W203 Lab 1: Exploratory Data Analysis of Corruption and Parking Violations

Chi Iong Ansjory, Tsung-Chin Han, Marcelo Queiroz

5/25/2018

Introduction

This analysis is motivated by the following question:

How is the amount of parking violations received by a UN diplomat related to the country corruption index?

Imagine that we 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 ocials in Manhattan. Until 2002, diplomatic immunity protected UN diplomats from parking enforcement actions, so diplomats actions were constrained by cultural norms alone. In 2002, enforcement authorities acquired the right to conscate diplomatic license plates of violators, after which diplomatic behavior was constrained by both cultural norms and the legal penalties of unpaid tickets. 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.

Cultural norms are unwritten yet well accepted behaviors that are established through years of society development. Different historic backgrounds and geographical particularities drive different set of standard behaviors and what is considered normal in one region, may be seen as unacceptable in a different area. However, when different cultures are invited to interact, we may have a rare opportunity to analyze the differences in a normalized environment. One good example of these opportunities happens at the United Nations (UN) office in New York, NY, USA, where diplomats from all major countries in the world interact not only on state interests, but also with more mundane tasks like having lunch, informal talks and even finding parking spots in Manhattan area. Maybe those day-to-day activities are even a better cultural indicator, since they have no bounds driven by commercial and/or diplomatic interests of one's country. Encouraged by the World Bank, we analyzed data from parking violations made by UN diplomats in Manhattan area and compare two scenarios: pre and post 2002, when local enforcement authorities acquired the right to confiscate diplomatic license plates, forcing the diplomats to deal with not only the cultural norms they were used before, but also with legal penalties of unpaid tickets.

Setup

Set up the working directory by putting data file and Rmd file in the same directory.

Load all necessary libraries for the R functions.

```
library(car)
```

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

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

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(scales)
library("ggplot2")
library("cowplot")
```

```
##
## Attaching package: 'cowplot'
```

```
## The following object is masked from 'package:ggplot2':
##
##     ggsave
```

Load the data set into R.

```
load("Corrupt.RData")
```

Data Validation and Selection

The raw dataset consists of 'r nrow(FMcorrupt)' observations of 'r ncol(FMcorrupt)' variables that shows economic, cultural and geographical aspects of countries that attended UN meetings pre and post the 2002 law change. A glimpse of that data is shown below:

```
str(FMcorrupt)
```

```
## '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 ...
```

For a better understanding of this data set, we provide a variable description bellow:

Variable Name	Description
wbcode	Country 3 letter code
prepost	**Pre:** data pre-2002, **Pos:** data post-2002
violations	Unpaid New York City parking violations
fines	Unpaid fines due to violations (USD)
mission	Number of missions in NY
staff	Number of staff people in NY
spouse	Number of diplomat family people in NY
gov_wage_gdp	Not specified in enunciate.
pctmuslim	Percent of Muslims in population
majoritymuslim	Is a country with majority Muslim? (Dummy)
trade	total trade with the US in 1998
cars_total	total number of diplomatic licensed cars in NY
cars_personal	Number of personal-use cars with diplomatic plates in NY
cars_mission	Number of professional-use cars with diplomatic plates in NY
pop1998	Country's population in 1998
gdppcus1998	Not specified in enunciate.
ecaid	Not specified in enunciate.
milaid	Not specified in enunciate.
region	Region code variable. To be studied.
corruption	Country Corruption index
totaid	Not specified in enunciate.
r_africa	Country belongs to Africa? (Dummy)
r_middleeast	Country belongs to Middle East? (Dummy)
r_europe	Country belongs to Europe? (Dummy)
r_southamerica	Country belongs to South America? (Dummy)
r_asia	Country belongs to Asia? (Dummy)
country	Country's name
distUNplz	Not specified in enunciate.

However, provided that we want to answer the question **"How is the amount of parking violations received by a UN diplomat related to the country corruption index?"**, some transformation were made to the data to best fit our research demands:

- First we dropped the observations where violations and/or fines data were not available
- Second we used the columns related to the geographical position to populate a new column, containing the name of the region for ploting purposes.
- Africa is region 6:

```
subcase_africa = corrupt$r_africa == 1 & !is.na(corrupt$r_africa)
corrupt_africa = corrupt[subcase_africa, ]
head(subset(corrupt_africa, select = c ("wbcode", "region", "country")), 10)
```

```
##    wbcode region      country
## 2     AGO      6       ANGOLA
## 3     AGO      6       ANGOLA
```

```
## 19    BDI      6      BURUNDI
## 20    BDI      6      BURUNDI
## 23    BEN      6        BENIN
## 24    BEN      6        BENIN
## 25    BFA      6 BURKINA-FASO
## 26    BFA      6 BURKINA-FASO
## 46    BWA      6     BOTSWANA
## 47    BWA      6     BOTSWANA
```

- Asia is region 4:

```
subcase_asia = corrupt$r_asia == 1 & !is.na(corrupt$r_asia)
corrupt_asia = corrupt[subcase_asia, ]
head(subset(corrupt_asia, select = c ("wbcode", "region", "country")), 10)
```

```
##    wbcode region     country
## 10    ARM      4     ARMENIA
## 11    ARM      4     ARMENIA
## 17    AZE      4  AZERBAIJAN
## 18    AZE      4  AZERBAIJAN
## 27    BGD      4  BANGLADESH
## 28    BGD      4  BANGLADESH
## 44    BTN      4      BHUTAN
## 45    BTN      4      BHUTAN
## 56    CHN      4 CHINA (PRC)
## 57    CHN      4 CHINA (PRC)
```

- South America is region 2:

```
subcase_sa = corrupt$r_southamerica == 1 & !is.na(corrupt$r_southamerica)
corrupt_sa = corrupt[subcase_sa, ]
head(subset(corrupt_sa, select = c ("wbcode", "region", "country")), 10)
```

```
##    wbcode region   country
## 8     ARG      2 ARGENTINA
## 9     ARG      2 ARGENTINA
## 38    BOL      2   BOLIVIA
## 39    BOL      2   BOLIVIA
## 40    BRA      2    BRAZIL
## 41    BRA      2    BRAZIL
## 54    CHL      2     CHILE
## 55    CHL      2     CHILE
## 64    COL      2
## 65    COL      2
```

- Middle East is region 7:

```
subcase_me = corrupt$r_middleeast == 1 & !is.na(corrupt$r_middleeast)
subcase_me = corrupt[subcase_me, ]
head(subset(subcase_me, select = c ("wbcode", "region", "country")), 10)
```

```
##     wbcode region country
## 6      ARE      7
## 7      ARE      7
## 31     BHR      7 BAHRAIN
## 32     BHR      7 BAHRAIN
## 72     CYP      7  CYPRUS
## 73     CYP      7  CYPRUS
```

```
## 90     EGY      7   EGYPT
## 91     EGY      7   EGYPT
## 146    IRN      7    IRAN
## 147    IRN      7    IRAN
```

- Europe is region 3:

```
subcase_europe = corrupt$r_europe == 1 & !is.na(corrupt$r_europe)
corrupt_europe = corrupt[subcase_europe, ]
head(subset(corrupt_europe, select = c ("wbcode", "region", "country")), 10)
```

```
##    wbcode region            country
## 4     ALB      3            ALBANIA
## 5     ALB      3            ALBANIA
## 15    AUT      3            AUSTRIA
## 16    AUT      3            AUSTRIA
## 21    BEL      3            BELGIUM
## 22    BEL      3            BELGIUM
## 29    BGR      3           BULGARIA
## 30    BGR      3           BULGARIA
## 33    BIH      3 BOSNIA-HERZEGOVINA
## 34    BIH      3 BOSNIA-HERZEGOVINA
```

- North and Central America is region 1:

```
head(subset(corrupt[corrupt$region == 1,], select = c("wbcode", "region", "country")), 10)
```

```
##     wbcode region             country
## 50     CAN      1
## 51     CAN      1
## 82     DOM      1  DOMINICAN REPUBLIC
## 83     DOM      1  DOMINICAN REPUBLIC
## 136    HTI      1               HAITI
## 137    HTI      1               HAITI
## 154    JAM      1
## 155    JAM      1
## 313    TTO      1 TRINIDAD AND TOBAGO
## 314    TTO      1 TRINIDAD AND TOBAGO
```

- Oceania is region 5:

```
head(subset(corrupt[corrupt$region == 5,], select = c("wbcode", "region", "country")), 10)
```

```
##      wbcode region          country
## 13      AUS      5        AUSTRALIA
## 14      AUS      5        AUSTRALIA
## 103     FJI      5             FIJI
## 104     FJI      5             FIJI
## 243     NZL      5      NEW ZEALAND
## 244     NZL      5      NEW ZEALAND
## 256     PNG      5 PAPUA NEW GUINEA
## 257     PNG      5 PAPUA NEW GUINEA
## NA     <NA>     NA             <NA>
## NA.1   <NA>     NA             <NA>
```

```
subcase_oceania = corrupt$region == 5
corrupt_oceania = corrupt[subcase_oceania,]
```

- Creating the new variable:

```
reg_labels = c("North and Central America", "South America", "Europe",
               "Asia", "Oceania", "Africa", "Middle East")
reg_levels = c(1,2,3,4,5,6,7)
corrupt$region_label <- factor(corrupt$region, levels = reg_levels, labels = reg_labels)
```

- Third we dropped columns that will not be usefull for our analisys, to

```
new_corrupt <- subset( corrupt, select = -c(mission, gov_wage_gdp, pctmuslim,
                                            majoritymuslim, gdppcus1998, ecaid,
                                            milaid, cars_total, totaid, r_africa,
                                            r_middleeast, r_europe, r_southamerica,
                                            r_asia, distUNplz), drop = FALSE)
```

When performing a last check in the data we found that some observation still miss their region, trade and number of cars. We fixed the region variable, while ignoring the tradee and car count since they may not be essencial for our analysis.

```
summary(new_corrupt)
```

```
##     wbcode             prepost            violations
##  Length:298         Length:298         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
##
##      fines                staff            spouse            trade
##  Min.   :     0.00   Min.   : 2.00   Min.   : 0.000   Min.   :0.000e+00
##  1st Qu.:    65.41   1st Qu.: 6.00   1st Qu.: 3.000   1st Qu.:8.911e+07
##  Median :   579.72   Median : 9.00   Median : 6.000   Median :5.194e+08
##  Mean   :  5579.60   Mean   :11.81   Mean   : 7.758   Mean   :1.025e+10
##  3rd Qu.:  2999.05   3rd Qu.:14.00   3rd Qu.:10.000   3rd Qu.:4.796e+09
##  Max.   :186163.17   Max.   :86.00   Max.   :81.000   Max.   :3.290e+11
##                                                       NA's   :4
##  cars_personal     cars_mission        pop1998             region
##  Min.   : 0.000   Min.   :  0.000   Min.   :5.308e+05   Min.   :1.000
##  1st Qu.: 1.000   1st Qu.:  2.000   1st Qu.:3.815e+06   1st Qu.:3.000
##  Median : 2.000   Median :  3.000   Median :8.852e+06   Median :4.000
##  Mean   : 5.324   Mean   :  5.144   Mean   :3.655e+07   Mean   :4.372
##  3rd Qu.: 6.000   3rd Qu.:  6.000   3rd Qu.:2.341e+07   3rd Qu.:6.000
##  Max.   :64.000   Max.   :116.000   Max.   :1.242e+09   Max.   :7.000
##  NA's   :20       NA's   :20                            NA's   :2
##    corruption         country                region_label
##  Min.   :-2.58299   Length:298         Africa       :92
##  1st Qu.:-0.41515   Class :character   Europe       :70
##  Median : 0.32696   Mode  :character   Asia         :50
##  Mean   : 0.01364                      South America:36
##  3rd Qu.: 0.72025                      Middle East  :30
##  Max.   : 1.58281                      (Other)      :18
##                                        NA's         : 2
```

```
new_corrupt[is.na(new_corrupt$region_label),]
```

```
##     wbcode prepost violations     fines staff spouse trade cars_personal
```

```
## 345    ZAR     pre  38.485016 2085.2803     6      1    NA           NA
## 346    ZAR     pos   1.308244  117.7419     6      1    NA           NA
##     cars_mission  pop1998 region corruption country region_label
## 345           NA 47700500     NA   1.582807   ZAIRE         <NA>
## 346           NA 47700500     NA   1.582807   ZAIRE         <NA>
```

```
new_corrupt["345", "region"] <- 6
new_corrupt["346", "region"] <- 6
new_corrupt["345", "region_label"] <- "Africa"
new_corrupt["346", "region_label"] <- "Africa"
```

As a final sanity test, each country must have exactly 2 observations, one before and one after the 2002 law enforcement. With a 298 observations data set that also must have exactly 149 observations for before and after 2002:

```
max(table(new_corrupt$wbcode))
```

```
## [1] 2
```

```
min(table(new_corrupt$wbcode))
```

```
## [1] 2
```

```
table(new_corrupt$prepost)
```

```
##
## pos pre
## 149 149
```

Finally, we have our working data set, with 298 observations of 14 variables. This means we can analyze data of 149 countries for insights on how the traffic violations are correlated with a country's corruption index and how this is affected by the introduction of legal penalties.

Variable Name	Description
wbcode	Country 3 letter code
prepost	**Pre:** data pre-2002, **Pos:** data post-2002
violations	Unpaid New York City parking violations
fines	Unpaid fines due to violations (USD)
staff	Number of staff people in NY
spouse	Number of diplomat family people in NY
trade	Total trade with the US in 1998
cars_personal	Number of personal-use cars with diplomatic plates in NY
cars_mission	Number of professional-use cars with diplomatic plates in NY
pop1998	Country's population in 1998
region	Region code
corruption	Country Corruption index
country	Country's name
region_label	Region name

Exploratory Analysis

As stated before, in 2002 the New Youk authorities acquired the right to confiscate diplomatic license plates. Those plates are issued for diplomatic missions staff cars and for cars of personal use of the main diplomats and their families. Until 2002, therefore, the diplomats and their families where not required to pay for their traffic tickets, i.e., their only incentive for paying was the cultural norms of their society. However after

2002 an extra incentive was introduced, now a legal and economical one, which impacted the diplomatic community and their traffic habits.

Before the enforcement of paying tickets a diplomatic mission had a median of 51.65 unpaid violations:

```
summary(new_corrupt[new_corrupt$prepost == "pre",]$violations)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   17.22   51.65  198.07  189.59 3392.96
```

```
pre_corrupt <- new_corrupt[new_corrupt$prepost == "pre",]
```

After it this number dropped to 1.31:

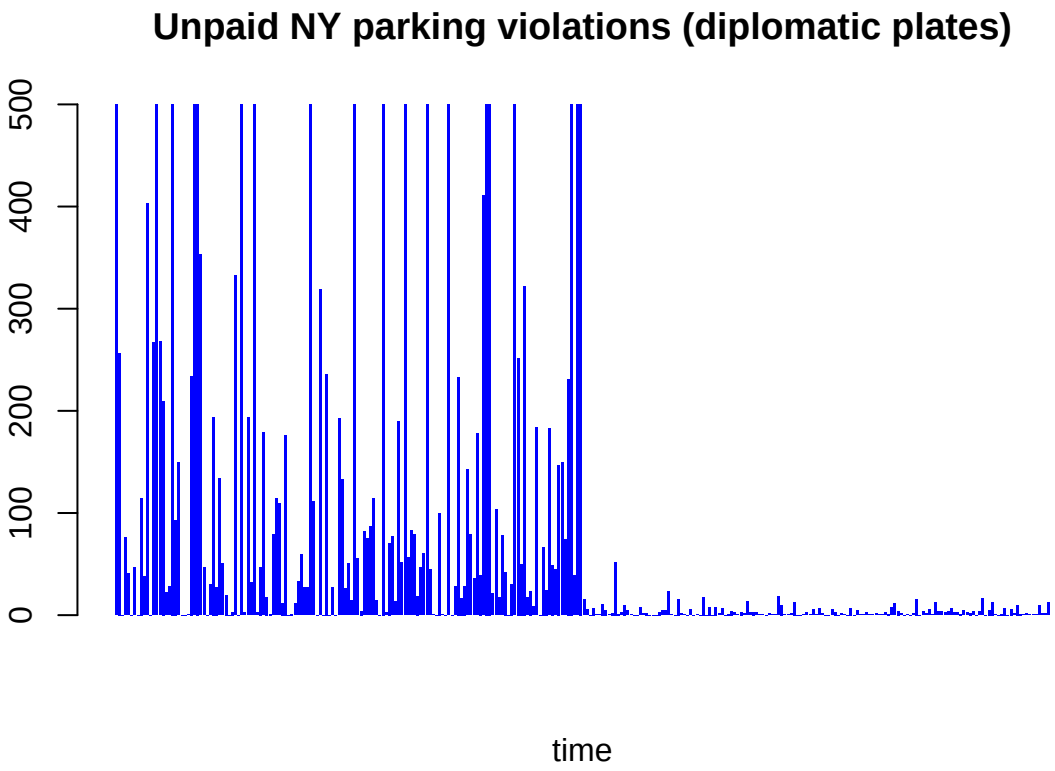
```
summary(new_corrupt[new_corrupt$prepost == "pos",]$violations)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.3271  1.3082  3.6877  4.5789 52.0027
```

```
pos_corrupt <- new_corrupt[new_corrupt$prepost == "pos",]
```

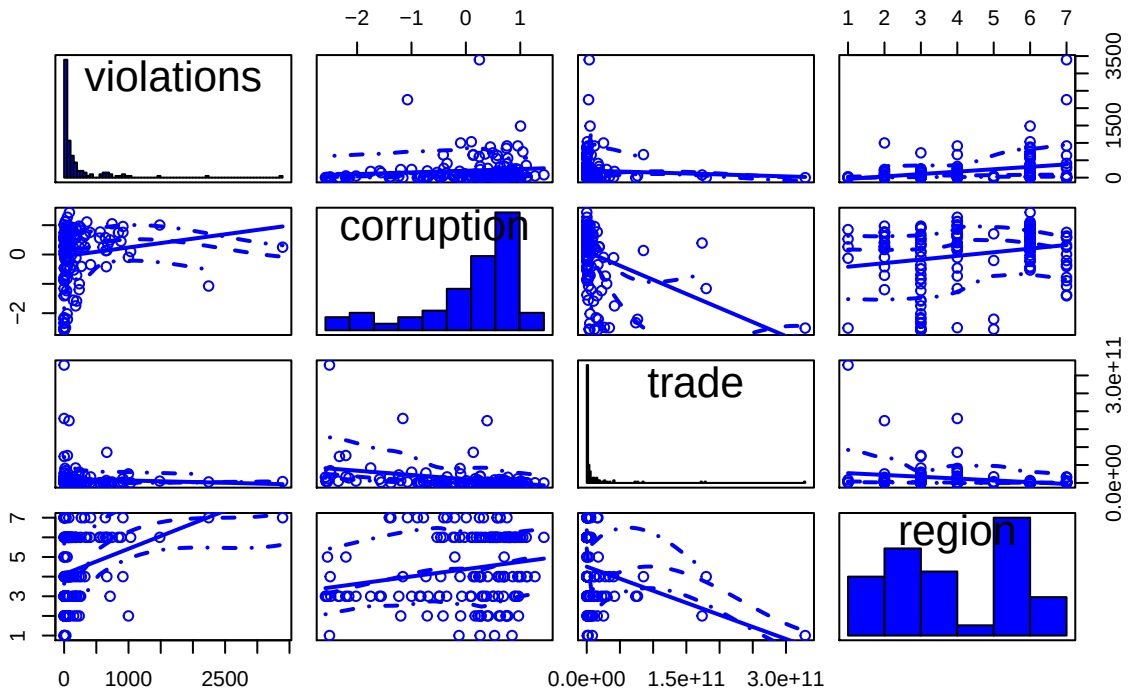
This big difference can be easier noted with a barplot:

```
pp_corrupt <- arrange(new_corrupt, desc(new_corrupt$prepos))
barplot(pp_corrupt$violations, col = "blue", ylim = c(0,500), main = "Unpaid NY parking violations (dip
```

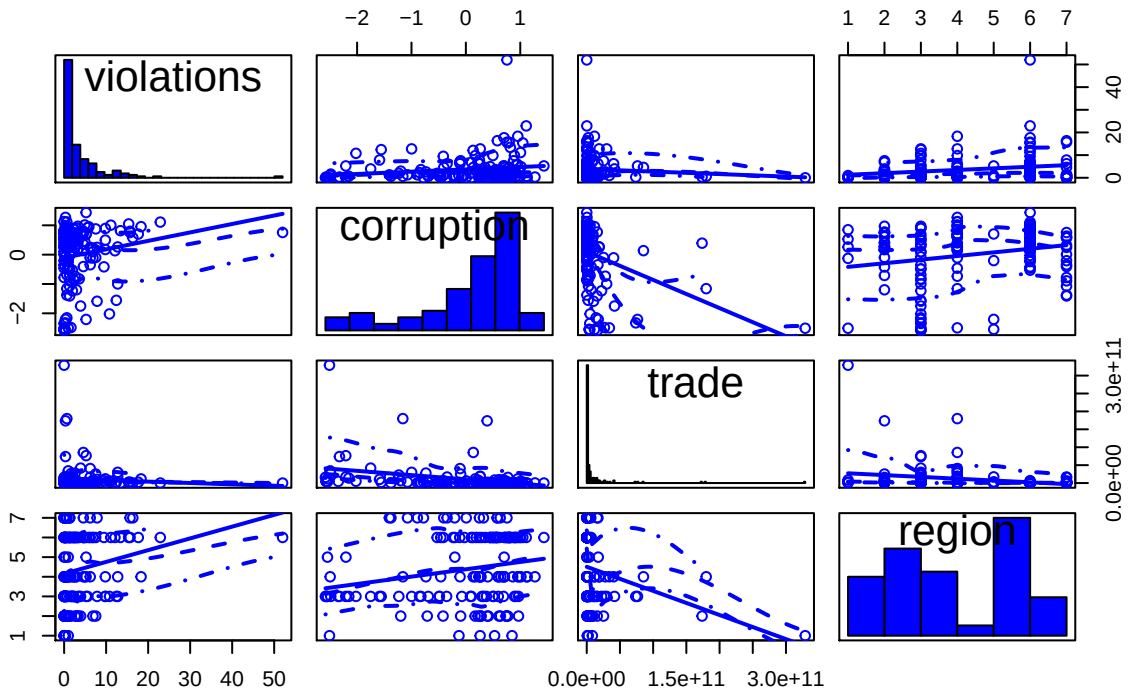


We begin the scatterplot matrix. We want to get a high level overview. The general pattern between pre-2002 and post-2002 are very similar:

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




```
scatterplotMatrix(~ violations + corruption + trade + region,
                  data = pos_corrupt,
                  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(new_corrupt$violations, new_corrupt$corruption, use="complete.obs")
```

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

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

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

Apart from looking at the violations, we notice the corruption index and trade may have some negative relationship.

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

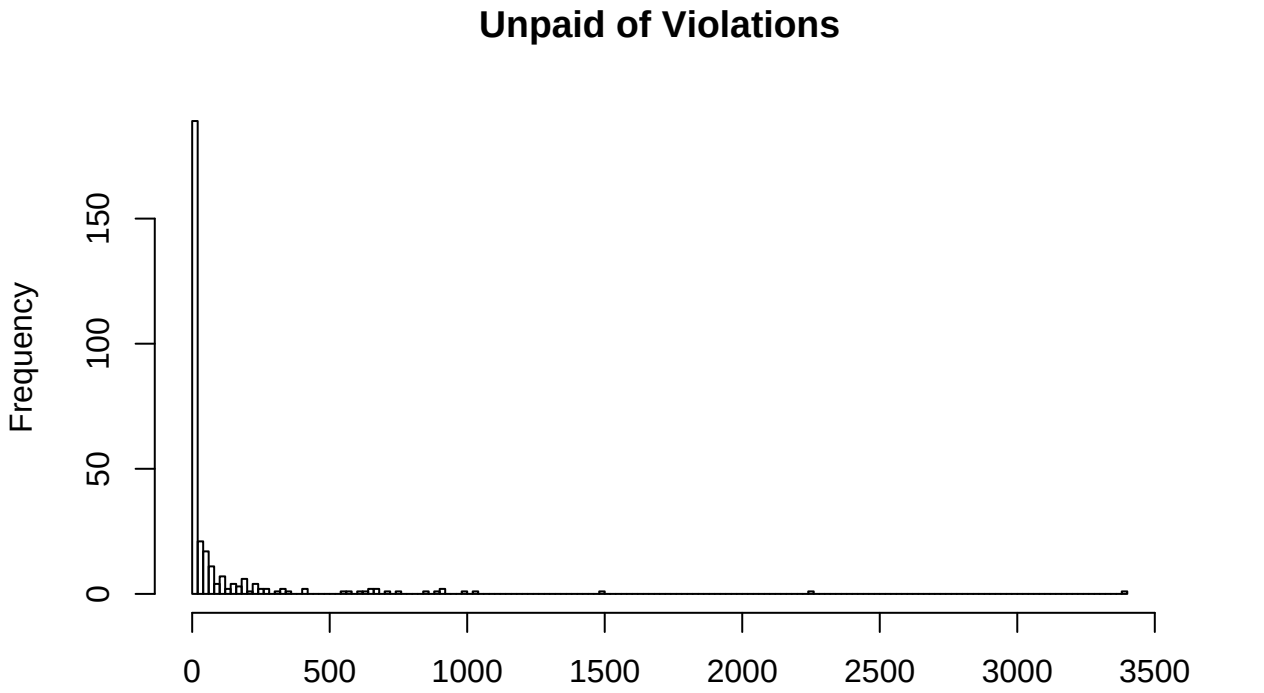
```
## [1] -0.3381395
```

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(new_corrupt$violations)
```

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

```
hist(new_corrupt$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.

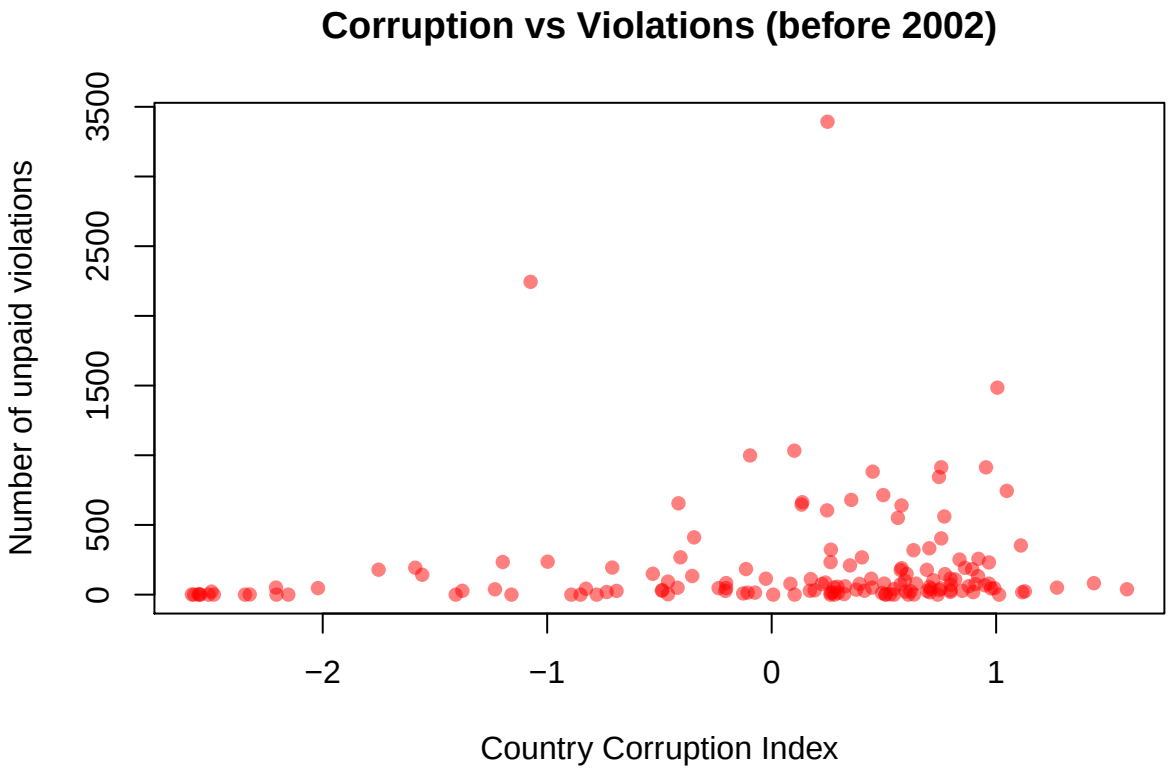
Parking violations and corruption

It's easy to see that after the law enforcement all diplomats started paying their tickets, with a few exception cases. Therefore, we want to base our analysis on the time pre-enforcement. The question we want to answer is:

"How is the amount of parking violations received by a UN diplomat related to the country corruption index?"

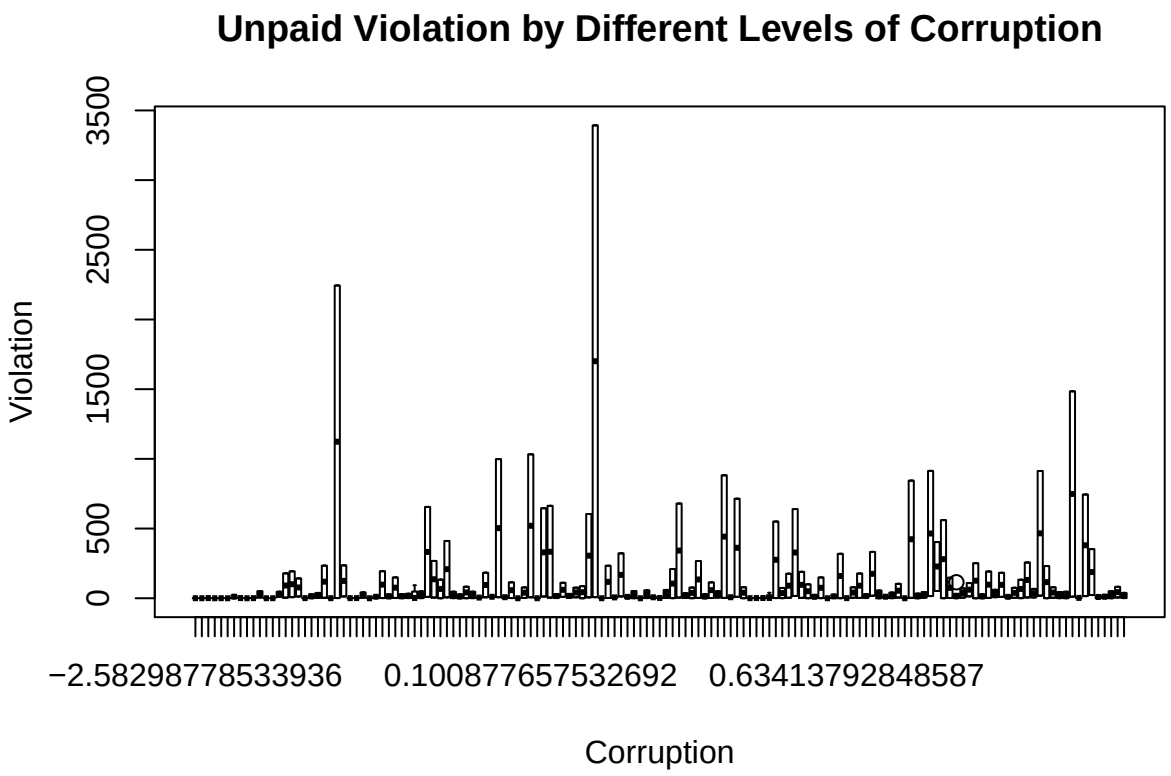
One of the available variables in our dataset was the corruption index for each country in 1998. We will consider that index is relatevelly stable through years and its analisys can be extended to the years after 2002. The corruption index is negative when a country have low corruption cases, and may be greater then 0 if a particular country have more corruption cases reported. To correlate that index with parking violations the first plot made was a scatterplot relating number of violations and corruption index:

```
plot(pre_corrupt$corruption, pre_corrupt$violations,
     main = "Corruption vs Violations (before 2002)", xlab = "Country Corruption Index",
     ylab = "Number of unpaid violations", col = alpha("red",0.5), pch = 16, alpha = 50)
```



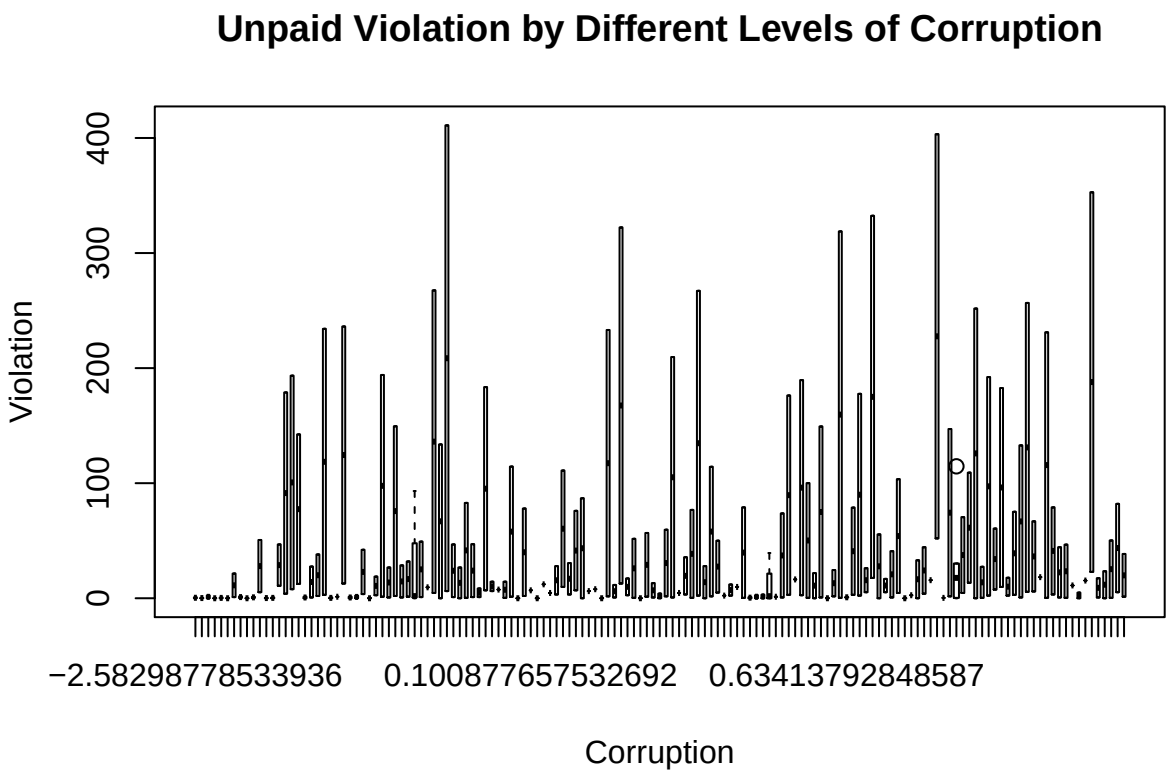
This plot tells us that there is little or no linear relationship bettween violations and corruption variables. Earlier we also know the correlation between two is 0.078, which does not have much magnitude to show there is a linear relationship.

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



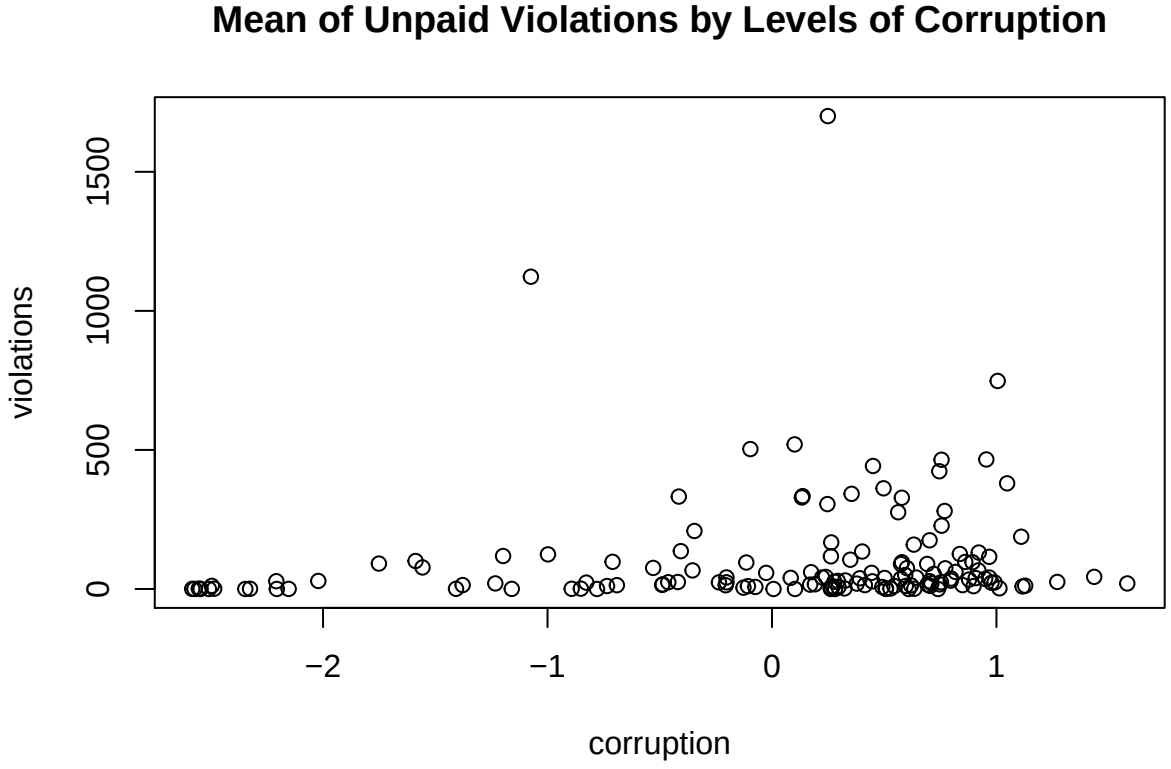
The relationship does not appear linear and too much noises. We noticed that the majority of violations is under 500. We want to see what happen if only looking at violations under 500.

```
Sub2<-subset(new_corrupt, new_corrupt$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(new_corrupt$violations, new_corrupt$corruption, mean, na.rm=T)
plot(sort(unique(new_corrupt$corruption)), violations_mean, xlab = "corruption", ylab = "violations",
main = "Mean of Unpaid Violations by Levels of Corruption")
```

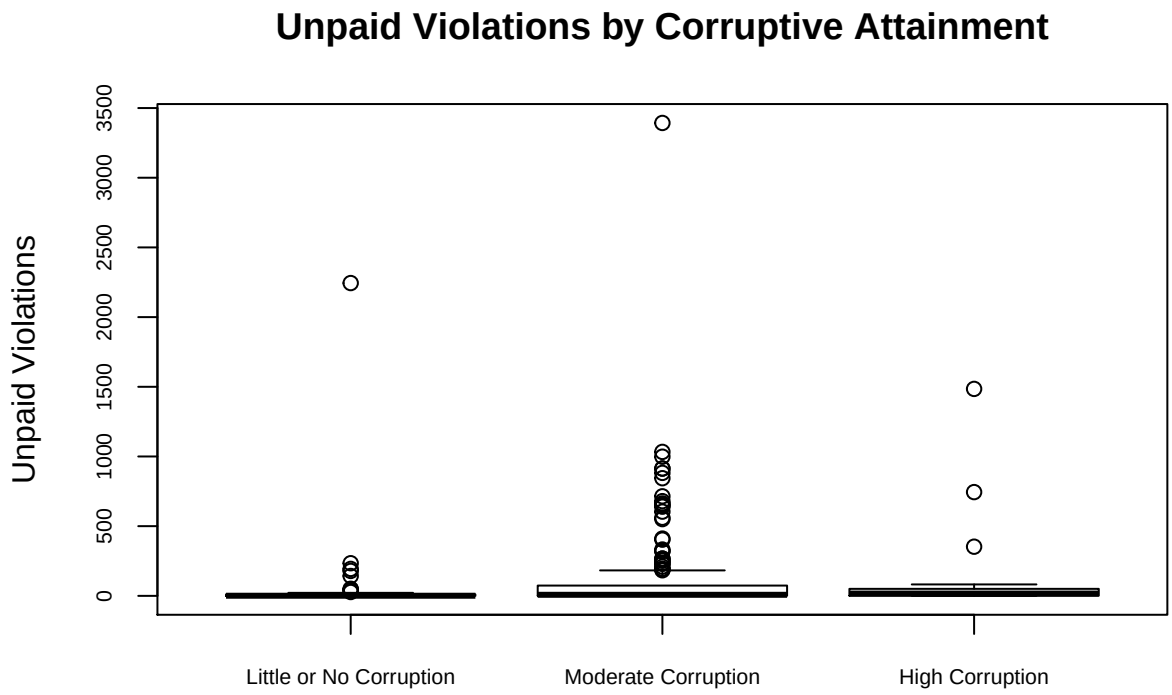


We plot the mean of violations for each levels of corruption for better assess the relationship. Violation above around 500 might be deemed as outliers. Also, the corruption index between -1 and 1 contain more unpaid violations. To focus our attention on levels of corruption, we might speculate and want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

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

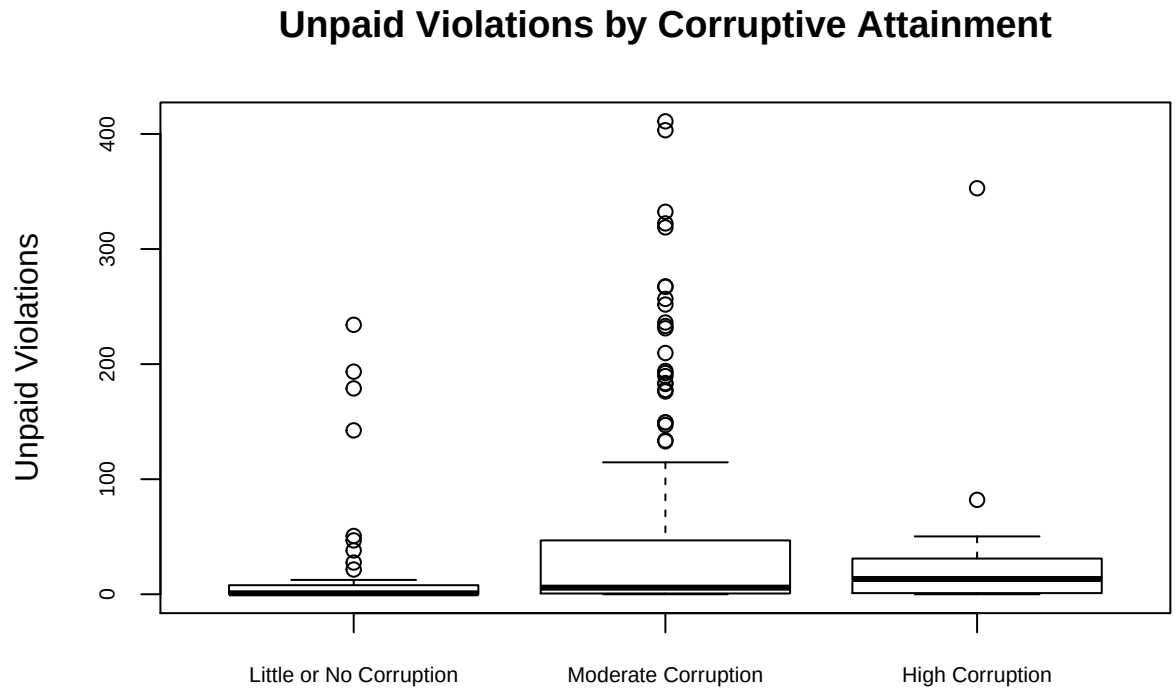
```
## Little or No Corruption     Moderate Corruption         High Corruption
##                      46                     234                      18
```

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



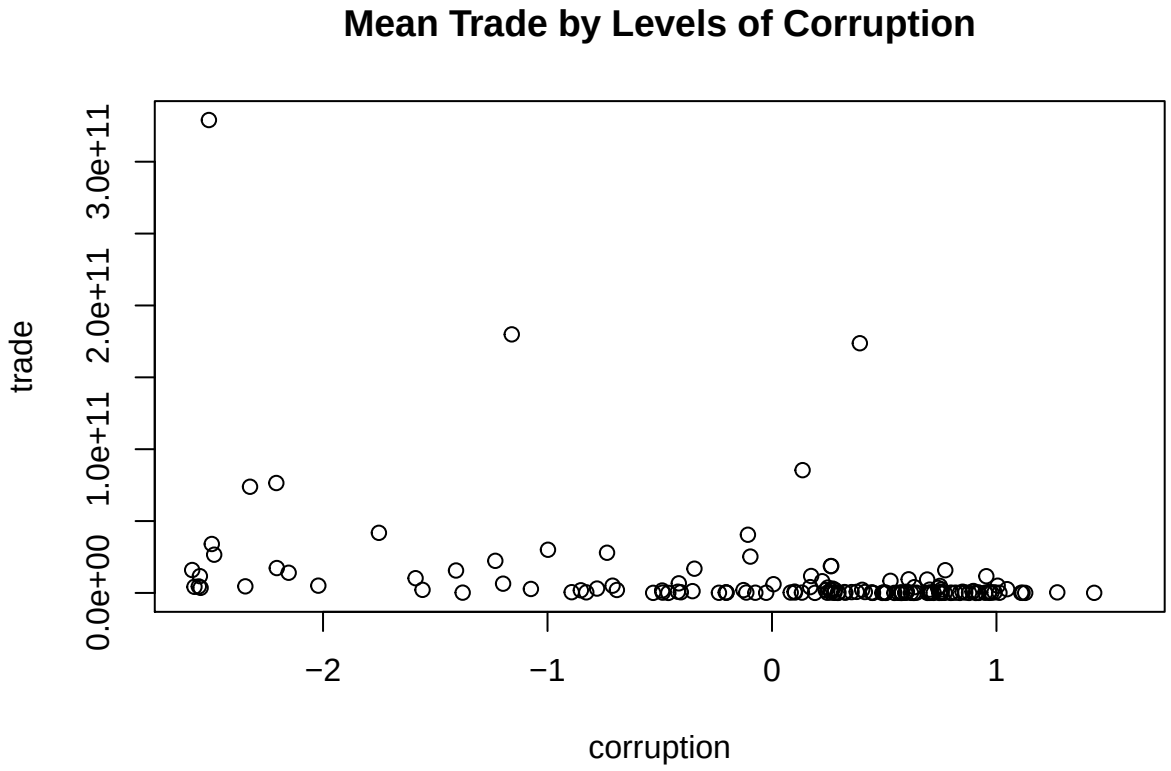
The resulting boxplot shows the unpaid violations for each group. It could tells us different story. If we only look at unpaid violations under 500, the results might gives us more granular level of detailed relationship.

```
### 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")
```



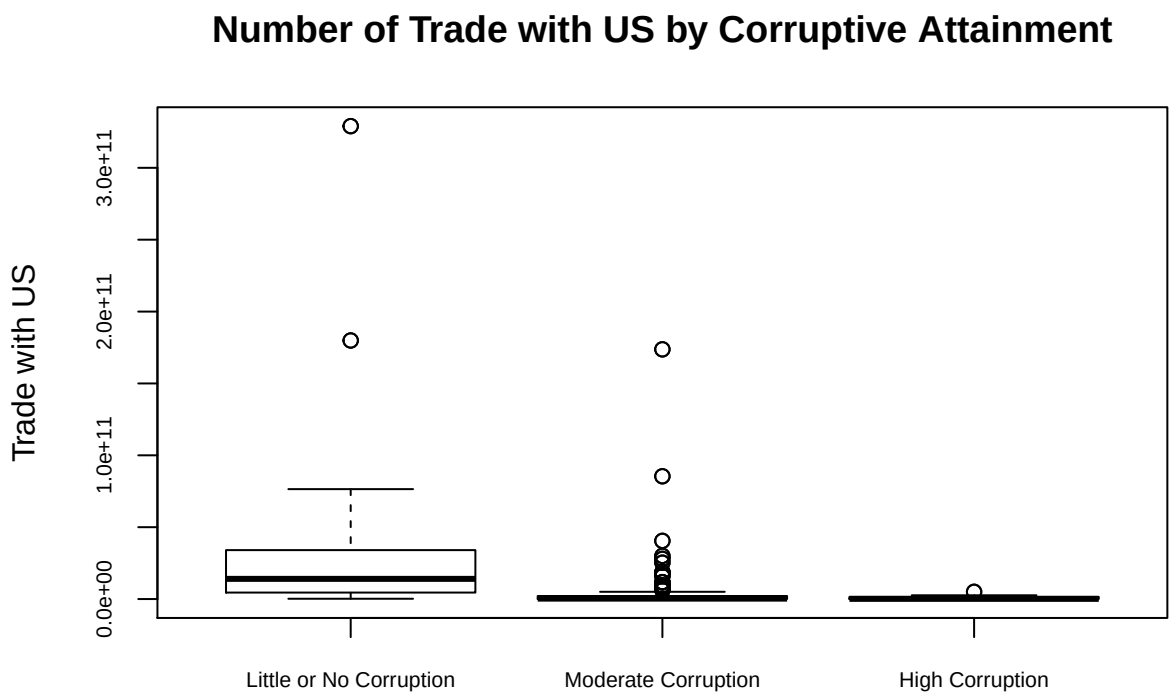
Trade and corruption

```
trade_mean2<-by(new_corrupt$trade, new_corrupt$corruption, mean, na.rm=T)
plot(sort(unique(new_corrupt$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. Binning into three corruption intervals could help us understand more detailed.

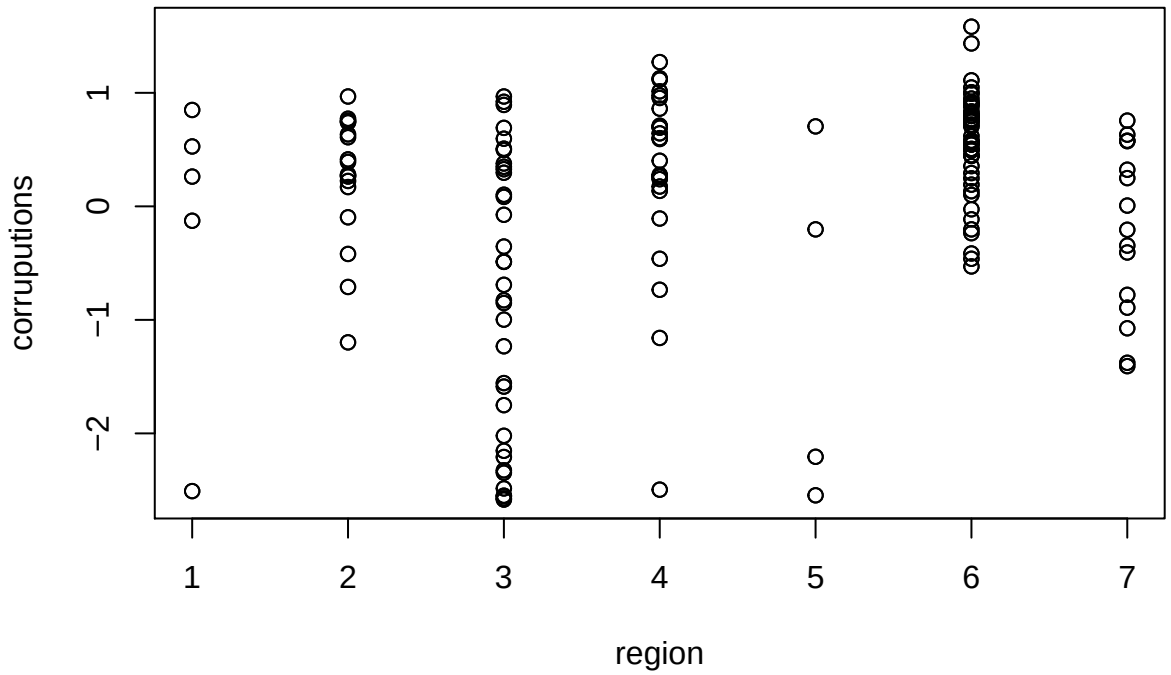
```
corruption_bin = cut(new_corrupt$corruption, breaks = c(-3,-1,1, Inf),
labels = c("Little or No Corruption", "Moderate Corruption", "High Corruption"))
boxplot(trade ~ corruption_bin, data = new_corrupt, cex.axis = .7,
main = "Number of Trade with US by Corruptive Attainment", ylab = "Trade with US")
```



Region and corruption

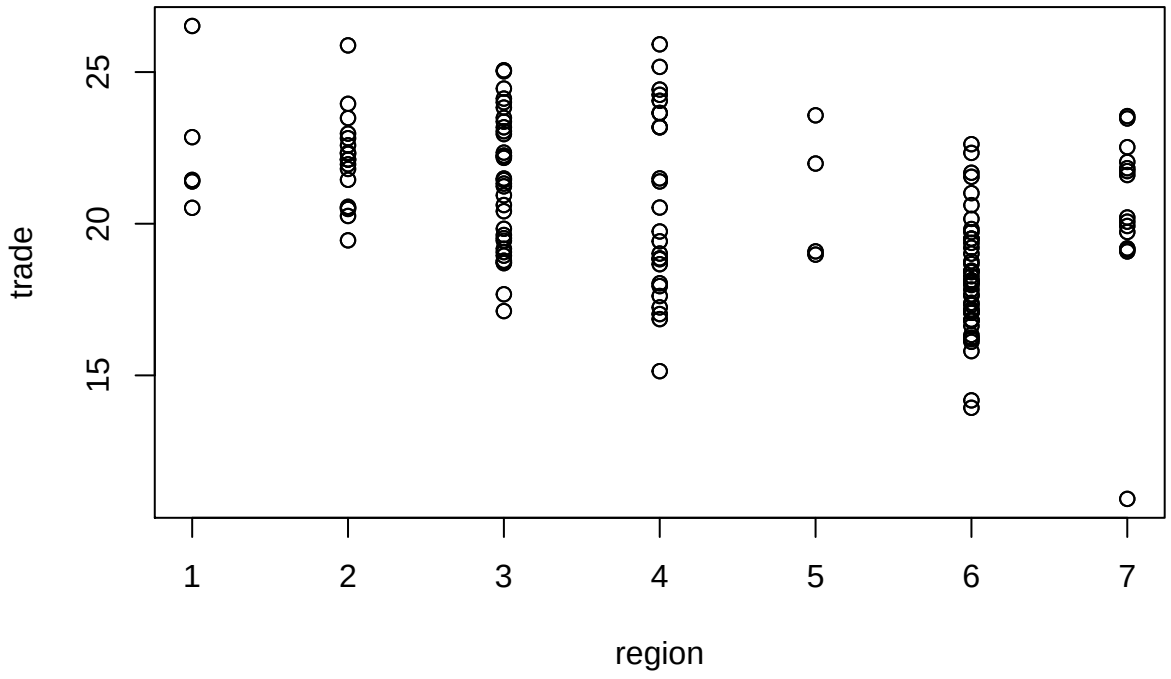
Now, trying to plot the corruptions against each region. Region 6 (Africa) has the most distribution of corruptions 4 between 0 and 1.

```
plot(new_corrupt$region, new_corrupt$corruption, xlab = "region", ylab = "corruputions")
```



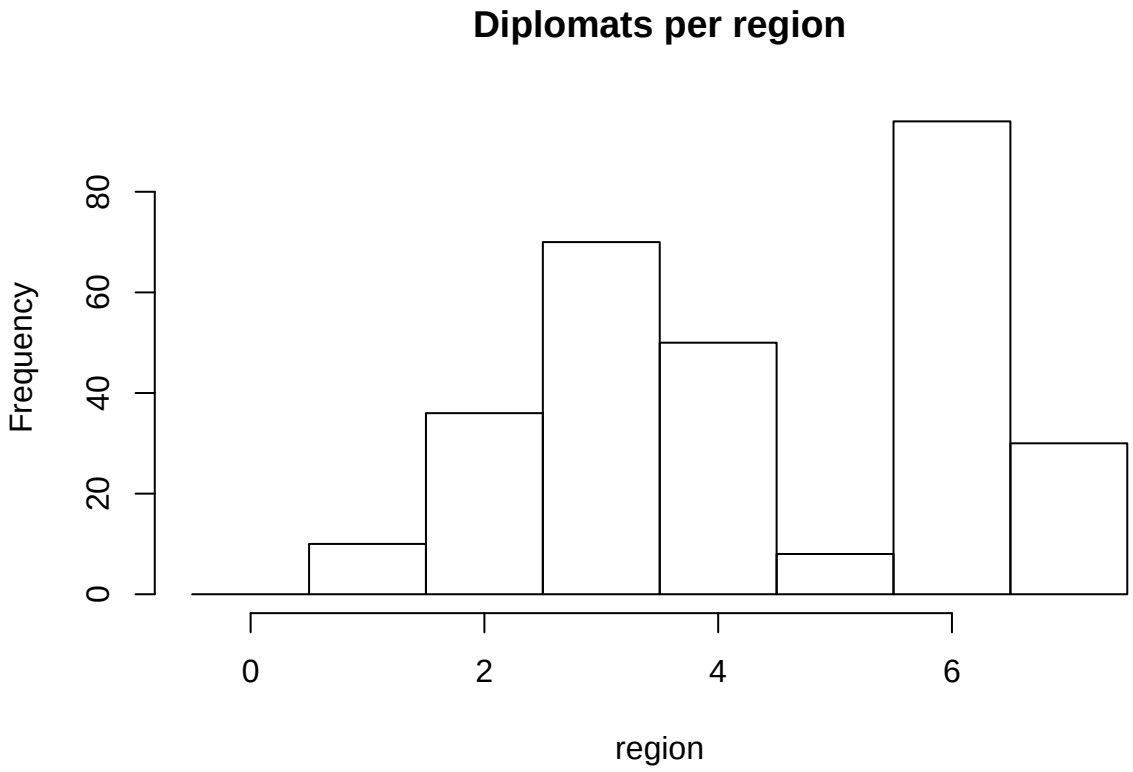
Region 6 (Africa) tends to have lower total trade with United States.

```
plot(new_corrupt$region, log(new_corrupt$trade), xlab = "region", ylab = "trade")
```



Region 6 (Africa) has the most UN diplomats presence.

```
hist(new_corrupt$region, breaks = 0:8 - 0.5, xlab = "region", main = "Diplomats per region")
```



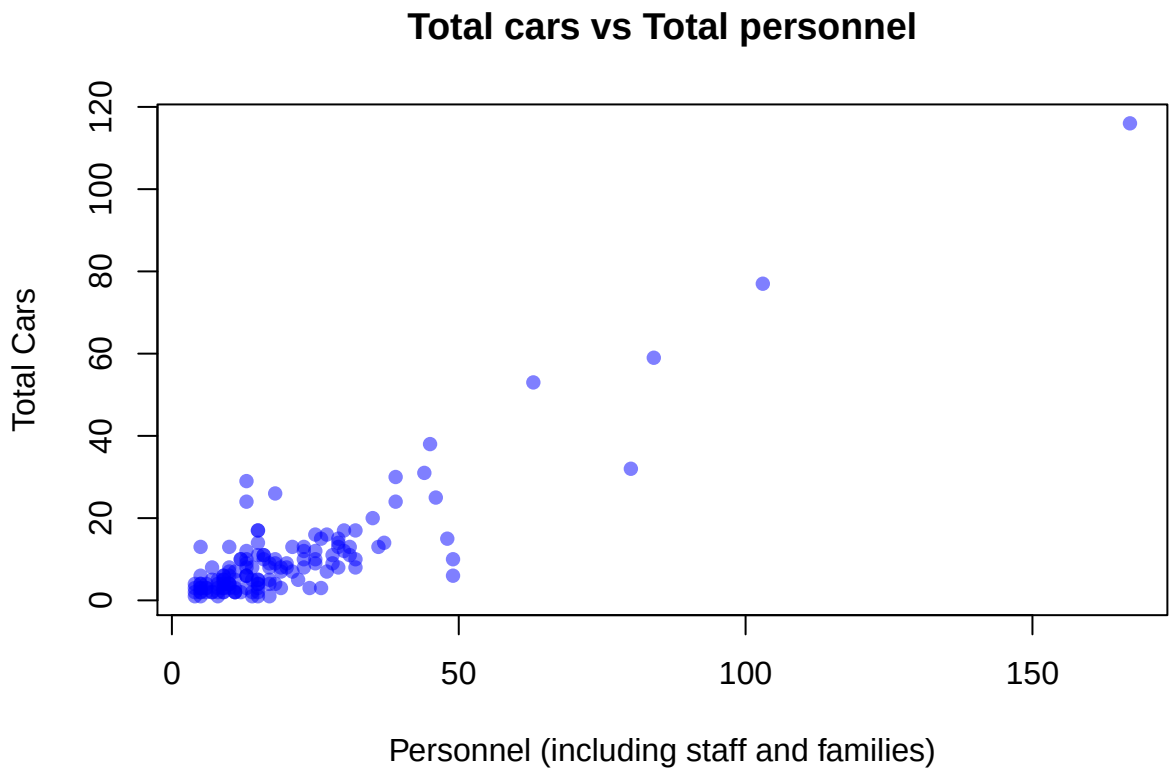
Number of Cars

As seen, most of the countries seems to lie on the lower part of the violations, with some outliers. However we may see some trend where higher corruption index correlates with higher amount of unpaid tickets. Two hypothesis were raised in order to clarify this relationship:

- Countries with more diplomats may have more cars.
- Countries with more cars assigned to their missions may have more tickets.

To test the first hypothesys we may see the distribution of total personnel vs number of cars.

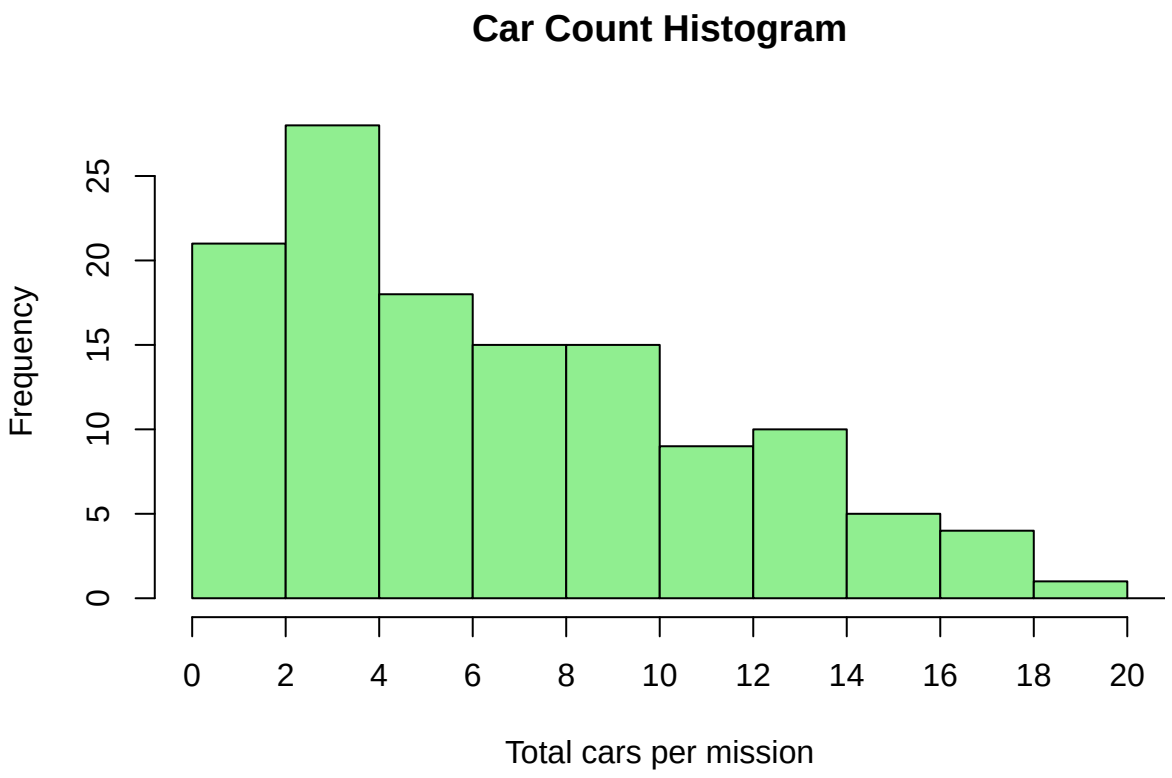
```
plot(pre_corrupt$staff + pre_corrupt$spouse,
     pre_corrupt$cars_mission + pre_corrupt$cars_personal,
     main = "Total cars vs Total personnel",
     xlab = "Personnel (including staff and families)",
     ylab = "Total Cars", col = alpha("blue",0.5), pch = 16, alpha = 50)
```



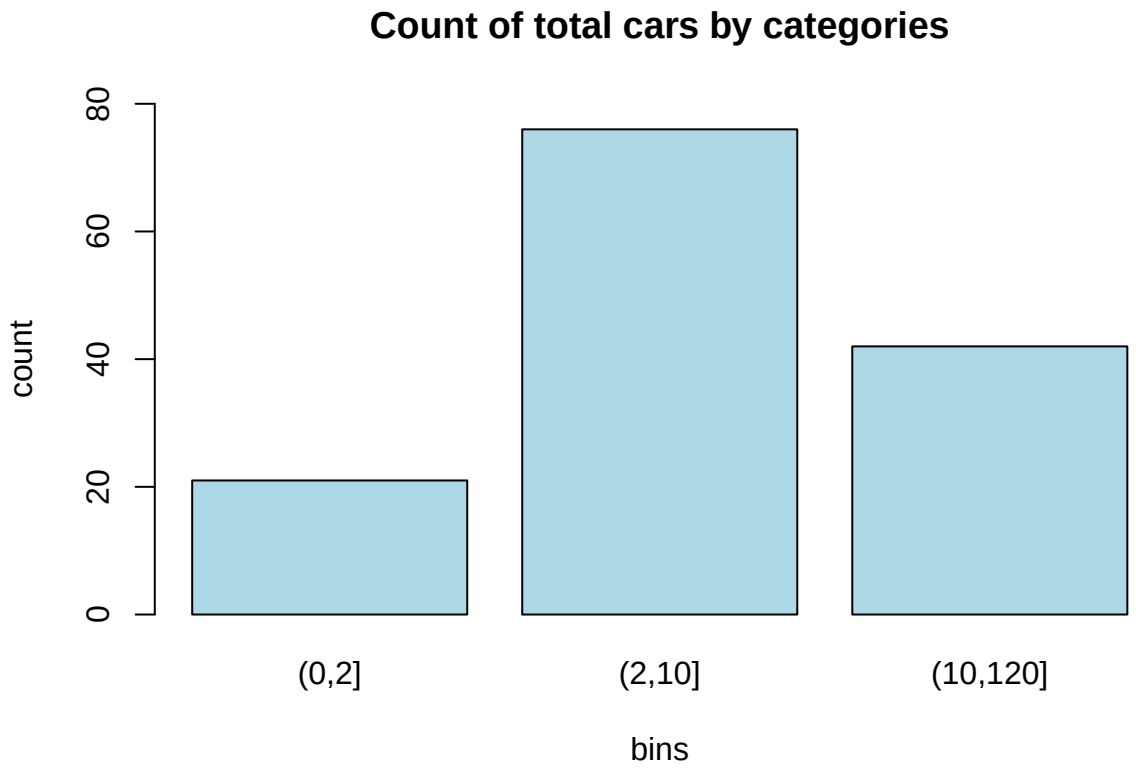
It is clear that the correlation is true, and we can highlight the case of Russia, with 167 persons assigned to thei diplomatic mission somehow and 116 vehicles.

To test the second hypothesis we identified that more than 20% of the countries have either one or two cars. In order to improve visualization the total car count were tranformed in a new categorized variable.

```
car_bins = seq(0,120, by = 2)
y1 <- hist(pre_corrupt$cars_mission + pre_corrupt$cars_personal,
           breaks = car_bins, xlim = c(0,20), col = "light green",
     main = "Car Count Histogram", xlab = "Total cars per mission", axes = F)
axis(1, at = seq(0,20, by = 2))
axis(2)
```




```
pre_corrupt$cat_car <- cut(pre_corrupt$cars_personal + pre_corrupt$cars_mission,
                           c(0,2,10,120))
barplot(table(pre_corrupt$cat_car),col = "light blue",
        main = "Count of total cars by categories",
        ylim = c(0,80), ylab = "count", xlab = "bins")
```



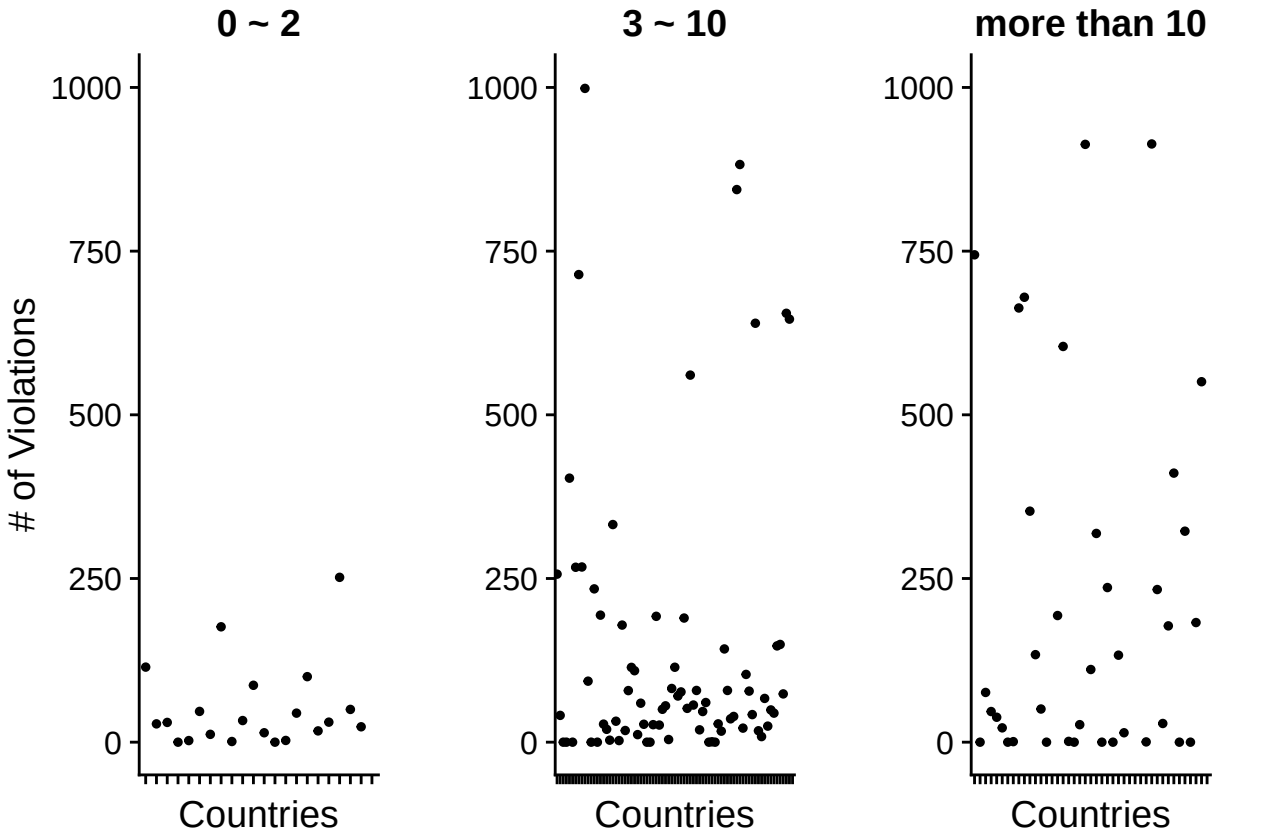
When verified how small missions, with 2 or less cars compares to the others, we see that they tend to have less tickets:

```
up_to_two <- ggplot(pre_corrupt[pre_corrupt$cat_car == "(0,2]",],
  aes(x = wbcode, y = violations)) +
  geom_point(size = 1) + theme(axis.text.x=element_blank()) +
  labs(x="Countries", y = "# of Violations", title = "0 ~ 2") +
  scale_y_continuous(limits = c(0,1000))

three_to_ten <- ggplot(pre_corrupt[pre_corrupt$cat_car == "(2,10]",],
  aes(x = wbcode, y = violations)) +
  geom_point(size = 1) + theme(axis.text.x=element_blank()) +
  labs(x="Countries", y = "", title = "3 ~ 10") +
  scale_y_continuous(limits = c(0,1000))

ten_plus <- ggplot(pre_corrupt[pre_corrupt$cat_car == "(10,120]",],
  aes(x = wbcode, y = violations)) +
  geom_point(size = 1) + theme(axis.text.x=element_blank()) +
  labs(x="Countries", y = "", title = "more than 10") +
  scale_y_continuous(limits = c(0,1000))

plot_grid(up_to_two,three_to_ten,ten_plus, ncol = 3, align = "hv")
```

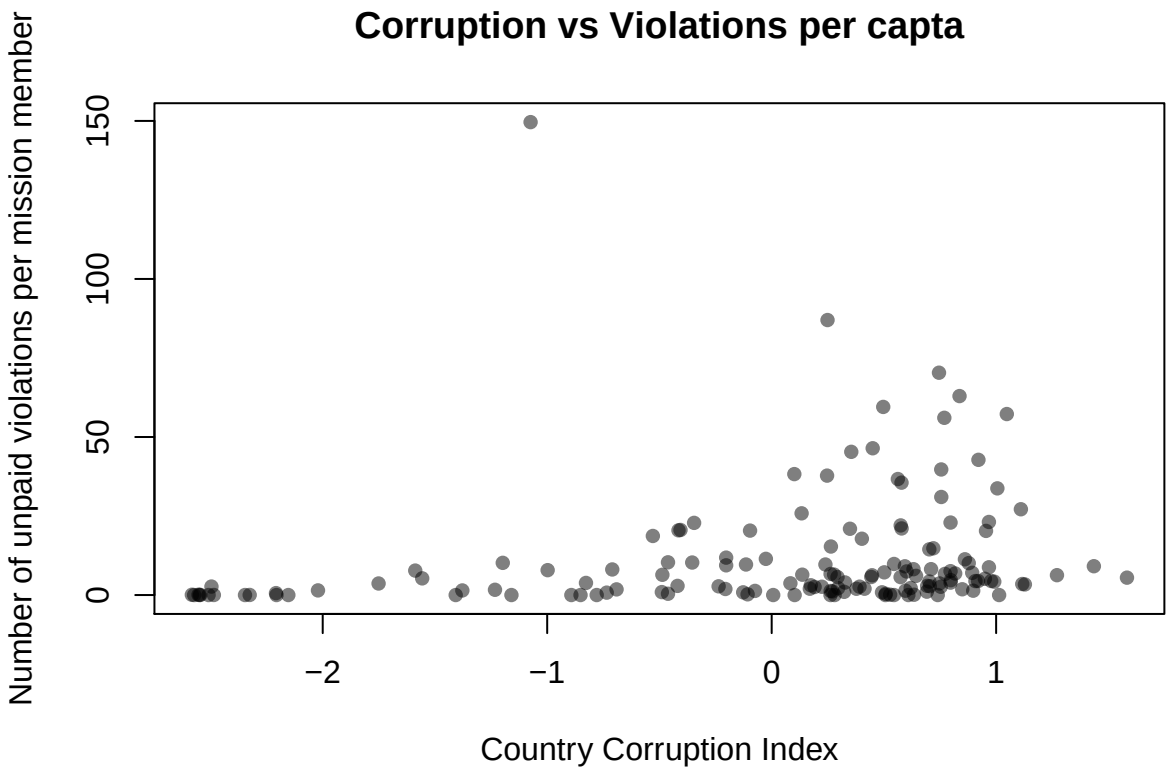


One idea that could explain the lower amount of tickets issued for small missions is that UN might have a limited amount of parking spots and those may be sufficient for one or two cars.

Tickets per diplomat

Another way to approach the dataset and minimize impact from the size of the missions in NY, is to normalize the number of tickets per personnel present in the city.

```
pre_corrupt$esp_violations <- pre_corrupt$violations/(pre_corrupt$staff + pre_corrupt$spouse)
plot(pre_corrupt$corruption, pre_corrupt$esp_violations,
     main = "Corruption vs Violations per capta", xlab = "Country Corruption Index",
     ylab = "Number of unpaid violations per mission member",
     col = alpha("black",0.5), pch = 16, alpha = 50)
```

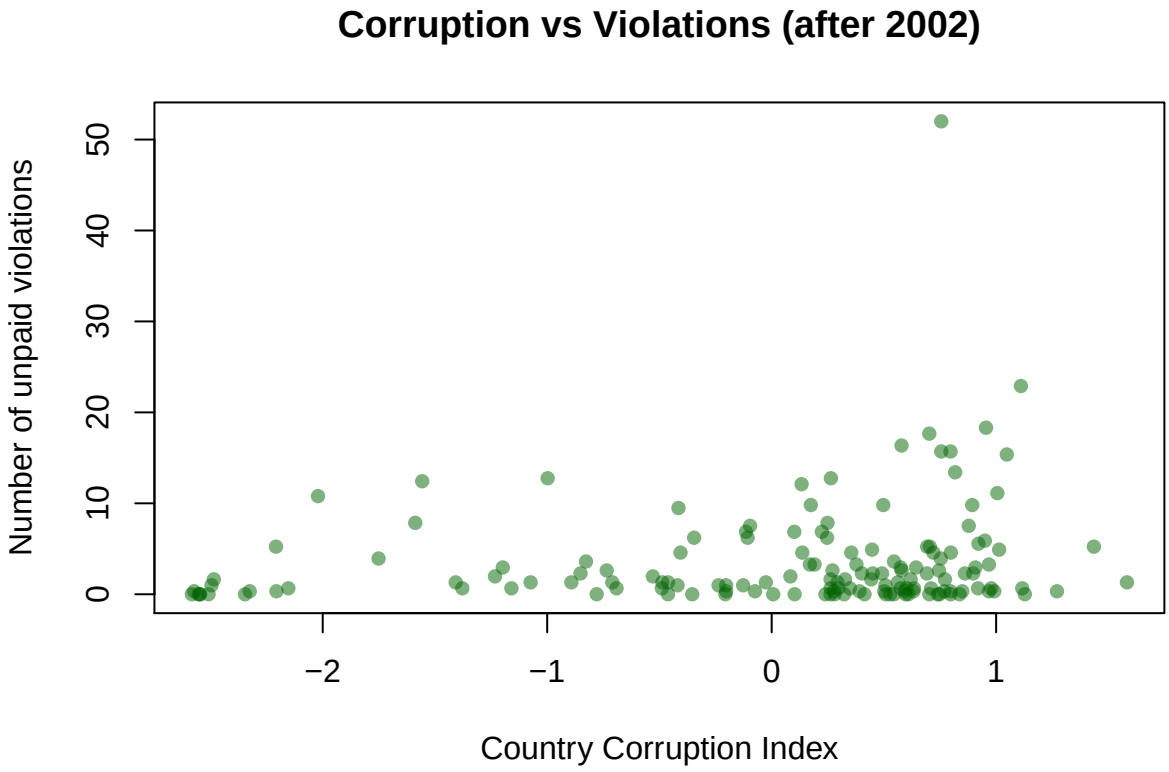


With the plot we can see that the correlation is verified, although there are countries that, even with a positive corruption index, the unpaid parking tickets are still low when compared with their peers.

Analyzing the impact of the 2002 change in the previous analysis

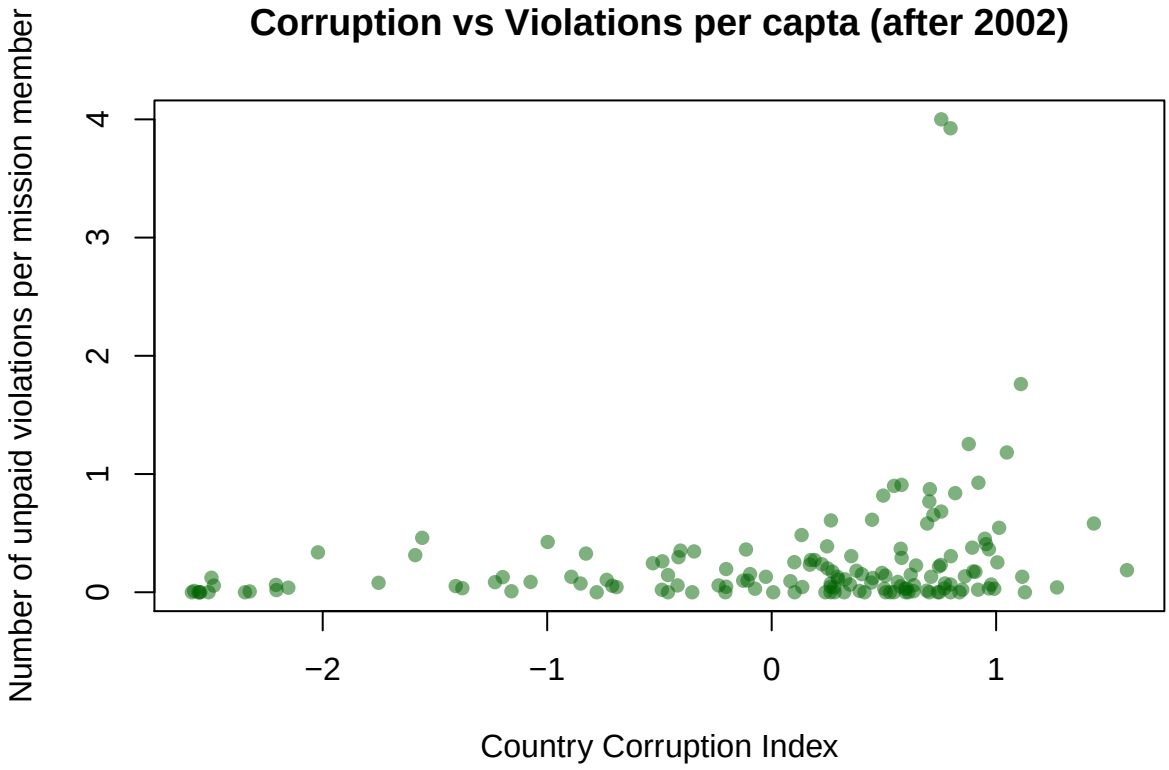
After verifying that the corruption index is positively correlated to the unpaid parking tickets in the previous section, we decided to apply the same analysis to the dataset after the license plates began to be confiscated. At a first sight, the correlation persist, but in a lower absolute number of violations. Where before we had a few countries with numbers around 1000, now we have up to 50 violations.

```
plot(pos_corrupt$corruption, pos_corrupt$violations,
     main = "Corruption vs Violations (after 2002)",
     xlab = "Country Corruption Index",
     ylab = "Number of unpaid violations",
     col = alpha("dark green",0.5), pch = 16, alpha = 50)
```



The same applies to the per capta violations, where we had a maximum of 4 violations per capta against roughly 150 before.

```
pos_corrupt$esp_violations <- pos_corrupt$violations/(pos_corrupt$staff + pos_corrupt$spouse)
plot(pos_corrupt$corruption, pos_corrupt$esp_violations,
     main = "Corruption vs Violations per capta (after 2002)",
     xlab = "Country Corruption Index",
     ylab = "Number of unpaid violations per mission member",
     col = alpha("dark green",0.5), pch = 16, alpha = 50)
```



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 violations 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.

The major contribution to diplomatic behavior is cultural norm and legal penalties don't alter the behavior but only suppress the violations. The diplomats from Africa have the significant contribtions to the violations indicated by the Country corruption index between 0 and 1, and lower total trade with the United States. The amount of African diplomats is the most among diplomats from other regions.

We may see some trend where higher corruption index correlates with higher amount of unpaid tickets. Small missions, with 2 or less cars compares to the others, we see that they tend to have less tickets. There are countries that, even with a positive corruption index, the unpaid parking tickets are still low when compared with their peers.