 Blair Brown : 1   
## 81 : 3 RB : 6 3rd Qu.:108.75 Bobby Rainey : 1   
## 14 : 2 TE : 6 Max. :121.00 Bralon Addison : 1   
## (Other):49 (Other):13 (Other) :60   
## Team Type.of.Hit Velocity..m.s. G.s   
## TEN : 7 Helmet-body :34 Min. :0.270 Min. : 0.640   
## WAS : 5 Helmet-helmet :16 1st Qu.:1.597 1st Qu.: 8.002   
## CAR : 4 Helmet-helmet :16 Median :2.925 Median : 24.790   
## KC : 4 Mean :3.505 Mean : 33.844   
## NYJ : 4 3rd Qu.:5.053 3rd Qu.: 48.735   
## DEN : 3 Max. :8.140 Max. :186.960   
## (Other):39   
## Force..Newtons. Momentum   
## Min. : 8.1 Min. : 25.11   
## 1st Qu.: 102.2 1st Qu.:160.16   
## Median : 315.9 Median :282.31   
## Mean : 431.7 Mean :345.62   
## 3rd Qu.: 621.6 3rd Qu.:503.40   
## Max. :2384.3 Max. :862.84   
##

## Exploritory Plots

Box Plot for Astros Cheating Scandal for Home vs away for Jose Altuve:



The third dataset involves multiple locations of Ebola outbreaks. The statistics are up to data as to the point I collected the .csv file.

EBOLA is one if not the most dangerous viruses. It will liquify your organs, eliminate all connective tissue and you will bleed out from every openning. This virus is something that you do not want to play with. Which is something that I have.

The dataset provides total number of cases, treated vs cured vs deaths and several other columns.

download.file("https://raw.githubusercontent.com/calthaus/Ebola/master/West%20Africa%20(PLOS%20Curr%202014)/Ebola\_outbreak\_West\_Africa\_data.csv", "EBOLA\_outbreak")  
ebola.df <-read.csv("EBOLA\_outbreak")  
summary(ebola.df)

## Date Total\_Cases Total\_Death Guinea\_Cases   
## 01 Apr 2014: 1 Min. : 49.0 Min. : 29.0 Min. : 49.0   
## 01 Aug 2014: 1 1st Qu.: 163.5 1st Qu.: 107.0 1st Qu.:197.0   
## 01 Jun 2014: 1 Median : 253.0 Median : 174.0 Median :291.0   
## 01 May 2014: 1 Mean : 700.9 Mean : 405.7 Mean :310.5   
## 02 Jul 2014: 1 3rd Qu.:1071.0 3rd Qu.: 646.0 3rd Qu.:413.0   
## 03 Jun 2014: 1 Max. :2615.0 Max. :1427.0 Max. :607.0   
## (Other) :52 NA's :11 NA's :11 NA's :5   
## Guinea\_Death SierraLeone\_Cases SierraLeone\_Death Liberia\_Cases   
## Min. : 29.0 Min. : 16.0 Min. : 5.00 Min. : 33.0   
## 1st Qu.:122.0 1st Qu.:142.8 1st Qu.: 48.25 1st Qu.: 149.5   
## Median :193.0 Median :419.5 Median :201.50 Median : 289.0   
## Mean :217.6 Mean :434.0 Mean :183.50 Mean : 399.8   
## 3rd Qu.:310.0 3rd Qu.:697.5 3rd Qu.:289.00 3rd Qu.: 587.8   
## Max. :406.0 Max. :910.0 Max. :392.00 Max. :1082.0   
## NA's :5 NA's :30 NA's :30 NA's :36   
## Liberia\_Death Nigeria\_Cases Nigeria\_Death   
## Min. : 24.00 Min. : 1.00 Min. :0.000   
## 1st Qu.: 92.25 1st Qu.: 3.75 1st Qu.:1.000   
## Median :142.50 Median :12.00 Median :2.000   
## Mean :223.82 Mean : 9.50 Mean :2.333   
## 3rd Qu.:315.75 3rd Qu.:13.50 3rd Qu.:4.000   
## Max. :624.00 Max. :16.00 Max. :5.000   
## NA's :36 NA's :46 NA's :46   
## Comment   
## :55   
## Arrival of index patient in Nigeria : 1   
## First confirmed cases in Sierra Leone : 1   
## First new cases in Liberia since 6 April: 1   
##   
##   
##

## Exploritory Plots

Box Plot for Astros Cheating Scandal for Home vs away for Jose Altuve:

