Project 3

Rohit Kamat rgk359

Introduction: For this project I will use the “Baseball” data set which is built into R under the vcd package. The Baseball data set gives baseball players baseball statistics during the 1986 season and during their career.

```
#Import Baseball Data
Baseball <- read.csv("/Volumes/USB20FD/Computational Biology/Baseball.csv")
```

The column content are as follows:

- **name1**: First Name of the baseball Player
- **name2**: Last Name of the baseball Player.
- **atbat86**: Number of times at bat a baseball player faced during the 1986 baseball season.
- **hits86**: Number of hits a baseball player had during the 1986 baseball season.
- **homer86**: Number of homeruns a baseball player had during the 1986 baseball season.
- **runs86**: Number of runs a baseball player had during the 1986 baseball season.
- **rbi86**: Number of rbi’s(runs batted in) a baseball player had during the 1986 baseball season.
- **walks86**: Number of walks a baseball player had during the 1986 baseball season.
- **years**: Number of years the baseball player has played in the major leagues.
- **atbat**: Baseball player’s career times at bat.
- **hits**: Baseball player’s career hits.
- **homeruns**: Baseball player’s career home runs.
- **runs**: Baseball player’s career runs.
- **rbi**: Baseball player’s career rbi’s.
- **walks**: Baseball player’s career walks.
- **league86**: Baseball player’s league during the 1986 baseball season.
- **div86**: Baseball player’s division during the 1986 baseball season.
- **team86**: Baseball player’s team during the 1986 baseball season.
- **posit86**: Baseball player’s position during the 1986 baseball season.
- **outs86**: Baseball player’s number of putouts during the 1986 baseball season.
- **assist86**: Baseball player’s number of assist during the 1986 baseball season.
- **error86**: Baseball player’s number of errors during the 1986 baseball season.
- **sal87**: Baseball player’s annual salary on opening day (in USD 1000).
- **league87**: Baseball player’s league in 1987.

- **team87**: Baseball player's team in 1987.

Problems

Question 1 During the 1986 baseball season, was there a correlation for times at bat and homeruns? Was the correlation between times at bat and homeruns scored similiar between the two divisions in the NL League?

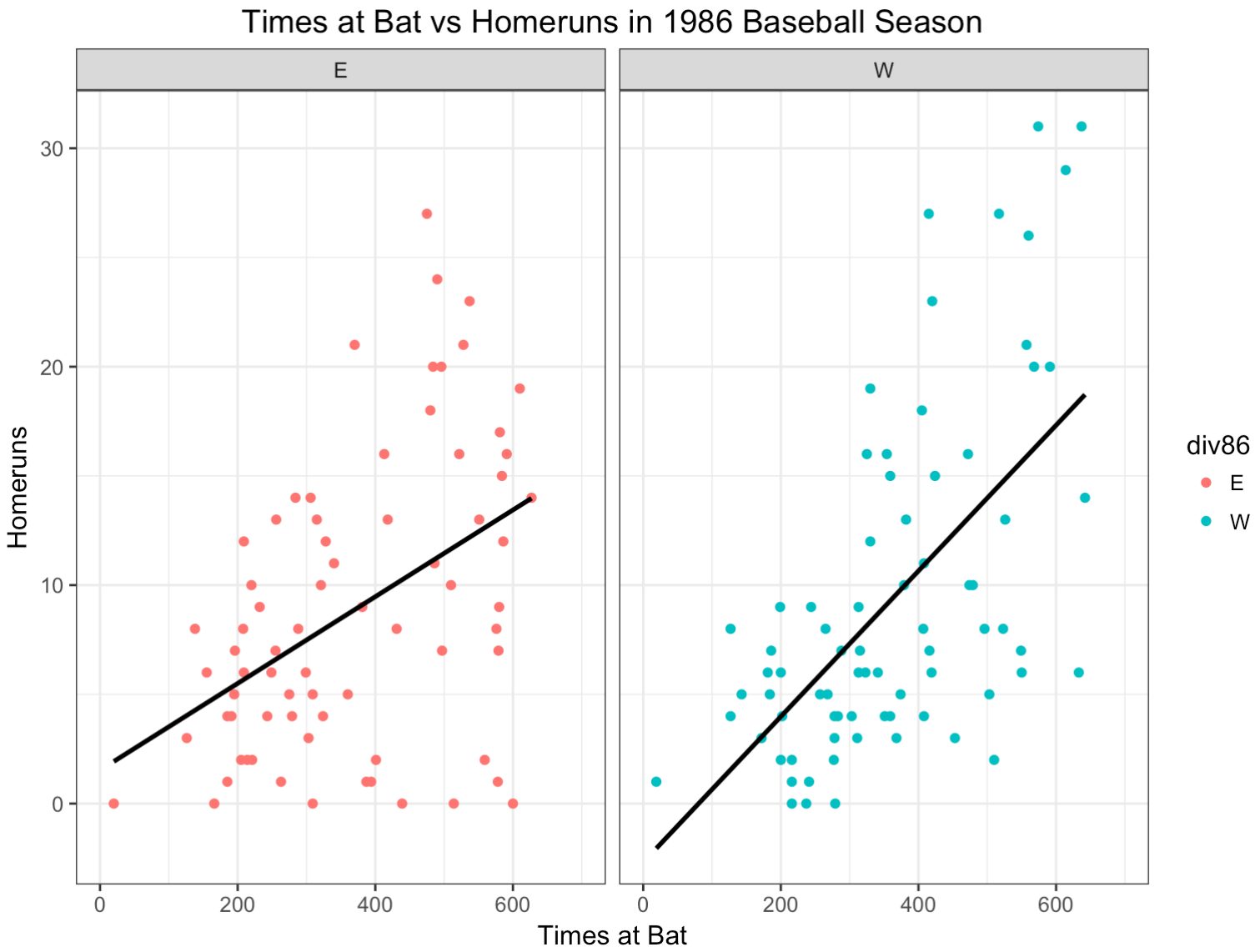
```
#Filter the Baseball data among players that played in the NL in 1986
Baseball %>% filter(league86=="N") %>%
#Arrange the players based on what division in 1986
arrange(div86) %>%
#My data will consist at looking at at bats, homeruns, and the divisions in 1986
select(atbat86,homer86,div86)->Baseball2
head(Baseball2)
```

```
##   atbat86 homer86 div86
## 1     185       1     E
## 2     496      20     E
## 3     321      10     E
## 4     418      13     E
## 5     413      16     E
## 6     196       7     E
```

```
#Perform the Linear Regression with number at bats interacting with each division
old_base<- lm(homer86 ~ atbat86 + div86 + div86*atbat86, Baseball2)
summary(old_base)
```

```
##
## Call:
## lm(formula = homer86 ~ atbat86 + div86 + div86 * atbat86, data = Baseball2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.4410  -4.1947  -0.4739   3.6154  16.0402
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.531496   1.921135   0.797   0.4267
## atbat86         0.019849   0.004839   4.102 6.86e-05 ***
## div86W         -4.216408   2.747557  -1.535   0.1271
## atbat86:div86W  0.013494   0.007008   1.925   0.0562 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.198 on 143 degrees of freedom
## Multiple R-squared:  0.2968, Adjusted R-squared:  0.282
## F-statistic: 20.12 on 3 and 143 DF,  p-value: 6.194e-11
```

```
#Perform a ggplot of homeruns based on number of bats of both division
ggplot(Baseball2, aes(x=atbat86, y=homer86, color=div86)) + geom_point()  + xlim(0,700) + ggtitle("Times at Bat v
s Homeruns in 1986 Baseball Season") + theme(plot.title = element_text(hjust = 0.5)) + xlab("Times at Bat") + yla
b("Homeruns") + facet_wrap(~div86) + geom_smooth(aes(group=div86), method=lm, color='black', se=F)
```



So I performed a linear regression to compare the relationship between at bats and homeruns between the East and West division. I first filtered the Baseball data so that I could get the baseball players within the NL League, seperating the players based on the two division, and then focusing on the variables of interest: at bats, homeruns, and division in the 1986 baseball season. Then I performed a linear regression, with divisions, categorical variable, interacting with number at bats, quantitative variable, to determine the dependent variable homeruns.

Since we are comparing the relationship of at bats and homeruns based on the two divisions, I need to perform a linear regression with each division as the categorical variable interacting with the quantitative variable of number at bats to determine the amount of homeruns scored.

Based on the linear regression, the number at bats did have a moderate positive linear relationship as increasing the number of times at bat by 100 would increase the amount of homeruns by 1.985, t(143)=4.102, p<.05. However there was not a statistical difference in the slope of homeruns over number of bats in the West division compared to East division with t(143)=1.925 and p=.0562>.05. Based on the ggplot both divisions had a similiar positive linear relationship between times at bat and homeruns, though the West division seemed to have players that batted more and scored more homeruns that players in the East division.

I concluded that based on the 1986 baseball season, there was a statistically similiar positive, moderate, linear relationship on times at bat predicting homeruns between the two division in the NL League.

Question 2 Between the divisions of the AL League, which characteristics in baseball: hits, runs, rbi (runs batted in), and walks were the most distinguished during the 1986 baseball season?

```
#Filter data among players in the AL league
Baseball%>% filter(league86=="A") %>%
#Group Al League Players based on divisions
arrange(div86) %>%
#Focus on hits, runs, rbi, and walks during 1986 season among the divisions
select(div86, hits86,runs86,rbi86,walks86) -> Baseball3

#Perform a PCA Analysis
Baseball3 %>% select(-div86) %>% # remove the division column
scale() %>%  #scale to 0 the mean variance
prcomp() -> #do pca
pca

pca # display the pca
```

```
## Standard deviations (1, .., p=4):
## [1] 1.8108691 0.6712288 0.4671202 0.2280437
##
## Rotation (n x k) = (4 x 4):
##                PC1         PC2         PC3         PC4
## hits86  -0.5159228 -0.37841238  0.43818120 -0.63136750
## runs86  -0.5342607 -0.07940797  0.38250660  0.74963229
## rbi86   -0.4984106 -0.30428004 -0.81135413  0.02655285
## walks86 -0.4471868  0.87058148 -0.05822729 -0.19677768
```

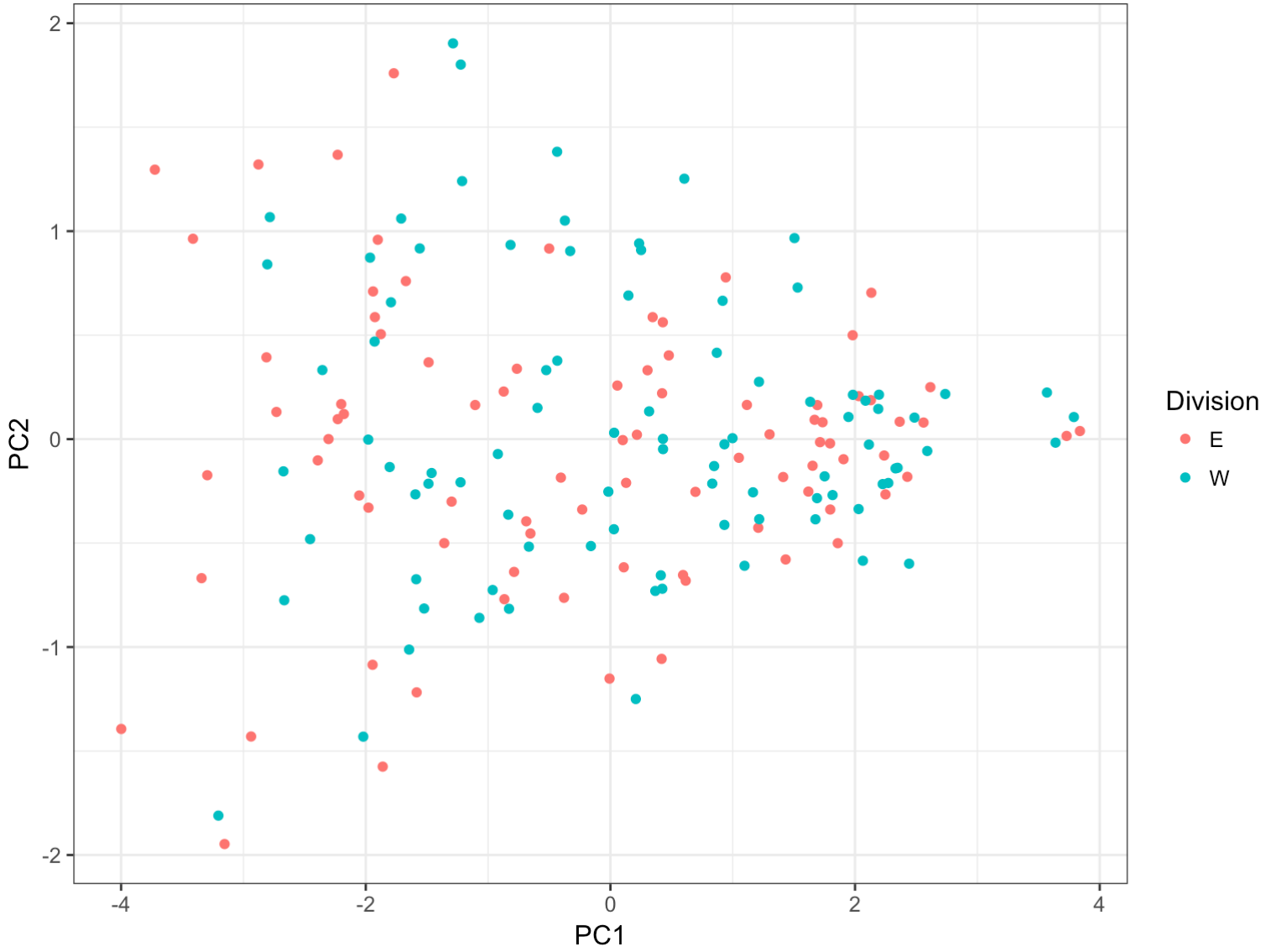
```
head(pca$x) #Look at PC1 and PC2
```

```
##             PC1        PC2         PC3         PC4
## [1,] -2.3922853 -0.1021936  0.48513364  0.58753648
## [2,]  2.0298873  0.2067437  0.47421428  0.03682271
## [3,] -0.5006467  0.9163373 -0.68459214 -0.17719826
## [4,]  1.7973324 -0.3386502  0.02322145  0.04343153
## [5,]  0.6951797 -0.2536477 -0.41295403 -0.46192248
## [6,] -1.9442727 -1.0854202 -0.64735028 -0.27118758
```

```
#Add the division back to the pca data
pca_data <- data.frame(pca$x, Division=Baseball3$div86)
head(pca_data)
```

```
##          PC1        PC2         PC3         PC4 Division
## 1 -2.3922853 -0.1021936  0.48513364  0.58753648        E
## 2  2.0298873  0.2067437  0.47421428  0.03682271        E
## 3 -0.5006467  0.9163373 -0.68459214 -0.17719826        E
## 4  1.7973324 -0.3386502  0.02322145  0.04343153        E
## 5  0.6951797 -0.2536477 -0.41295403 -0.46192248        E
## 6 -1.9442727 -1.0854202 -0.64735028 -0.27118758        E
```

```
#Perform the ggplot
ggplot(pca_data, aes(x=PC1, y=PC2, color=Division)) + geom_point()
```



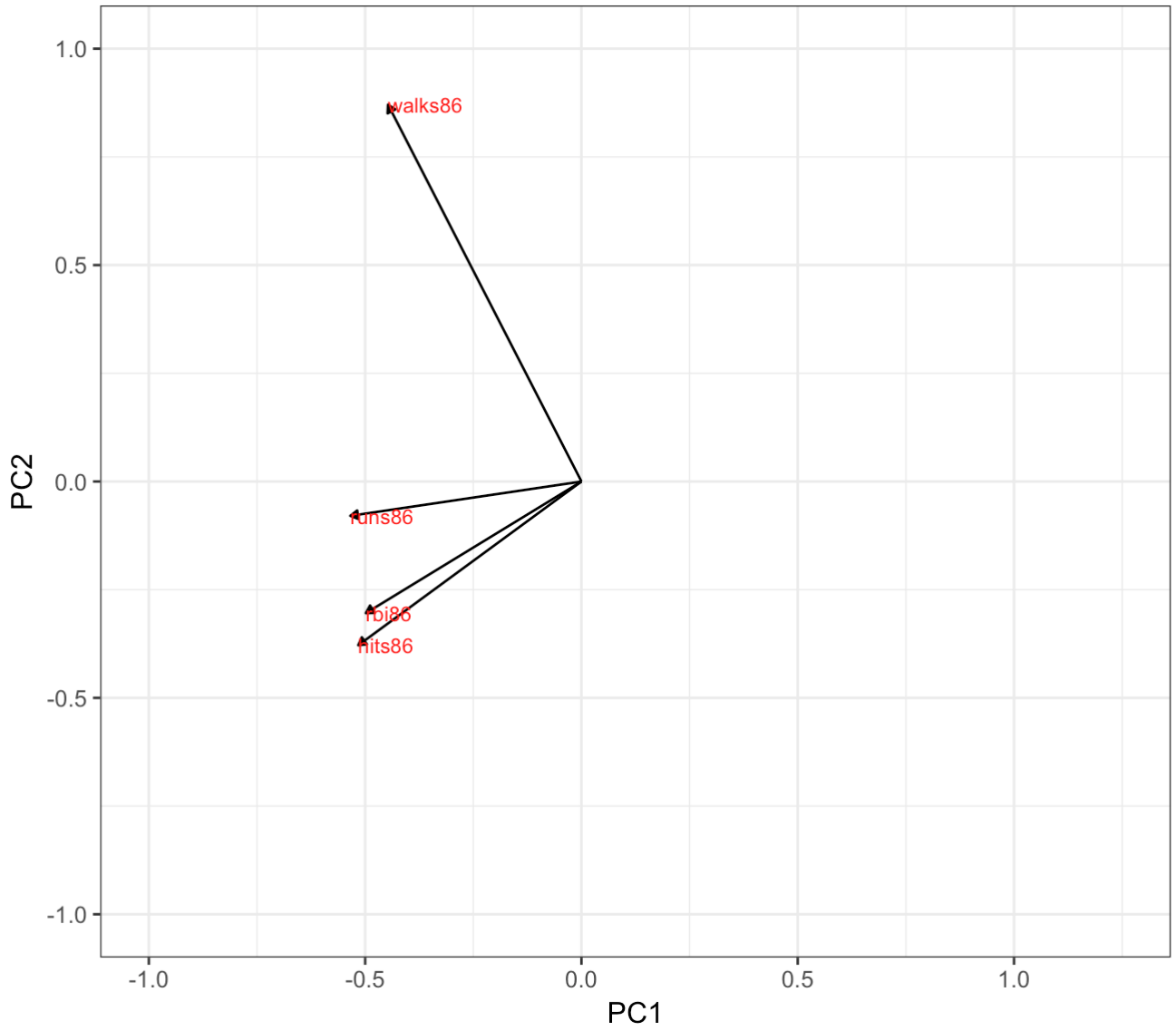

```
##Look at the rotation matrix
pca$rotation
```

```
##                PC1         PC2         PC3         PC4
## hits86  -0.5159228 -0.37841238  0.43818120 -0.63136750
## runs86  -0.5342607 -0.07940797  0.38250660  0.74963229
## rbi86   -0.4984106 -0.30428004 -0.81135413  0.02655285
## walks86 -0.4471868  0.87058148 -0.05822729 -0.19677768
```

```
#capture the rotation matrix in a data frame
rotation_data <- data.frame(pca$rotation, variable=row.names(pca$rotation))

# define a arrow style
arrow_style <- arrow(length = unit(0.05, "inches"),type = "closed")

# now plot, using geom_segment() for arrows and geom_text for labels
ggplot(rotation_data) +
  geom_segment(aes(xend=PC1, yend=PC2), x=0, y=0, arrow=arrow_style) +
  geom_text(aes(x=PC1, y=PC2, label=variable), hjust=0, size=3, color='red') +
  xlim(-1.,1.25) +
  ylim(-1.,1.) +
  coord_fixed() # fix aspect ratio to 1:1
```




```
#Show the percent variance for the 4 Principal Componets

percent <- 100*pca$sdev^2/sum(pca$sdev^2)
percent
```

```
## [1] 81.981169 11.263701  5.455032  1.300098
```

To compare the four characteristics between the divisions in the AL league, a Principal Component Analysis (PCA) was performed. The data was first filtered so that I could look at the division, hits, runs, rbi, and walks among baseball players in the 1986 AL League. Then a PCA analysis was then performed on the data set. Then a ggplot was performed showing the relationship between PC1 and PC2 for both the East and West Division. Then a rotation matrix plot was performed for comparing PC1 and PC2 graph. Then I calculated the percent variance for the four principal components.

Since I was trying to find a pattern of which variables were most distinguished among several baseball characteristics, I performed a PCA on the data. I then plotted a PC1 vs. PC2 graph to see if there was a difference in pattern between the East division and the West division. Then a plot of the rotation matrix of PC1 vs. PC2 graph was made to see which variables contribute more for both PC1 and PC2.

Based on the PCA plot, the data points of East and West division seemed to be mixed in the plot and not separated. Based on the rotation matrix of PC1 vs PC2 graph, walks mainly contribute to PC2, while runs, rbi's, and hits contribute equally to PC1 and PC2. We use PC1 and PC2 for the PCA plot of the division and the rotation matrix plot because PC1 and PC2 contribute to 93.24% of the total variance of the data.

I conclude based on the PCA plot that there was not a difference of pattern of the four offensive characteristics between East and West divisions in the AL League during 1986 baseball season. Based on the rotation matrix graph, out of the four characteristics: homeruns, walk, runs, and rbi's, walks was the most distinguished characteristic with a significantly higher PC2 value compared to the other characteristics in both divisions.