Exam 1 Solutions

Stats and sports class

Fall 2020

Part I: Data wrangling and exploratory analysis (35 pts)

```
library(Lahman); library(tidyverse)
Fielding_1 <- Fielding %>%
  mutate(fielding_attempts = PO + A + E,
         fpct = (PO + A)/fielding_attempts) %>%
  filter(fielding_attempts >= 100, yearID >= 1970, yearID <= 2000)
```

Question 1

Identify the player/year with the lowest fielding percentage in any season in this time frame.

```
Fielding_1 %>%
  arrange(fpct) %>%
  head(1)
```

```
##    playerID yearID stint teamID lgID POS  G GS InnOuts PO  A  E DP PB WP SB CS
## 1 colesda01   1987     1    DET   AL  3B 36 31     810 31 63 17  5 NA NA NA NA
##   ZR fielding_attempts      fpct
## 1 NA               111 0.8468468
```

ANSWER: The player `colesda01` boasted the lowest fielding percentage (Darnell Coles)

Question 2

ANSWER: Identify the outfielder (`POS == "OF"`) with the lowest fielding percentage in any season in this time frame.

```
Fielding_1 %>%
  filter(POS == "OF") %>%
  arrange(fpct) %>%
  head(1)
```

```
##    playerID yearID stint teamID lgID POS  G GS InnOuts  PO A  E DP PB WP SB CS
## 1 braggg101   1986     1    ML4   AL  OF 56 54    1426 116 5 12  0 NA NA NA NA
##   ZR fielding_attempts      fpct
## 1 NA               133 0.9097744
```

ANSWER: The player `braggg101` had the lowest field percentage among outfielders in this time frame (Glen Braggs)

Question 3

A coach wants to identify *perfect* fielders – that is, those whose `fpct` is 100 percent. What percent of players at each position register as having perfect fielding percentages?

```
Fielding_1 %>%
  mutate(is_perfect = (fpct == 1)) %>%
  group_by(POS) %>%
  summarise(ave_pos_perfect = mean(is_perfect))
```

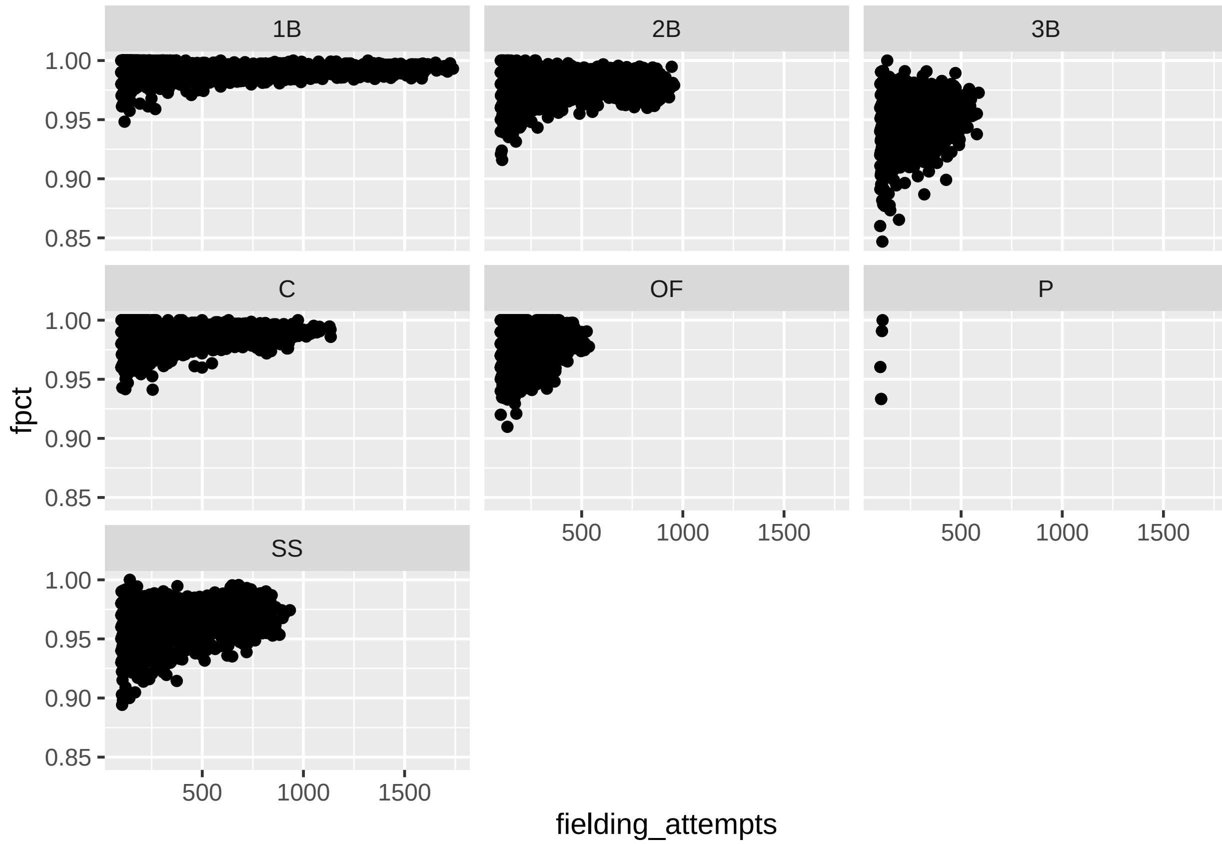
```
## # A tibble: 7 x 2
##   POS   ave_pos_perfect
##   <chr>           <dbl>
## 1 1B           0.0914
## 2 2B           0.0149
## 3 3B           0.000896
## 4 C            0.0410
## 5 OF           0.0365
## 6 P            0.2
## 7 SS           0.000723
```

ANSWER: The table above shows the percent of players at each position that field perfectly in a given season. Pitchers (20 percent) and first basemen (9 percent) are the highest, while shortstops and third basemen (less than 1 percent) are lowest.

Question 4

Make a chart of fielding percentage (y-axis) versus fielding attempts (x-axis), faceted by position. Describe the general trend of what happens as attempts goes up. What does this indicate about fielding percentage?

```
ggplot(Fielding_1, aes(fielding_attempts, fpct)) +
  geom_point() +
  facet_wrap(~POS)
```

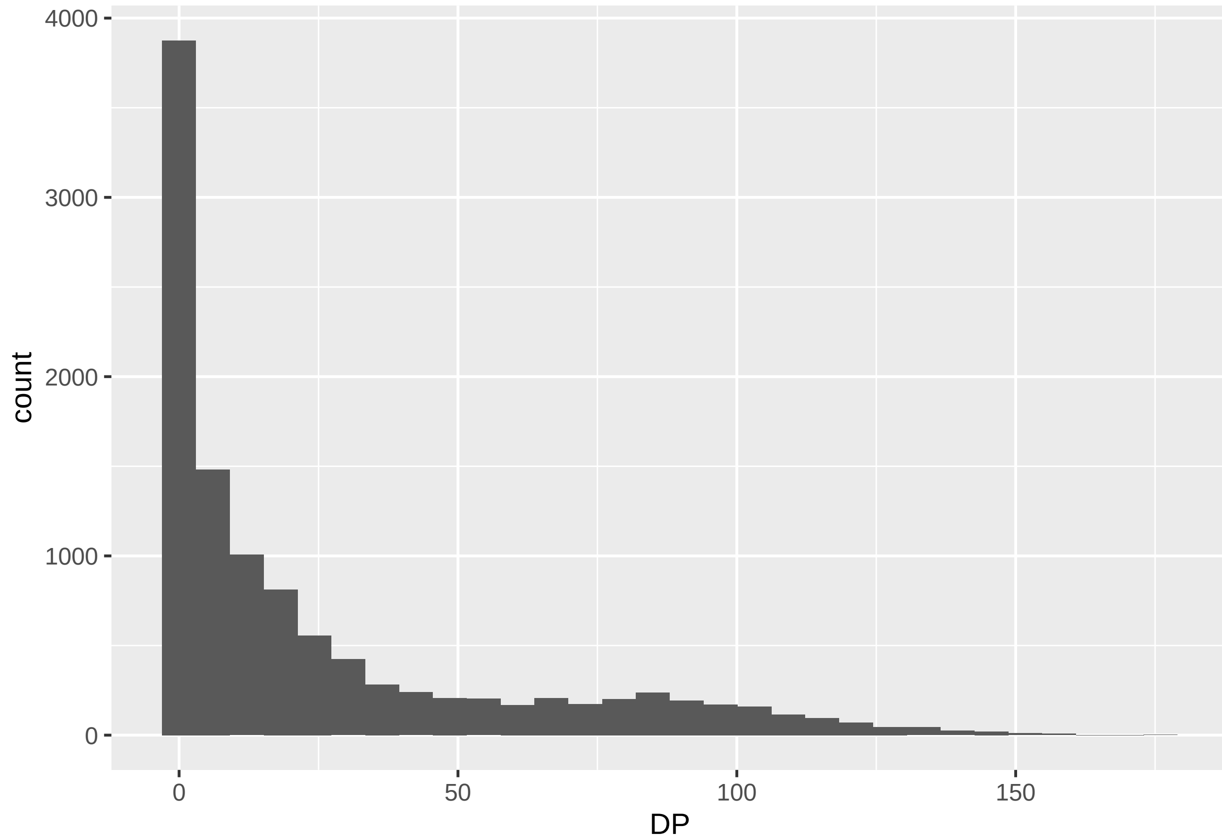


ANSWER: As attempts rise, the likelihood of really high (1) or really low (less than 0.9) fielding percentages drop. Most players tend to be pulled towards some type of positional average with an increasing number of attempts.

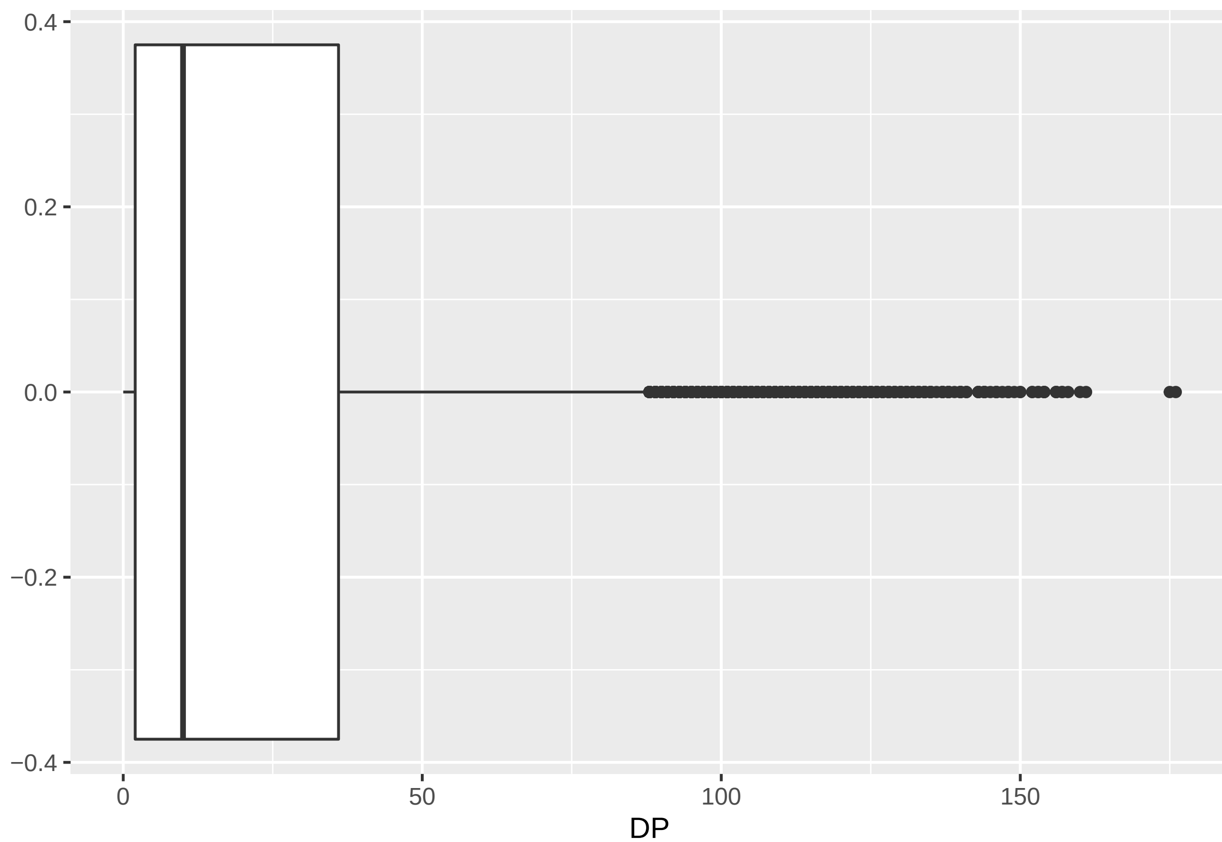
Question 5

Describe the center, shape, and spread of the double plays (DP) variable

```
ggplot(Fielding_1, aes(DP)) +
  geom_histogram()
```




```
ggplot(Fielding_1, aes(DP)) +
  geom_boxplot()
```



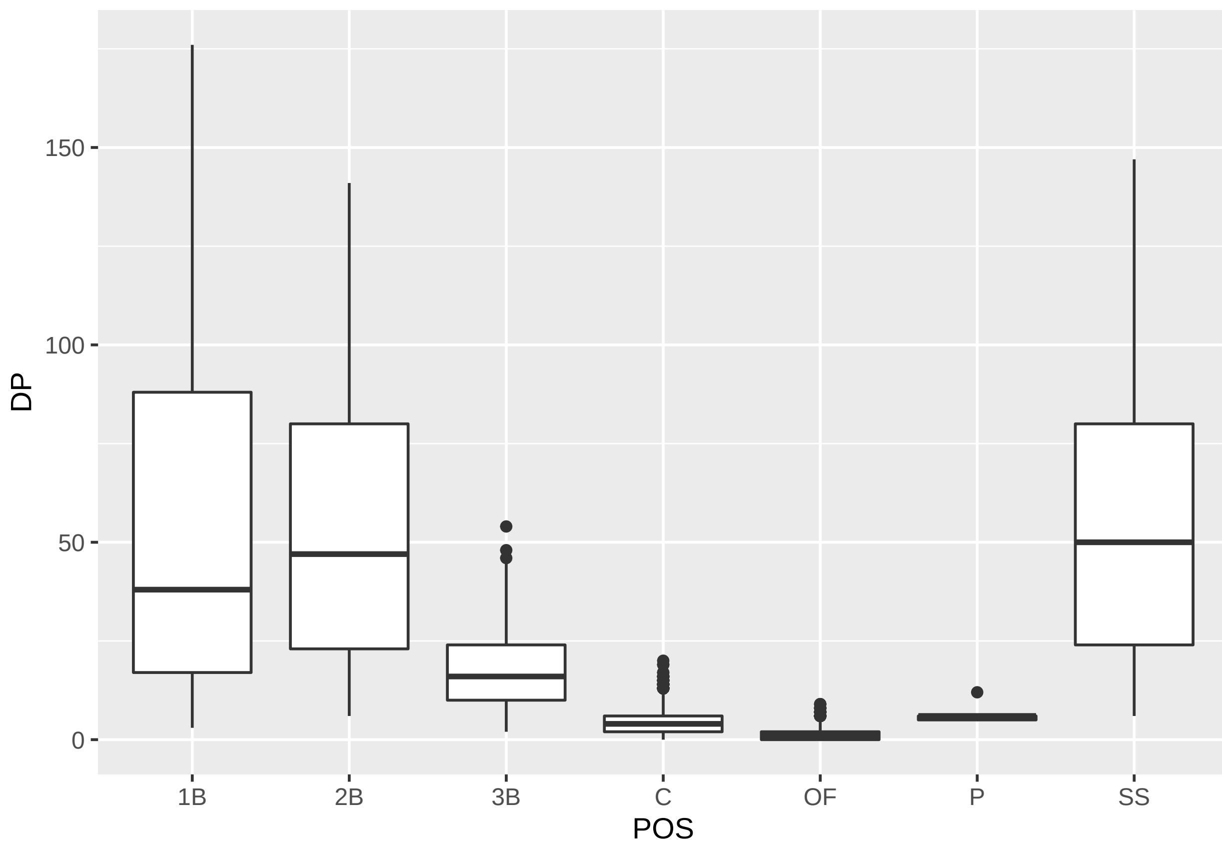
ANSWER: The distribution is strongly skewed right, ranging from 0 to around 170. The median is about 10 double plays per year.

Question 6

Use a visualization to compare the distribution of double plays (`DP`) turned by players at each position.

ANSWER: Various answers will work, Boxplots preferred over histograms for side by side comparisons

```
ggplot(Fielding_1, aes(x = POS, y = DP)) +
  geom_boxplot()
```



Question 7

For each player, calculate his average (average of each season) fielding percentage across this time frame. Which 5 players have the lowest average fielding percentage?

```
Fielding_1 %>%
  group_by(playerID) %>%
  summarise(ave_pct = mean(fpct)) %>%
  arrange(ave_pct) %>%
  head(5)
```

```
## # A tibble: 5 x 2
##   playerID  ave_pct
##   <chr>       <dbl>
## 1 alvarga01   0.873
## 2 bussera01   0.898
## 3 veraswi01   0.907
## 4 hartji01    0.908
## 5 hiattph01   0.909
```

ANSWER: The five players listed above have the lowest average season fielding percentages

Part II (35 pts)

```
nfl_kick <- read.csv("https://raw.githubusercontent.com/statsbylopez/StatsSports/master/Data/nfl_fg.csv
nfl_kick <- nfl_kick %>%
```

```r
  mutate(Distance_sq = Distance^2)
head(nfl_kick)
```

```
##   Team Year GameMinute Kicker Distance ScoreDiff Grass Temp Success Distance_sq
## 1  PHI 2005          3  Akers       49         0 FALSE   72       0        2401
## 2  PHI 2005         29  Akers       49        -7 FALSE   72       0        2401
## 3  PHI 2005         51  Akers       44        -7 FALSE   72       1        1936
## 4  PHI 2005         14  Akers       43        14  TRUE   82       0        1849
## 5  PHI 2005         60  Akers       23         0  TRUE   75       1         529
## 6  PHI 2005         39  Akers       34        -3  TRUE   68       1        1156
```

```r
fit_1 <- glm(Success ~ Distance_sq + Distance + Grass + Year,
             data = nfl_kick, family = "binomial")

fit_2 <- glm(Success ~ Distance + Grass + Year,
             data = nfl_kick, family = "binomial")
```

Question 1

Interpret the coefficient on Distance in `fit_2`, on the log odds scale

```r
library(broom)
tidy(fit_2)
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept) -104.      17.3         -6.02 1.78e-  9
## 2 Distance      -0.105    0.00318    -33.0  1.78e-238
## 3 GrassTRUE     -0.155    0.0549      -2.82 4.76e-  3
## 4 Year           0.0548   0.00863      6.35 2.13e- 10
```

ANSWER: For each additional increase in yards, the log odds of a field goal drop by 0.105, given a model with GRASS and Year

Question 2

Interpret the coefficient on Grass in `fit_2`, on the odds scale

ANSWER: Kicks on grass have an 0.856 times (or 14.4 percent lower) odds of going in, relative to kicks not on grass, given a model with distance and year.

Question 3

Which model would you recommend? Use two justifications

```r
AIC(fit_1)
```

```
## [1] 8705.486
```

```r
AIC(fit_2)
```

```
## [1] 8706.263
```

ANSWER: Fit 1 has a lower AIC

ANSWER: The coefficient on the squared term in Fit 1 is significant

Question 4

Using each of the models, estimate the probability of a successful field given the following conditions:

- 50 yards
- not on Grass
- Kicked in 2013

How important is the squared term in terms of changing the probability of this successful field goal?

ANSWER: You could do this by hand, or use R

```
new_data <- data.frame(Distance = 50, Distance_sq = 2500, Grass = TRUE, Year = 2013)
predict(fit_1, new_data, type = "response")
```

```
##         1
## 0.6628001
```

```
predict(fit_2, new_data, type = "response")
```

```
##         1
## 0.6622738
```

ANSWER: The squared term on distance impacts the likelihood of this field goal by less than a tenth of a percent

Question 5

Using the distance and surface info above, estimate the likelihood of the same field goal being made in 2030, and comment on the appropriateness of this estimate.

```
new_data <- data.frame(Distance = 50, Distance_sq = 2500, Grass = TRUE, Year = 2030)
predict(fit_1, new_data, type = "response")
```

```
##         1
## 0.8313757
```

ANSWER: The estimated likelihood is 83 percent. This seems like extrapolation, given that we don't have any data from the 2020's.

Question 6 (open ended)

Using a combination a code and intuition, assess "field goal success" as a yearly measure of kicker aptitude. Consider the three ways we've discussed measuring a metric.

ANSWER: Part I: Field goal sucess is linked strongly to scoring, as it leads directly to points.

Part II: Field goal success rates are mostly controlled by kickers, but there are other factors (distance, surface) that do impact success

Part III: Is field goal success repeatable? Multiple answers accepted

Part III (30 pts)

Return to the Lahman data.

```
Batting_1 <- Batting %>%
  filter(yearID >= 1970, AB >= 500) %>%
  mutate(K_rate = SO/(AB + BB),
         BB_rate = BB/(AB + BB),
         BA = H/AB,
```

```
         HR_rate = HR/(AB + BB),
         X1B = H - X2B - X3B - HR,
         TB = X1B + 2*X2B + 3*X3B + 4*HR,
         RC = (H + BB)*TB/(AB + BB))

Batting_1 <- Batting_1 %>%
  arrange(playerID, yearID) %>%
  group_by(playerID) %>%
  mutate(RC_next = lead(RC)) %>%
  filter(!is.na(RC_next)) %>%
  ungroup()
```

The following code creates categories for hitters based on the number of stolen bases they record in a season.

```
Batting_1 <- Batting_1 %>%
  mutate(SB_category = case_when(SB > 25 ~ "Fast",
                                 SB > 5 ~ "Moderate",
                                 SB <= 5 ~ "Slow"))
```

Question 1

Fit a regression model of runs created as a function of stolen base category, and interpret the coefficient for `Moderate` speed

```
fit_rc <- lm(RC ~ SB_category, data= Batting_1)
tidy(fit_rc)
```

```
## # A tibble: 3 x 5
##   term                estimate std.error statistic    p.value
##   <chr>                  <dbl>     <dbl>     <dbl>      <dbl>
## 1 (Intercept)            88.5      0.880    101.    0
## 2 SB_categoryModerate     3.38     1.05       3.23 0.00124
## 3 SB_categorySlow         5.13     1.10       4.68 0.00000299
```

ANSWER: Moderate speed players create 3.38 more runs than slow speed players

Question 2

What is the estimated difference in runs created between runners in the `Moderate` and `Slow` categories?

ANSWER: The difference between 5.13 and 3.38 is 1.75 runs

Question 3

Fit a series of multiple regression models trying to estimate the link between runs created in the future and runs created in the season. Consider the inclusion of three additional variables: `HR_Rate`, `K_rate`, and `BB_rate`, in addition to `RC`. Pick the best possible model using the AIC criterion.

ANSWER: Answers will vary

Example

```
fit_final <- lm(RC_next ~ HR_rate + K_rate + BB_rate + RC, data = Batting_1)
Batting_1 <- Batting_1 %>%
  mutate(predict_RC = predict(fit_final),
         resid_RC = resid(fit_final))
```

Question 4

Using your model in Question 3, estimate the residual for the very first row of the data set (`Batting_1`). Did this player create more or fewer runs than the model expected?

ANSWER: Answers will vary. Positive numbers: more runs created than the model expected

```
Batting_1 %>% head(1) %>% print.data.frame()
```

```
##    playerID yearID stint teamID lgID   G  AB   R   H X2B X3B HR RBI SB CS  BB
## 1 abreubo01   1999     1    PHI   NL 152 546 118 183  35  11 20  93 27  9 109
##    SO IBB HBP SH SF GIDP    K_rate   BB_rate        BA    HR_rate X1B  TB
## 1 113   8   3  0  4   13 0.1725191 0.1664122 0.3351648 0.03053435 117 300
##         RC RC_next SB_category predict_RC resid_RC
## 1 133.7405 133.074        Fast   114.6681 18.40582
```

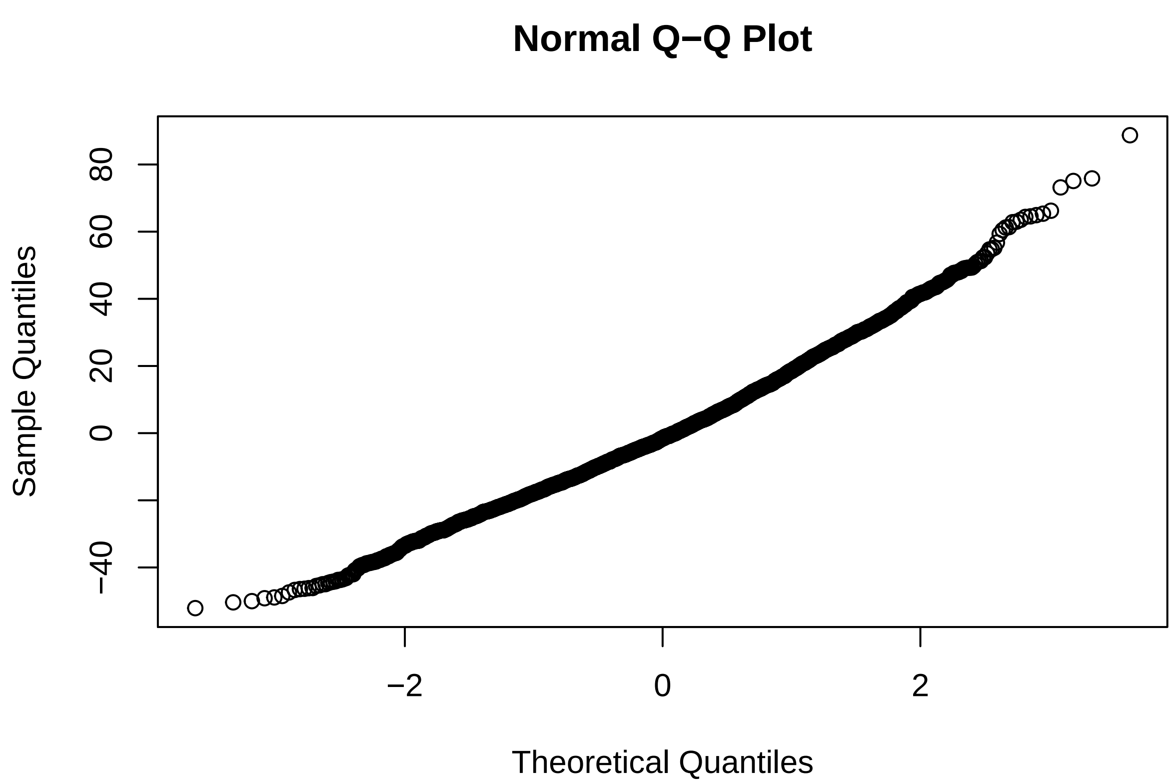
ANSWER: In the model above, the player created 19.4 more runs than expected

Question 5

Check your assumptions for fitting a multiple regression model (using your model in Q3)

ANSWER: Answers will vary – checking normality of residuals and scatter plots (assumption of linearity)

```
qqnorm(fit_final$residuals)
```



Question 6

For your best model, estimate the mean absolute error (MAE) when applying your model to the `Batting_1` data set. Interpret this number.

ANSWER: Answers will vary

```
Batting_1 %>%
  summarise(mae = mean(abs(predict_RC - RC_next)))
```

```
## # A tibble: 1 x 1
##     mae
##   <dbl>
## 1  14.6
```

ANSWER: Above the estimated MAE is 14.6 runs. That is, the model is typically off by 14.6 runs created.