Lab 1_EDA_Tsung-Chin Han

Han, Tsung-Chin

May 25, 2018

Backgroud

Imagine that you have been hired by the World Bank to study the effect of cultural norms and legal enforcement in controlling corruption by analyzing the parking behavior of United Nations officials in Manhattan. Until 2002, diplomatic immunity protected UN diplomats from parking enforcement ac- tions, so diplomats actions were constrained by cultural norms alone. In 2002, enforcement authorities acquired the right to confiscate diplomatic license plates of violators, after which diplomatic behavior was constrained by both cultural norms and the legal penalties of unpaid tickets.

Data

You are given a dataset for a selection of UN diplomatic missions, Corrupt.R. The dependent (or target) variable in this data is named violations.

The labels of some of the variables are listed below; the rest of the variables should be self-explanatory.

(1) corruption: Country corruption index, 1998
(2) violations: Unpaid New York City parking violations
(3) trade: total trade with the United States (1998 US$)

Objective

The World Bank would like to know what if any relationship there is between corruption and parking violations both pre and post 2002 and if there are any other relevant explanatory variables.

Setup

Setup the working directory and load the given file and library.

```
getwd()
```

```
## [1] "C:/Users/Ken/Desktop/UC Berkeley MIDS/4 - 2018 Summer/5 - Statistics for Data Science/lab/1"
```

```
setwd("C:/Users/Ken/Desktop/UC Berkeley MIDS/4 - 2018 Summer/5 - Statistics for Data Science/lab/1")
getwd()
```

```
## [1] "C:/Users/Ken/Desktop/UC Berkeley MIDS/4 - 2018 Summer/5 - Statistics for Data Science/lab/1"
```

```
#load library
library(car)
```

```
## Warning: package 'car' was built under R version 3.4.4
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 3.4.4
```

```
#load the data and rename data for analysis
load("Corrupt.Rdata")
Data<-FMcorrupt
```

Data Selection

We know we have 364 obersvation and 28 variables. Data type includes floating numbers, integer, and charecter.

```
View(Data) #brifely check what the data looks like

#Dimension of data and obersvation
dim(Data)
```

```
## [1] 364  28
```

```
str(Data)
```

```
## 'data.frame':    364 obs. of  28 variables:
##  $ wbcode        : chr  "AFG" "AGO" "AGO" "ALB" ...
##  $ prepost       : chr  "" "pre" "pos" "pre" ...
##  $ violations    : num  NA 744.38 15.37 256.63 5.56 ...
##  $ fines         : num  NA 40294 1208 13970 610 ...
##  $ mission       : int  NA 1 1 1 1 1 1 1 1 1 ...
##  $ staff         : int  NA 9 9 3 3 3 3 19 19 4 ...
##  $ spouse        : int  NA 4 4 3 3 2 2 10 10 1 ...
##  $ gov_wage_gdp  : num  NA 1.3 1.3 1.3 1.3 ...
##  $ pctmuslim     : num  NA 0.01 0.01 0.7 0.7 ...
##  $ majoritymuslim: int  NA 0 0 1 1 1 1 0 0 -1 ...
##  $ trade         : num  NA 2.61e+09 2.61e+09 2.72e+07 2.72e+07 ...
##  $ cars_total    : int  NA 24 24 4 4 13 13 15 15 3 ...
##  $ cars_personal : int  NA 3 3 0 0 6 6 14 14 1 ...
##  $ cars_mission  : int  NA 21 21 4 4 7 7 1 1 2 ...
##  $ pop1998       : num  NA 11739390 11739390 3101330 3101330 ...
##  $ gdppcus1998   : num  NA 731 731 1008 1008 ...
##  $ ecaid         : num  NA 92.3 92.3 62.8 62.8 ...
##  $ milaid        : num  NA 0 0 2.2 2.2 ...
##  $ region        : int  NA 6 6 3 3 7 7 2 2 4 ...
##  $ corruption    : num  NA 1.048 1.048 0.921 0.921 ...
##  $ totaid        : num  NA 92.3 92.3 65 65 ...
##  $ r_africa      : int  NA 1 1 0 0 0 0 0 0 0 ...
##  $ r_middleeast  : int  NA 0 0 0 0 1 1 0 0 0 ...
##  $ r_europe      : int  NA 0 0 1 1 0 0 0 0 0 ...
##  $ r_southamerica: int  NA 0 0 0 0 0 0 1 1 0 ...
##  $ r_asia        : int  NA 0 0 0 0 0 0 0 0 1 ...
##  $ country       : chr  "AFGANISTAN" "ANGOLA" "ANGOLA" "ALBANIA" ...
##  $ distUNplz     : num  0.445 1.554 1.554 1.775 1.775 ...
```

```
summary(Data$violations)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
##    0.000    0.654    5.724  100.879   51.915 3392.961       66
```

```r
summary(Data$corruption)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
## -2.58299 -0.46186  0.32292 -0.00932  0.71516  1.58281       61 
```

```r
summary(Data$prepost)
```

```
##    Length     Class      Mode 
##       364 character character 
```

Given the questions, the data we are concerned contains both pre and post 2002.

the subset data of pre and post 2002 have 302 observations.

```r
# check levels of prepost column, any other values?
table(Data$prepost)
```

```
## 
##     pos pre 
##  62 151 151 
```

```r
# subset only pre & post 2002 dataset and get rid of NAs
prepost_Data<-subset(Data, Data$prepost == "pre" | Data$prepost == "pos")
nrow(prepost_Data)
```

```
## [1] 302
```

```r
summary(prepost_Data)
```

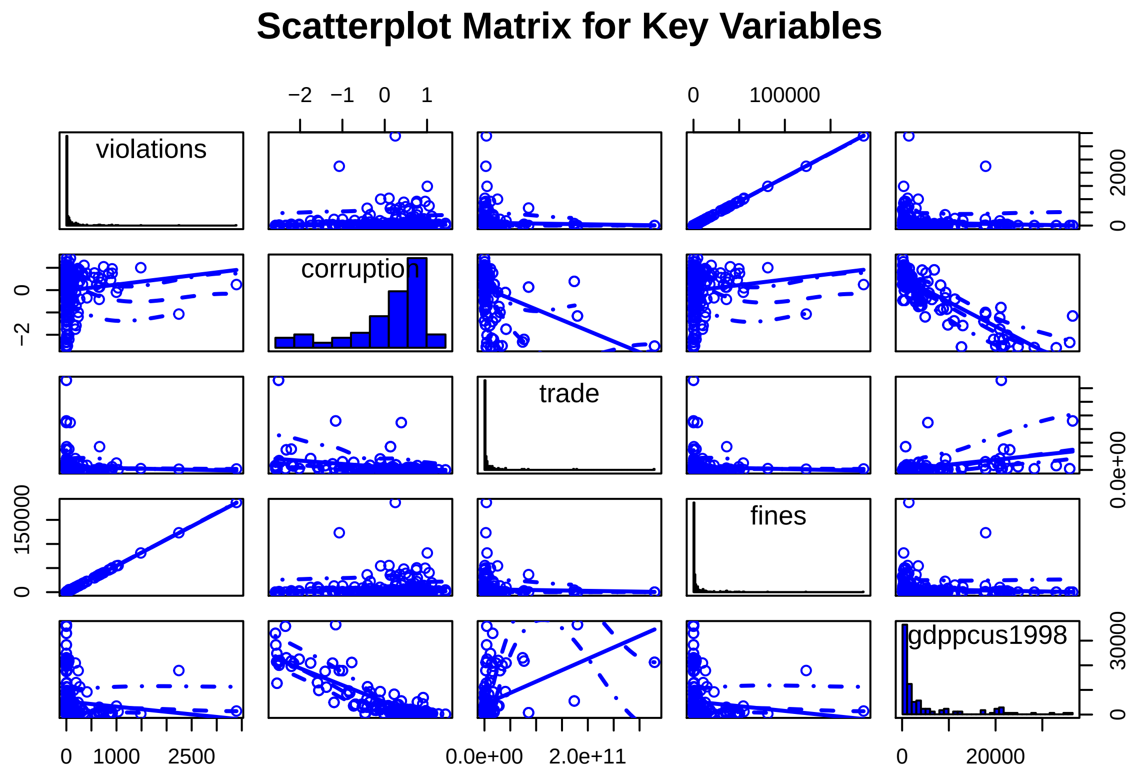
```
##     wbcode            prepost            violations      
##  Length:302         Length:302         Min.   :   0.000  
##  Class :character   Class :character   1st Qu.:   0.654  
##  Mode  :character   Mode  :character   Median :   5.724  
##                                        Mean   : 100.879  
##                                        3rd Qu.:  51.915  
##                                        Max.   :3392.961  
##                                        NA's   :4         
##      fines              mission           staff           spouse      
##  Min.   :     0.00   Min.   :0.0000   Min.   : 0.00   Min.   : 0.000  
##  1st Qu.:    65.41   1st Qu.:1.0000   1st Qu.: 5.00   1st Qu.: 3.000  
##  Median :   579.72   Median :1.0000   Median : 9.00   Median : 5.000  
##  Mean   :  5579.60   Mean   :0.9868   Mean   :11.65   Mean   : 7.656  
##  3rd Qu.:  2999.05   3rd Qu.:1.0000   3rd Qu.:14.00   3rd Qu.:10.000  
##  Max.   :186163.17   Max.   :1.0000   Max.   :86.00   Max.   :81.000  
##  NA's   :4                                                            
##   gov_wage_gdp      pctmuslim      majoritymuslim        trade          
##  Min.   : 0.100   Min.   :0.0000   Min.   :-1.0000   Min.   :0.000e+00  
##  1st Qu.: 1.300   1st Qu.:0.0060   1st Qu.: 0.0000   1st Qu.:9.532e+07  
##  Median : 1.900   Median :0.0500   Median : 0.0000   Median :5.443e+08  
##  Mean   : 2.828   Mean   :0.2766   Mean   : 0.2416   Mean   :1.034e+10  
##  3rd Qu.: 3.625   3rd Qu.:0.5400   3rd Qu.: 1.0000   3rd Qu.:4.904e+09  
##  Max.   :11.800   Max.   :0.9990   Max.   : 1.0000   Max.   :3.290e+11  
##  NA's   :118      NA's   :4        NA's   :4         NA's   :6          
##    cars_total      cars_personal      cars_mission        pop1998         
##  Min.   :  1.00   Min.   : 0.000   Min.   :  0.000   Min.   :5.308e+05  
##  1st Qu.:  3.00   1st Qu.: 1.000   1st Qu.:  2.000   1st Qu.:3.775e+06
##  Median :  7.00   Median : 2.000   Median :  3.000   Median :8.257e+06
##  Mean   : 10.47   Mean   : 5.324   Mean   :  5.144   Mean   :3.613e+07
##  3rd Qu.: 12.00   3rd Qu.: 6.000   3rd Qu.:  6.000   3rd Qu.:2.319e+07
##  Max.   :116.00   Max.   :64.000   Max.   :116.000   Max.   :1.242e+09
##  NA's   :24       NA's   :24       NA's   :24
##   gdppcus1998          ecaid             milaid             region
##  Min.   :   95.45   Min.   :   0.00   Min.   :   0.000   Min.   :1.000
##  1st Qu.:  413.61   1st Qu.:   0.00   1st Qu.:   0.000   1st Qu.:3.000
##  Median : 1416.04   Median :   8.70   Median :   0.200   Median :4.000
##  Mean   : 5223.74   Mean   :  49.27   Mean   :  33.048   Mean   :4.347
##  3rd Qu.: 5142.80   3rd Qu.:  40.30   3rd Qu.:   0.775   3rd Qu.:6.000
##  Max.   :36485.64   Max.   :1026.10   Max.   :3120.000   Max.   :7.000
##                     NA's   :8         NA's   :8          NA's   :2
##    corruption           totaid             r_africa       r_middleeast
##  Min.   :-2.582988   Min.   :   0.000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:-0.451213   1st Qu.:   0.325   1st Qu.:0.0000   1st Qu.:0.00000
##  Median : 0.322920   Median :   9.000   Median :0.0000   Median :0.00000
##  Mean   :-0.007721   Mean   :  82.320   Mean   :0.3046   Mean   :0.09934
##  3rd Qu.: 0.717707   3rd Qu.:  42.950   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   : 1.582807   Max.   :4069.100   Max.   :1.0000   Max.   :1.00000
##                      NA's   :8
##     r_europe      r_southamerica       r_asia         country
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Length:302
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   Class :character
##  Median :0.0000   Median :0.0000   Median :0.0000   Mode  :character
##  Mean   :0.2318   Mean   :0.1192   Mean   :0.1722
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##    distUNplz
##  Min.   : 0.0000
##  1st Qu.: 0.2219
##  Median : 0.2956
##  Mean   : 0.5493
##  3rd Qu.: 0.4608
##  Max.   :15.0552
##  NA's   :10
```

looking at the subset, we note there are some missing values. corruption index has no missing values, whereas violations have two missing values. This is reasonably small fraction of our cases.

Exploratory Analysis

We begin the scatterplot matrix. We want to get a high level overview.

```
scatterplotMatrix(~ violations+ corruption+ trade+ fines+ gdppcus1998,
                  data = prepost_Data,
                  diagonal=list(method="histogram", breaks="FD"),
                  main = "Scatterplot Matrix for Key Variables")
```



Violations is our dependent variable to look at. Interestingly, there seems little or no relationship between parking violations and corruption index. Rather, violations seems to have strong postive relationship with fines.

```
cor(prepost_Data$violations, prepost_Data$corruption, use="complete.obs")
```

```
## [1] 0.07884143
```

```
cor(prepost_Data$violations, prepost_Data$fines, use="complete.obs")
```

```
## [1] 0.999899
```

Aprat from looking at the violations, we notice the corruption index and trade may have some negative relationship. What captures our eyes is that corruption index seems to have strong negative relationship with gdppcus1998. Also, trade and gdppcus1998 have some postive relationship.

```
cor(prepost_Data$corruption, prepost_Data$trade, use="complete.obs")
```

```
## [1] -0.3389331
```

```
cor(prepost_Data$corruption, prepost_Data$gdppcus1998, use="complete.obs")
```

```
## [1] -0.8663537
```

```
cor(prepost_Data$trade, prepost_Data$gdppcus1998, use="complete.obs")
```

```
## [1] 0.4100351
```

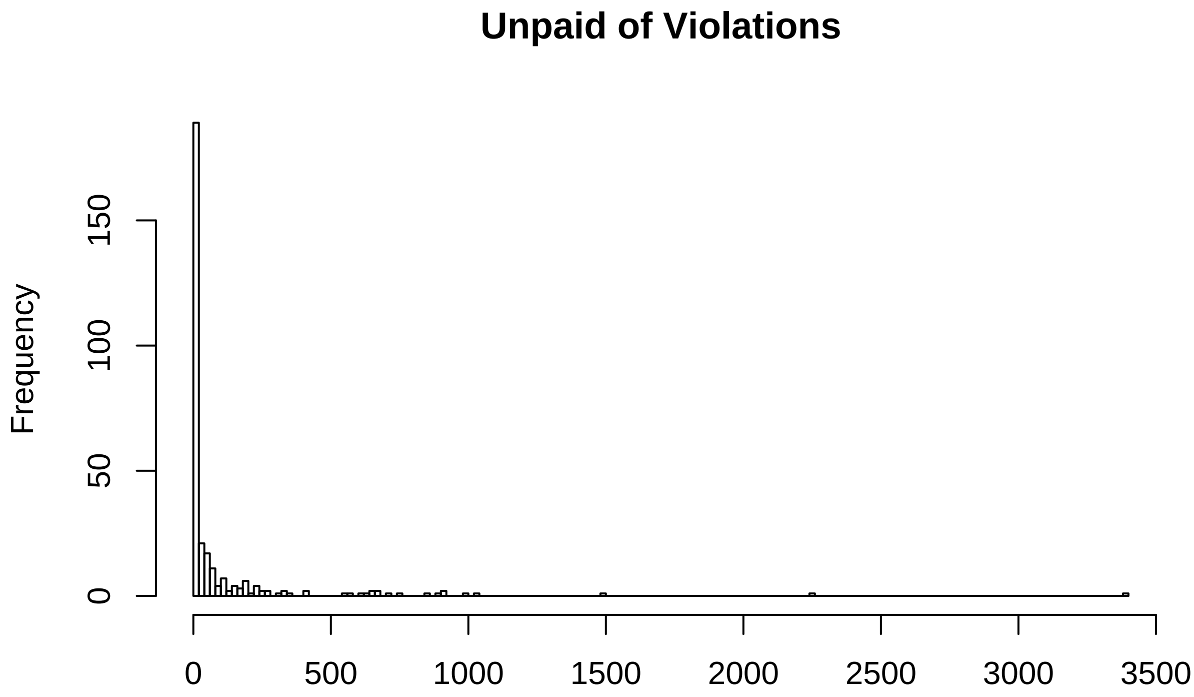
Overall, the plot suggests that violations and corruption may not really related. The fines variable is what we can dig further to see if it affects the bivariate relationships.

Since our outcome variable is parking violations (violations). we summarize and create a histogram.

```
summary(prepost_Data$violations)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
##    0.000    0.654    5.724  100.879   51.915 3392.961        4
```

```
hist(prepost_Data$violations, breaks="FD", main="Unpaid of Violations", xlab=NULL)
```



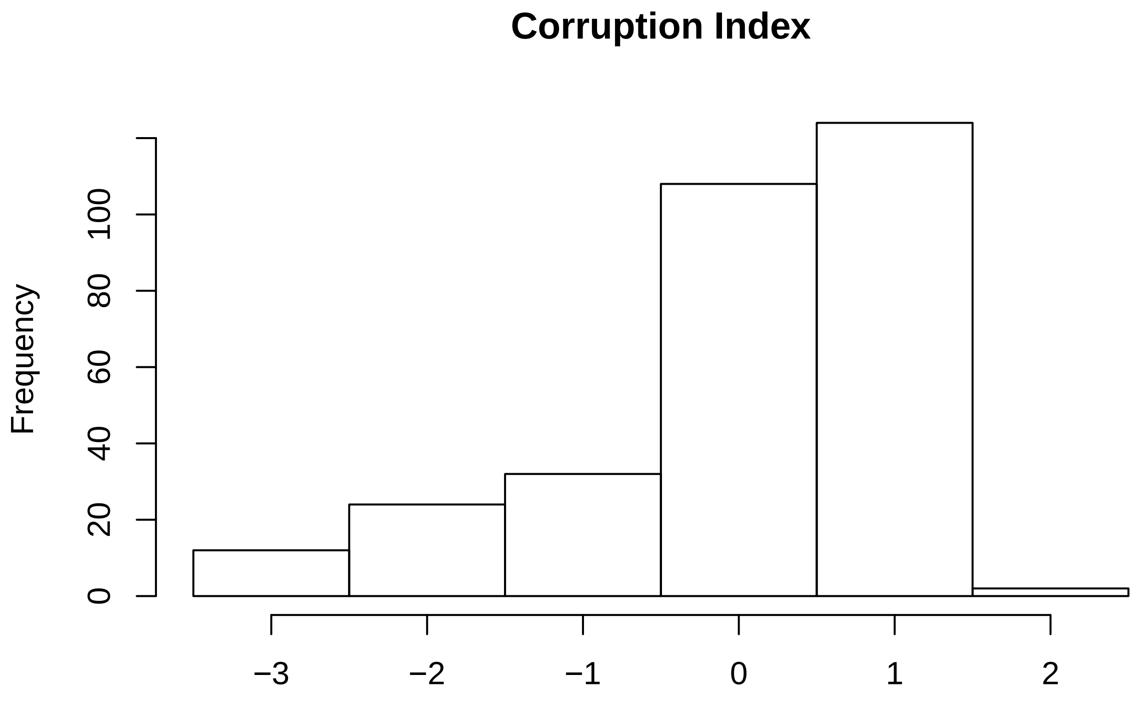
Visually, the histogram shows to have a postive skew. The vast majority of the data is less than $500 unpaid.

Next, we check out the corrunption variable. This is our main input variable.

```
summary(prepost_Data$corruption)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max. 
## -2.582988 -0.451213  0.322920 -0.007721  0.717707  1.582807
```

```
hist(prepost_Data$corruption, breaks=-4:2+0.5, main="Corruption Index", xlab=NULL)
```



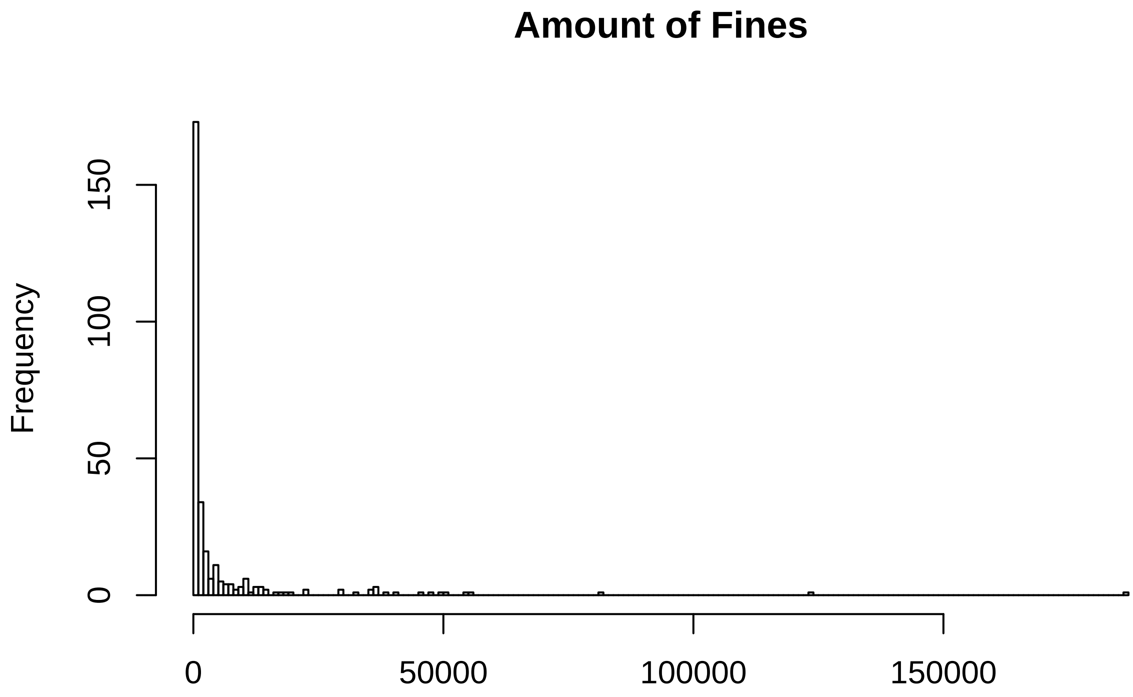
## First of all, the corruption index is a floating numbers between -3 and 2. The distribution seems more dispersed and the distribution of the data seems to have a negative skew.

Now we examine the fines variable.

```
summary(prepost_Data$fines)
```

```
##      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.       NA's 
##      0.00      65.41     579.72    5579.60    2999.05  186163.17          4
```

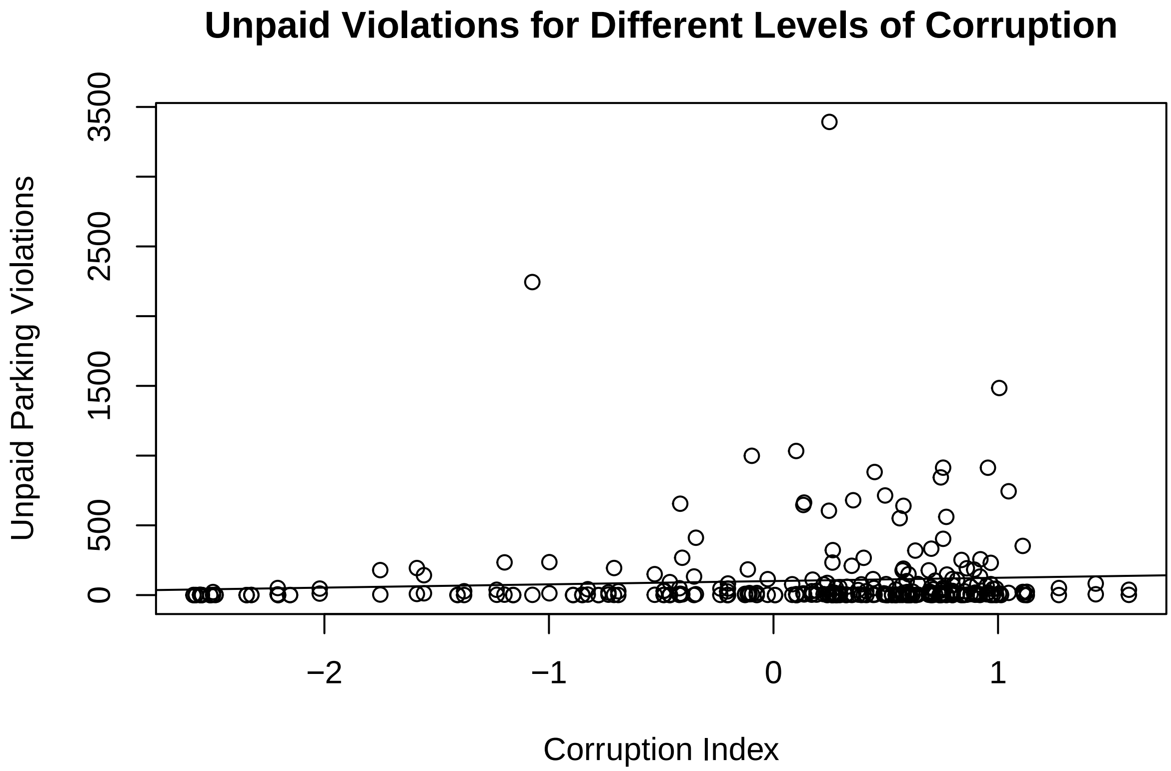
```
hist(prepost_Data$fines, breaks = "FD", main="Amount of Fines", xlab=NULL)
```



## Note that the distribution seems to have similar shape of violations, also the positive skew and may have closely related to violations variable.

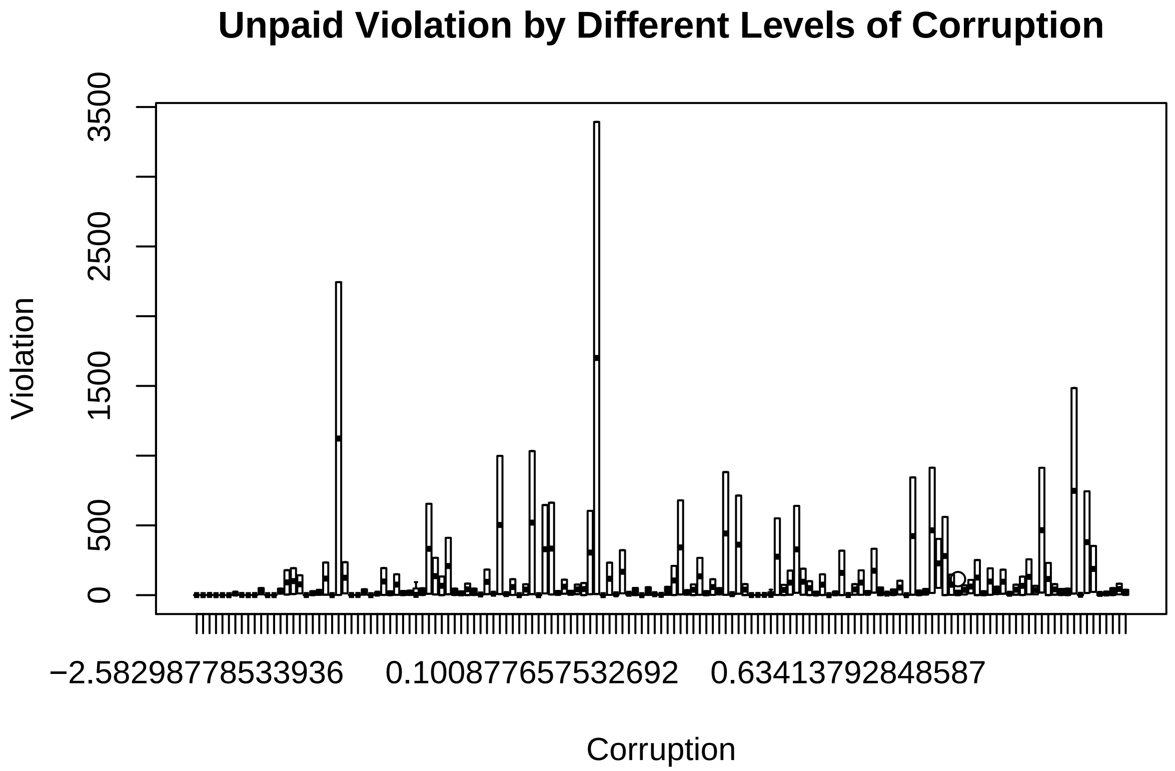
We want to understand what bivariate relationship exists between our main variable of interest, violations and corruption. We begin with a scatterplot, adding jitter to make sure points don't overlap. Also, we add OLS line, assuming we look to see some linear relationships.

```
plot(jitter(prepost_Data$corruption, factor=2),jitter(prepost_Data$violations,factor=2),
     xlab = "Corruption Index", ylab = "Unpaid Parking Violations",
     main = "Unpaid Violations for Different Levels of Corruption")

abline(lm(prepost_Data$violations ~ prepost_Data$corruption))
```



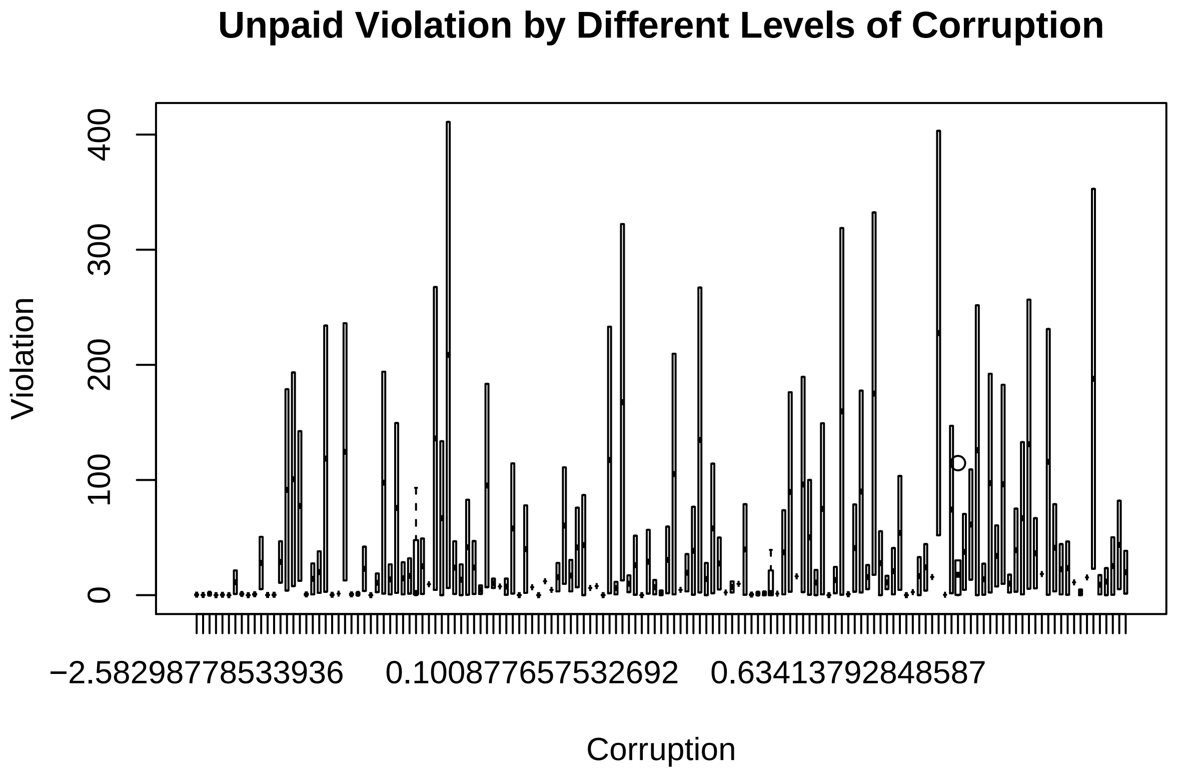
## This plot tells us that there is little or no linear relationship bettween violations and corruption variables. Earlier we also know the correlation between two is 0.078, which does not have much magnitude to show there is a linear relationship.

```
boxplot(violations ~ corruption, data = prepost_Data,
        main = "Unpaid Violation by Different Levels of Corruption",
        xlab = "Corruption", ylab = "Violation")
```



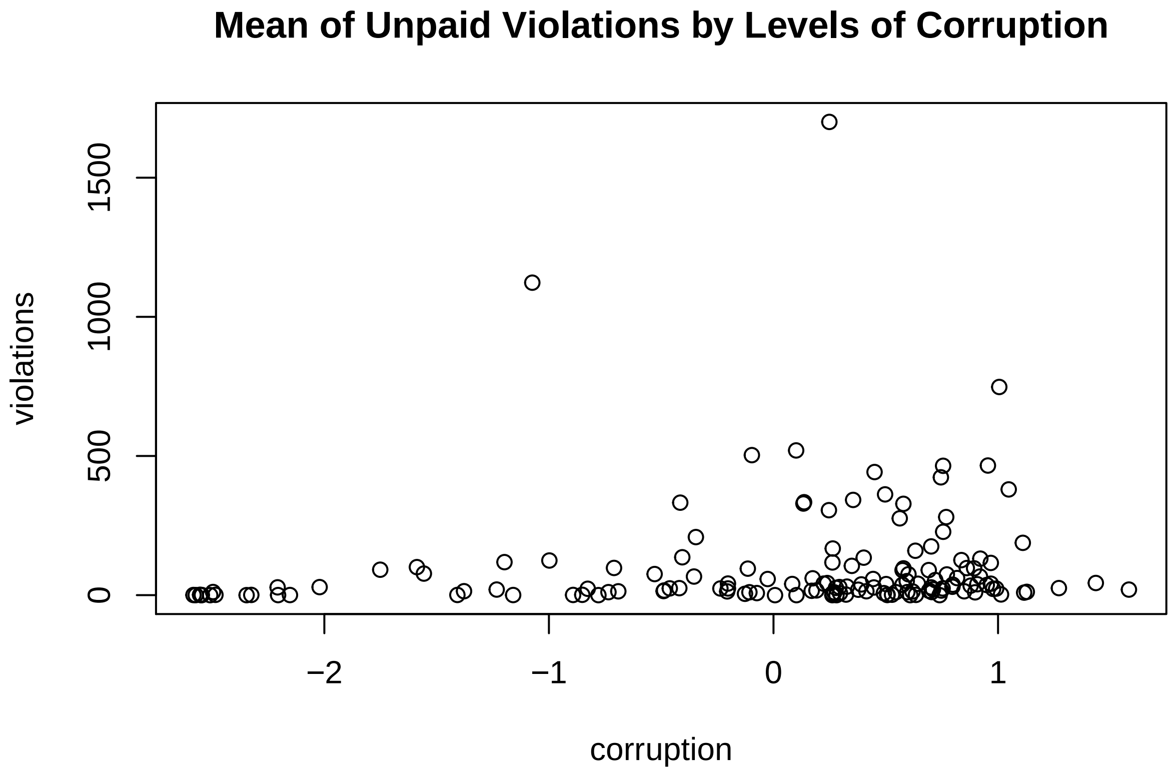
## The relationship does not appear linear and too much noises. We noticed that the majority of violations is under 500. We want to see what happen if only looking at violations under 500.

```
Sub2<-subset(prepost_Data, prepost_Data$violations<=500)

boxplot(violations ~ corruption, data = Sub2,
        main = "Unpaid Violation by Different Levels of Corruption",
        varwidth=TRUE,
        xlab = "Corruption", ylab = "Violation")
```



## If we only look at violations under 500, the results still look little or no linear relationship.

```
violations_mean<-by(prepost_Data$violations, prepost_Data$corruption, mean, na.rm=T)

plot(sort(unique(prepost_Data$corruption)), violations_mean,
     xlab = "corruption", ylab = "violations",
     main = "Mean of Unpaid Violations by Levels of Corruption")
```

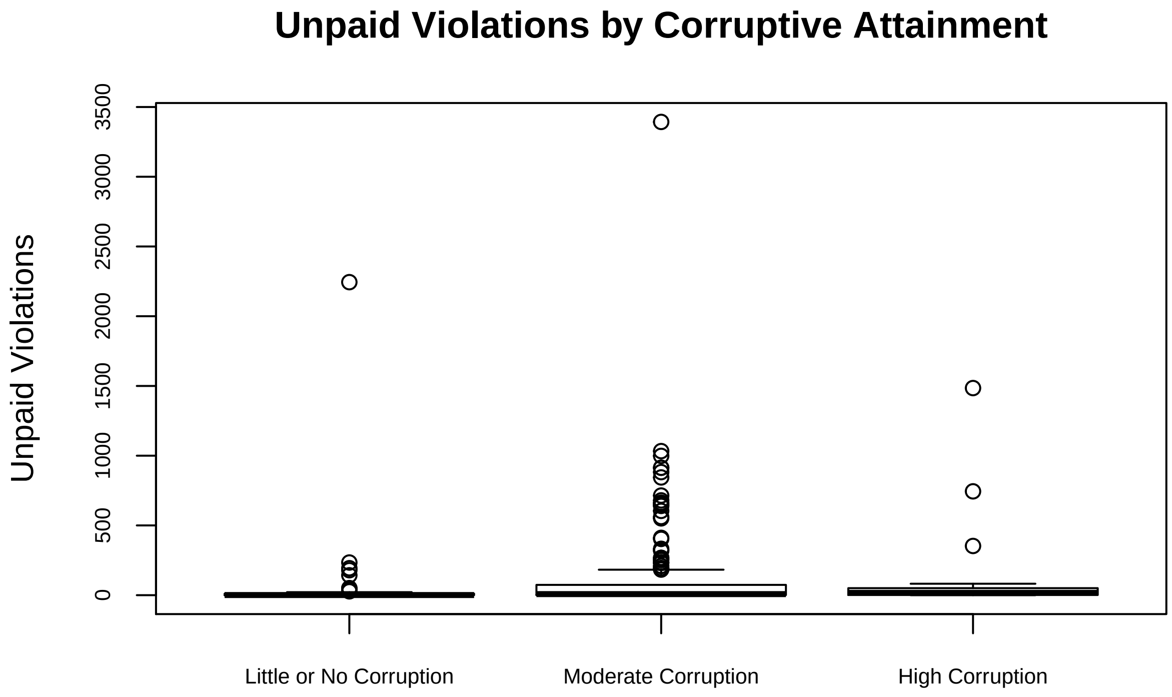


## We plot the mean of violations for each levels of corruption for better assess the relationship. ## Violation above around 500 might be deemed as outliers. Also, the corruption index between -1 and 1 contain more unpaid violations. ## To focus our attention on levels of corruption, we might speculate and want to bin our corruption variable into intervals.

```
corruption_bin = cut(prepost_Data$corruption, breaks = c(-3,-1,1, Inf),
                     labels = c("Little or No Corruption", "Moderate Corruption", "High Corruption"))
summary(corruption_bin)
```

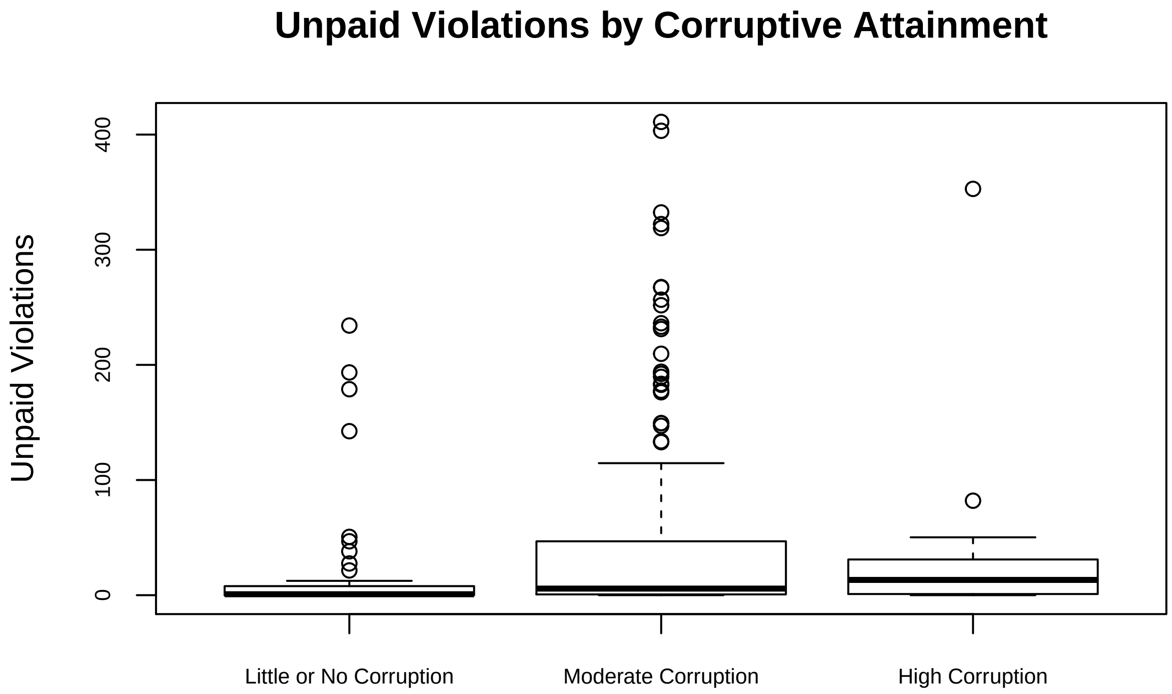
```
## Little or No Corruption     Moderate Corruption         High Corruption
##                      50                     234                      18
```

```
boxplot(violations ~ corruption_bin, data = prepost_Data, cex.axis = .7,
        main = "Unpaid Violations by Corruptive Attainment", ylab = "Unpaid Violations")
```



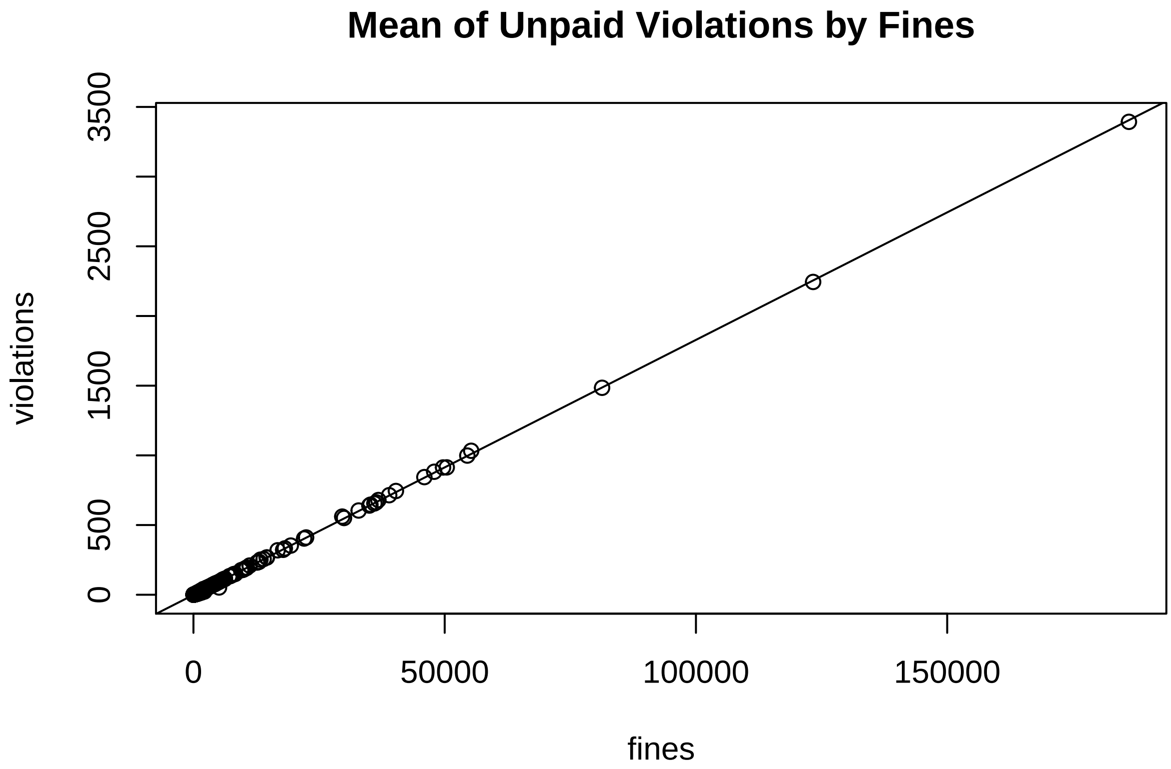
## The resulting boxplot shows the unpaid violations for each group. It could tells us different story. If we only look at unpaid violations under 500, the results might gives us more granular level of detailed relationship.

```
### if we only look at unpaid violations under 500

corruption_bin = cut(Sub2$corruption, breaks = c(-3,-1,1, Inf),
                     labels = c("Little or No Corruption", "Moderate Corruption", "High Corruption"))

boxplot(violations ~ corruption_bin, data = Sub2, cex.axis = .7,
        main = "Unpaid Violations by Corruptive Attainment", ylab = "Unpaid Violations")
```



Finally, we want to examine how our fines variable relates to violations and corruption. This will help us understand if fines would have confounding effects of our study.

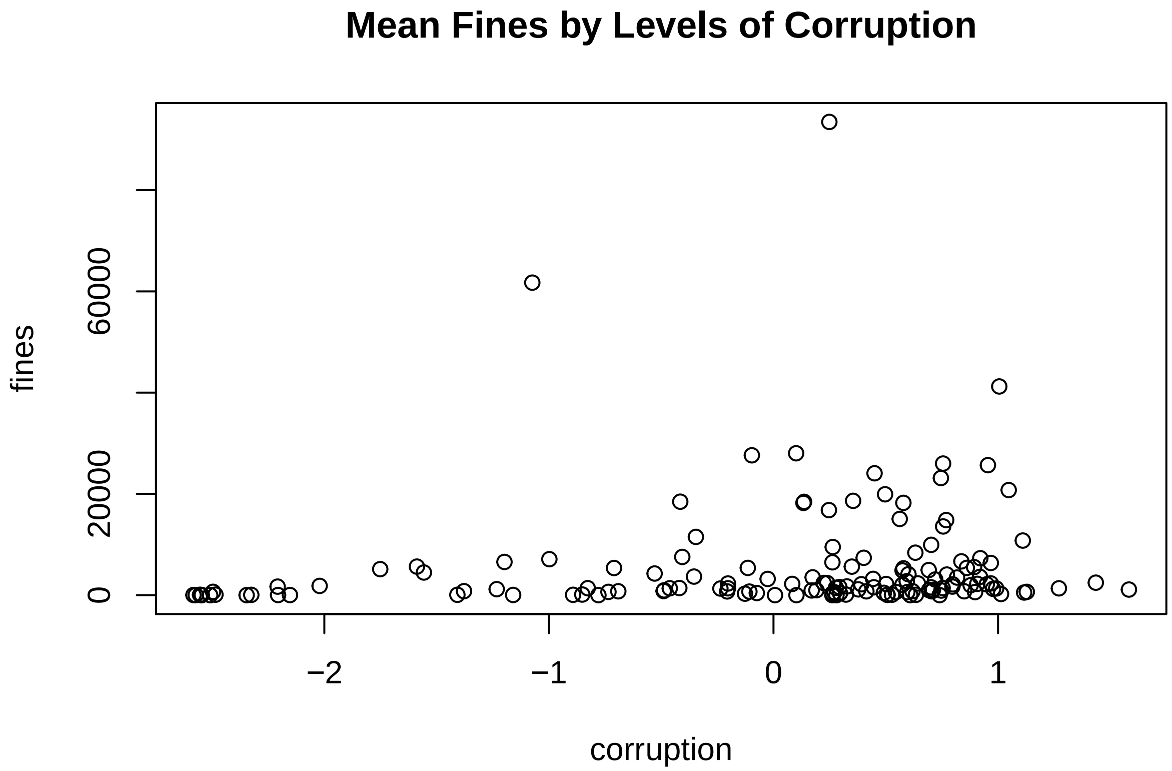
```
fines_mean<-by(prepost_Data$violations, prepost_Data$fines, mean, na.rm=T)

plot(sort(unique(prepost_Data$fines)), fines_mean,
     xlab = "fines", ylab = "violations",
     main = "Mean of Unpaid Violations by Fines")
abline(lm(prepost_Data$violations ~ prepost_Data$fines))
```



## We observe there is a strong postive relationship bewtween fines and violations. We can see much clear If we draw the possible OLS line. The results could suggest us that fine varibale would be a driver of unpaid violations.

Now, we also want to look at the relationship between fines and corruption.

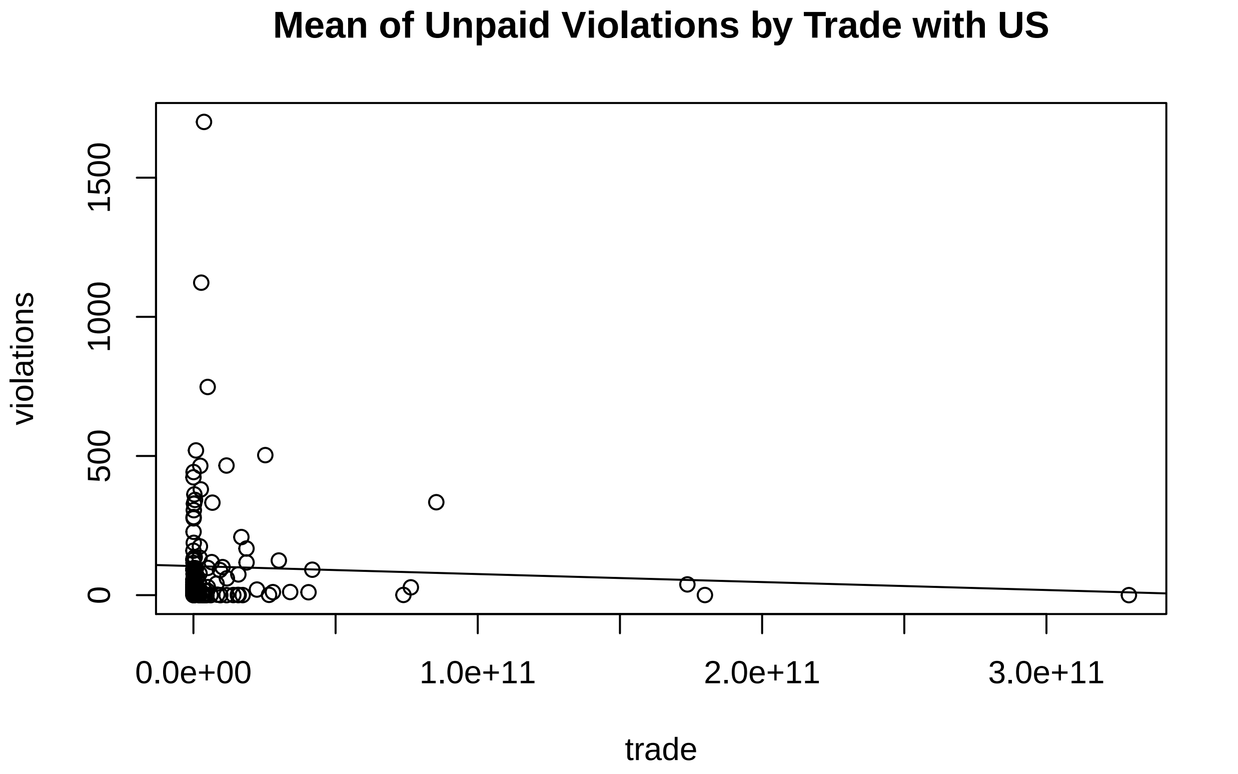
```
fines_mean2<-by(prepost_Data$fines, prepost_Data$corruption, mean, na.rm=T)

plot(sort(unique(prepost_Data$corruption)), fines_mean2,
     xlab = "corruption", ylab = "fines",
     main = "Mean Fines by Levels of Corruption")
```



## The results look much like the relationship between violations and corruption. not much linearly correlated.

similary, if we look at trade data compared to violations

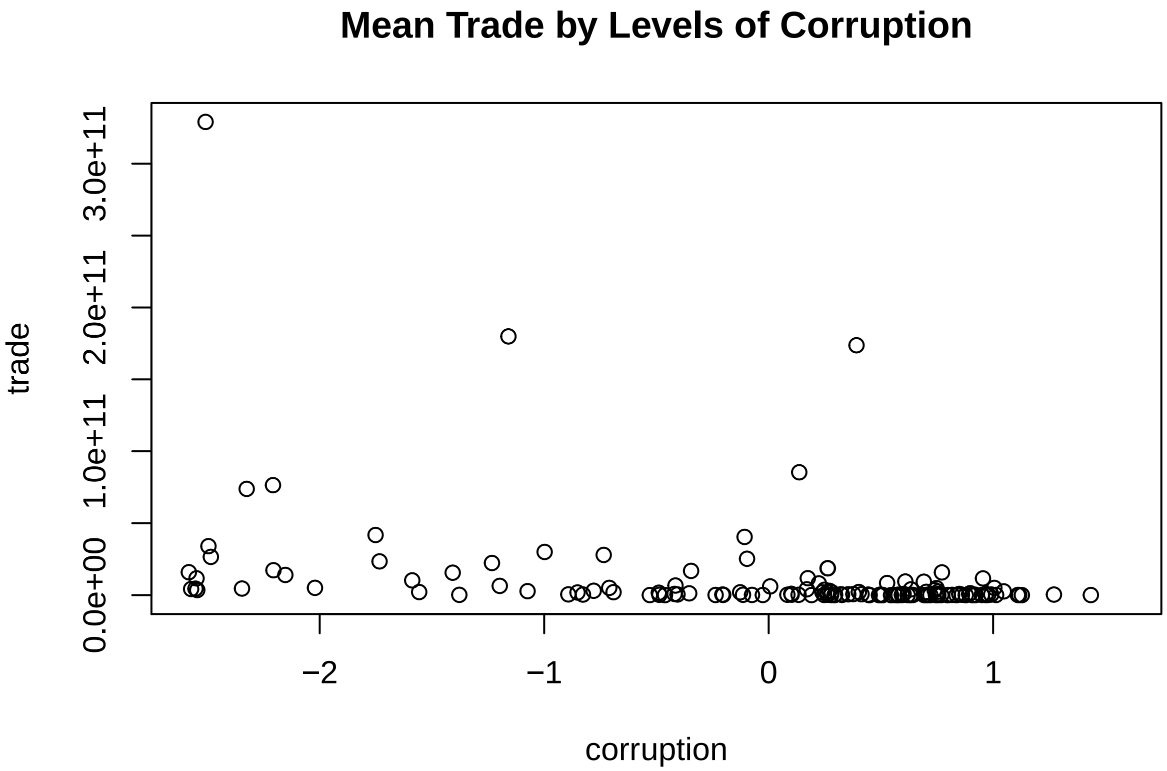
```
trade_mean<-by(prepost_Data$violations, prepost_Data$trade, mean, na.rm=T)

plot(sort(unique(prepost_Data$trade)), trade_mean,
     xlab = "trade", ylab = "violations",
     main = "Mean of Unpaid Violations by Trade with US")
abline(lm(prepost_Data$violations ~ prepost_Data$trade))
```



## Trade does not appear a clear relationship with violations. Perhaps, what we can see now is that lower trade with the US result a wide range of violations, and higher trade with the US only have fewer violations and are under 500 unpaid.

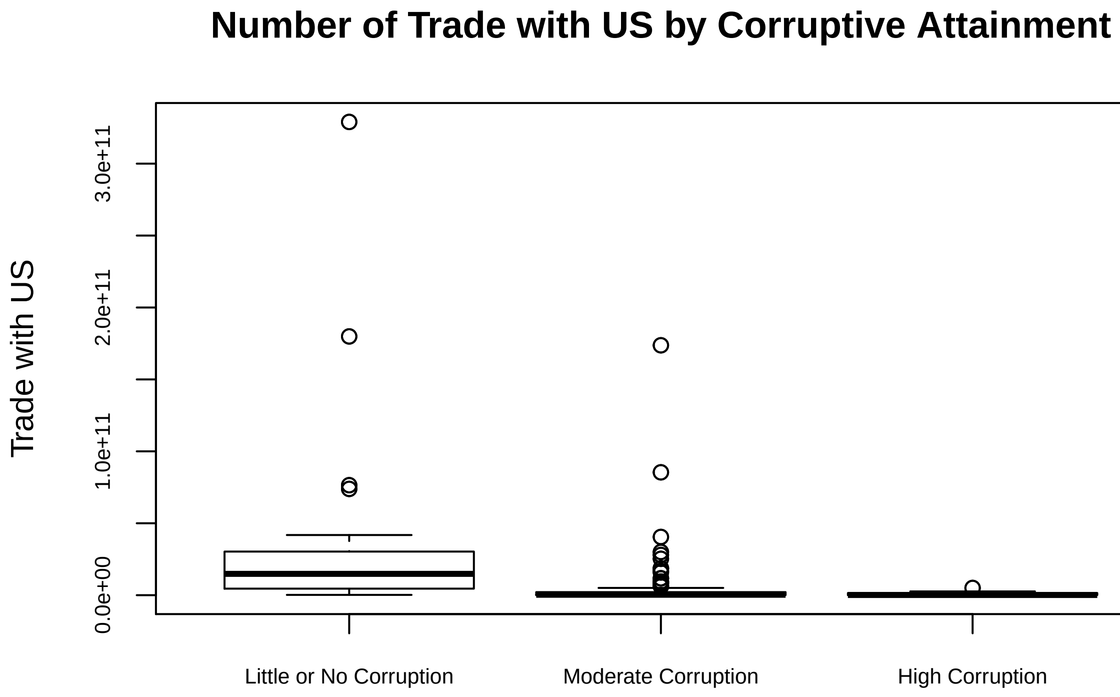
Then, we look at the trade v.s. corruption.

```
trade_mean2<-by(prepost_Data$trade, prepost_Data$corruption, mean, na.rm=T)

plot(sort(unique(prepost_Data$corruption)), trade_mean2,
     xlab = "corruption", ylab = "trade",
     main = "Mean Trade by Levels of Corruption")
```



## Similary, in the grpah, we don't see a very clear relationship between trade and corruption. However, we could see most of the lower trade fall between the corruption index -1 and 1.

```
corruption_bin = cut(prepost_Data$corruption, breaks = c(-3,-1,1, Inf),
                     labels = c("Little or No Corruption", "Moderate Corruption", "High Corruption"))

boxplot(trade ~ corruption_bin, data = prepost_Data, cex.axis = .7,
        main = "Number of Trade with US by Corruptive Attainment", ylab = "Trade with US")
```



## Binning into three corruption intervals could help us understand more detailed.

Conclusion

Having an overview of the dependent variables (violations) and the given questions (any relationship between corruption and parking violations), we do not see a clear linear relationship between each varibales. Perhaps, the results tell us there could be a different relationship other than linear between vioaltions and corruption.

Besides, we found the distribution of violations are skewed to the right and distribution of corruption are skewed to the left. Also, we know the vast majority of the unpaid violations happened under around 500. Unfortunately, when we look into the subset of the data, there still little or no relationship.

what's interesting is that we found fines have much stronger relationship with parking violations. This could suggest us that the fines varible could be instead the driver of violations. This can be our further studies.

In addition, we also examine the trade variable. Again, trade varible does not seem to have much relationship with violations either. However, there might be some interesting interrelations with corruption index if we look at the trade levels that falls between our binning intervals of corruption. In the future, we may be able to model from there.