Outliers

Bryana Benson
Conner Bryan

```
library(readr)
library(dplyr)
library(ggplot2)
```

```
data <- read_csv("cricket.csv")
head(data)
```

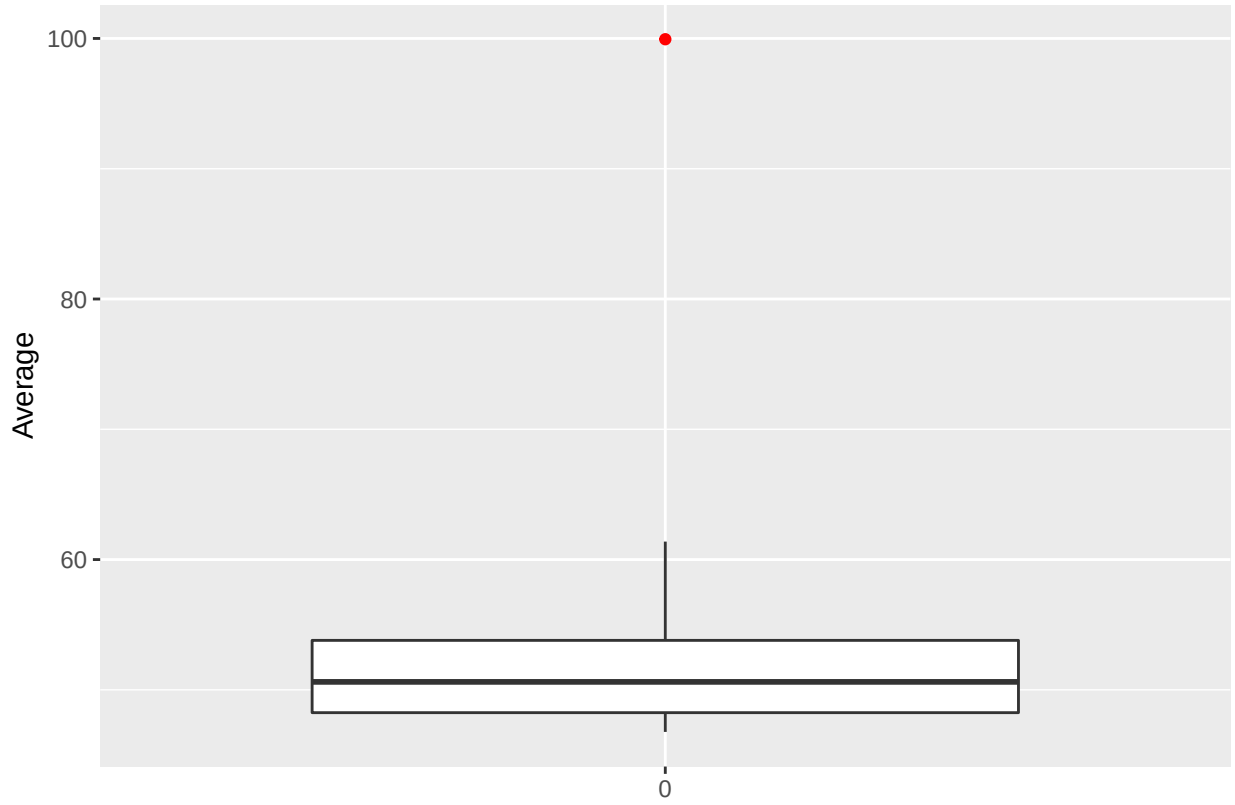
```
## # A tibble: 6 x 4
##   Batsman         Country     Period    Average
##   <chr>           <chr>       <chr>       <dbl>
## 1 Bradman, D G    Australia   1928-1948    99.9
## 2 Smith, S P D*   Australia   2010-        61.4
## 3 Sutcliffe, H    England     1924-1935    60.7
## 4 Barrington, K F England     1955-1968    58.7
## 5 Weekes, E D C   West Indies 1948-1958    58.6
## 6 Hammond, W R    England     1927-1947    58.5
```

```
str(data)
```

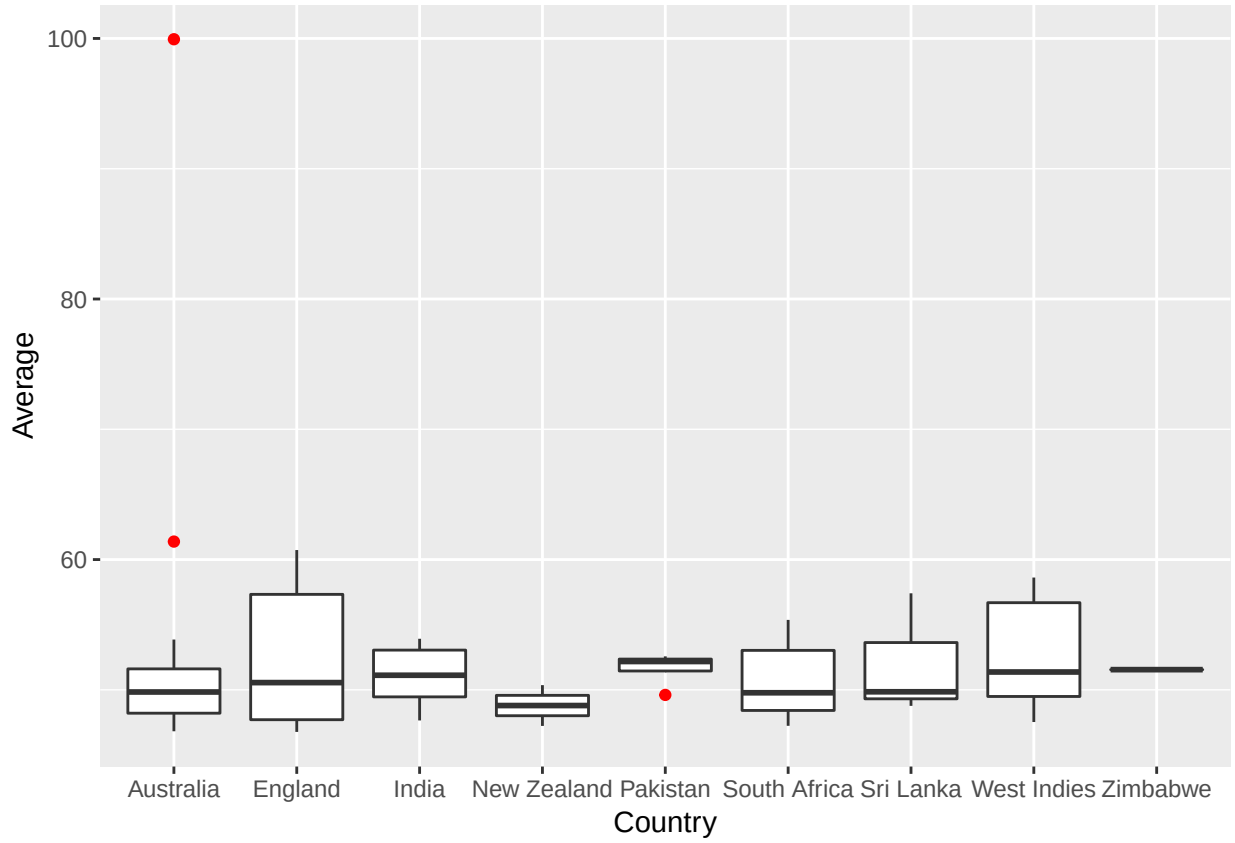
```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 60 obs. of  4 variables:
##  $ Batsman: chr  "Bradman, D G" "Smith, S P D*" "Sutcliffe, H" "Barrington, K F" ...
##  $ Country: chr  "Australia" "Australia" "England" "England" ...
##  $ Period : chr  "1928-1948" "2010-" "1924-1935" "1955-1968" ...
##  $ Average: num  99.9 61.4 60.7 58.7 58.6 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Batsman = col_character(),
##   ..   Country = col_character(),
##   ..   Period = col_character(),
##   ..   Average = col_double()
##   .. )
```

Is Bradman an outlier?

```
ggplot(data = data ,aes(factor(0),Average)) +
  geom_boxplot(outlier.colour='red') + xlab("") + ylab("Average")
```

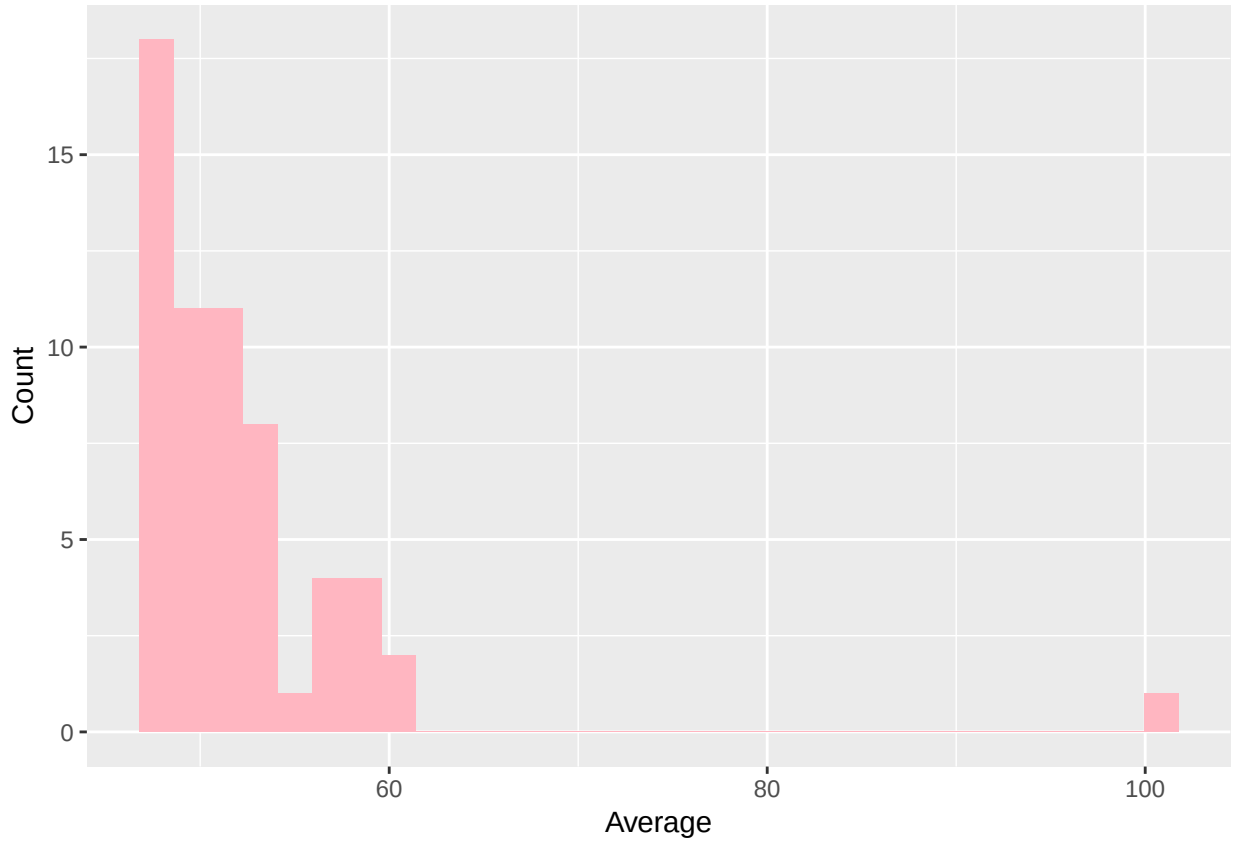



```
ggplot(data = data ,aes(Country,Average)) +
  geom_boxplot(outlier.colour='red') + xlab("Country") + ylab("Average")
```




```
ggplot(data = data, aes(Average)) +
  geom_histogram(fill = "light pink")  +
  xlab("Average") +
  ylab("Count")
```

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




```
#Conducting a Shapiro-Wilk test for normality
#A small p-value indicates a low probability that data is normally distributed
shapiro.test(data$Average)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Average
## W = 0.55016, p-value = 2.782e-12
```

The p-value for the Shapiro-Wilk test for Normality is less than .0000. This indicates that the data is normally distributed.

```
#Using Residual analysis for testing outlier status
lmAverage<-lm(Average ~ Country, data=data) # linear regression model
data$Predictions <- predict(lmAverage, newdata=data)
data$Residuals <- data$Average - data$Predictions
head(data)
```

```
## # A tibble: 6 x 6
##   Batsman         Country     Period    Average Predictions Residuals
##   <chr>           <chr>       <chr>       <dbl>       <dbl>     <dbl>
## 1 Bradman, D G    Australia   1928-1948    99.9        53.4     46.5
## 2 Smith, S P D*   Australia   2010-        61.4        53.4      7.96
## 3 Sutcliffe, H    England     1924-1935    60.7        52.5      8.24
## 4 Barrington, K F England     1955-1968    58.7        52.5      6.18
## 5 Weekes, E D C   West Indies 1948-1958    58.6        52.5      6.15
## 6 Hammond, W R    England     1927-1947    58.5        52.5      5.97
```

A linear regression model was created predicting the Average based on Country. Those Average Predictions were stored in the dataset. The residuals were also stored in the dataset.

```
#Test to see if Residuals are normally distributed
shapiro.test(data$Residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Residuals
## W = 0.59065, p-value = 1.138e-11
```

The p-value of the Shapiro-Wilks test is less than .0000 which suggests the residuals are normally distributed and suggests no trend or pattern.

```
###Using manipulated data for testing/visualizing
noBradman <- filter(data, Batsman != "Bradman, D G")
```

Exluded Brandman from the dataset.

```
shapiro.test(noBradman$Average) #running test without Bradman gives less certainty
```

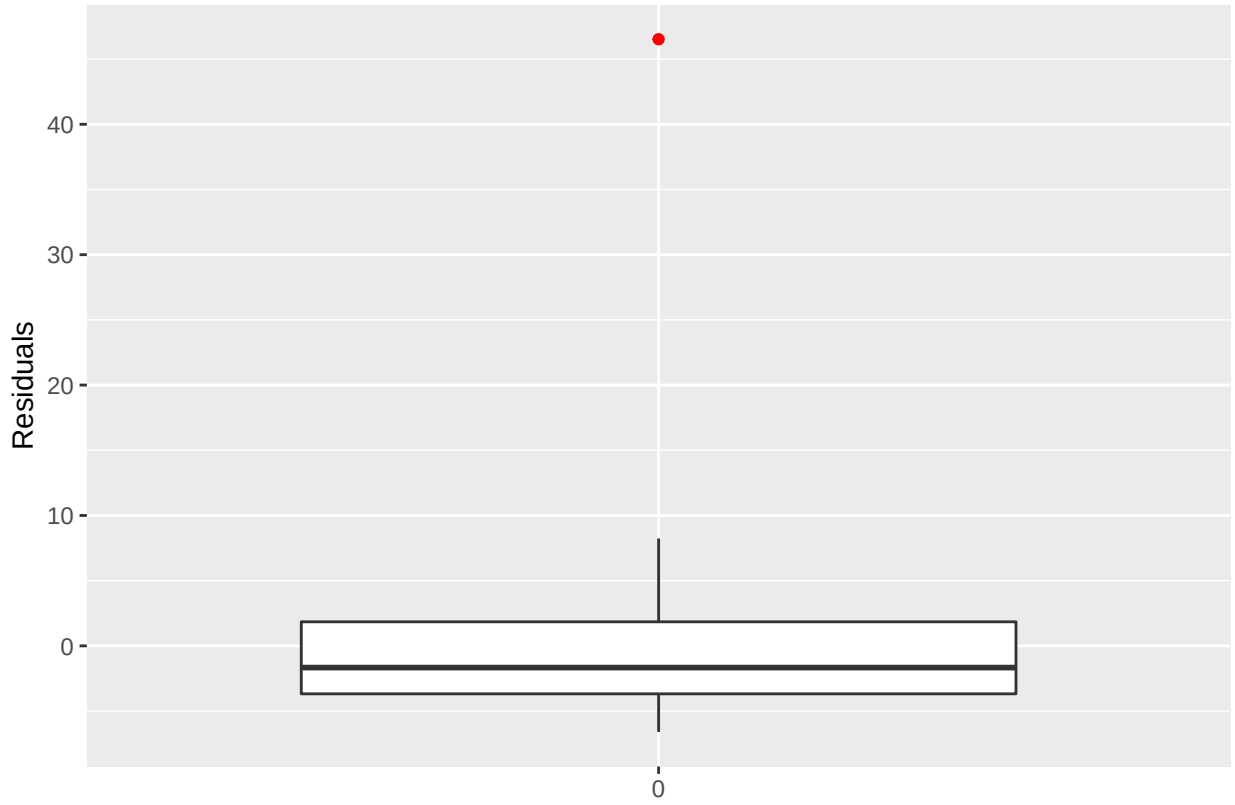
```
##
##  Shapiro-Wilk normality test
##
## data:  noBradman$Average
## W = 0.89728, p-value = 0.0001192
```

Running the Shapiro-Wilks test without Bradman results in less certainty, but is still statistically significant for normality.

```
shapiro.test(noBradman$Residuals) #running residual test without Bradman
```

```
##
##  Shapiro-Wilk normality test
##
## data:  noBradman$Residuals
## W = 0.9364, p-value = 0.00412
```

```
ggplot(data,aes(factor(0),Residuals)) +
  geom_boxplot(outlier.colour='red') + xlab("") + ylab("Residuals") #plotting residuals
```




```
#Calculating how far Bradman is from "average"
avg_Average<-mean(data$Residuals) #the average of all residuals
data[1,6] - (avg_Average + (3*sd(data$Residuals)))
```

```
##   Residuals
## 1  24.77986
```

```
#the difference between Bradman's residual and 3 deviations above the mean
```

CONCLUSIONS

Bradman is an outlier. A quick boxplot seems to confirm this fact. Even when segmenting out results by country, Bradman appears as an outlier in Australia. When Bradman is removed from the sample, it conforms more closely to a normal distribution. Even an analysis of the residuals with Country as the explanatory variable reveals that Bradman is far removed from any sort of standard deviation of the group.