# March Madness Project

*Sophie Moore*

*5/6/2019*

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------------------- tidyv
```

```
## v ggplot2 3.1.0     v purrr   0.3.0
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

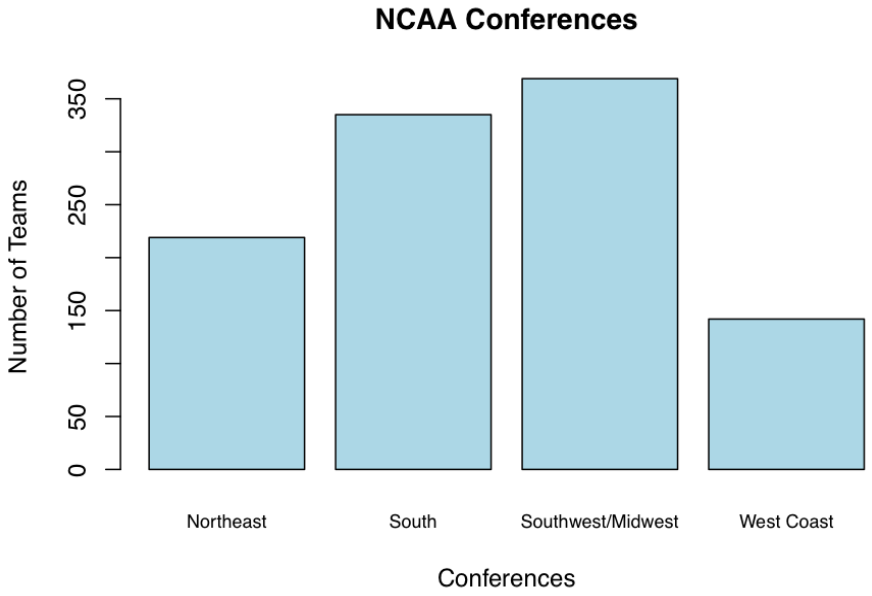
```
## -- Conflicts ------------------------------------------------------------------------------ tidyverse_c
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
bball <- read.csv("https://query.data.world/s/fwxjnf67s32hnzuyqy6ksedjlseb5m", header=TRUE, stringsAsFa
ncaabball1 = bball %>% filter(!is.na(`ncaa_result`))
ncaabball2 = ncaabball1 %>% filter(`ncaa_result`!= "Playing First Round")
ncaabball3 = ncaabball2 %>% filter(`ncaa_result`!= "Playing First Four")
ncaabball=ncaabball3 %>% filter(`ap_final`!=30)
head(ncaabball)
```

```
##    school conf rk  w  l    wl   srs  sos pts_for pts_vs pts_total ap_pre
## 1 alabama  SEC 13 24  8 0.750 16.74 6.81    75.8   65.1     140.9     18
## 2 alabama  SEC 16 27  8 0.771 16.93 7.67    75.9   66.7     142.6     24
## 3 alabama  SEC 23 23 10 0.697 15.47 6.59    73.0     NA        NA     18
## 4 alabama  SEC 26 26  9 0.743 13.52 7.29      NA     NA        NA     17
## 5 alabama  SEC 27 23 10 0.697 14.30 8.70      NA     NA        NA      7
## 6 alabama  SEC 28 26  9 0.743 15.95 5.77      NA     NA        NA     30
##   ap_high ap_final pts_diff            ncaa_result ncaa_numeric  season
## 1      11       21     10.7        Lost First Round            1 2004-05
## 2       5        8      9.2       Lost Second Round            2 2001-02
## 3      18       20       NA       Lost Second Round            2 1994-95
## 4       9       13       NA       Lost Second Round            2 1991-92
## 5       6       19       NA Lost Regional Semifinal            8 1990-91
## 6      19       23       NA Lost Regional Semifinal            8 1989-90
##                  coaches year
## 1  Mark Gottfried (24-8) 2004
## 2  Mark Gottfried (27-8) 2001
## 3    David Hobbs (23-10) 1994
## 4  Wimp Sanderson (26-9) 1991
## 5 Wimp Sanderson (23-10) 1990
## 6  Wimp Sanderson (26-9) 1989
```

Conference

```
ncaabball$conf=factor(ncaabball$conf, ordered = TRUE)
ncaabball$newconf=NA
ncaabball$newconf[ncaabball$conf %in% c("A-10", "Big East", "CBA", "ECACM", "Ivy","MAAC", "MAC", "Metro
ncaabball$newconf[ncaabball$conf %in% c("AAWU", "MW City", "MW College", "MWC", "Pac-10", "Pac-12", "Pa
ncaabball$newconf[ncaabball$conf %in% c("A-Sun","AAC","ACC", "Big South", "CAA", "CUSA", "SEC", "Southe
ncaabball$newconf[ncaabball$conf %in% c("Big 10", "Big 7", "Big 8", "Big Sky", "Big 12","BIAA", "Big 10
counts <- table(ncaabball$newconf)
barplot(counts, main= "NCAA Conferences", xlab= "Conferences", ylab= "Number of Teams", cex.names=.75,
```

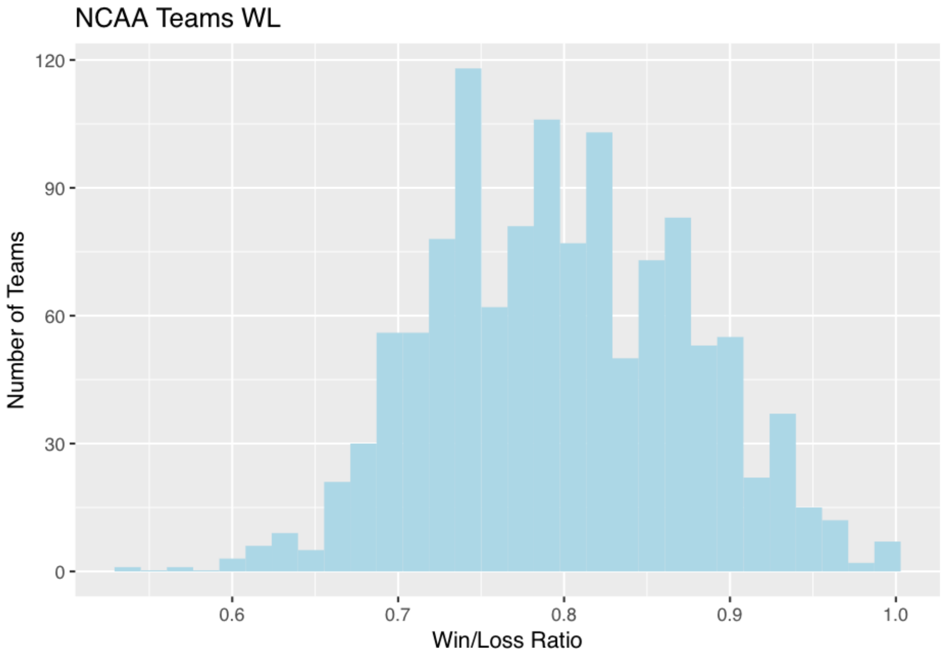


```
table(counts)
```

```
## counts
## 142 219 335 369
##   1   1   1   1
```

WL (a team's win percentage)

```
ggplot(ncaabball, aes(wl)) + geom_histogram(bins = 30, fill=" light blue") + labs(title = "NCAA Teams W
```



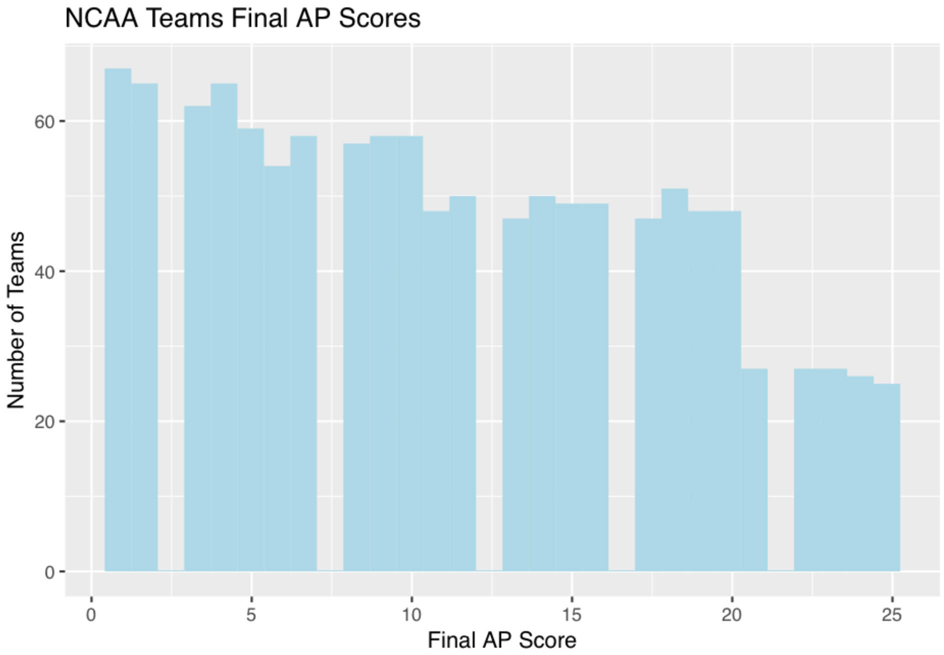
```
summary(ncaabball$wl)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.5420  0.7420  0.7940  0.7979  0.8570  1.0000
```

AP Final

```
ggplot(ncaabball, aes(ap_final)) + geom_histogram(fill="light blue") + labs(title = "NCAA Teams Final AI
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

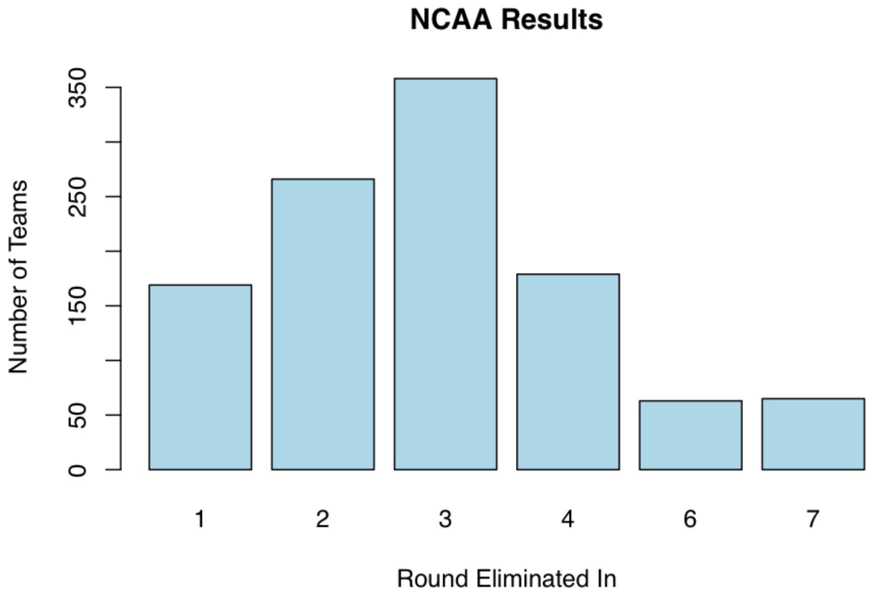


```
summary(ncaabball$ap_final)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     1.0     5.0    11.0    11.3    17.0    25.0
```

NCAA result

```
ncaabball$ncaa_result=factor(ncaabball$ncaa_result, levels=c("Lost First Four","Lost Opening Round","Lo
ncaabball$numericresult=NA
ncaabball$numericresult[ncaabball$ncaa_result %in% c("Lost First Four", "Lost Opening Round", "Lost Fir
ncaabball$numericresult[ncaabball$ncaa_result %in% c("Lost Second Round")]<-2
ncaabball$numericresult[ncaabball$ncaa_result %in% c("Lost Third Round","Lost Regional Semifinal")]<-3
ncaabball$numericresult[ncaabball$ncaa_result %in% c("Lost Regional Final")]<-4
ncaabball$numericresult[ncaabball$ncaa_result %in% c("Lost Regional Final (Final Four)","Lost National 
ncaabball$numericresult[ncaabball$ncaa_result %in% c( "Lost National Final")]<-6
ncaabball$numericresult[ncaabball$ncaa_result %in% c("Won National Final")]<-7
counts2 <- table(ncaabball$numericresult)
barplot(counts2, main= "NCAA Results", xlab = "Round Eliminated In", ylab= "Number of Teams", col = "li
```

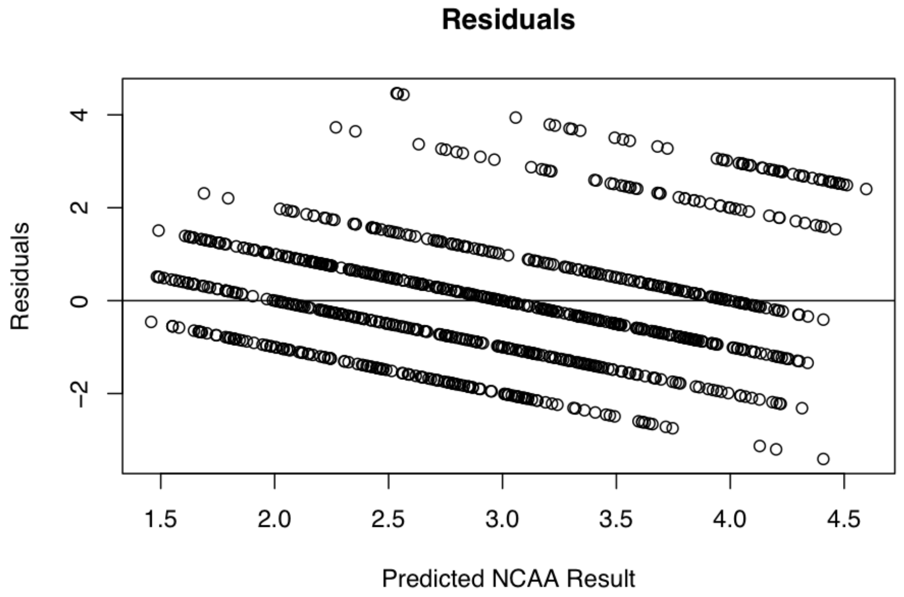


```r
table(counts2)
```

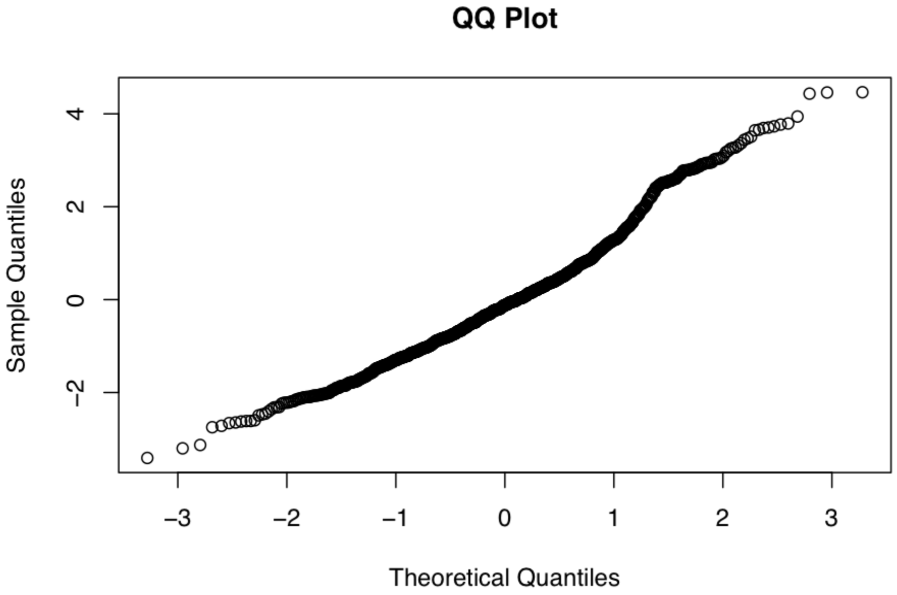
```
## counts2
##  63  65 169 179 266 358
##   1   1   1   1   1   1
```

Multiple Regression

```r
numericresultline=lm(numericresult~wl+ap_final+newconf, data=ncaabball)
result.res = resid(numericresultline)
plot(predict(numericresultline), result.res, ylab = "Residuals", xlab="Predicted NCAA Result", main = "
abline(0,0)
```



```r
qqnorm(result.res, main= "QQ Plot")
```



```r
summary(numericresultline)
```

```
##
```

```
## Call:
## lm(formula = numericresult ~ wl + ap_final + newconf, data = ncaabball)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4085 -0.9739 -0.1203  0.7433  4.4651
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.933633   0.736882   2.624  0.00883 **
## wl                         2.651817   0.814762   3.255  0.00117 **
## ap_final                  -0.088313   0.009112  -9.692  < 2e-16 ***
## newconfSouth               0.099650   0.125376   0.795  0.42692
## newconfSouthwest/Midwest -0.092474   0.123155  -0.751  0.45291
## newconfWest Coast         -0.017764   0.154205  -0.115  0.90831
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.373 on 956 degrees of freedom
##   (260 observations deleted due to missingness)
## Multiple R-squared:  0.2379, Adjusted R-squared:  0.2339
## F-statistic: 59.68 on 5 and 956 DF,  p-value: < 2.2e-16
```

```r
confint(numericresultline)
```

```
##                               2.5 %      97.5 %
## (Intercept)                0.4875395  3.37972639
## wl                         1.0528892  4.25074499
## ap_final                  -0.1061953 -0.07043156
## newconfSouth              -0.1463934  0.34569387
## newconfSouthwest/Midwest -0.3341588  0.14921149
## newconfWest Coast         -0.3203833  0.28485471
```

$$\widehat{\text{NCAA Result}} = 1.933633 + 2.651817 \times \text{WL} - 0.088313 \times \text{AP Final}$$