W203 Lab 1: EDA

Daniel VanLunen, Nishant Velagapudi, Michelle Cutler, Eleanor Proust

Introduction

Our research question is whether or not there is a relationship between the number of unpaid parking violations per diplomat and the level of corruption within a country.

We are interested in evaluating this relationship both pre and post 2002 (when enforcement of parking violations began). We are also interested in exploring what other variables might correlate strongly with corruption, parking violations, or both - confounding a hypothetically simple relationship.

The following code loads the dataset into R.

```
root <- "D:/dropbox/Dropbox/MIDS work/W203/week 2/lab"
load(paste(root,"/Corruption_EDA/Corrupt.Rdata",sep=""))
```

The Corruption Dataset: What types of variables does it contain? How many observations are there?

First, we will review the first few observations, the fields names and those fields' datatypes, and a couple of observations with the *str* command.

```
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 ...
```

The dataset has 364 observations and 28 fields. 3 fields are character strings, 12 are numeric, and 13 are integer. It appears that some of the integer fields are binary flags instead of counts (i.e., fields with the r_ prefix appear to be region indicators, mission appears to be a mission indicator, and majority muslim is an indicator for if a majority of the mission country's population is Muslim). As discussed later, each country with any non-NA values has two data points (one for each period: pre and post).

The fields generally fall into 4 categories:

1. Enforcement Indicator: The prepost field indicates whether the mission's data is before enforcement began.
2. Violation Measures: Violations gives the number of unpaid parking violations and fines likely displays the fines associated with those violations.
3. Mission Descriptors: Fields that describe the mission (e.g., staff, spouse, distUNplz).
4. Country Descriptors: Fields that describe the country the mission comes from (e.g., trade, region, totaid).

Evaluate the data quality. Are there any issues with the data? Explain how you handled these potential issues.

We will use this section to examine missing data, but leave a discussion of odd values, for the univariate analysis.

The following code counts missing values.[1]

```
colSums(FMcorrupt=="" | is.na(FMcorrupt))
```

```
##         wbcode        prepost     violations          fines        mission
##              0             62             66             66             62
##          staff         spouse   gov_wage_gdp       pctmuslim majoritymuslim
##             62             62            180             66             66
##          trade     cars_total  cars_personal   cars_mission        pop1998
##             68             86             86             86             42
##    gdppcus1998          ecaid         milaid         region     corruption
##             42             68             68             64             61
##         totaid       r_africa   r_middleeast       r_europe r_southamerica
##             68             42             42             42             42
##         r_asia        country       distUNplz
##             42             60             33
```

Given that violations is the key dependent variable, let's explore its 66 missing values first.

```
head(FMcorrupt[is.na(FMcorrupt$violations),],n = 5)
```

```
##    wbcode prepost violations fines mission staff spouse gov_wage_gdp
## 1     AFG                 NA    NA      NA    NA     NA           NA
## 12    ATG                 NA    NA      NA    NA     NA           NA
## 37    BLZ                 NA    NA      NA    NA     NA           NA
## 42    BRB                 NA    NA      NA    NA     NA           NA
## 43    BRN                 NA    NA      NA    NA     NA           NA
##    pctmuslim majoritymuslim trade cars_total cars_personal cars_mission
```

[1] Note that we count missing values by column counting the value as missing if it is NA or if it is "". Checking for "" is necessary for some of the character fields that not only have NA missing values, but "" values as well.

```
## 1          NA              NA    NA          NA             NA            NA
## 12         NA              NA    NA          NA             NA            NA
## 37         NA              NA    NA          NA             NA            NA
## 42         NA              NA    NA          NA             NA            NA
## 43         NA              NA    NA          NA             NA            NA
##    pop1998 gdppcus1998 ecaid milaid region corruption totaid r_africa
## 1       NA          NA    NA     NA     NA         NA     NA       NA
## 12      NA          NA    NA     NA     NA         NA     NA       NA
## 37      NA          NA    NA     NA     NA         NA     NA       NA
## 42      NA          NA    NA     NA     NA         NA     NA       NA
## 43      NA          NA    NA     NA     NA         NA     NA       NA
##    r_middleeast r_europe r_southamerica r_asia           country distUNplz
## 1            NA       NA             NA     NA        AFGANISTAN 0.4451198
## 12           NA       NA             NA     NA ANTIGUA & BARBUDA 0.7626809
## 37           NA       NA             NA     NA            BELIZE 0.1712945
## 42           NA       NA             NA     NA          BARBADOS 0.1712945
## 43           NA       NA             NA     NA            BRUNEI 0.1134206
```

Note that these data points are not only missing violations, they are also missing meaningful information about the missions: among these rows mission, staff, and spouse are NA or 0 while cars_total, cars_personal, cars_mission are always NA.[2] Given our objective of finding a relationship between corruption and violations, and that these observations don't have information about the violations or or the missions, we would exclude these points in our analysis.

Now we can update our missing counts to exclude these 66 rows.

```r
colSums(FMcorrupt[!is.na(FMcorrupt$violations),]=="" |
         is.na(FMcorrupt[!is.na(FMcorrupt$violations),]))
```

```
##         wbcode         prepost     violations          fines        mission
##              0               0              0              0              0
##          staff          spouse    gov_wage_gdp      pctmuslim majoritymuslim
##              0               0            114              4              4
##          trade      cars_total  cars_personal   cars_mission        pop1998
##              4              20             20             20              0
##    gdppcus1998           ecaid         milaid         region     corruption
##              0               4              4              2              0
##         totaid        r_africa   r_middleeast       r_europe r_southamerica
##              4               0              0              0              0
##         r_asia         country      distUNplz
##              0              22              6
```

By removing missing violations, a number of the other missing values disappear. Key variables like corruption, prepost, staff, spouse no longer have missing values.

We also see that after the 66 NA violations are removed, each country has two observations: one pre and one post (while we have checked each of these tuples, we include only 5 below).

```r
head(FMcorrupt[!is.na(FMcorrupt$violations),c("wbcode","country","prepost")],n=5)
```

```
##   wbcode country prepost
## 2    AGO  ANGOLA     pre
## 3    AGO  ANGOLA     pos
## 4    ALB ALBANIA     pre
## 5    ALB ALBANIA     pos
```

[2] Though the above code only selects a few rows to avoid a data dump, we did separately export the dataset to examine all 66 rows with NA violations and found the same results.

```
## 6    ARE             pre
```

Next we separate the data into pre and post datasets that exclude the observations with missing values.

```
pre = FMcorrupt[!is.na(FMcorrupt$violations) & FMcorrupt$prepost=="pre",]
pre = pre[with(pre, order(wbcode)),] # order to make comparable with post
post = FMcorrupt[!is.na(FMcorrupt$violations) & FMcorrupt$prepost=="pos",]
post = post[with(post, order(wbcode)),] # order to make comparable with pre
```

Nearly all the variables have the same value for the same country in the both the pre and post periods. The only exceptions are violations, fines, and distUNplz as seen from the code below. This tells us that the data is roughly contemporaneous for the pre-enforcement period (at least with respect to the economic data), but historical data for the post enforcement period.

```
# Check for equality in all columns except prepost
all.equal(pre[,!names(pre) %in% c("prepost")],
         post[,!names(post) %in% c("prepost")])
```

```
## [1] "Attributes: < Component \"row.names\": Mean relative difference: 0.005831931 >"
## [2] "Component \"violations\": Mean relative difference: 0.9823298"
## [3] "Component \"fines\": Mean relative difference: 0.969022"
## [4] "Component \"distUNplz\": Mean relative difference: 0.7374599"
```

There are still 20 missing values for the cars fields. We could look into imputing their values (by regression or otherwise) for these rows so the rest of the data on the rows could still be utilized.

The remaining fields with missing values (region, trade, ecaid, milaid, totaid, pctmuslim, majoritymuslim, gov_wage_gdp, and even distUNplz) could likely be filled by doing online research or comparing with other fields in the data (e.g. region/country is determined by wbcode).

Beyond missing values, there are a number of things that we would like to know more about with this dataset, but cannot find by exploring the data. For example, were the pre and post periods the same length? If violations is 10 years of violations for pre and 5 years for post, the reduced length of time might make enforcement look especially effective. Also, are personal diplomatic cars exempt from enforcement in the pre period, are spouses exempt? Are the fines from the violations only or some other kind of fine?

Explain whether any data processing or preparation is required for your data set.

There are two key preprocessing steps that are required for the dataset: (1) Remove the rows with missing values of violations as discussed above and (2) separate the dataset into pre and post datasets. If needed, other missing values can potentially be filled in using online research or imputed.

Given that the data is the same in both sets, for all but three of the variables we can combine our two data sets so we are able to compare how violations in the pre-enforcement period affect violations in the post period more easily.

```
combined=pre
combined$violations_post=post$violations
combined$fines_post=post$fines
combined$distUNplz_post=post$distUNplz
```

Univariate Analysis of Key Variables

In our univariate analysis below, we separately use pre and post datasets to avoid double counting for values that are the same in both periods and to understand if there is a difference in the univariate distribution across periods for the others.

We will begin with violations as it is the dependent variable per our questions.

```
boxplot(pre$violations,post$violations,
       names=c("Pre Enforcement", "Post Enforcement"), main="Violations")
```


```
head(pre$violations)
```

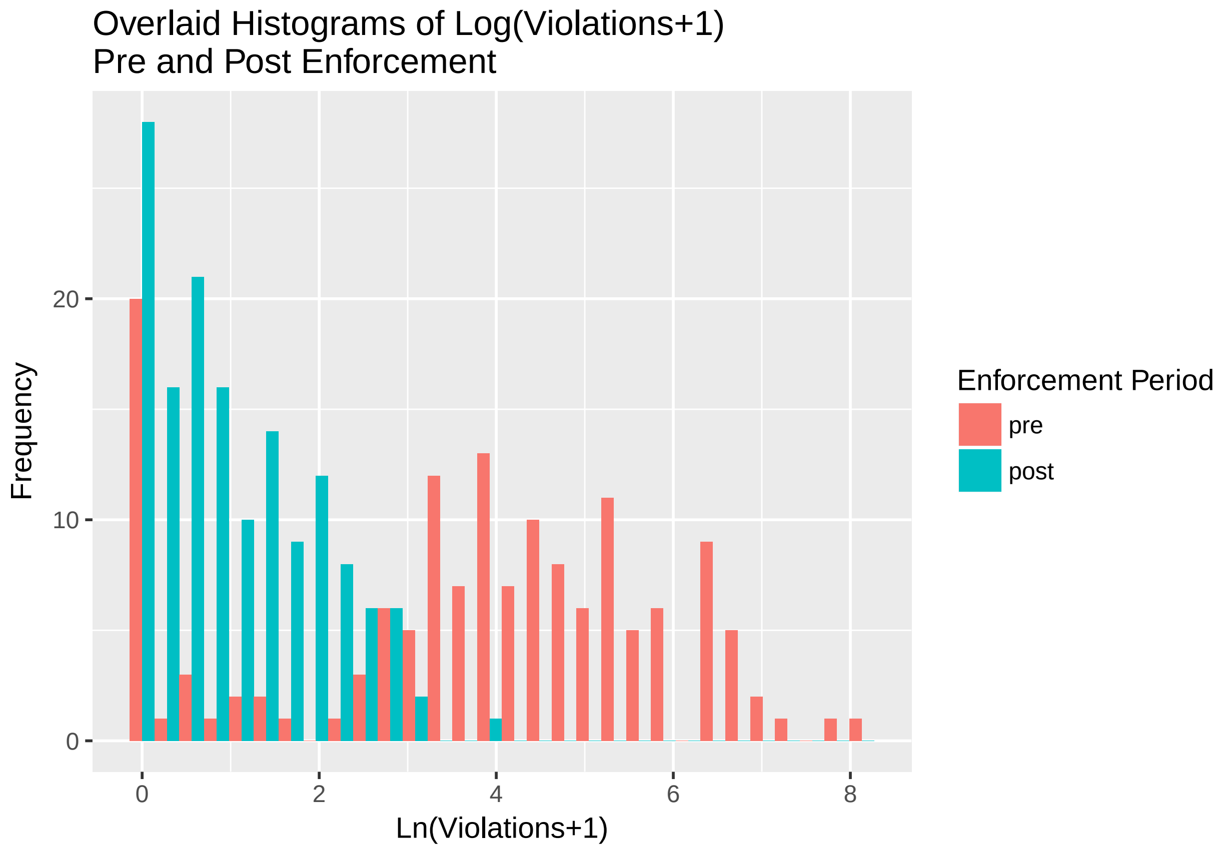
```
## [1] 744.38123 256.63431   0.00000  75.95727  40.91565   0.00000
```

Three things are apparent: (i) violations has values that are not whole numbers, (ii) the distribution is very skewed, and (iii) the enforcement period has a significant impact on the univariate distribution.

(i) We would expect violations ("Unpaid New York City parking violations") to be a whole number. However, the Fisman and Miguel paper notes that it is violations per diplomat, which explains the values. We do question the choice to provide the data in this format given staff numbers are consistent across the two variable sets, and it is unlikely that staff numbers did not change across 149 consulates.

(ii) Applying a concave transform will reduce the positive skewness of the variable so we can get a better sense of the distribution. Below, we show a histogram of the values + 1 (adding 1 because log(0) is undefined). This was also the approach of Fisman and Miguel.

```
ggplot(melt(data.frame(pre= log(pre$violations + 1),
                       post=log(post$violations + 1))),
      aes(value, fill = variable)) +
 geom_histogram(position = "dodge") +
 ggtitle("Overlaid Histograms of Log(Violations+1) \nPre and Post Enforcement") +
 labs(y="Frequency", x="Ln(Violations+1)") +
 scale_fill_discrete(name="Enforcement Period")
```



Both the pre and post periods have large spikes around zero, but otherwise the two data sets appear to follow different trends. The post data is almost monotonic in that the amount of violations across the histogram become less frequent as the log violation decreases.

In contrast the pre-enforcement data is, excluding the zero values, follows a slightly more normal distribution. This suggests that in the post enforcement data, there was more incentive for missions to clear the violations, or alternatively not incur them in the first place.

There are a number of data points with 0 violations. Though a value of 0 is certainly possible, it is worth exploring these data points to see if a zero value appears justifiable.

```
head(combined[combined$violations==0 & combined$cars_total>5 & combined$staff>10,
        !names(combined) %in% c("prepost","mission",
                               "majoritymuslim","cars_personal","cars_mission",
                               "r_africa","r_middleeast","r_europe",
                               "r_southamerica","r_asia","country",
                               "pop1998","trade","pctmuslim","gdppcus1998",
                               "gov_wage_gdp","ecaid","milaid","totaid",
                               "region","country")],n=5)
```

```
##     wbcode violations fines staff spouse cars_total corruption distUNplz
## 50     CAN          0     0    24     13         14 -2.5084674 0.2956051
## 80     DNK          0     0    17     12         13 -2.5728226 0.2956051
## 122    GRC          0     0    21     10         13 -0.8515174 0.2218524
## 150    ISR          0     0    15     10         16 -1.4075348 0.1712945
## 158    JPN          0     0    47     33         32 -1.1592430 0.2462278
##     violations_post fines_post distUNplz_post
## 50        0.0000000    0.00000      0.2956051
```

```
## 80        0.3270609   31.07079      0.2956051
## 122       2.2894266  230.57796      0.2218524
## 150       1.3082438  127.55376      0.1712945
## 158       0.6541219   68.68279      0.2462278
```

There are missions with more than 5 cars total and more than 10 staff, but still that still have 0 violations. Given the low values of distUNplz for both time periods, these could potentially be explained by staff walking to the plaza instead of parking nearby. Without driving, they would be less likely to get violations. Some of these data points have a wbcode, but are missing the full country field. We will remain aware of these points and will revisit the issue should we see unexpected behavior that might tie to these unusual points.

We also have 20 data points with a null "cars_total" value after removing null violations rows. We will exclude these points when examining any of the cars attributes below (but would try to impute the values in analyses if the cars values were significant).

We note that the log of fines appear to follow a more normal distribution in the post enforcement period than it did in the post period.

Next let's look at the corruption index. We note that the corruption variable is constant in both datasets.

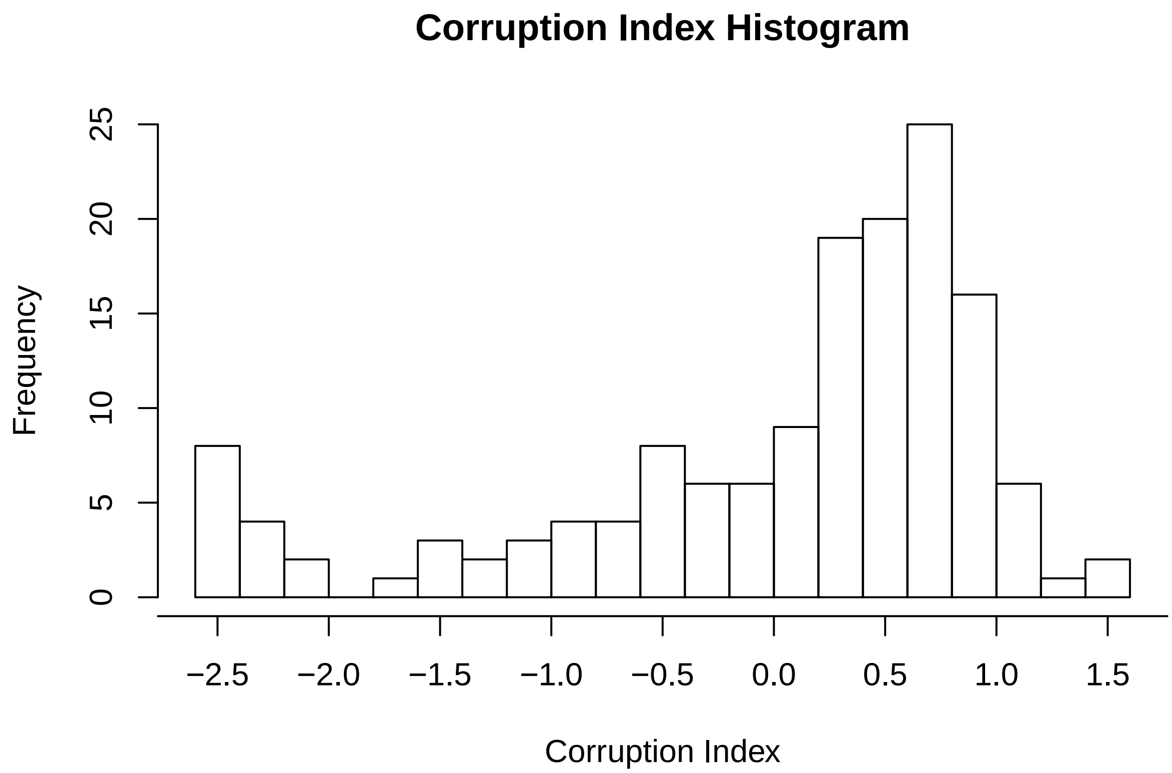
```r
summary(combined$corruption)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.58299 -0.41515  0.32696  0.01364  0.72025  1.58281
```

```r
var(combined$corruption,na.rm=T)
```

```
## [1] 1.028566
```

```r
hist(combined$corruption,breaks = 20,
    main="Corruption Index Histogram",xlab = "Corruption Index",xaxt="n")
axis(side=1, at=(seq(-4,4,.5)))
```



The relationship appears to be bimodal with a peak around -2.5 and another around 0.5. The peak at 0.5 is much larger than the peak at the negative amount. Also, the index seems to be close to normalized with 0 mean and unit variance. This variable will likely not require transformation.

Next, we need to identify whether high values mean more corruption or less corruption in order to interpret our results later. Let's look at the 6 countries with the highest and lowest index.

```
ordered<-combined[order(combined$corruption),c("wbcode","corruption")]
head(ordered[!is.na(ordered$corruption),],6)
```

```
##     wbcode corruption
## 52     CHE  -2.582988
## 80     DNK  -2.572823
## 101    FIN  -2.553532
## 295    SWE  -2.548720
## 243    NZL  -2.545430
## 50     CAN  -2.508467
```

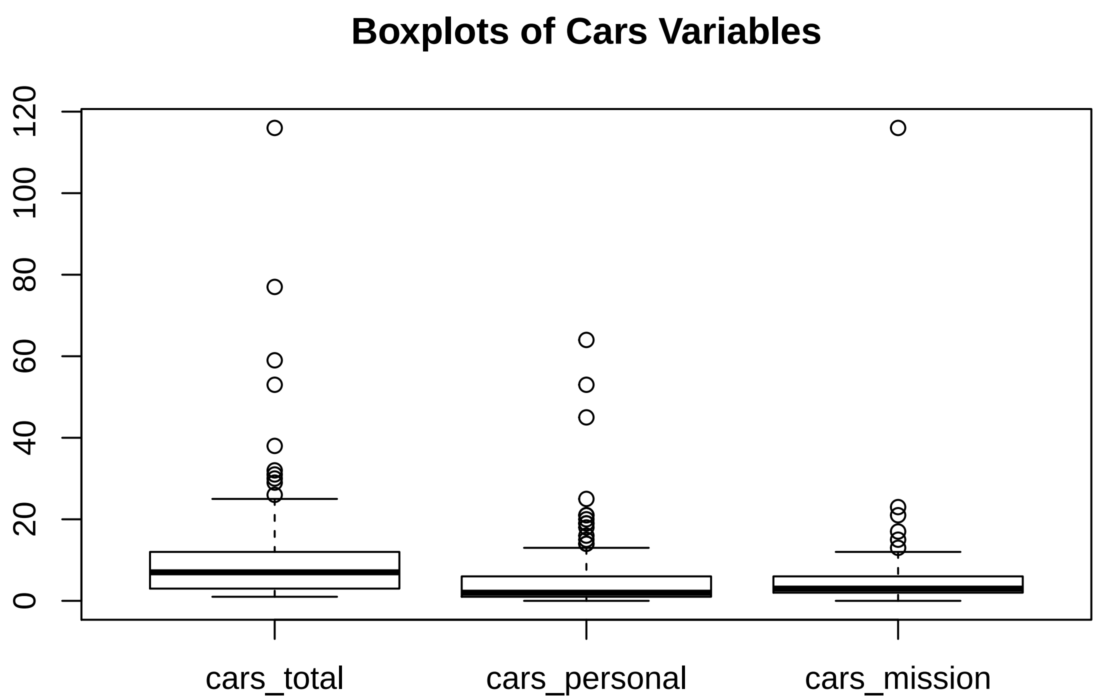
```
tail(ordered[!is.na(ordered$corruption),],6)
```

```
##     wbcode corruption
## 60     CMR   1.110222
## 308    TJK   1.116204
## 310    TKM   1.128060
## 166    KHM   1.270946
## 178    LBR   1.435285
## 345    ZAR   1.582807
```

It appears that higher values imply more corruption[3].

The number of cars and people could also have a large impact because more cars can get more tickets.

```
boxplot(combined[,c("cars_total","cars_personal","cars_mission")])
title("Boxplots of Cars Variables")
```




```
summary(combined[,c("cars_total","cars_personal","cars_mission")])
```

```
##    cars_total      cars_personal     cars_mission
##  Min.   :  1.00   Min.   : 0.000   Min.   :  0.000
##  1st Qu.:  3.00   1st Qu.: 1.000   1st Qu.:  2.000
##  Median :  7.00   Median : 2.000   Median :  3.000
##  Mean   : 10.47   Mean   : 5.324   Mean   :  5.144
##  3rd Qu.: 12.00   3rd Qu.: 6.000   3rd Qu.:  6.000
##  Max.   :116.00   Max.   :64.000   Max.   :116.000
##  NA's   :10       NA's   :10       NA's   :10
```

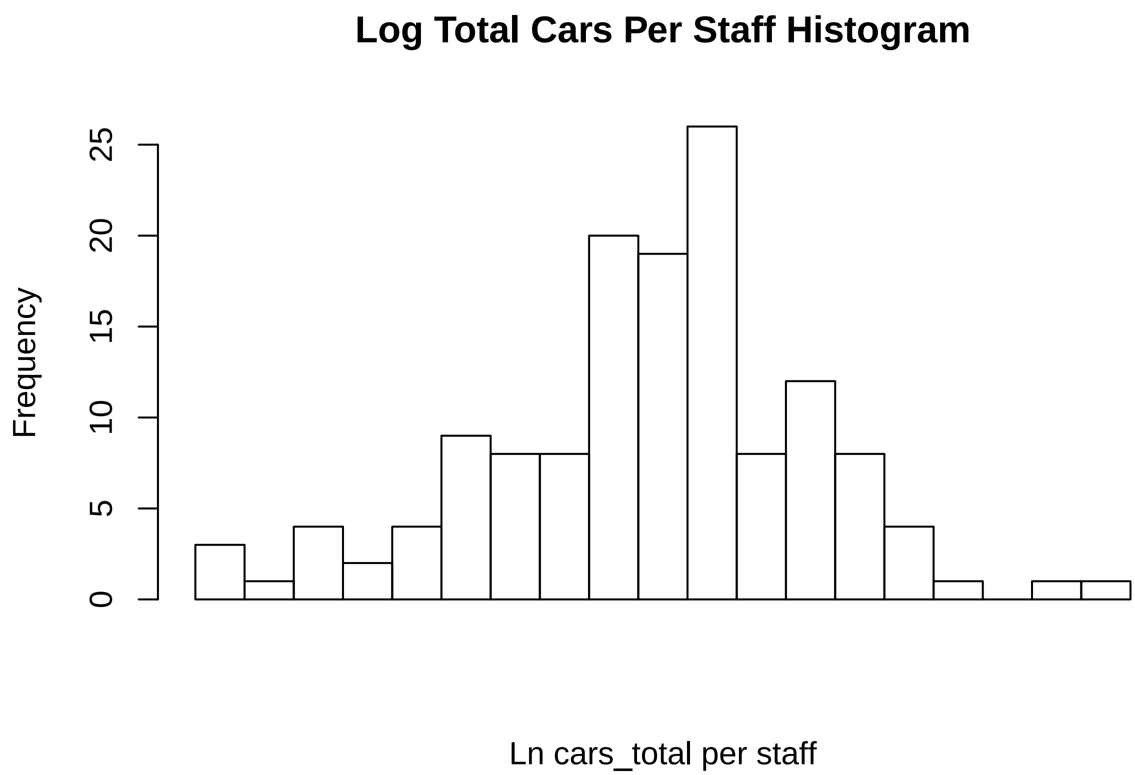
```
all.equal(combined$cars_personal+combined$cars_mission,combined$cars_total)
```

```
## [1] TRUE
```

cars_total=cars_personal+cars_mission. These distributions are also very skewed. To get a better sense of the distribution of cars, we use a log transform below. Given the violations is divided by the number of diplomats and cars are likely to relate to the number of diplomats, we create additional fields for cars per diplomat.

[3] When we compare to the online Transparency International Corruption Perceptions Index, Denmark and Finland have low corruption and Cameroon has high corruption

```
combined$cars_per_dip=(combined$cars_total/combined$staff)
post$cars_per_dip=(post$cars_total/post$staff)
hist(log(combined$cars_per_dip), breaks = 20,
    main="Log Total Cars Per Staff Histogram",xlab = "Ln cars_total per staff",xaxt="n")
```



The log gives us a distribution approximating a normal distribution. A log distribution is also logical in that the marginal effect of an extra car per staff member would likely be very small.

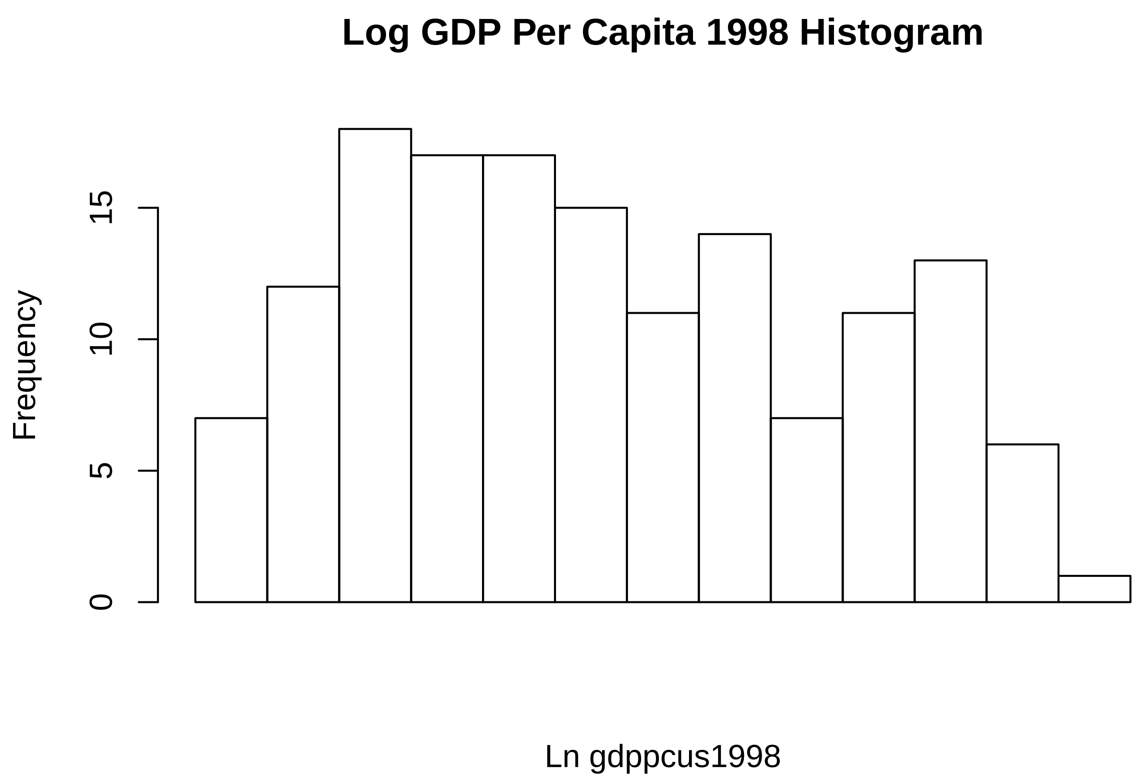
GDP per capita (gdppcus1998) may also be interesting in our analysis, as the wealth of a country may affect capacity to pay or behavioural norms.

```
summary(combined$gdppcus1998)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    95.45   412.07  1374.88  5044.09  4936.62 36485.64
```

These value are around what we would expect for a GDP measure. They are also highly skewed. Again a log transform makes the distribution more normal and less prone to outliers.

```
hist(log(combined$gdppcus1998),breaks = 20,
    main="Log GDP Per Capita 1998 Histogram",xlab = "Ln gdppcus1998",xaxt="n")
```



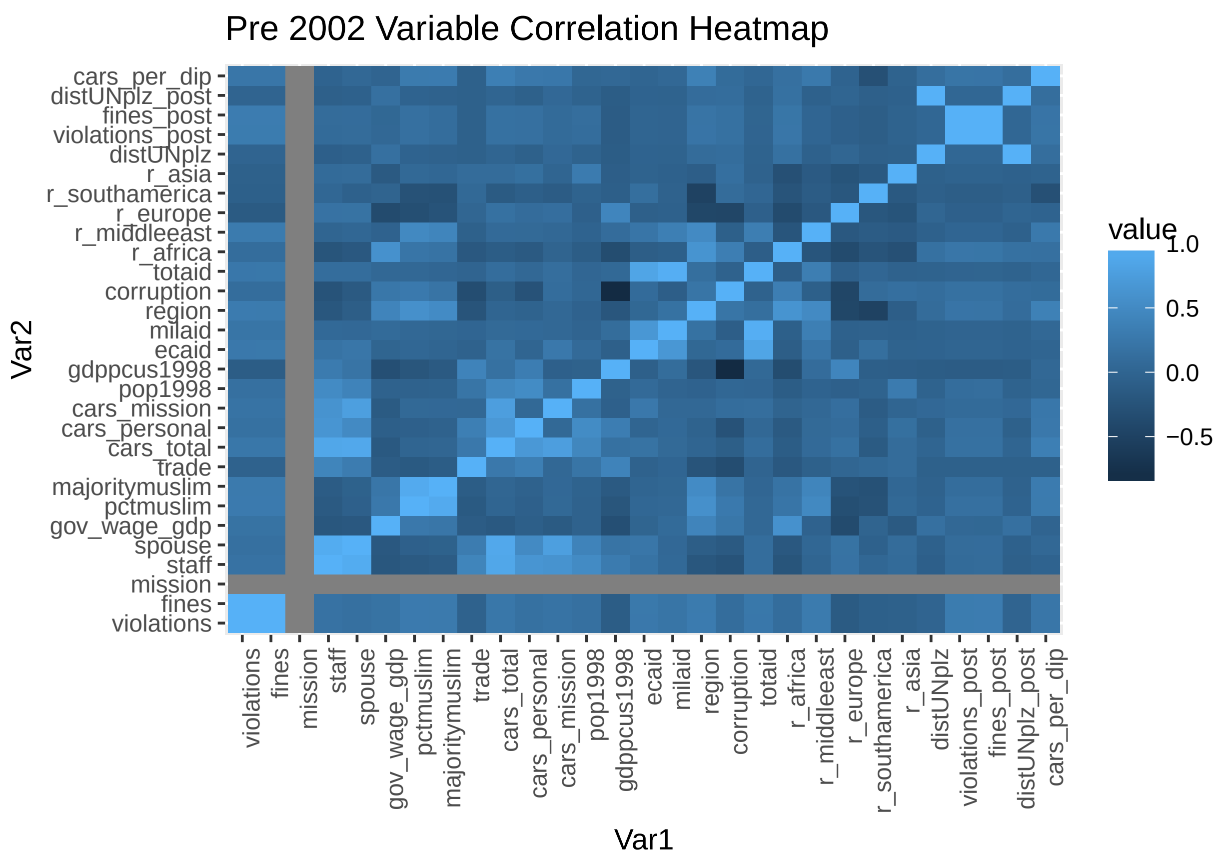
Analysis of Key Relationships

We'll start with a simple correlation heatmap between all of our non-categorical variables:

```
q <- qplot(x=Var1, y=Var2,
           data = melt(cor(combined[,-which(names(FMcorrupt) ==
                                             "country" |
                                             names(FMcorrupt) ==
                                             "wbcode" |
                                             names(FMcorrupt) ==
                                             "prepost")],
                           use="pairwise.complete.obs")),
           fill=value, geom="tile")
```

```
## Warning in cor(combined[, -which(names(FMcorrupt) == "country" |
## names(FMcorrupt) == : the standard deviation is zero
```

```
q + theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ggtitle("Pre 2002 Variable Correlation Heatmap")
```



We note that while there are some interesting relations there - it may not fully pick up the influence of non-linear relationships. There also may be some relation in the amount of violations pre enforcement to post enforcement as this may pick up behaviours of staff or behavioural norms of the embassy which could but may not necessarily so be static.

Explore how your outcome variable is related to the other variables in your dataset. Make sure to use visualizations to understand the nature of each bi-variate relationship.

Correlation of violations to all numeric variables:

Pre:

```
data.frame(cor(log(combined$violations+1),
               combined[,-which(names(combined) == "country" |
                                names(combined) == "wbcode" |
                                names(combined) == "prepost")],
               use="pairwise.complete.obs"))
```

```
## Warning in cor(log(combined$violations + 1), combined[, -
## which(names(combined) == : the standard deviation is zero

##   violations     fines mission      staff   spouse gov_wage_gdp pctmuslim
## 1  0.6073938 0.6058463      NA 0.08040103 0.112787    0.1163731 0.2130604
##   majoritymuslim      trade cars_total cars_personal cars_mission
## 1      0.1491716 -0.1751227  0.2002251    0.06759643    0.2144729
##     pop1998 gdppcus1998       ecaid      milaid    region corruption
```

```
## 1 0.1759187  -0.4315182 -0.02213396 -0.07083804 0.2880506  0.3937382
##        totaid r_africa r_middleeast   r_europe r_southamerica     r_asia
## 1 -0.05922228 0.287135   0.02724507 -0.1711545    -0.07928693 0.05681262
##     distUNplz violations_post fines_post distUNplz_post cars_per_dip
## 1 -0.07987979       0.4270613  0.4302603    -0.08463995     0.206716
```

Post:

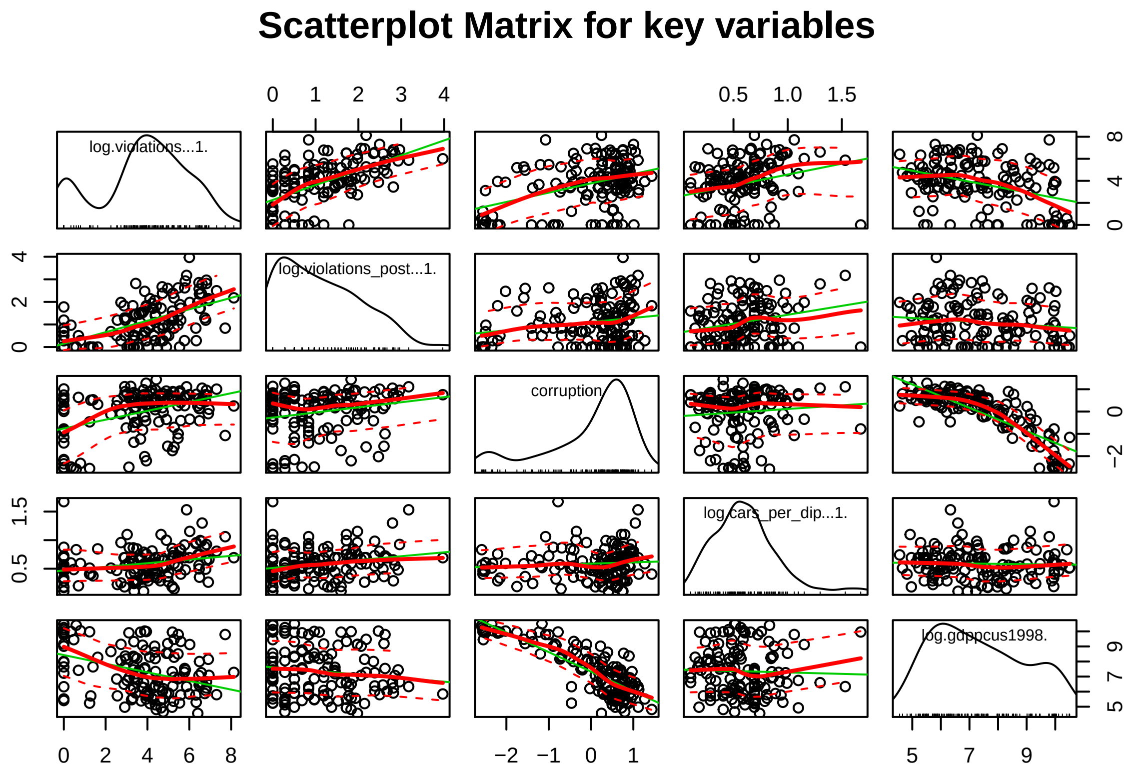
```
data.frame(cor(log(combined$violations_post+1),
               combined[,-which(names(combined) == "country" |
                                names(combined) == "wbcode" |
                                names(combined) == "prepost")],
               use="pairwise.complete.obs"))
```

```
## Warning in cor(log(combined$violations_post + 1), combined[, -
## which(names(combined) == : the standard deviation is zero
```

```
##   violations     fines mission     staff    spouse gov_wage_gdp pctmuslim
## 1  0.3842847 0.3842464      NA 0.1634696 0.1772967   0.07266612 0.2003943
##   majoritymuslim       trade cars_total cars_personal cars_mission
## 1      0.1439178 -0.08595856  0.2342673     0.2020832    0.1419614
##     pop1998 gdppcus1998       ecaid     milaid    region corruption
## 1 0.1859262  -0.1617409 0.003073435 0.01031459 0.2256507  0.1798094
##        totaid  r_africa r_middleeast    r_europe r_southamerica
## 1 0.008569667 0.2538617  -0.02007847 -0.06671115    -0.09662587
##        r_asia   distUNplz violations_post fines_post distUNplz_post
## 1 -0.01535367 -0.05728994       0.8345753  0.8277848     -0.0521851
##   cars_per_dip
## 1    0.2025694
```

We note that in this instance we are comparing non-transformed variables to the transformed explanatory value. So we may not capture the true underlying relationship. We will therefore compare these variables to the variables with the highest correlation via a scatterplot.

```
scatterplotMatrix( ~ log(violations+1) + log(violations_post+1) + corruption +
                     log(cars_per_dip+1) + log(gdppcus1998), data = combined,
                main = "Scatterplot Matrix for key variables")
```



In the pre-enforcement data, we see positive relationships between log of violations and the corruption index and log number of cars per diplomat. We see a negative linear relationship with the log of gdp.

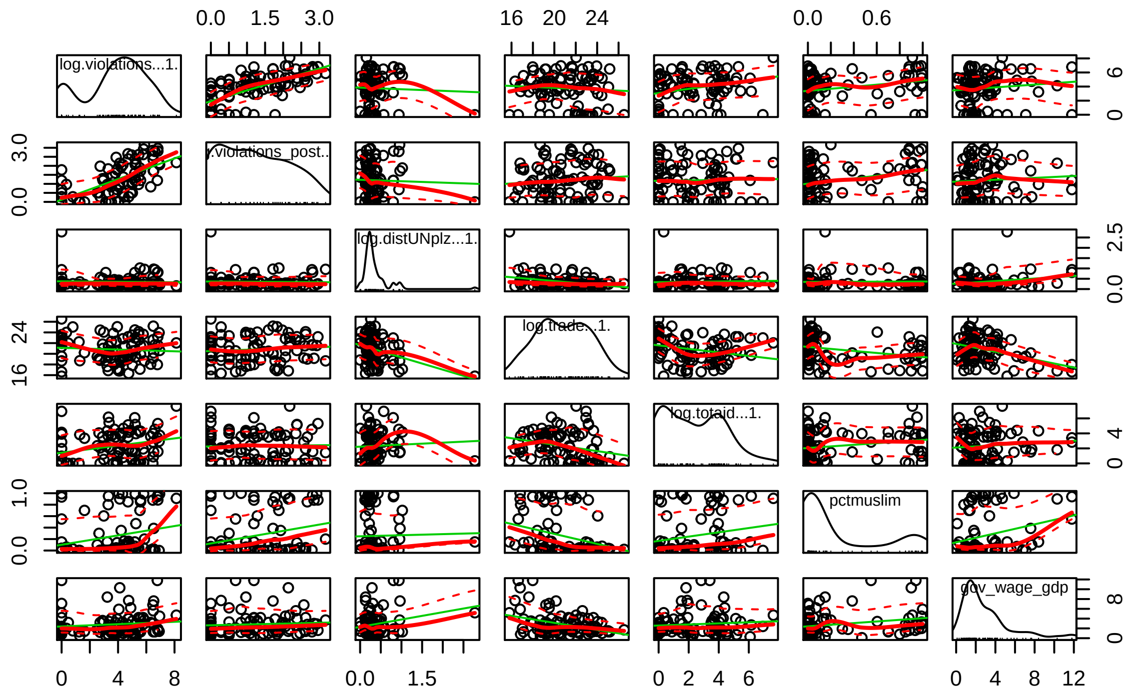
In the post enforcement violations we still see the relationship between corruption and violations but the relationship between cars per staff member and violations is less strong. The relationship with GDP has largely disappeared. This could be because the GDP variable is influential in the period that it is selected, and because we do not have a contemporaneous variable for GDP for the post enforcement period we cannot ascertain a relationship.

We also see a strong relationship between violations in both periods which we will discuss further below.

Scatterplot for non-key relationships

```
scatterplotMatrix( ~ log(violations+1) + log(violations_post+1) +
                     log(distUNplz+1) + log(trade+1) + log(totaid+1) +
                     pctmuslim + gov_wage_gdp, data = combined,
                main = "Scatterplot Matrix for other variables")
```

Scatterplot Matrix for other variables



There appear to be linearly positive relationships with percent Muslim and total aid. We explore the reasons for the percent muslim further in the secondary effects section. The DistUNPlz variable may also relate to violations, but we will need to transform the variable to truly understand it.

Post-enforcement period

Linear Relationship between violations and corruption

```
cor(combined$violations, combined$corruption)
```

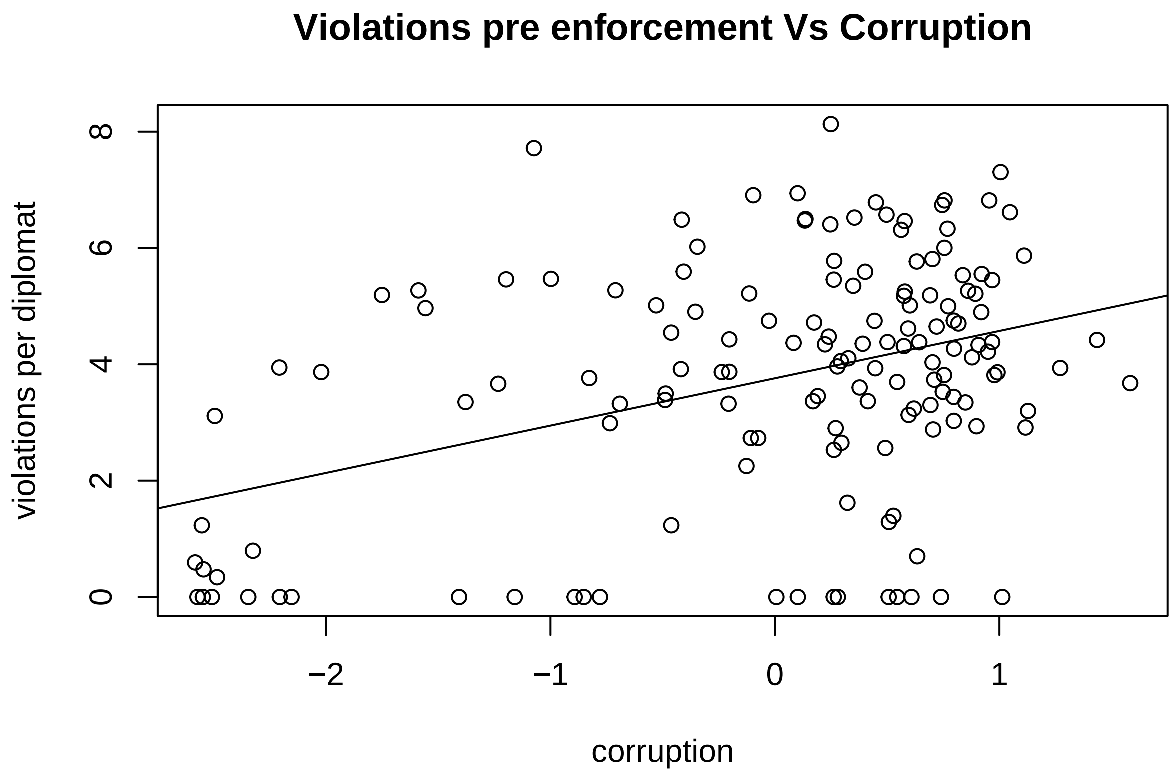
```
## [1] 0.1153411
```

The correlation coefficient appears to be a linear positive relationship, however the numerous zero values, and the skewed distribution of the violations data make the relationship unclear. We will therefore consider a log linear relationship.

```
cor(log(combined$violations+1), combined$corruption)
```

```
## [1] 0.3937382
```

```
plot(jitter(combined$corruption, factor=2),
     jitter(log(combined$violations+1), factor=2),
    xlab = "corruption", ylab = "violations per diplomat",
    main = "Violations pre enforcement Vs Corruption")
abline(lm(log(combined$violations+1) ~ combined$corruption))
```

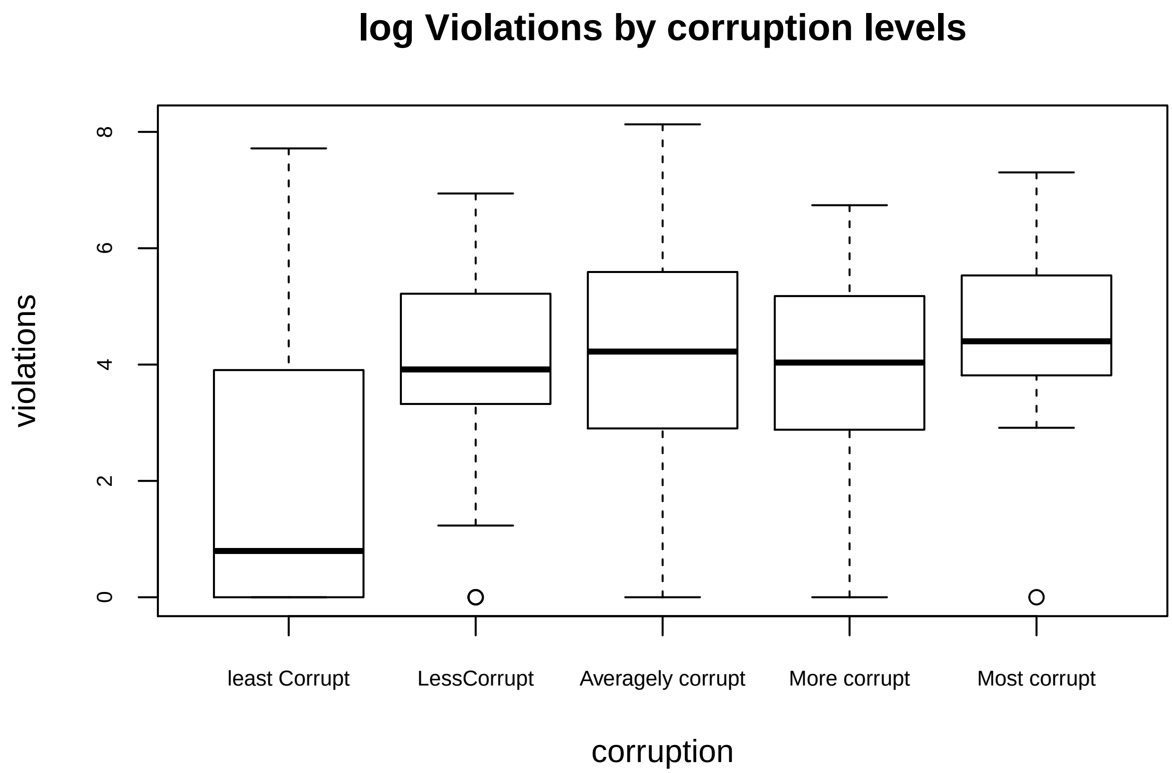


The relationship here is much stronger - the correlation of corruption to an approximate percentage distribution is ~0.4, which is considerably stronger. This implies that a unit of corruption increase has a reasonably strong relationship to a percentage increase in the number of violations.

```
corr_bin = cut(combined$corruption,
               breaks = c(-3,-0.8,0.2,0.5,0.75, Inf),
               labels =c("least Corrupt", "LessCorrupt",
                         "Averagely corrupt", "More corrupt",
                         "Most corrupt"))
summary(corr_bin)
```

```
##     least Corrupt       LessCorrupt Averagely corrupt      More corrupt
##                27                33                26                29
##      Most corrupt
##                34
```

```
boxplot(log(violations+1) ~ corr_bin, data = combined,
      cex.axis = .7,
       main = "log Violations by corruption levels",
       xlab = "corruption", ylab = "violations")
```



We have attempted to split the data reasonably evenly between plots - grouping by quintiles.

Interestingly - the box plot shows there is not substantial difference in violations between the top four approximate quintiles - though the most corrupt has a very short lower whisker. We see that is the least corrupt countries that is driving the overall trend.

Post-enforcement period

```
cor(combined$violations_post, combined$corruption)
```

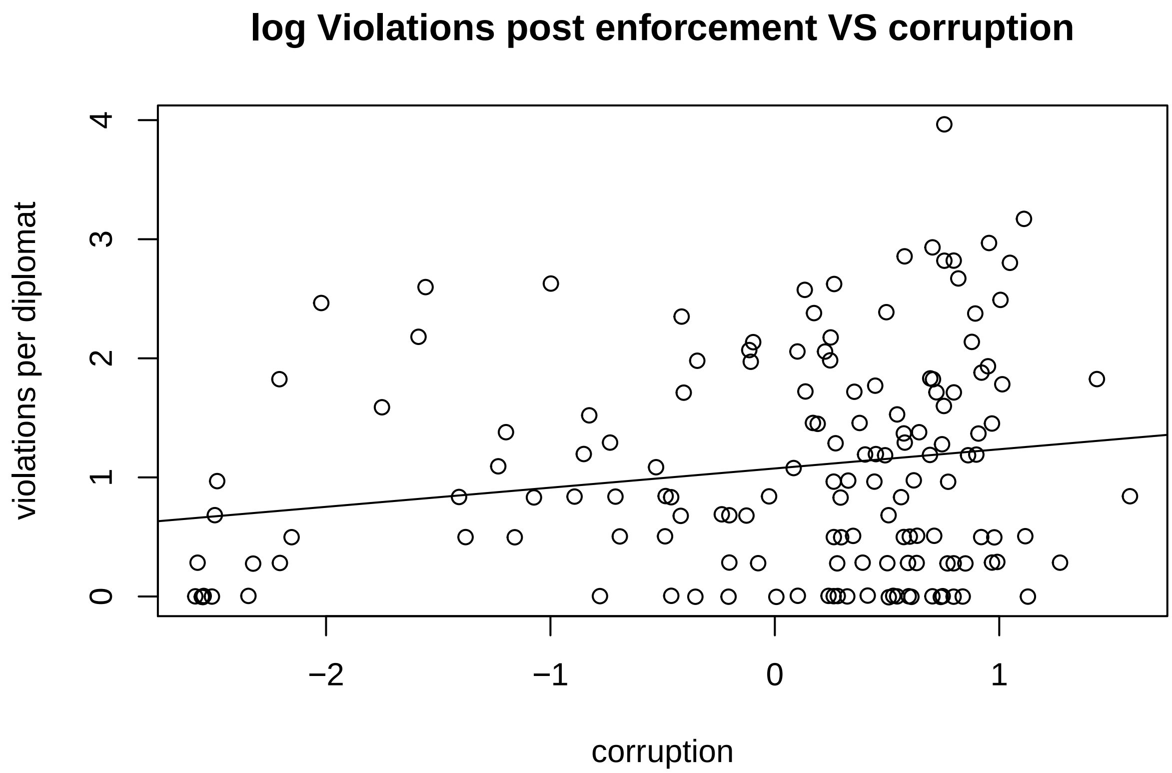
```
## [1] 0.165015
```

The linear relationship appears stronger in the post enforcement period, however, as the violations variable is not normally distributed, we will use the log transformation for a more reliable result.

```
cor(log(combined$violations_post+1), combined$corruption)
```

```
## [1] 0.1798094
```

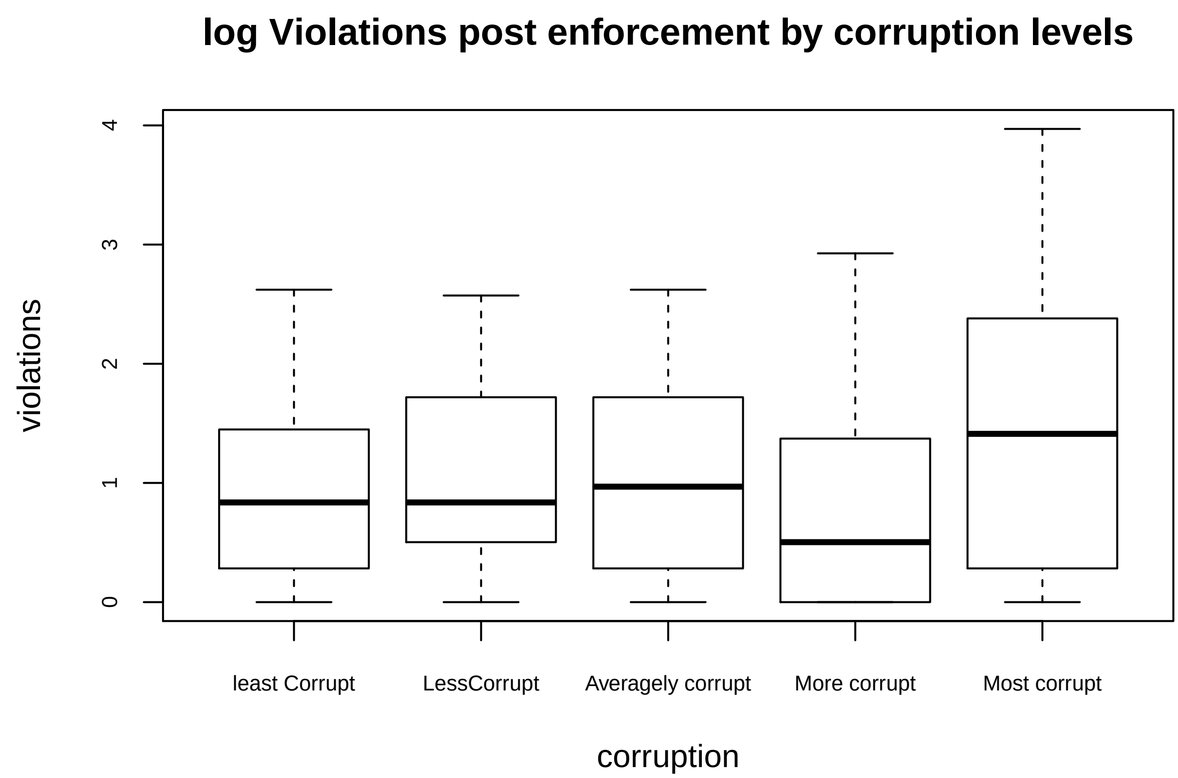
```
plot(jitter(combined$corruption, factor=2), jitter(log(combined$violations_post+1), factor=2),
    xlab = "corruption", ylab = "violations per diplomat",
    main = "log Violations post enforcement VS corruption")
abline(lm(log(combined$violations_post+1) ~ combined$corruption))
```



We still see a somewhat positive relationship for the log transformation, but it is a weaker relationship than that observed pre-enforcement. Post-enforcement, corrupt countries were perhaps less able to rely on their diplomatic immunity.

We can see a small positive relationship between corruption and violations in both the pre and the post data though this becomes, more muddled as corruption increases. The data appears heteroskedastic. We note that we have transformed the variables “cars”, “aid”, “population” and “gdp” to reflect that the data is highly dispersed.

```
boxplot(log(violations_post+1) ~ corr_bin, data = combined,
      cex.axis = .7,
       main = "log Violations post enforcement by corruption levels",
       xlab = "corruption", ylab = "violations")
```



The box-plot confirms what we suspected above - that post enforcement, relative to the least corrupt group, the other groups were less likely to offend in the post enforcement period, excepting the most corrupt countries.

```
summary(combined$violations)
```

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

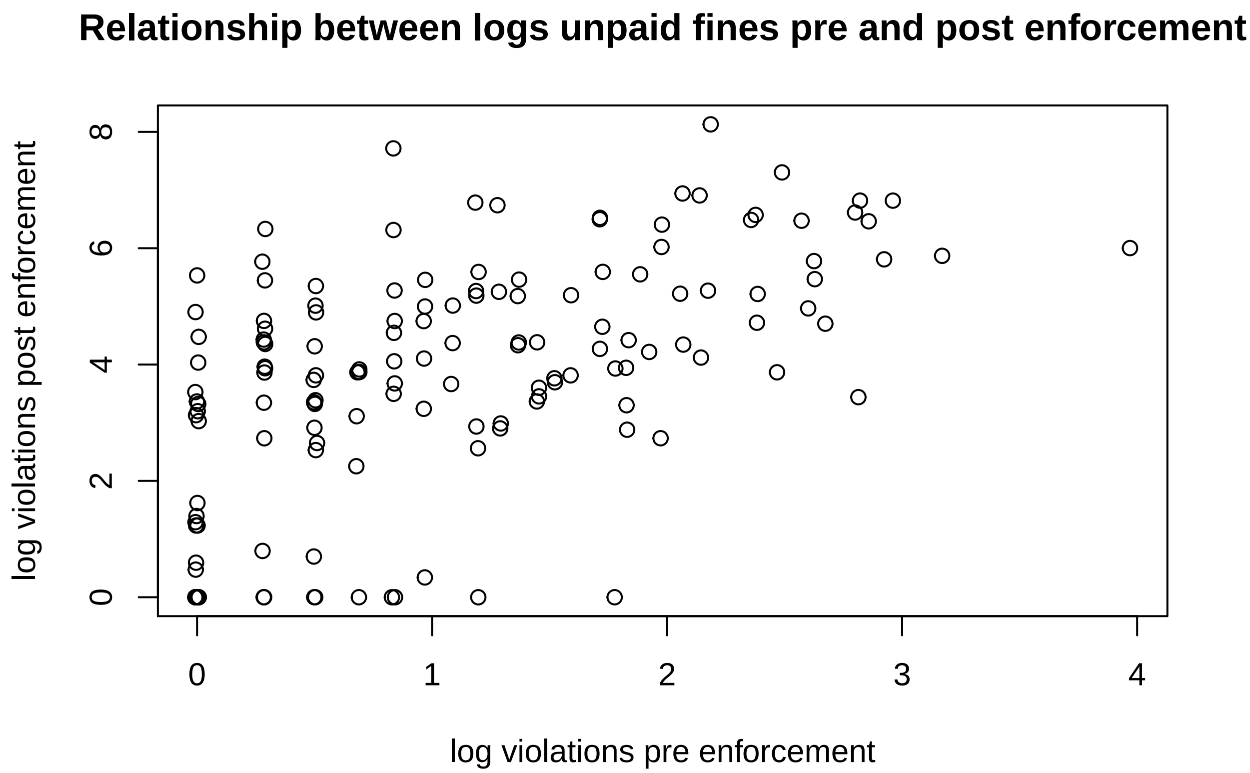
```
summary(combined$violations_post)
```

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

```
cor(log(combined$violations+1), log(combined$violations_post+1))
```

```
## [1] 0.5741702
```

```
plot(jitter(log(combined$violations_post+1), factor=2),
     jitter(log(combined$violations+1),factor=2),
    xlab = "log violations pre enforcement", ylab = "log violations post enforcement",
    main = "Relationship between logs unpaid fines pre and post enforcement")
```



Relationship between the violations data in the post and pre enforcement periods

The correlation coefficient is 0.6 which suggests that despite the time interval and the difference in magnitude of the violations, behaviour of individual staff or behaviours that are passed around the embassy as a whole appears to be somewhat conserved across time periods. This correlation is overstated, as the impact of corruption is being captured, as the corruption index does not change between the periods.

We can transform corruption into a binary variable (<0 or >0) and check to see if there is a statistically significant difference between violations observed for "corrupt" vs "non-corrupt" nations.

T-test over both periods

```
### Relationship between the violations data in the post and pre enforcement periods
#pre
t.test(FMcorrupt[!is.na(FMcorrupt$violations) & FMcorrupt$prepost=="pre" &
                   FMcorrupt$corruption>0,"violations"],
       FMcorrupt[!is.na(FMcorrupt$violations) & FMcorrupt$prepost=="pre" &
                   FMcorrupt$corruption<0,"violations"])
```

```
##
##  Welch Two Sample t-test
##
## data:  FMcorrupt[!is.na(FMcorrupt$violations) & FMcorrupt$prepost ==  and FMcorrupt[!is.na(FMcorrupt$
## t = 1.3867, df = 121.27, p-value = 0.1681
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -38.77452 220.10419
```

```
## sample estimates:
## mean of x mean of y
##  229.1037  138.4388
```

```r
#post
t.test(FMcorrupt[!is.na(FMcorrupt$violations) & FMcorrupt$prepost=="pos" &
                   FMcorrupt$corruption>0,"violations"],
       FMcorrupt[!is.na(FMcorrupt$violations) & FMcorrupt$prepost=="pos" &
                   FMcorrupt$corruption<0,"violations"])
```

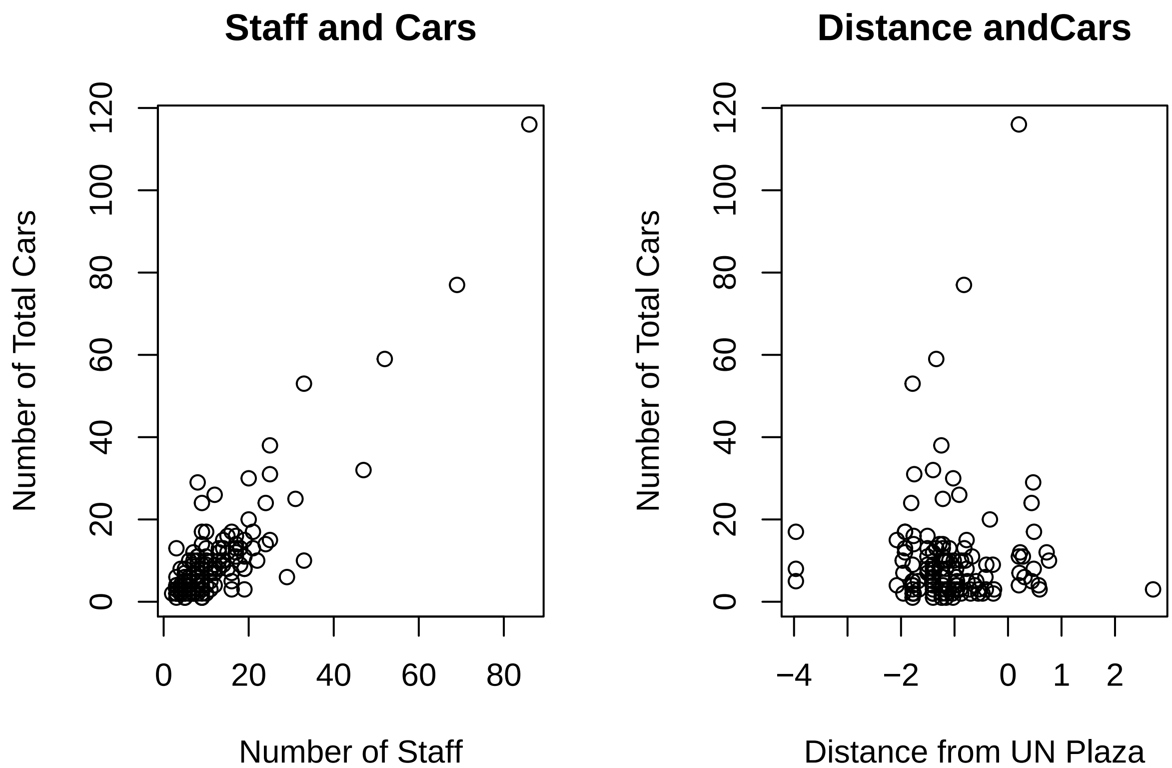
```
##
##  Welch Two Sample t-test
##
## data:  FMcorrupt[!is.na(FMcorrupt$violations) & FMcorrupt$prepost ==  and FMcorrupt[!is.na(FMcorrupt$
## t = 2.0548, df = 146.38, p-value = 0.04168
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.06628704 3.40547985
## sample estimates:
## mean of x mean of y
##  4.281828  2.545945
```

We see that pre 2002, there is p-value of .14, but post 2002, we have a p-value of .04168 when comparing corrupt to non-corrupt country violations. Indicating that corruption may have a more significant impact when there is enforcement.

Analysis of Secondary Effects

Here we explore which variables in the dataset appear to have influence on other variables in the dataset which may demonstrate an indirect influence on the outcome variable, violations. We have seen that cars and corruption have direct influences on the number of violations. Do these variables influence other variables, or are they influenced by other variables? The relationships we found are noted below. All of the relationships use the 'Pre' dataset because these variables do not change between the 'Pre' and 'Post' datasets.

```r
par(mfrow = c(1,2))
plot(pre$staff, pre$cars_total, xlab = "Number of Staff", ylab = "Number of Total Cars",
    main = "Staff and Cars")
plot((log(pre$distUNplz)),(pre$cars_total), xlab = "Distance from UN Plaza",
    ylab = "Number of Total Cars", main = "Distance andCars")
```



Number of cars has a direct positive relationship with number of staff. Also, a lower distUNplz is correlated with a lower number of cars.

The relationships between cars, distUNplz, and number of staff make sense. With lower distances between the country's office and the UN plaza, staff are less likely to require cars. Additionally, if a country has less staff, they will require less cars.

Secondary effects impacting our key variable

We are most interested in what may be biasing our estimate of corruption and some of the other variables.

```
data.frame(cor(combined$violations_post+1,
               combined[,-which(names(combined) == "country" |
                                  names(combined) == "wbcode" |
                                  names(combined) == "prepost")],
               use="pairwise.complete.obs"))
```

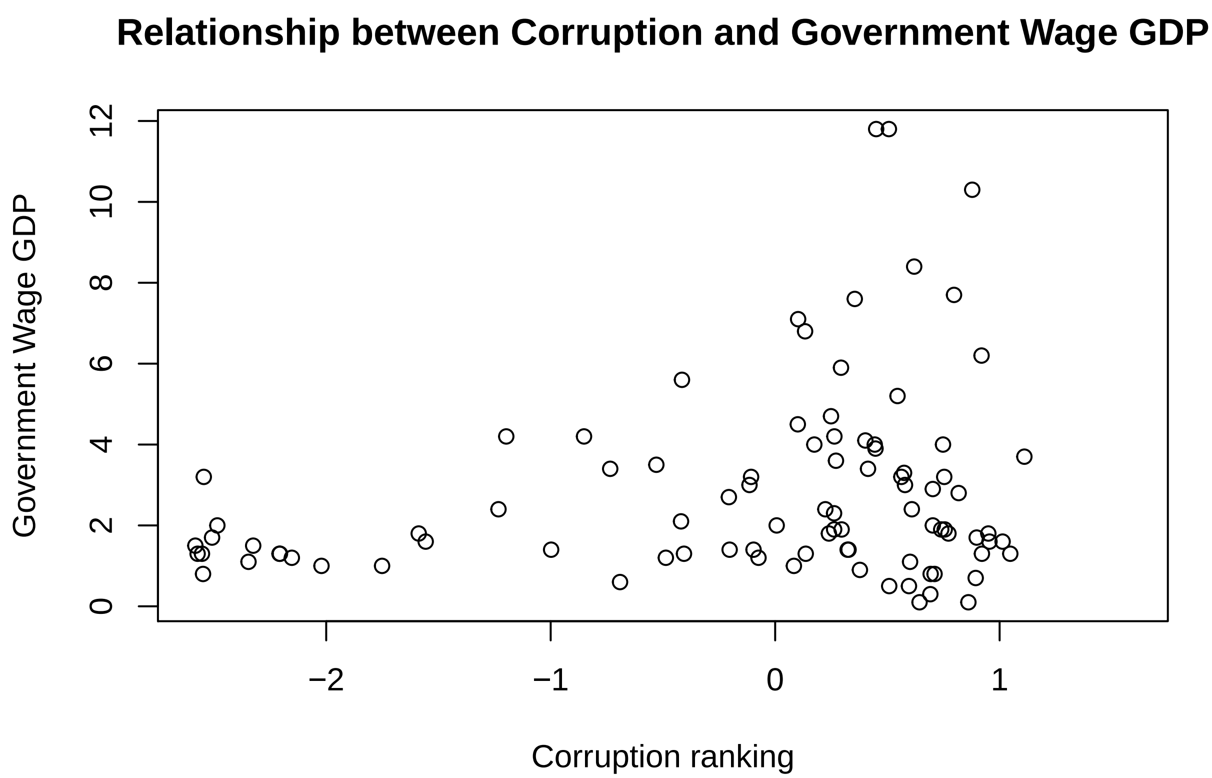
```
## Warning in cor(combined$violations_post + 1, combined[, -
## which(names(combined) == : the standard deviation is zero

##   violations     fines mission      staff     spouse gov_wage_gdp
## 1   0.305556 0.3055133      NA 0.08193433 0.09216729   0.03386871
##    pctmuslim majoritymuslim       trade cars_total cars_personal
## 1 0.1547556      0.1139144 -0.06480323  0.1568631     0.1469039
##   cars_mission   pop1998 gdppcus1998        ecaid       milaid   region
## 1     0.084786 0.1144714  -0.1290432 -0.004933038 -0.01274556 0.207567
##   corruption      totaid  r_africa r_middleeast    r_europe r_southamerica
```

```
## 1   0.165015 -0.01099632 0.2318156  0.003497099 -0.06967283     -0.1044817
##        r_asia  distUNplz violations_post fines_post distUNplz_post
## 1 -0.0361534 0.02879205               1  0.9962667     0.03214421
##    cars_per_dip
## 1    0.2115899
```

Looking above, we see that gdpppcus1998 has a negative relationship with corruption. It also has a negative relationship to our dependent variable, which will have an impact on that relationship. Some of the positive relationship between corruption and violations may be due to the relationship between corruption and GDP: the corruption coefficient could be capturing some of the correlation with GDP.

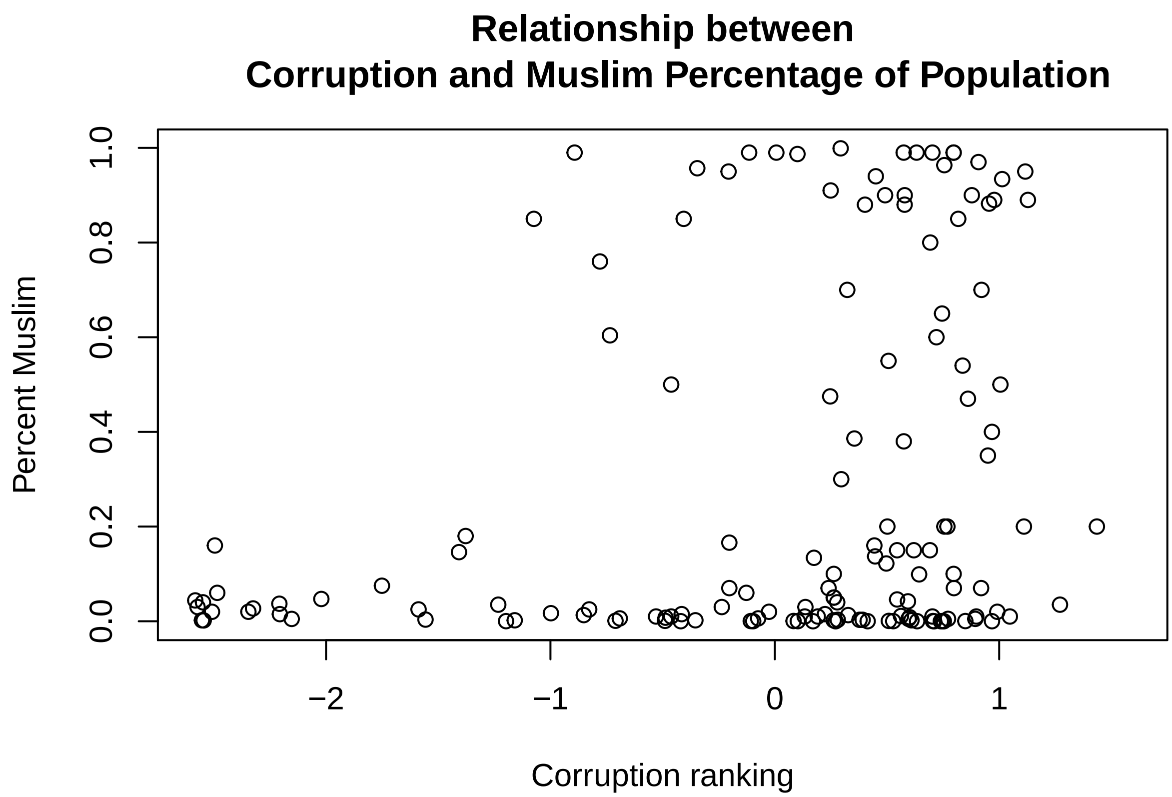
```
plot(pre$corruption,pre$gov_wage_gdp, xlab = "Corruption ranking",
    ylab = "Government Wage GDP",
    main = "Relationship between Corruption and Government Wage GDP")
```



While lower gov_wage_gdp is associated with corruption levels low and high, higher gov_wage_gdp tends to occur in countries that rank as more corrupt.

```
plot(pre$corruption,pre$pctmuslim, xlab = "Corruption ranking",
    ylab = "Percent Muslim", main = "Relationship between
    Corruption and Muslim Percentage of Population")
```

```
## Warning in title(...): font width unknown for character 0x9
```



A similar trend is seen with pctmuslim. Low pctmuslim countries appear in all ranges of corruption, but high pctmuslim countries trend towards the higher end of the corruption range.

However, the correlation between corruption and Muslim percentage appears less strong when you compare it to the dummy variables for region Middle East and Africa, both of which have large Muslim populations. We would expect to see Middle Eastern countries having a strong positive relationship to corruption, if a muslim population was the cause of corruption. Given the variable for Africa is strongly positive this suggests that it is location of high percentage muslim countries in Africa that is related to the corruption level. The level of corruption in Africa may be due to lower GDP levels as well. We would need to regress fully to clean the relationship between violations and corruption independent of the other variables.

Conclusion

We can conclude that there is an apparent difference in how less corrupt and (some) more corrupt countries rely on diplomatic immunity. The coefficient between the violations post and pre-enforcement supports the idea that cultural norms do have an effect on the size of violations.

The direct correlation coefficients of .11 and .16 between violations and corruption is not particularly large. However, they are significant enough to suggest that corruption has some relationship to violations. After transforming the violations logarithmically, so the dependent variable is ~normally distributed we see correlation coefficients of 0.39 and 0.17. This suggests that in the pre-enforcement period, there was a reasonably strong relationship between score in the corruption index and a percentage increase in the violations score.

Without regressing the other variables however it is difficult to determine whether that relationship is. We can see a negative relationship between per capita GDP and corruption in the pre-enforcement time frame. We also see a positive relationship between population and violations and a negative relationship between violations and gdp. This shows that there are several potentially confounding variables in the observed relationship between violations and corruption.

Even the relationship between cars per staff number and corruption may be problematic. If we examine the relationship between the corruption index and cars - we see that more corrupt countries tend to have more cars per staff member and more cars overall. Thus, it seems possible that corrupt countries just have more cars and thus incur more violations.

Given that we do not have updated population, economic, or staff data for the two periods - our data set cannot be considered complete. It is not ideal to compare post enforcement data to GDP data of only 1998. In the post enforcement period, the GDP of that time period independent of 1998 may also have an effect - as the contemporaneous data appears to in 1998.

We conclude that we would need to use more advanced statistical techniques to come to further conclusions. Our exploratory data analysis shows us that corruption may play a part in the number of violations committed, but that there are also several key factors that require further investigation in determining relationship strength/causality.