Major League Baseball

Project Appendix

Max Shalom - STA 3000

Setup Code

```
library(readr)
stats <- read_csv("stats.csv")
summary(stats)

library(dplyr)
library(ggplot2)
```

Code Block A

Calculating an additional statistic of OPS+

```
# Calculate league averages for OBP and SLG by year which is needed for the OPS+ formula
league_stats <- stats %>%
  group_by(year) %>%
  summarise(lgOBP = mean(on_base_percent),
            lgSLG = mean(slg_percent))

# Add OPS+ data by using the formula
stats <- stats %>%
  left_join(league_stats, by = "Year") %>%
  mutate(OPSplus = 100*((on_base_percent/lgOBP)+(slg_percent/lgSLG)-1))

# Create scatterplot of OPS+ by year with each player plotted
ggplot(stats, aes(x = year, y = OPSplus)) +
  geom_point()
```

Code Block B

Creating a correlation matrix

```r
install.packages("corrplot")
library(corrplot)

correlation_matrix <- cor(stats[, c("batting_avg", "on_base_percent", "slg_percent",
"woba", "xwoba", "sweet_spot_percent", "barrel_batted_rate", "hard_hit_percent")])
corrplot(correlation_matrix, method = "number")
```

Code Block C

Line chart to show different statistics over time

```r
stats_by_year <- stats %>%
  filter(year >= 2000) %>%
  group_by(year) %>%
  summarise(
    batting_avg = mean(batting_avg, na.rm = TRUE),
    on_base_percent = mean(on_base_percent, na.rm = TRUE),
    xslg = mean(xslg, na.rm = TRUE),
    xwoba = mean(xwoba, na.rm = TRUE)
  )

ggplot(stats_by_year, aes(x = year)) +
  geom_line(aes(y = batting_avg, color = "Batting Average")) +
  geom_line(aes(y = on_base_percent, color = "On-Base Percentage")) +
  geom_line(aes(y = xslg, color = "XSLG")) +
  geom_line(aes(y = xwoba, color = "XwOBA")) +
  labs(title = "Batting Statistics Over Time",
       x = "Year",
       y = "Statistic")
```

Code Block D

```r
# Calculate the correlation between barrel_batted_rate and home_run
correlation <- cor(stats$barrel_batted_rate, stats$home_run)

# Create a scatterplot of barrel_batted_rate vs. home_run
ggplot(stats, aes(x = barrel_batted_rate, y = home_run)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Barrel Batted Rate vs. Home Run",
       x = "Barrel Batted Rate",
       y = "Home Run") +
  annotate("text", x = 2, y = 50, label = paste("Correlation:", round(correlation, 2)))
```

Code Block E

```r
# Calculate the correlation between barrel_batted_rate and home_run
correlation <- cor(stats$barrel_batted_rate, stats$home_run)

# Fit a linear model to the data
model <- lm(home_run ~ barrel_batted_rate, data = stats)

# show the correlation and model summary
print(paste("Correlation:", round(correlation, 2)))
summary(model)
```