An Exploratory Analysis of Corruption and Parking Violations

Kenneth Chen, Shiraz Chakraverty, Praba Santhanakrishnan

May 28, 2018

1. Introduction

In this lab assignment, we have access to a unique social experiment to understand relationships between culture and corruption. We define these concepts here and then provide the operational definition to guide our measurements. Our goal is to perform basic exploratory data analysis, based on the following constructs,

i. Every region and country has some form of corruption, a prevailing diplomatic relationship with the UN and by extention, The United States.*

ii. Diplomatic attitudes are tuned into the following:*

- Economic development of nation.
- Current world events, how they effect their nation.*
- Level of crime in nation*
- Population of nation, per capita income and crime index.

iii. The Clinton-Schumer amendment of October 2002 happens about 13 months after WTC terrorist attacks. This change is visible in the dataset as pre and post records,gives a numerical measure of effects of enforcement.

We are motivated to identify strong relationships between various factors (independent variables) and voilations (dependent variable).Ultimately the construct we want to evaluate is the effect of diplomatic culture, aid and economic metrics, population, corruption and parking violations. We would like to explore relationships there variables have to one another and draw some observations based on them

1.1 R Environment Setup

The following packages are required prior to running this project in your Rstudio environment, by running installed.packages() at your R console, you can confirm your list of packages.*

To install the following packages, simply run install.packages('pachage-name')

- List of Packages
- car - Companion to Applied Regression
- Hmisc - Harrell Miscellaneous
- tinytex - To build pdf renders using knit
- tidyverse - To perform more advanced data transformations
- dplyr - data transformations (part of tidyverse)
- corrr - Performing Correlation in R
- knitr - For R markdown tables, graphs and rendering features.
- ggplot2 - For advanced features for descriptive graphs (line, box, dot,etc)

Package library : https://cran.r-project.org/web/packages/

1.2 The Dataset (Summary View)

This section describes the dataset, variable types, number of observations, schema, dimensions. We also delve into data quality, issues, handling of issues we found. Finally we address data processing and preparation.

```
# Load the data
load("Corrupt.Rdata")
df_un = data.frame(FMcorrupt)

# Convert to tidyverse object, tibble for additional sql
# style functionality
tb_un = dplyr::as_tibble(df_un)
```

Dataset size, shape, data gaps, schema and features

- Dataset has 364 rows and 28 columns.
- Shape dimensions are (364, 28).
- Data gaps : blanks(Na represents a blank) ranging from 33 to 180, staggered across variables. We address these on a case by case basis.
- Schema and features:

```
# Show to dimensions ( rows x columns ) of dataset
dim(tb_un)
```

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

```
# Show summary statistics of all fields(variables) in table
str(tb_un)
```

```
## Classes 'tbl_df', 'tbl' and '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 data table is composed of the following variables (variables are fields):

- Volume of parking violations : Maximum number at 3393, average of 100.
- Total number of diplomats(from each country) : MAximum 86, average of 11.
- Individual country corruption index : -2.5 to a maximum of 1.5
- Fines computed in USD: Maximum of 186163, average of 5579 USD.
- Government wages index : 180 NA records, over 35% of dataset is blank, we have to drop this field from analysis.
- Trade with the US:
- Breakdown of Vehicles : official, personal and total
- Population of Country (as of 1998)
- GDP of country (as of 1998)
- Aid to country : military, economic and total US aid
- Country corruption index
- Continent identification : five variables marking each countries geographical location
- Name of the country and country code
- Proportion of Muslim population

1.3 Data Quality assessment and underlying issues

This section shows the quality of the records, issues we found and steps we took to prepare it for exploratory data analysis.

```
# filter for the four columns that have a lot of NA values,
# for cars and diplomat wage index
tb_view_na = select(filter_all(tb_un, any_vars(is.na(.))), prepost,
    corruption, violations, gdppcus1998, totaid, gov_wage_gdp,
    cars_total)

# create a nicely formatted markdown table, the matrix
# slicing shows the first 10 rows for cars and diplomat wage
# blanks.
kable(head(tb_view_na[1:10, ]), caption = "Rows with blank columns values")
```

Table 1: Rows with blank columns values

prepost	corruption	violations	gdppcus1998	totaid	gov_wage_gdp	cars_total
	NA	NA	NA	NA	NA	NA
pre	-0.7794677	0.00000	21143.5391	NA	NA	13
pos	-0.7794677	0.00000	21143.5391	NA	NA	13
	NA	NA	NA	NA	NA	NA
pre	0.7555962	403.28247	344.9218	21.1	NA	8
pos	0.7555962	52.00269	344.9218	21.1	NA	8

The above table shows us top 10 rows of 190, where some columns have a blank or NA in several rows.

```
# [Vertical slicing 1st pass] Update the base tibble by
# removing the columns for diplomat wage index and cars data.
```

```r
tb_un_clean = dplyr::select(tb_un, -gov_wage_gdp, -cars_personal,
    -cars_mission, -cars_total)

# filter from rows with no violation data and store them for
# a view tibble
bad_data_vl = dplyr::filter(tb_un_clean, is.na(violations))

# filter for rows with no pre/post tagging and store them for
# a view tibble
bad_data_pp = dplyr::filter(tb_un_clean, prepost == "")

# create a nicely formatted markdown table, the matrix
# slicing shows the first 10 rows for the first seven
# columns.
kable(head(bad_data_pp[1:10, 1:7]), caption = "Rows with blank pre/post tagging")
```

Table 2: Rows with blank pre/post tagging

wbcode	prepost	violations	fines	mission	staff	spouse
AFG		NA	NA	NA	NA	NA
ATG		NA	NA	NA	NA	NA
BLZ		NA	NA	NA	NA	NA
BRB		NA	NA	NA	NA	NA
BRN		NA	NA	NA	NA	NA
CPV		NA	NA	NA	NA	NA

```r
# create a nicely formatted markdown table, the matrix
# slicing shows the first 10 rows for the first seven
# columns.
kable(head(bad_data_vl[1:10, 1:7]), caption = "Rows with blank violations")
```

Table 3: Rows with blank violations

wbcode	prepost	violations	fines	mission	staff	spouse
AFG		NA	NA	NA	NA	NA
ATG		NA	NA	NA	NA	NA
BLZ		NA	NA	NA	NA	NA
BRB		NA	NA	NA	NA	NA
BRN		NA	NA	NA	NA	NA
CPV		NA	NA	NA	NA	NA

```r
# [Horizontal slicing 2nd pass]update the base tibble by
# removing the rows where violations is empty.
tb_un_clean = dplyr::select(filter(tb_un_clean, prepost != ""),
    everything())

# [Horizontal slicing 3rd pass] Update the base tibble by
# removing rows with prepost blank.
tb_un_clean = dplyr::select(filter(tb_un_clean, !is.na(violations)),
    everything())
```

```
# create a view only tibble to validate post processing
# status of all rows that have NA in at least 1 column
tb_view_na = select(filter_all(tb_un_clean, any_vars(is.na(.))),
    wbcode, prepost, corruption, violations, gdppcus1998, totaid)

# create a nicely formatted markdown table, the matrix
# slicing shows the first 10 rows for cars and diplomat wage
# blanks.
kable(head(tb_view_na[1:10, ]), caption = "Rows with blank columns values post processing")
```

Table 4: Rows with blank columns values post processing

wbcode	prepost	corruption	violations	gdppcus1998	totaid
ARE	pre	-0.7794677	0.0000000	21143.54	NA
ARE	pos	-0.7794677	0.0000000	21143.54	NA
BIH	pre	0.3488850	209.6420593	1075.86	149.4
BIH	pos	0.3488850	0.6541219	1075.86	149.4
CHE	pre	-2.5829878	0.8102109	32975.70	0.0
CHE	pos	-2.5829878	0.0000000	32975.70	0.0

The above tables shows us a total of 10 rows with scattered NA values which we can still utilize as the primary set of independent and dependent variables we are interested in are minimally impacted.

1.4 Data processing and preparation

We performed the following modifications to make the data more uniform. Here are the changes,

1. Removed the 62 rows above where prepost is blank.

2. Removed the 4 rows where violations are blank, without this data, the record is not useful for our analysis.

3. Calculate average violations per nation to perform average analysis per diplomat.

4. Calculate revised trade in millions, population in millions as aid is presented in millions, this steps makes the unit for these to be the same.

5. Vertical slicing and severation of cars and diplomat wage index columns due to excessive blanks.

```
# Create calculated fields using tidyverse functions, round
# floats.
tb_un_revised = dplyr::select(mutate(tb_un_clean, corruption = round(corruption,
    2), avg_viols = round((violations/staff), 0), trade_mil = round((trade/1e+06),
    0), pop_mil = round((pop1998/1e+06), 0), gdp_1000s = round((gdppcus1998/1000),
    2)), everything())

# Create a base tibble with all analysis fields
computed = select(tb_un_revised, country, region, prepost, corruption,
    totaid, ecaid, milaid, avg_viols, trade_mil, pop_mil, gdp_1000s)

# Create a table view only tibble to show the computed fields
# with country reference.
tib_view = dplyr::select(computed, country, corruption, avg_viols,
    totaid, ecaid, trade_mil, pop_mil, gdp_1000s)
```

```r
# create a nicely formatted markdown table, show first 10
# rows and all columns
kable(head(tib_view[1:10, ]), caption = "Sample of revised fields")
```

Table 5: Sample of revised fields

country	corruption	avg_viols	totaid	ecaid	trade_mil	pop_mil	gdp_1000s
ANGOLA	1.05	83	92.3	92.3	2606	12	0.73
ANGOLA	1.05	2	92.3	92.3	2606	12	0.73
ALBANIA	0.92	86	65.0	62.8	27	3	1.01
ALBANIA	0.92	2	65.0	62.8	27	3	1.01
	-0.78	0	NA	NA	3030	3	21.14
	-0.78	0	NA	NA	3030	3	21.14

```r
# test dataset for any further na values

kable(filter_all(tib_view, any_vars(is.na(.))), caption = "Rows with NA values")
```

Table 6: Rows with NA values

country	corruption	avg_viols	totaid	ecaid	trade_mil	pop_mil	gdp_1000s
	-0.78	0	NA	NA	3030	3	21.14
	-0.78	0	NA	NA	3030	3	21.14
BOSNIA-HERZEGOVINA	0.35	35	149.4	97.5	NA	4	1.08
BOSNIA-HERZEGOVINA	0.35	0	149.4	97.5	NA	4	1.08
MONTENEGRO & SERBIA	0.97	39	NA	NA	47	11	0.94
MONTENEGRO & SERBIA	0.97	0	NA	NA	47	11	0.94
ZAIRE	1.58	6	22.4	22.4	NA	48	0.10
ZAIRE	1.58	0	22.4	22.4	NA	48	0.10

2. Univariate Analysis for key variables

Here we review at a glance some key descriptive features of all the variables we have been provided.

- Country and Country code (column name(s) : wbcode , country)

Here we also talk about the regions and boolean flags for each major region. Our goal is to view the depth of the dataset here. Hence we compute a grouped view of countries by region.

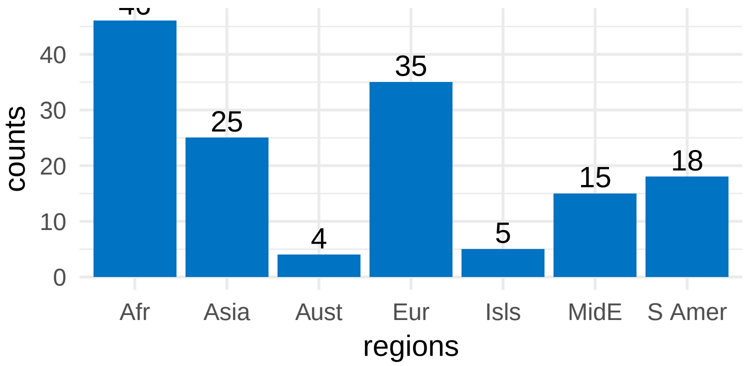
At a glance we observe the regions as following:

1 Caribbean Islands
2 south_americas
3 Europe
4 asia
5 Australia
6 Africa
7 middle east

Each of these continents have a boolean variable : Africa, Middle East, South America, Asia.

```r
# compute the number of countries by region. This is visually
# more useful.
```

```r
country_base <- filter(tb_un_clean, prepost == "pre") %>% group_by(region) %>%
    summarise(counts = n())
# Remove any NA rows

country_base = select(filter(country_base, !is.na(region)), everything())

# Make regions more readable

regions = c("Isls", "S Amer", "Eur", "Asia", "Aust", "Afr", "MidE")

country_base$regions = regions

# Create a bar plot to show the grouped total of countries by
# continental region.
ggplot(country_base, aes(x = regions, y = counts)) + geom_bar(fill = "#0073C2FF",
    stat = "identity") + geom_text(aes(label = counts), vjust = -0.3) +
    theme_minimal()
```



This is a text field where we found a total of 364 rows. There are 60 rows with no values.

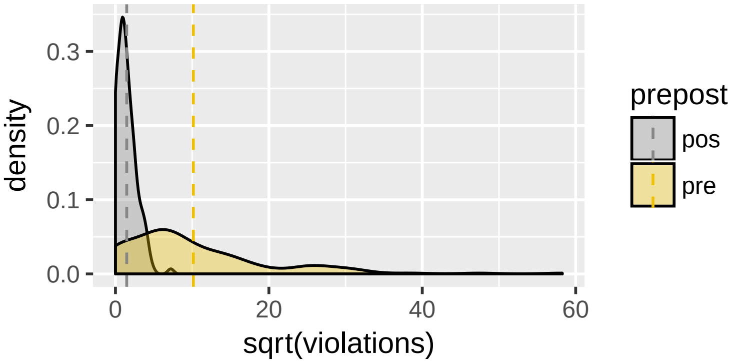
- Pre and Post 2002 records (column name : prepost)

This field tags the row for a pre or post parking enforcement summary of violations.

- Volume of parking violations (column : violations)

(a) Before enforcement : Very high mean, max, a lot of overall violations, however, post enforcement the distribution has a much smaller magnitude. We took a square root of the violations as there are a few very large values that make the graph very hard to review. We clearly see a major decline in the pre vs post number of violations. Also the mean is noteworthy.

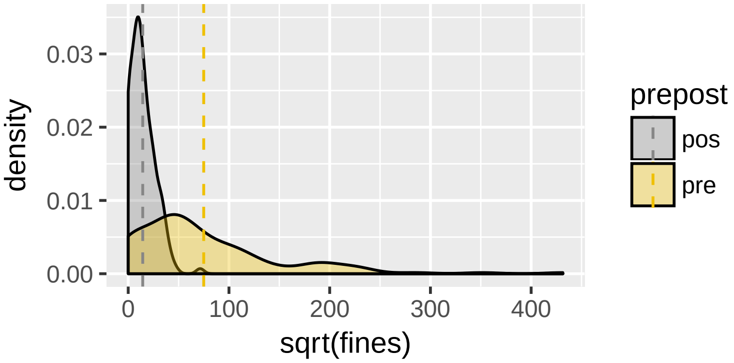
```r
# compute the mean of pre and post violations
v_mean <- tb_un_clean %>% group_by(prepost) %>% summarise(grp.mean = mean(sqrt(violations)))

# Using ggplot object to plot violations

v <- ggplot(tb_un_clean, aes(x = sqrt(violations)))

# Change the filled in color by pre-post and add a mean line
# Using transparent fill: alpha = 0.35
v + geom_density(aes(fill = prepost), alpha = 0.35) + geom_vline(aes(xintercept = grp.mean,
    color = prepost), data = v_mean, linetype = "dashed") + scale_color_manual(values = c("#868686FF",
    "#EFC000FF")) + scale_fill_manual(values = c("#868686FF",
    "#EFC000FF"))
```



- Fines computed in USD (column : fines)

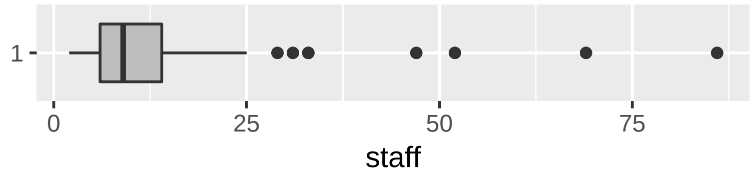
As fines are dependend on the number of violations, we see similar decline in distribution of fines owed after the enforcement. As fines have a very skewed distribution, visually the histogram is hard to review, hence we computer a square root to see the distribution better. We see see that missions have been fined a lot more before enforcement of fines however after the enforcement the missions have dramatically reduced fines owed.

```
# compute the mean of pre and post fines
v_mean <- tb_un_clean %>% group_by(prepost) %>% summarise(grp.mean = mean(sqrt(fines)))

# Using ggplot object to plot fines

v <- ggplot(tb_un_clean, aes(x = sqrt(fines)))

# Change the filled in color by pre-post and add a mean line
# Using transparent fill: alpha = 0.35
v + geom_density(aes(fill = prepost), alpha = 0.35) + geom_vline(aes(xintercept = grp.mean,
    color = prepost), data = v_mean, linetype = "dashed") + scale_color_manual(values = c("#868686FF",
    "#EFC000FF")) + scale_fill_manual(values = c("#868686FF",
    "#EFC000FF"))
```



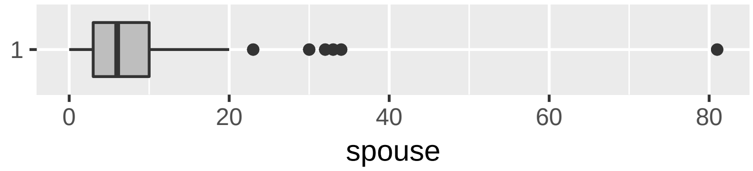
- Diplomatic mission details (columns : staff, spouse)
- Total number of diplomats(from each country) - Majority of missions have under 20 diplomats.

```
# Using ggplot object to plot staffing and family
# distribution

g_staff <- ggplot(tb_un_clean, aes(x = factor(1), y = staff)) +
    geom_boxplot(width = 0.6, fill = "grey") + scale_color_manual(values = c("#00AFBB",
    "#E7B800")) + labs(x = NULL)

g_staff + coord_flip()
```



- total number of family members - Most missions have under 20 family members.

```
# Using ggplot object to plot family distribution

g_sp <- ggplot(tb_un_clean, aes(x = factor(1), y = spouse)) +
    geom_boxplot(width = 0.6, fill = "grey") + scale_color_manual(values = c("#00AFBB",
    "#E7B800")) + labs(x = NULL)

g_sp + coord_flip()
```

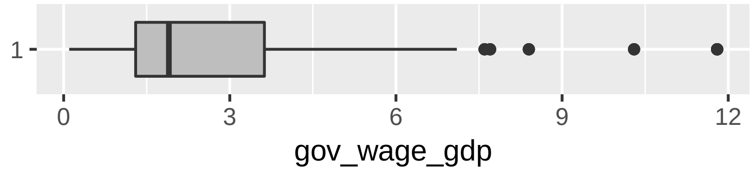


- Government wages index (column : gov_wage_gdp) Here we notice a most diplomats getting paid within 2-4 times the GDP of their country. We have to keep in mind that this index by itself is not helpful as GDP varis a lot by country. Also we decided to remove this field from our analysis as this has over 180 NA values.

We notice that government diplomat compensation varies a lot, from 10% to over 1100% of the GDP. The mean is 280%. Not all nations have a similar cost of living as does the US, so this major disparity between GDP and government diplomat wages is noteworthy. We will further evaluate this in this project.

```
# Using ggplot object to plot wage index distribution
g_wg <- ggplot(tb_un, aes(x = factor(1), y = gov_wage_gdp)) +
    geom_boxplot(width = 0.6, fill = "grey") + scale_color_manual(values = c("#00AFBB",
    "#E7B800")) + labs(x = NULL)

g_wg + coord_flip()
```

```
## Warning: Removed 180 rows containing non-finite values (stat_boxplot).
```

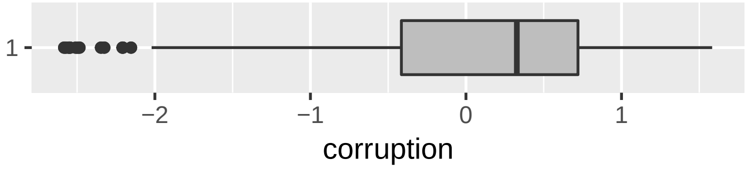


- Individual country corruption index (column : corruption)

We find this is a composite index that comprises of several underlying factors. This type of a variable -2.5 to a maximum of 1.5. We know this is a composite index where a higher number means more corruption. However, we don't understand the negative numbers. We address our concerns below in our summary table.

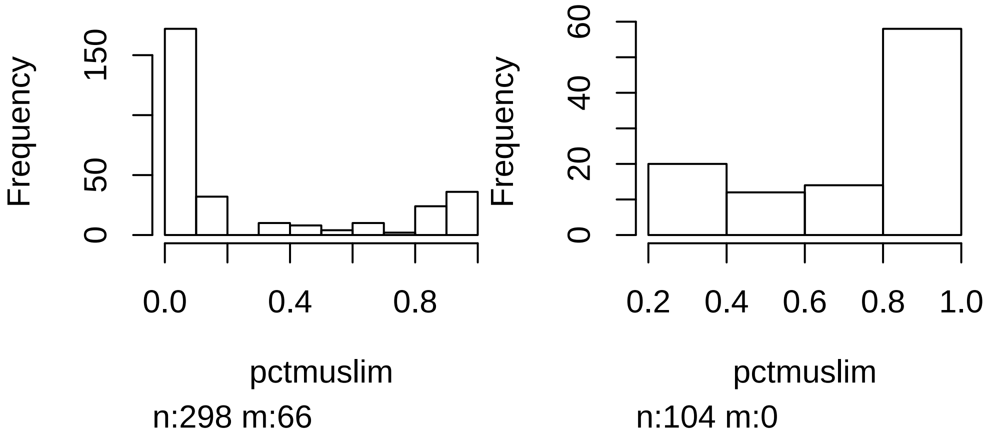
```
# Using ggplot object to plot corruption index distribution

co <- ggplot(tb_un_clean, aes(x = factor(1), y = corruption)) +
    geom_boxplot(width = 0.6, fill = "grey") + scale_color_manual(values = c("#00AFBB",
    "#E7B800")) + labs(x = NULL)
```

```
co + coord_flip()
```



- Proportion of Muslim population
- Percentage of Muslim population - We see in the 2 histograms the distribution. The first has all nations where we see over 150 nations with a 0. Hence we build a second histogram with at least 20% population muslim. This view shows us the distribution of over 75 nations with at least 60% muslim population.
- Majority Muslim population - this is a boolean 0 or 1 flag to indicate majority are muslim.

```
hist(select(tb_un, pctmuslim), breaks = 0:1 - 0.01, main = "Percentage of Muslim Population",
    xlab = NULL)
hist(select(filter(tb_un, pctmuslim > 0.2), pctmuslim), breaks = 0:1 -
    0.01, main = "At least 20% of population is Muslim ", xlab = NULL)
```



- Trade with the US: the trade relationships have a massive range from less than 100000 to several billions.

```
summary(tb_un_clean$trade)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's 
## 0.000e+00 8.911e+07 5.194e+08 1.025e+10 4.796e+09 3.290e+11         4
```

- Breakdown of Vehicles : official, personal and total
- Total number of cars
- Breakdown of person and official cars

```
print("Personal cars")
```

```
## [1] "Personal cars"
```

```
summary(tb_un$cars_personal)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   0.000   1.000   2.000   5.324   6.000  64.000      86
```

```
print("Mission cars")
```

```
## [1] "Mission cars"
```

```
summary(tb_un$cars_mission)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   2.000   3.000   5.144   6.000 116.000      86
```

```
print("Total cars")
```

```
## [1] "Total cars"
```

```
summary(tb_un$cars_total)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    3.00    7.00   10.47   12.00  116.00      86
```

- Population of Country (as of 1998) : We find a large range here from population into just under half a million to over billion people.

```
summary(tb_un_clean$pop1998)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 5.308e+05 3.815e+06 8.852e+06 3.655e+07 2.341e+07 1.242e+09
```

remove duplicates due to pre/post * GDP of country (as of 1998) : We notice here extremely poor nations with the lowest GDP as 95, a mean of about 5000 and as high as 36485. *We notice here too a huge disparity between nations. At the lowest end we see a GDP of only 95, average of 5236 and maximum of 36485. To equalize this a bit, we will compute a total compensation using the wage index by multiplying wage index to gdp, which together will give us a sense of total compensation. This allows us to use the variable better as the index while very useful does not help us understand the poverty or wealth of nations and their diplomats income.*

```
summary(tb_un_clean$gdppcus1998)
```

```
Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
```

95.45 412.07 1374.88 5044.09 4936.62 36485.64

```
kable(select(arrange(filter(tb_un_clean, prepost == "pre"), gdppcus1998),
    gdppcus1998, country)[1:10, ], caption = "Lowest GDP")
```

Table 7: Lowest GDP

gdppcus1998	country
95.44793	ETHIOPIA
101.49330	ZAIRE
105.59200	BURUNDI
123.56780	LIBERIA
137.54359	SIERRA LEONE
143.73100	GUINEA
144.01669	TAJIKISTAN
168.54730	MALAWI
182.60890	NIGER
187.97700	ERITREA

```
kable(select(arrange(filter(tb_un_clean, prepost == "pre"), desc(gdppcus1998)),
    gdppcus1998, country)[1:10, ], caption = "Highest GDP")
```

Table 8: Highest GDP

gdppcus1998	country
36485.64	JAPAN
35855.47	
32975.70	SWITZERLAND
28281.00	DENMARK
24806.11	
23025.60	UNITED KINGDOM
22482.70	AUSTRIA
21942.73	NETHERLANDS
21717.66	GERMANY
21413.84	FINLAND

- Aid to country :
- military : We notice that aid to have a massive range, while the mean is relatively small at 0.2 million, we find nations receiving no aid, over 75% of military aid walls below 0.775 million.
- economic : Here we see the mean at 49 million and about 75% of aid below 40 million. There are some nations reseiving very high amount of economic aid at 1026 million(to Columbia)
- total US aid : Here we find 75% of all aid below 42 million with the highest aid to Israel, Egypt and Colombo.

```
print("Economic aid")
```

[1] "Economic aid"

```
summary(tb_un_clean$ecaid)
```

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's 0.00 0.00 8.70 49.27 40.30 1026.10 4

```
print("Military aid")
```

[1] "Military aid"

```
summary(tb_un_clean$milaid)
```

```
Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
```

0.000 0.000 0.200 33.048 0.775 3120.000 4

```
print("Total aid")
```

[1] "Total aid"

```
summary(tb_un_clean$totaid)
```

```
Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
```

0.000 0.325 9.000 82.320 42.950 4069.100 4 *Index variable for 'distUNplz' - Insufficiant information about this column.

2.1 Anamolies

i) Violations : We found float or decimals in violations, which seemed like an error, typically a parking violation is a whole number and not given in fractions.

ii) Corruption index is a composite variable, meaning we don't fully understand why a country is -2 vs 0 vs 1. zero surely does not mean no corruption, -2 does not mean negetive corruption.

iii) Gov wage index is a simple measure of how many times the diplomats wages are of their country's GDP. Oddly, we found a huge range here, which is a function of the huge range of GDP. As we have over half of the dataset blank for this index, we unfortunately decided to not use this variable.

iv) One variable we could not understand is distUNplz : This also looks like an index, we could not extrapolate what this means, and as is did not find a significant correlation to violations. As we don't know what this could be made up of and we found almost no correlation, we have put this variable aside in the lack of more information.

2.2 Coding issues, Erroneous values

i) Missing (Na, blanks) : We found rows for country code, prepost, violations, country to have blanks. We have a very small dataset, with summary values for each country with means where we have blanks we don't have a way to know which country we are looking at. Hence we had to remove these rows.

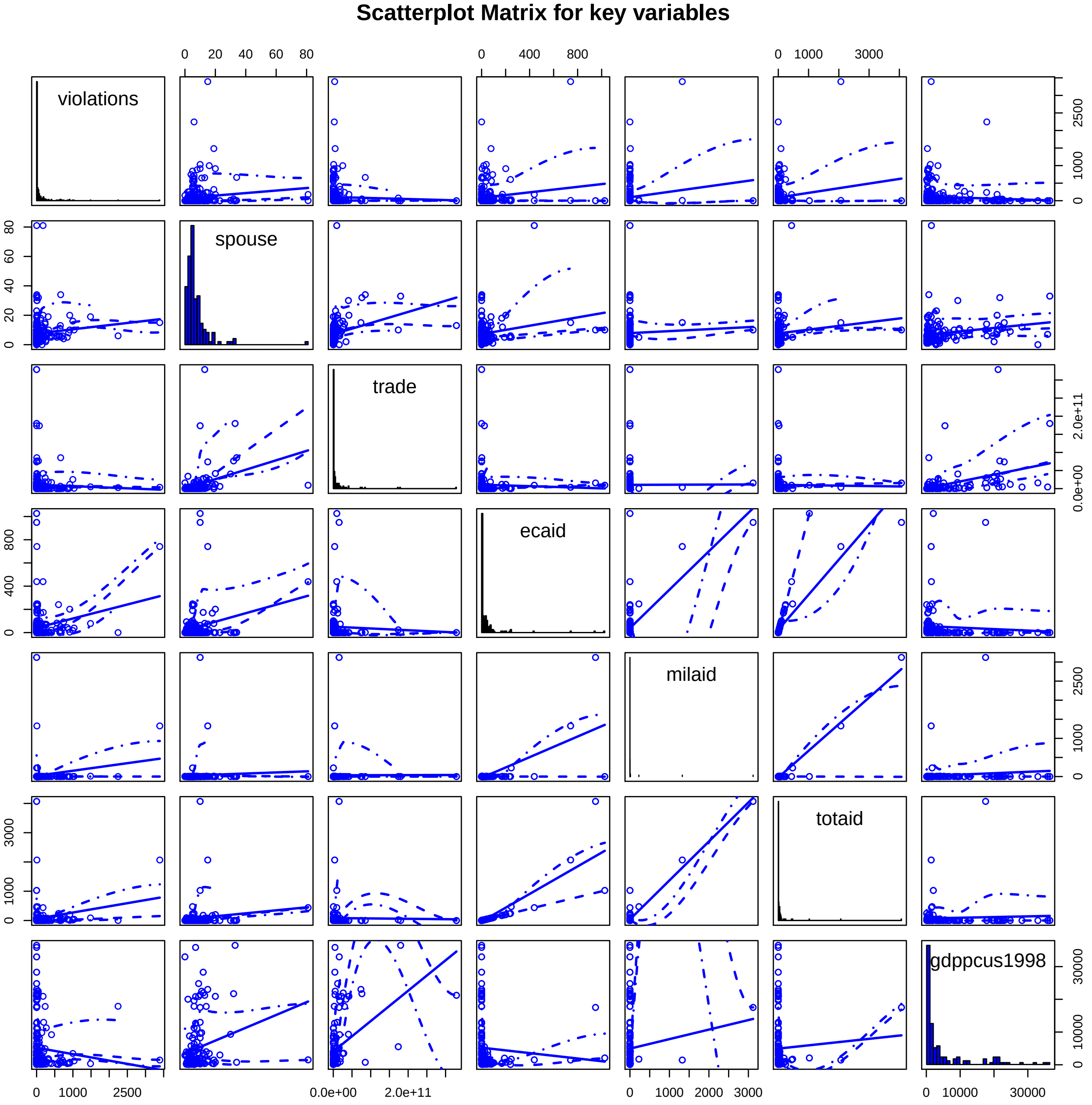
ii) Aid is in millions, we can safely assume, however we have population as an exact total number. We decided to convert population also in millions.

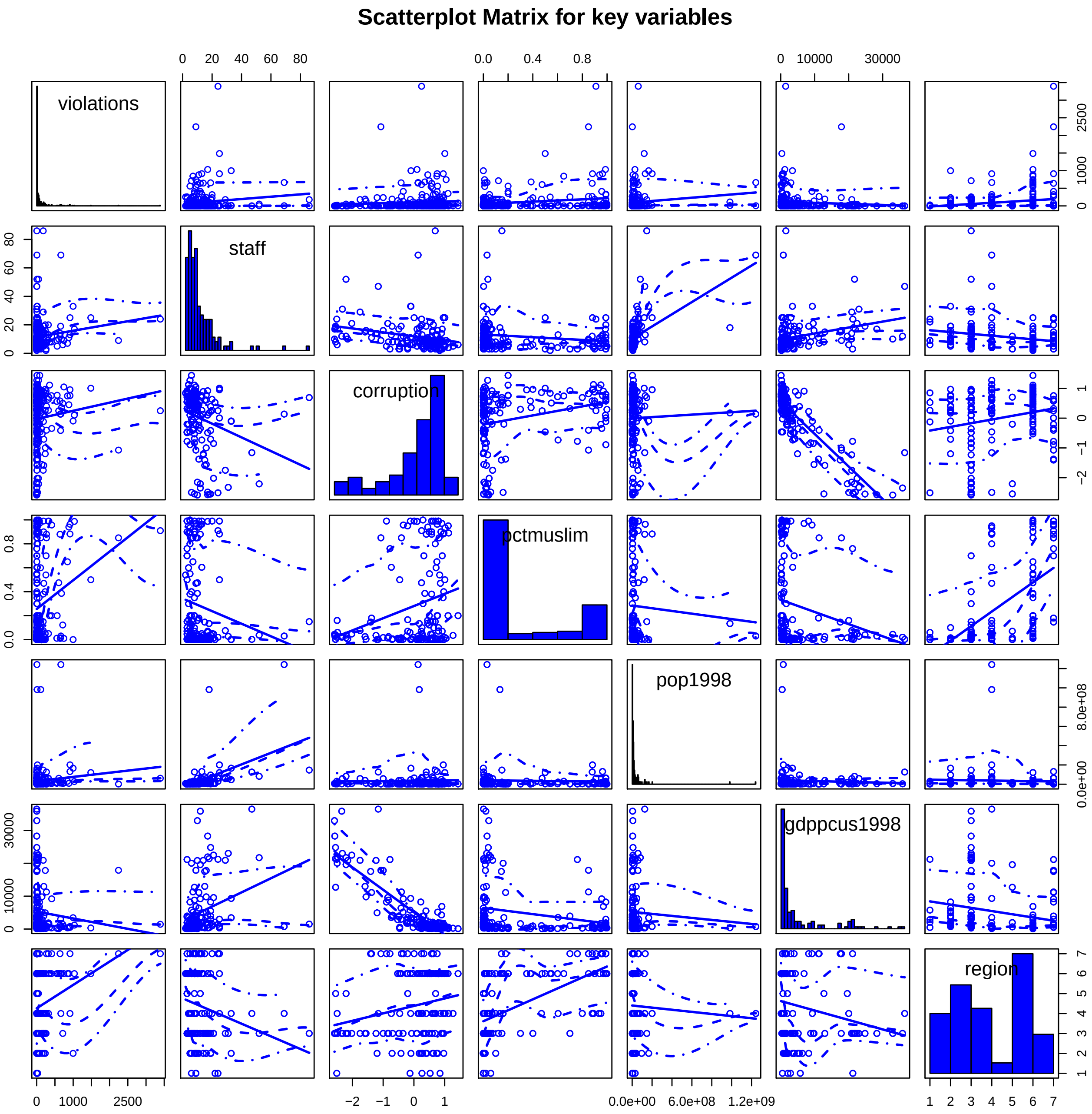
iii) Boolean variables : There are five variables to denote the continental location of countries. We decided to not use these as they confuse our correlation computation. We decided to use the region which has values 1-7 for each continent. This seemed to be consist.

iv) We have a redundant variable for country name and codes. We kep these but this simply took up space in the dataset. In a much larger dataset, we may have to choose to just use the code and save space/memory.

This section needs clarification, modification of variables to improve relationships ## Analysis of key relationships

** USE ME Our first step is preliminary check across all key variables such as violations, staff and corruption. Interestingly, we found that there is no immediate evidence that the more the number of diplomats, the higher the violations. Most of the violations appears clustered at the lower bounds of the staff number between 0 and 20. However we observed an interesting pattern between violations and corruption. The more corrupt the country is, i.e., indicated by the corruption index, the more likely we would see the violation events. **


Scatterplot Matrix for key variables
violations
spouse
trade
ecaid
milaid
totaid
gdppcus1998
0
20
40
60
80
0
400
800
0
1000
3000
0
1000
2500
0.0e+00
2.0e+11
0
1000
2000
3000
0
10000
30000
0
20
40
60
80
0
400
800
0
1000
3000
0
1000
2500
0.0e+00
2.0e+11
0
10000
30000



Specific questions we have identified for exploration: ** Need to filter responses and fill this one**

(a) Was there a relationship between corruption and parking violations? Y

(b) How does the number of diplomats contribute to the frequency of violations? Nothing

(c) Does the legislative change in October 2002 dramatically change volume of violations? Yes, show table and graph

(d) Does the ranking of corruption index (descending order) show a relationship to the volume of parking violations(per diplomat)? -revise table 11 and add corruption to show this, there is a relation here.

(e) Does the level of aid to the country or trade with country show relationship to the volume of parking violations(per diplomat) -No, refer to correlation absence

(f) Does the country gdp, diplomat wage have a relationship to corruption index? i.e. what could have a statistical correlation to a culture of engaging in negligent acts of corruption. - Yes there is, refer to scatterplots

(g) Does WTC attack impact on parking violations? - Yes. Refer to charts

** DELETE ME(h) Which country have the largest diplomatic footprint, including family and is there a relationship with violations ?**

** DELETE ME(i) What is the ralationship between GDP and diplomatic wage ?**

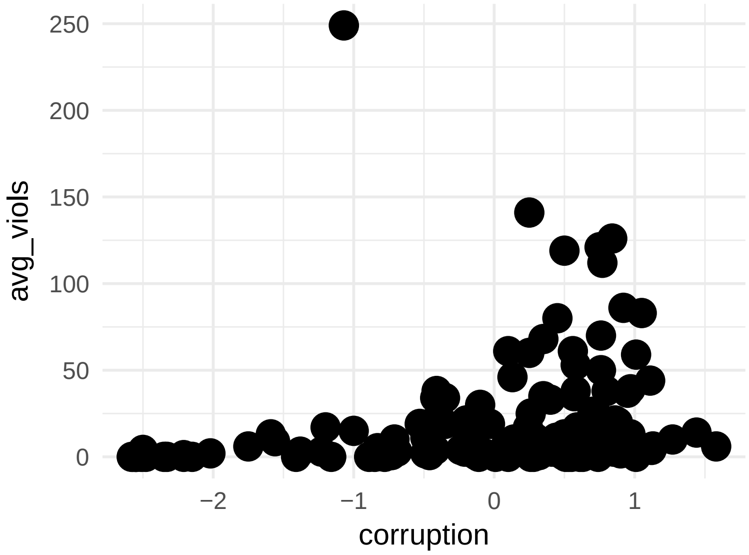
** DELETE ME(j) What is the relationship between economic aid, military aid and other country data like population, GDP?* **

** DELETE ME (j) Which country have more cars? If so, more cars means more staff?**

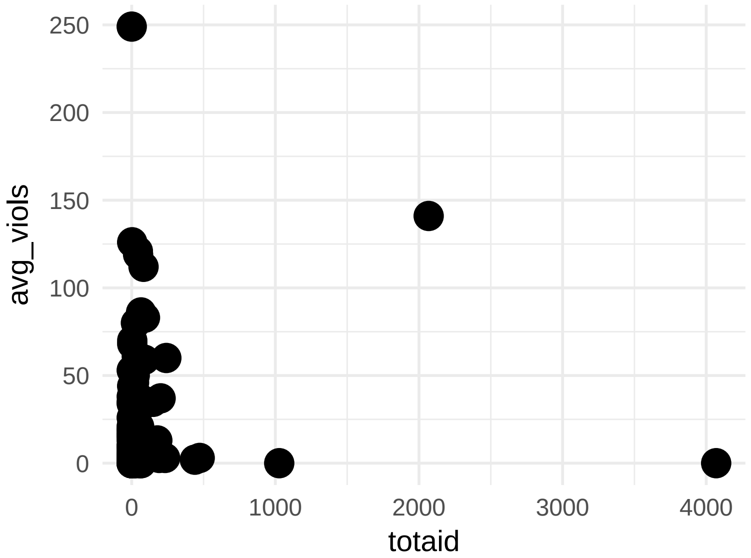
Keep the pre charts and remove post as they are influenced heavily by enforcement

Table 9: Correlation - Pre 2002

rowname	corruption	totaid	avg_viols	trade_mil	gdp_1000s
corruption	NA	-0.0407963	0.1830667	-0.3382683	-0.8619694
totaid	-0.0407963	NA	0.0855975	-0.0119217	0.0486442
avg_viols	0.1830667	0.0855975	NA	-0.1196477	-0.1518583
trade_mil	-0.3382683	-0.0119217	-0.1196477	NA	0.4111201
gdp_1000s	-0.8619694	0.0486442	-0.1518583	0.4111201	NA



```
## Warning: Removed 2 rows containing missing values (geom_point).
```



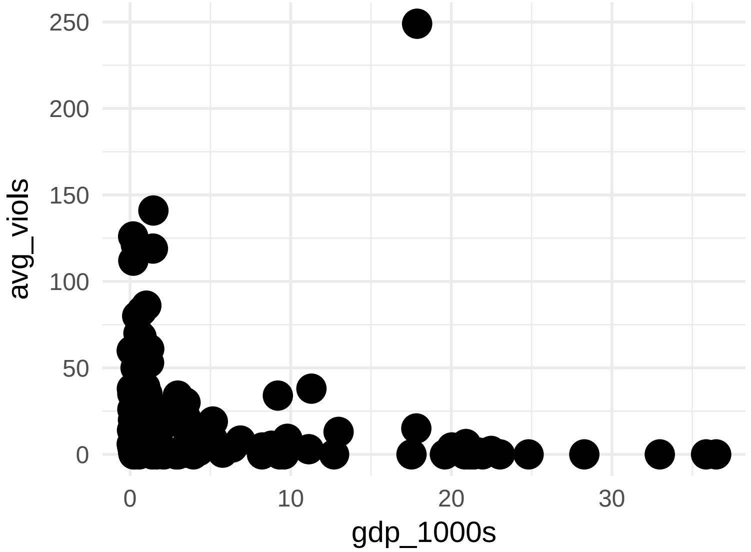
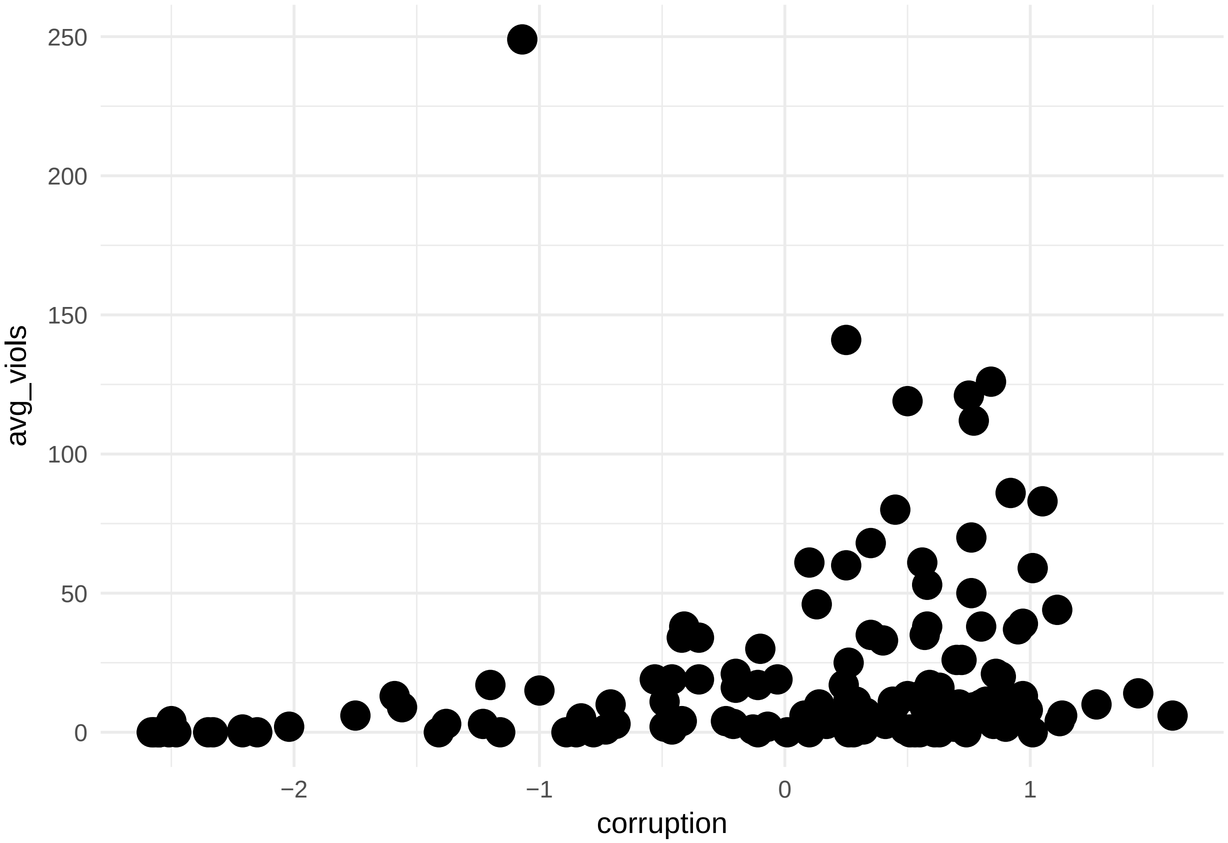
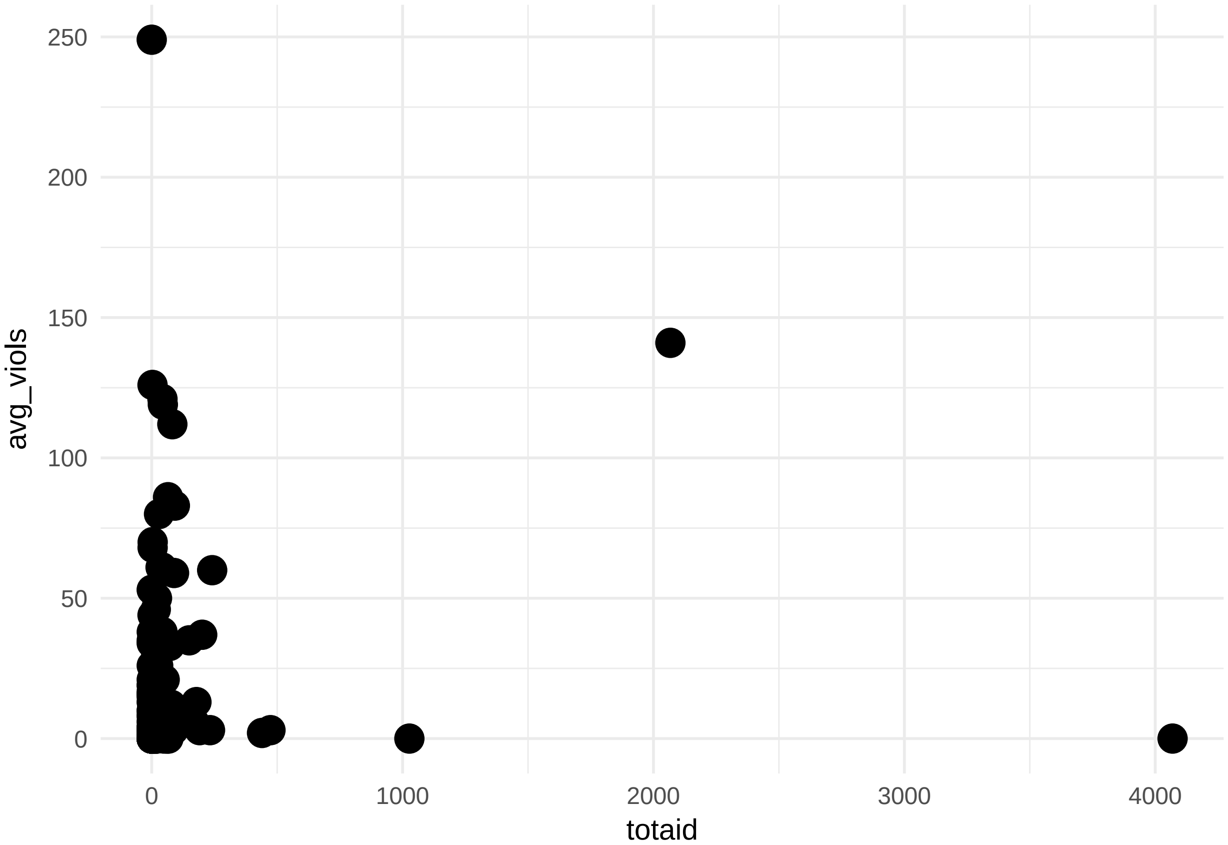


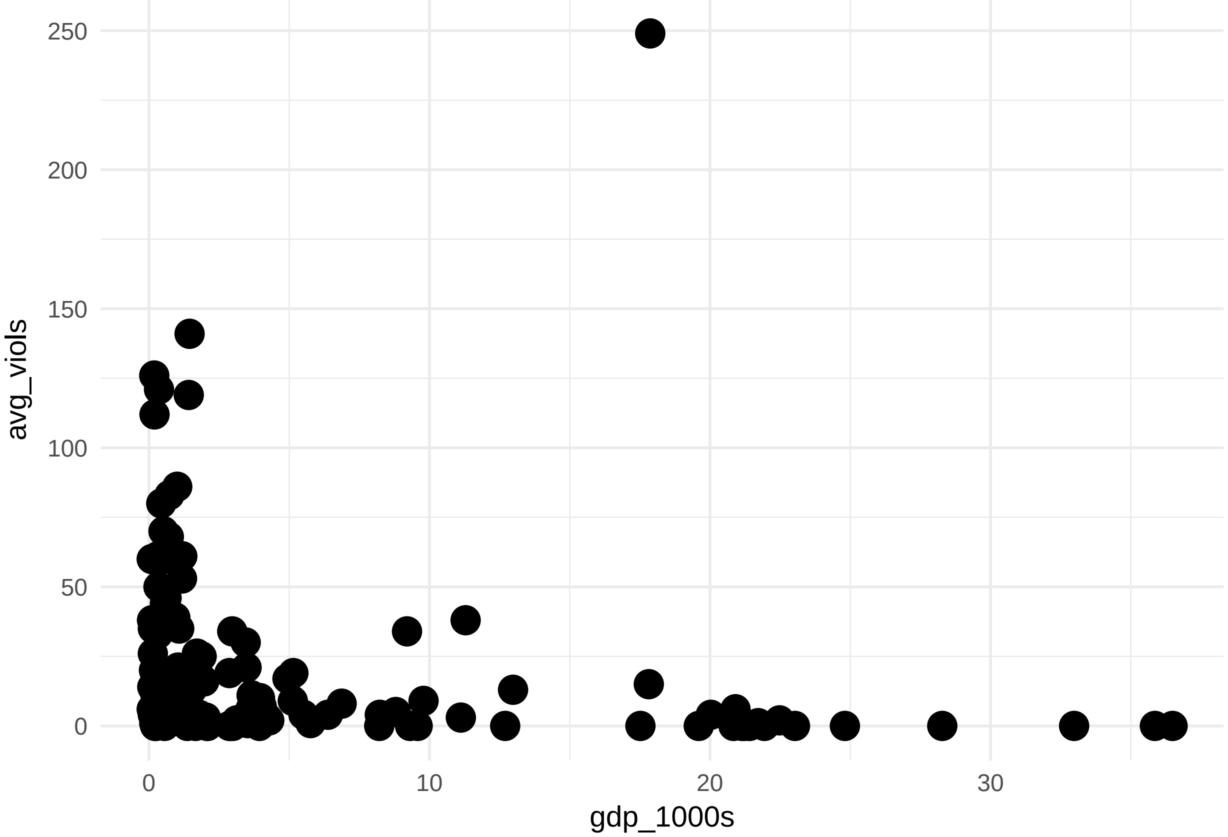
Table 10: Correlation - Post 2002

rowname	corruption	totaid	ecaid	milaid	avg_viols	trade_mil	pop_mil	gdp_1000s
corruption	NA	-0.0407963	0.0875437	-0.0996862	0.2110147	-0.3382683	0.0270581	-0.8619694
totaid	-0.0407963	NA	0.8429109	0.9638705	-0.0543492	-0.0119217	0.0154283	0.0486442
ecaid	0.0875437	0.8429109	NA	0.6691352	-0.0544377	-0.0393562	0.0704715	-0.0726040
milaid	-0.0996862	0.9638705	0.6691352	NA	-0.0481151	0.0030107	-0.0135790	0.1031294
avg_viols	0.2110147	-0.0543492	-0.0544377	-0.0481151	NA	-0.0929531	-0.0005111	-0.1534943
trade_mil	-0.3382683	-0.0119217	-0.0393562	0.0030107	-0.0929531	NA	0.2116664	0.4111201
pop_mil	0.0270581	0.0154283	0.0704715	-0.0135790	-0.0005111	0.2116664	NA	-0.0500077
gdp_1000s	-0.8619694	0.0486442	-0.0726040	0.1031294	-0.1534943	0.4111201	-0.0500077	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA



```
## Warning: Removed 2 rows containing missing values (geom_point).
```


avg_viols
250
200
150
100
50
0
0
1000
2000
3000
4000
totaid


avg_viols
250
200
150
100
50
0
0
10
20
30
gdp_1000s

Analysis of Secondary Effects (10 pts)

What secondary variables might have confounding effects on the relationships you have identified? Explain how these variables affect your understanding of the data.

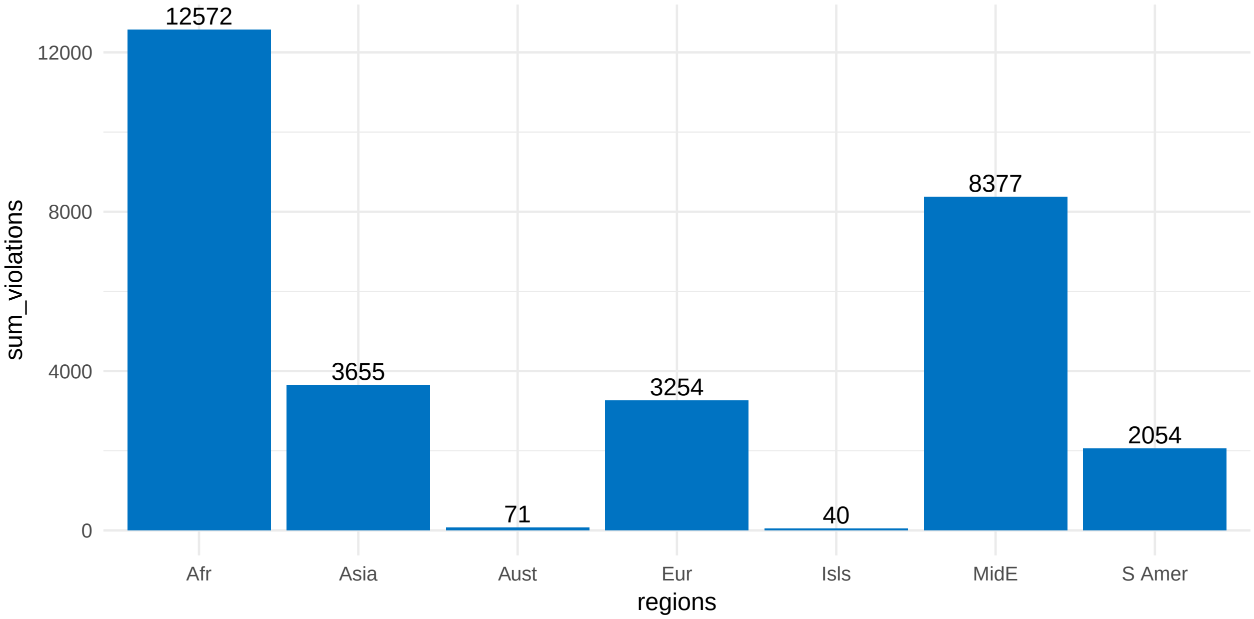
- Show the relation here with the top 25 offenders, with corrption and violations, pre and post 9/11 explain the secondard effect of a major world event and change of US law. ** Only show top 20 offenders and show corruption index**

EXAMPLE 1

Table 11: Parking Violation side by side (pre / post enforcement)

country	pre_2002_violations	pos_2002_violations
KUWAIT	249	0
EGYPT	141	0
CHAD	126	0
SUDAN	121	0
BULGARIA	119	2
MOZAMBIQUE	112	0
ALBANIA	86	2
ANGOLA	83	2
SENEGAL	80	0
PAKISTAN	70	1
IVORY COAST	68	0
MOROCCO	61	0
ZAMBIA	61	0
ETHIOPIA	60	1
NIGERIA	59	0
SYRIA	53	1
BENIN	50	7
ZIMBABWE	46	1
CAMEROON	44	3
MONTENEGRO & SERBIA	39	0
BAHRAIN	38	1
BURUNDI	38	0
MALI	38	1
INDONESIA	37	1
BOSNIA-HERZEGOVINA	35	0

EXAMPLE 2



(b) How does the number of diplomats contribute to the frequency of violations and which country diplomat offended the most parking violations?

As we observe that there are countries with the total number of diplomats at NA, we are interested in the average number of parking violations per individual diplomats. In order to do so, we divided the violations variable by the staff number in each country. However as there are some missing value in these two variables, we first created a subdata which do not have a missing value in two critical variables, i.e., violations and staff.

```
subcases_per_dip = !is.na(FMcorrupt$violations) & !is.na(FMcorrupt$staff)
FM_subcases_per_dip = FMcorrupt[subcases_per_dip, ]
FM_subcases_per_dip$vpd = (FM_subcases_per_dip$violations/FM_subcases_per_dip$staff)
summary(FM_subcases_per_dip)
```

```
##     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               mission      staff            spouse      
##  Min.   :     0.00   Min.   :1   Min.   : 2.00   Min.   : 0.000  
##  1st Qu.:    65.41   1st Qu.:1   1st Qu.: 6.00   1st Qu.: 3.000  
##  Median :   579.72   Median :1   Median : 9.00   Median : 6.000  
##  Mean   :  5579.60   Mean   :1   Mean   :11.81   Mean   : 7.758  
##  3rd Qu.:  2999.05   3rd Qu.:1   3rd Qu.:14.00   3rd Qu.:10.000  
##  Max.   :186163.17   Max.   :1   Max.   :86.00   Max.   :81.000  
##                                                                  
##   gov_wage_gdp      pctmuslim         majoritymuslim        trade          
##  Min.   : 0.100   Min.   :0.000000   Min.   :-1.0000   Min.   :0.000e+00  
##  1st Qu.: 1.300   1st Qu.:0.006375   1st Qu.: 0.0000   1st Qu.:8.911e+07  
##  Median : 1.900   Median :0.050000   Median : 0.0000   Median :5.194e+08  
##  Mean   : 2.828   Mean   :0.280317   Mean   : 0.2517   Mean   :1.025e+10
##  3rd Qu.: 3.625   3rd Qu.:0.547500   3rd Qu.: 1.0000   3rd Qu.:4.796e+09
##  Max.   :11.800   Max.   :0.999000   Max.   : 1.0000   Max.   :3.290e+11
##  NA's   :114      NA's   :4          NA's   :4         NA's   :4
##    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.:3.815e+06
##  Median :  7.00   Median : 2.000   Median :  3.000   Median :8.852e+06
##  Mean   : 10.47   Mean   : 5.324   Mean   :  5.144   Mean   :3.655e+07
##  3rd Qu.: 12.00   3rd Qu.: 6.000   3rd Qu.:  6.000   3rd Qu.:2.341e+07
##  Max.   :116.00   Max.   :64.000   Max.   :116.000   Max.   :1.242e+09
##  NA's   :20       NA's   :20       NA's   :20
##   gdppcus1998         ecaid             milaid              region
##  Min.   :   95.45   Min.   :   0.00   Min.   :   0.000   Min.   :1.000
##  1st Qu.:  412.07   1st Qu.:   0.00   1st Qu.:   0.000   1st Qu.:3.000
##  Median : 1374.88   Median :   8.70   Median :   0.200   Median :4.000
##  Mean   : 5044.09   Mean   :  49.27   Mean   :  33.048   Mean   :4.372
##  3rd Qu.: 4936.62   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   :4         NA's   :4          NA's   :2
##    corruption         totaid             r_africa        r_middleeast
##  Min.   :-2.58299   Min.   :   0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:-0.41515   1st Qu.:   0.325   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 0.32696   Median :   9.000   Median :0.0000   Median :0.0000
##  Mean   : 0.01364   Mean   :  82.320   Mean   :0.3087   Mean   :0.1007
##  3rd Qu.: 0.72025   3rd Qu.:  42.950   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   : 1.58281   Max.   :4069.100   Max.   :1.0000   Max.   :1.0000
##                     NA's   :4
##     r_europe      r_southamerica       r_asia         country
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Length:298
##  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   :0.2349   Mean   :0.1208   Mean   :0.1678
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##    distUNplz          vpd
##  Min.   : 0.0000   Min.   :  0.00000
##  1st Qu.: 0.2219   1st Qu.:  0.07722
##  Median : 0.2956   Median :  0.60506
##  Mean   : 0.5493   Mean   :  9.86292
##  3rd Qu.: 0.4608   3rd Qu.:  7.80324
##  Max.   :15.0552   Max.   :249.36491
##  NA's   :6
```

```r
min_vio = format(round(min(FM_subcases_per_dip$vpd), 2))
max_vio = format(round(max(FM_subcases_per_dip$vpd), 2))
```

Interestingly we found that violations per diplomat ranges from 0 to 249.36 which further confirms our previous analysis that the number of staff does not correlate to the number of violations. It would otherwise indicate that the average violation would be similar across the countries.

```r
FM_subcases_per_dip$country[FM_subcases_per_dip$vpd == max(FM_subcases_per_dip$vpd)]
```

```
## [1] "KUWAIT"
```

We found that the country that commited more parking violations in Manhattan NY was Kuwait with an outstanding violations of 249 violations per diplomats. We further investigated the variables for Kuwait.

```
FM_subcases_per_dip[FM_subcases_per_dip$country == "KUWAIT",
    ]
```

```
##     wbcode prepost  violations       fines mission staff spouse
## 171    KWT     pre 2244.284180 123319.1562       1     9      6
## 172    KWT     pos    1.308244    140.6362       1     9      6
##     gov_wage_gdp pctmuslim majoritymuslim      trade cars_total
## 171           NA      0.85              1 2751607552         17
## 172           NA      0.85              1 2751607552         17
##     cars_personal cars_mission pop1998 gdppcus1998 ecaid milaid region
## 171             5           12 2027000    17874.07     0      0      7
## 172             5           12 2027000    17874.07     0      0      7
##     corruption totaid r_africa r_middleeast r_europe r_southamerica r_asia
## 171  -1.073995      0        0            1        0              0      0
## 172  -1.073995      0        0            1        0              0      0
##     country distUNplz         vpd
## 171  KUWAIT  0.145854 249.3649089
## 172  KUWAIT  0.145854   0.1453604
```

To our surprise, violations of Kuwait pre and post 2002 was astonashing. Its pre violation stood at 2244.2841797 while its post violations stood at 1.3082438 . The violations per diplomat therefore significantly reduced from 249.3649089 to 0.1453604 while all other variables remains the same.

Results

Upon thoroughly checking the data given, we came to a conclusion that parking violations in Manhattan New York did not have any relationship with the majority of the variables in the dataset such as staff, spouse, pctmuslim (muslim population), ecaid, milaid (economic and military aid) so on and so forth. However, we found a linear relationship between the number of parking violations and the corruption of the country indicated by the index. The correlation between these two variables (violations Vs corruption) before 2002 was 0.183 (pre) and became 0.211 after 2002 (pos).

Another strong relationship between these two variables was confirmed after we investigated the top 25 offending countries: Kuwait, Egypt, Chad, Sudan, etc pre and pos 2002 situation. Considering the fact that World Trade Center attack happened in the US and corruption of the country belongs to those corresponding countries, we expect to see these two physical locations as two separate variables. However the fact that a significant drop of the parking violations from those top offending countries pre and pos 2002 showed the otherwise. Our data analysis therefore showed that there were some relationship between top offending countries and WTC attack in terms of political climate, social stigma and strong aversion to those countries after WTC attack on 2001 in the US and law enforcement in New York after 2001.

Bibliography and R packages used in this project