Lab 0

Alexander Brandt
SID: 24092167

September 1 2017

1 Loading the Data

We load the USArrests in R, as well the file from the data/ directory.

```
library("dplyr")
library("tidyverse")

setwd("/Users/Alex/Dropbox/STAT_215A/stat215a/lab0")

# Load the data from the assignment
statecoord <- read.table("data/stateCoord.txt")
# Load the library's data on US arrests
data("USArrests")
```

2 Manipulating the Data

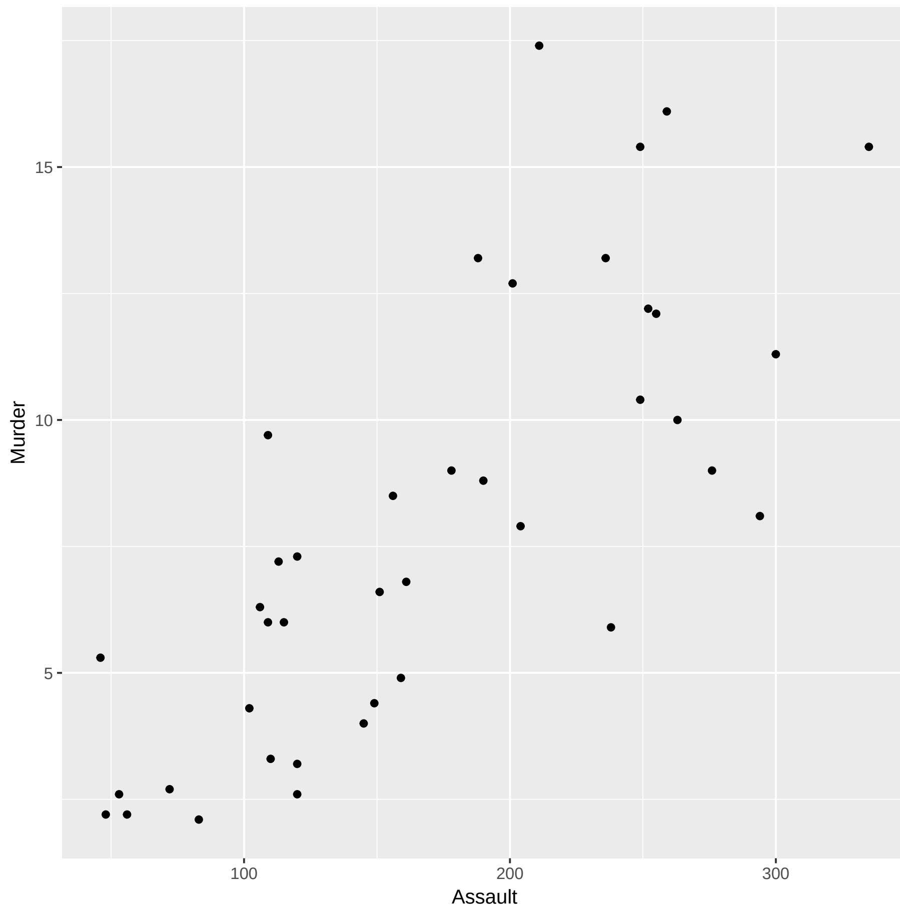
Next we merge the datasets by introducing a new column which I will call "state" – this is because dplyr doesn't really play well with column names natively.

```
# Populate both with a common column so they can be merged (dplyr doesn't really like to me
# by row names
USArrests$state <- rownames(USArrests)
statecoord$state <- rownames(statecoord)
full_dataset <- inner_join(USArrests, statecoord, by = "state")
```

3 Visualizing the Data

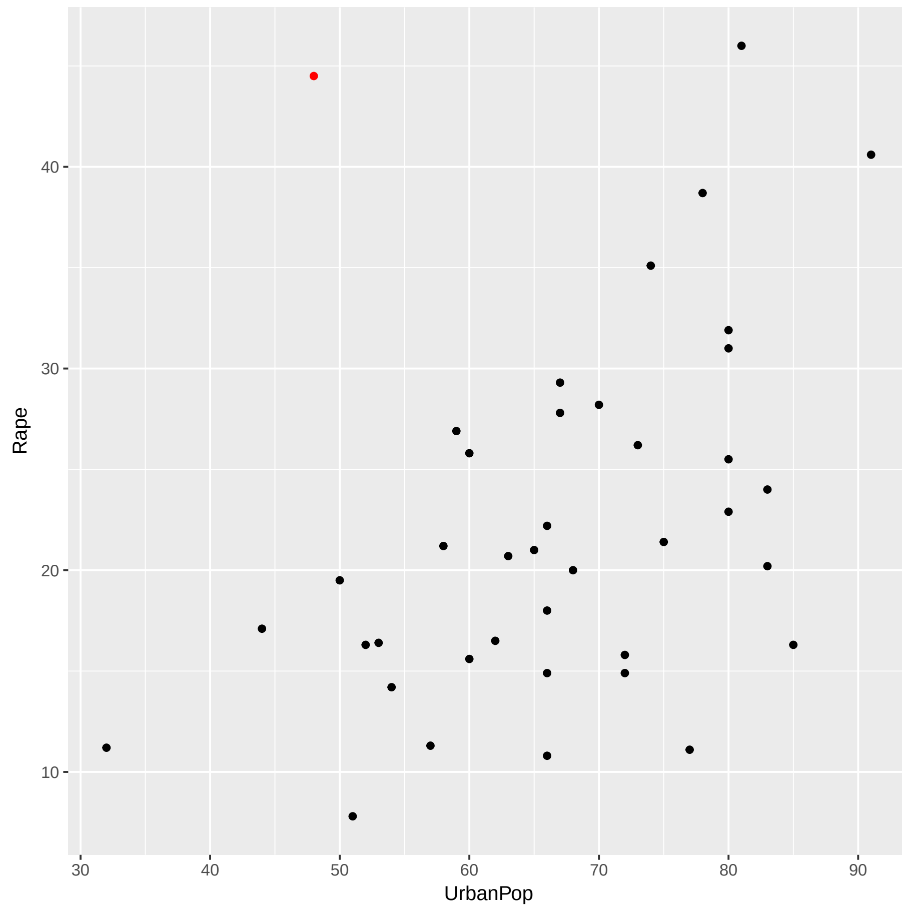
First we plot "Murder" vs. "Assault" – there seems to be a slightly positive trend:

```
p <- ggplot(full_dataset, aes(Assault, Murder, ))
p + geom_point()
```



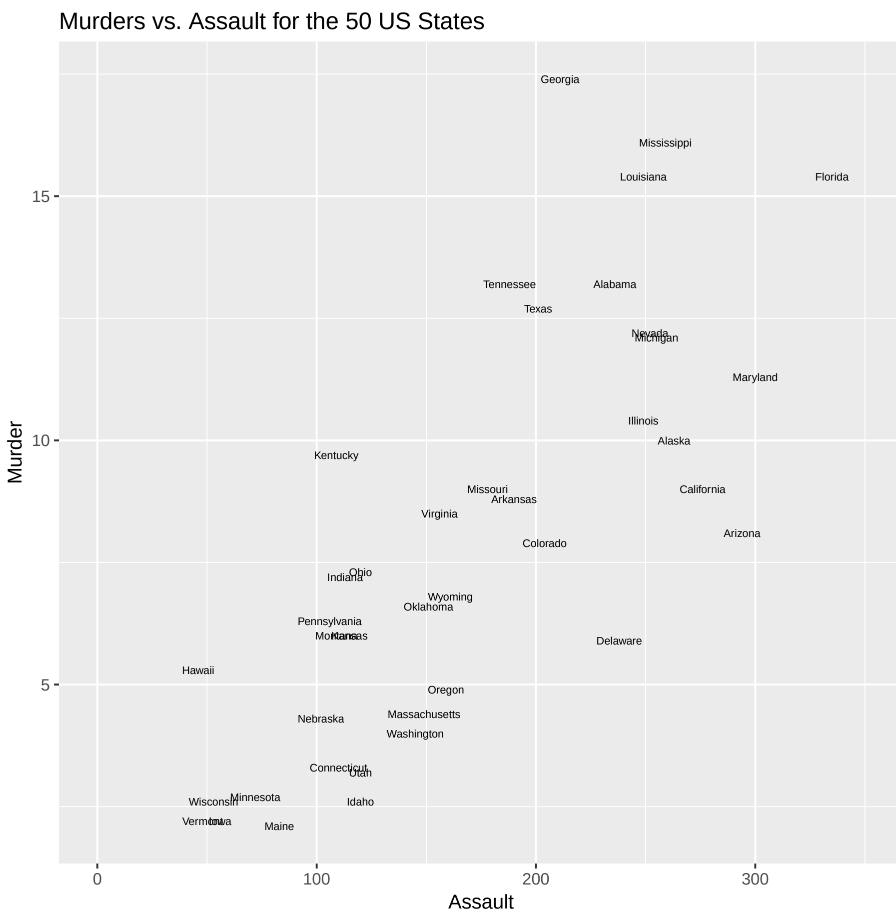
Next we plot “Rape” vs. urban population. We mark the outlier with a red dot.

```
point_colors <- rep('black', times = nrow(full_dataset))
point_colors[full_dataset$UrbanPop < 50 & full_dataset$Rape > 40] <- 'red'
p <- ggplot(full_dataset, aes(UrbanPop, Rape, ))
p + geom_point(color = point_colors)
```



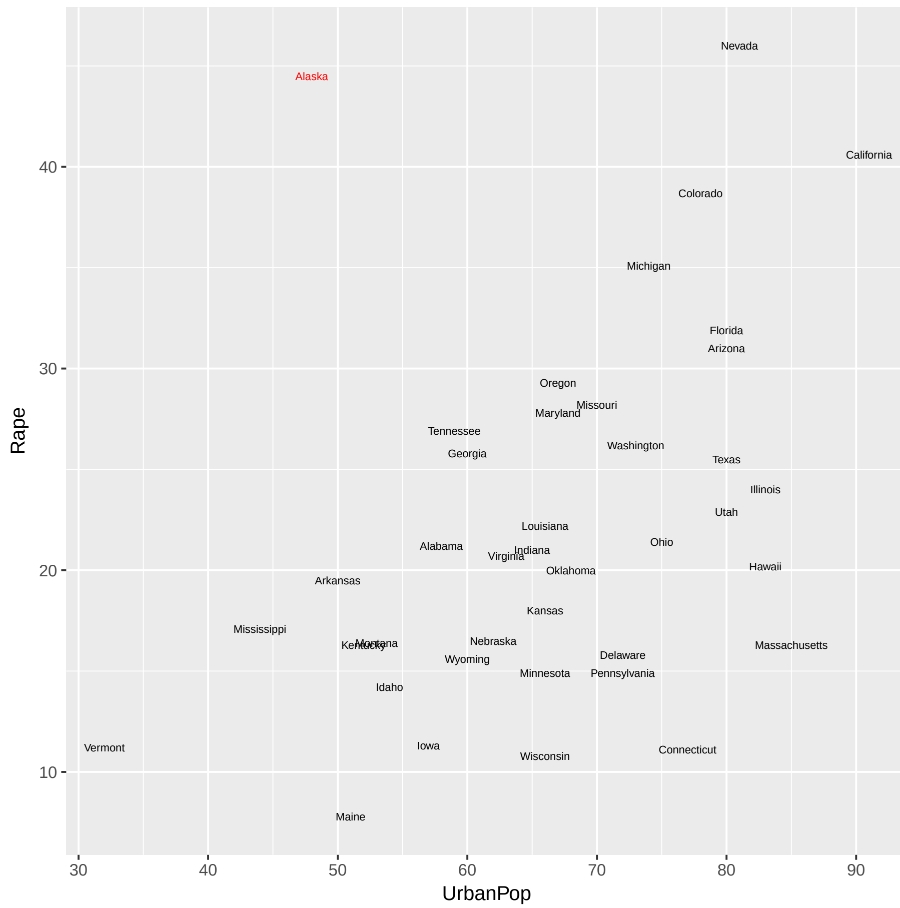
Then we remake these plots with names. First “Murder” vs. “Assault”:

```
p <- ggplot(full_dataset, aes(Assault, Murder, label = full_dataset$state))
p + geom_text(size=2) +
  labs(title = "Murders vs. Assault for the 50 US States") +
  xlim(0,350)
```

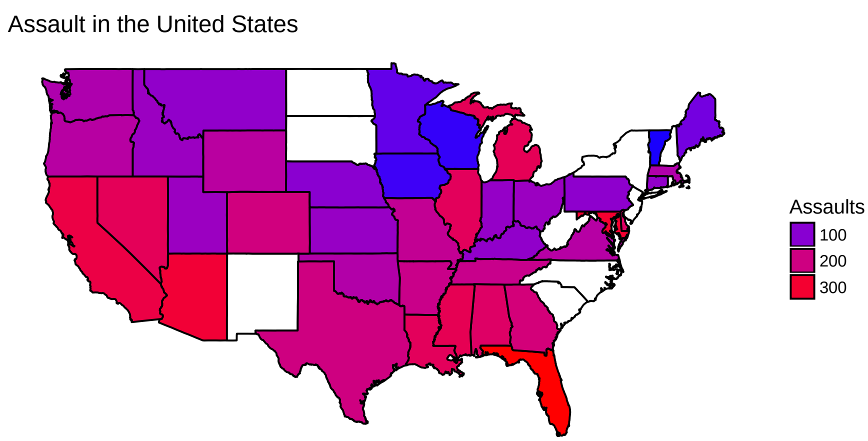


Then "Rape" vs. urban population:

```
p <- ggplot(full_dataset, aes(UrbanPop, Rape, label = full_dataset$state))
p + geom_text(size=2, color = point_colors)
```



I decided to try the challenge exercise:



4 Regression

First we fit a linear regression using the lm function:

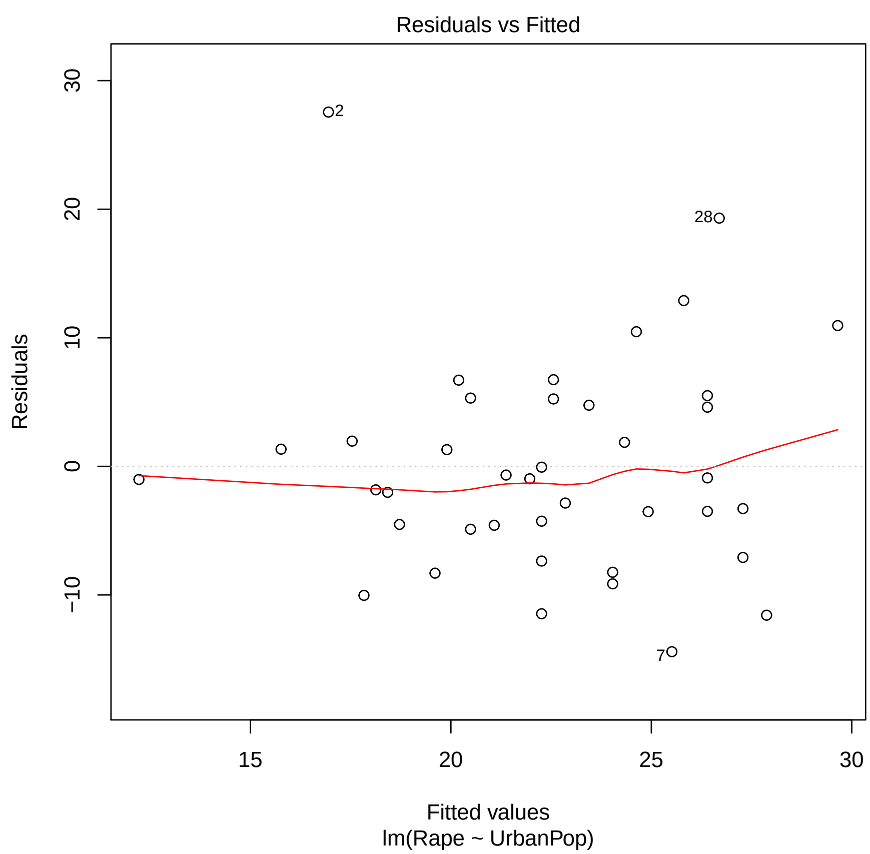
```
fit <- lm(Rape ~ UrbanPop, data=full_dataset)
summary(fit)

##
## Call:
## lm(formula = Rape ~ UrbanPop, data = full_dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4128  -4.6588  -0.9933   4.8767  27.5544
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7653     7.3041   0.379  0.70709
## UrbanPop      0.2954     0.1076   2.745  0.00918 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.576 on 38 degrees of freedom
## Multiple R-squared:  0.1655,Adjusted R-squared:  0.1436
## F-statistic: 7.538 on 1 and 38 DF,  p-value: 0.009178
```

Next we plot the predicted values versus the residuals:

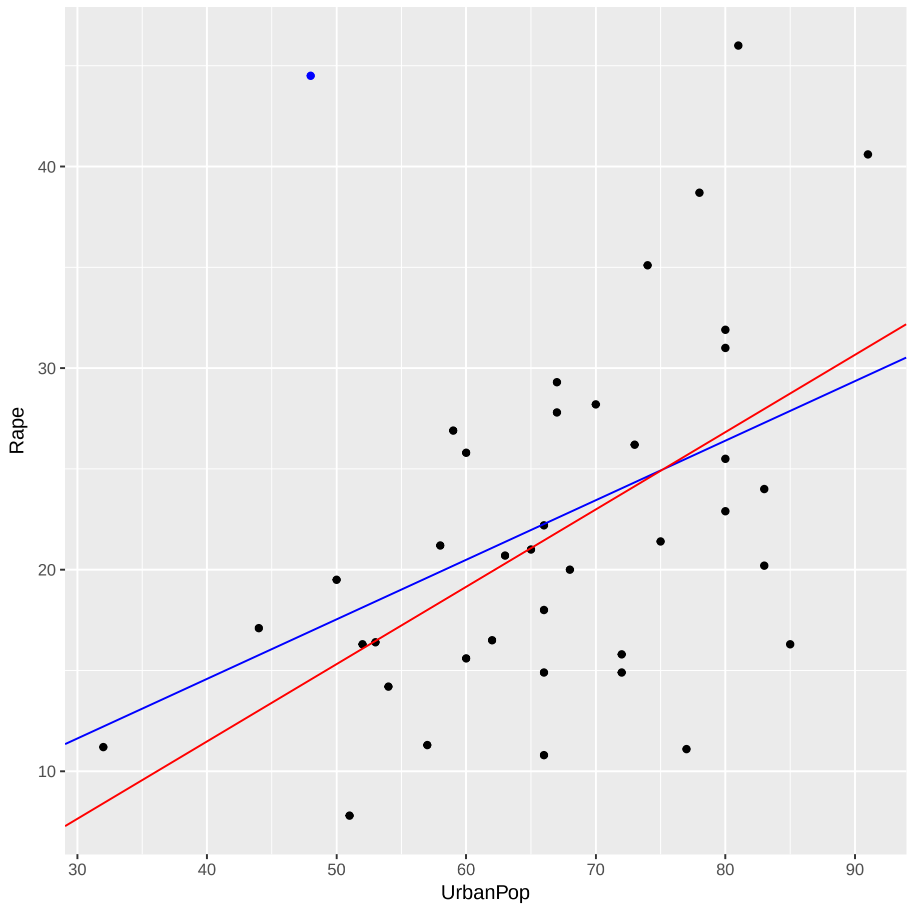
```r
plot(fit, which = 1,)
```



There seems to be a spreading of the residuals as the fitted values increase.

```r
# Fit with the outlier (Alaska) removed
no_outlier <- full_dataset[c(-2),]
fit2 <- lm(Rape ~ UrbanPop, data=no_outlier)

point_colors[point_colors=='red'] <- 'blue'
xvals <- seq(min(full_dataset$UrbanPop), max(full_dataset$UrbanPop))
p <- ggplot(full_dataset, aes(UrbanPop, Rape, label = full_dataset$state))
p + geom_point(color = point_colors) +
  geom_abline(intercept=coefficients(fit)[1], slope=coefficients(fit)[2], col='blue') +
  geom_abline(intercept=coefficients(fit2)[1], slope=coefficients(fit2)[2], col='red')
```



When we compare the lines, we see the data set with the removed data point has a better fit (judged by p-value and R^2 value).

```r
summary(fit)

##
```

```
## Call:
## lm(formula = Rape ~ UrbanPop, data = full_dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4128  -4.6588  -0.9933   4.8767  27.5544
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7653     7.3041   0.379  0.70709
## UrbanPop      0.2954     0.1076   2.745  0.00918 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.576 on 38 degrees of freedom
## Multiple R-squared:  0.1655,Adjusted R-squared:  0.1436
## F-statistic: 7.538 on 1 and 38 DF,  p-value: 0.009178
```

```
summary(fit2)
```

```
##
## Call:
## lm(formula = Rape ~ UrbanPop, data = no_outlier)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.572  -3.948  -0.066   4.631  18.793
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.86452    6.43793   -0.60 0.551983
## UrbanPop     0.38359    0.09424    4.07 0.000237 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.296 on 37 degrees of freedom
## Multiple R-squared:  0.3093,Adjusted R-squared:  0.2906
## F-statistic: 16.57 on 1 and 37 DF,  p-value: 0.0002369
```