W203 Lab1: EDA of Parking Violations at the UN

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.
```

```
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

```
# 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

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

# 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

```
# 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.

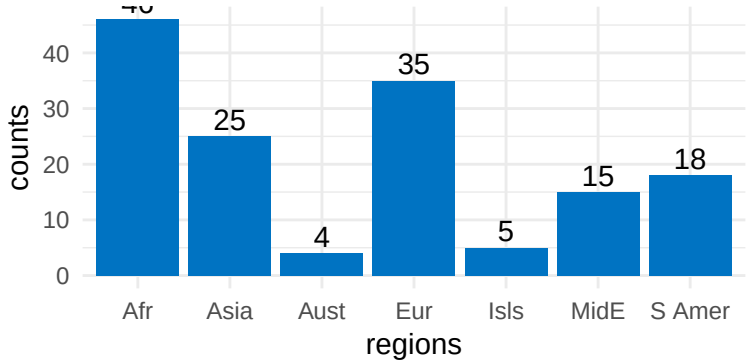
At a glance we observe the regions as following:

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.

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

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.

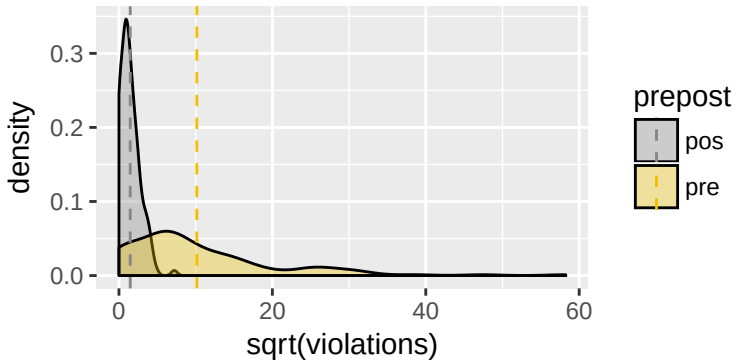
- Pre and Post 2002 records (column name : prepost)

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.

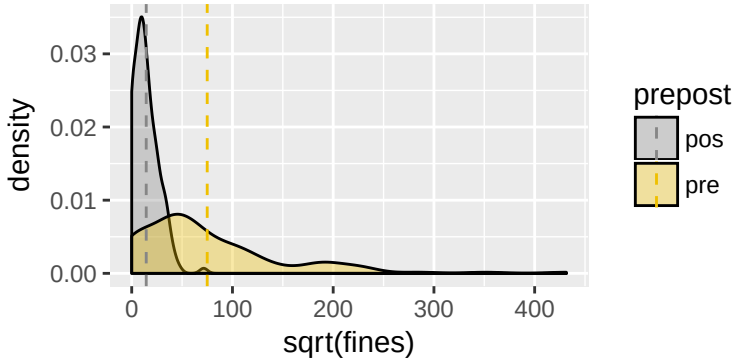
```
# compute the mean of pre and post violations
v_mean <- tb_un_clean %>% group_by(prepost) %>% summarise(grp.mean = mean(sqrt(violations)))

# 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)

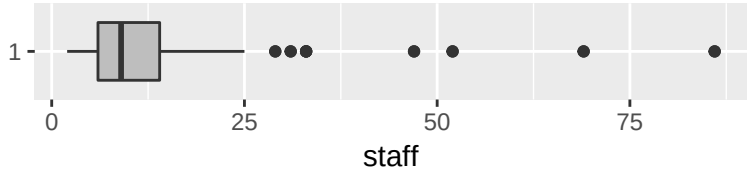
As fines are dependend on the number of violations, we see similar decline in distribution of fines owed after the enforcement. As fines have a very skewed distribution, visually the histogram is hard to review, hence we computer a square root to see the distribution better. We see see that missions 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"))
```



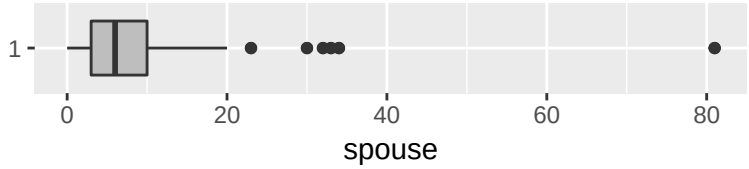
- Diplomatic mission details (columns : staff, spouse)
- Total number of diplomats(from each country) - Majority of missions have under 20 diplomats.

```
# 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()
```

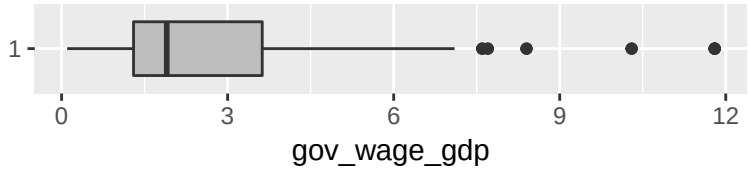


- Government wages index (column : gov_wage_gdp) Here we notice a most diplomats getting paid within 2-4 times the GDP of their country. We have to keep in mind that this index by itself is not helpful as GDP varis a lot by country. Also we decided to remove this field from our analysis as this has over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

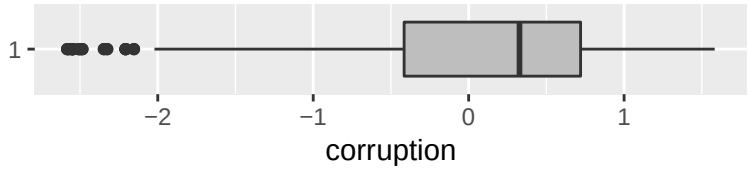


- Individual country corruption index (column : corruption)

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.

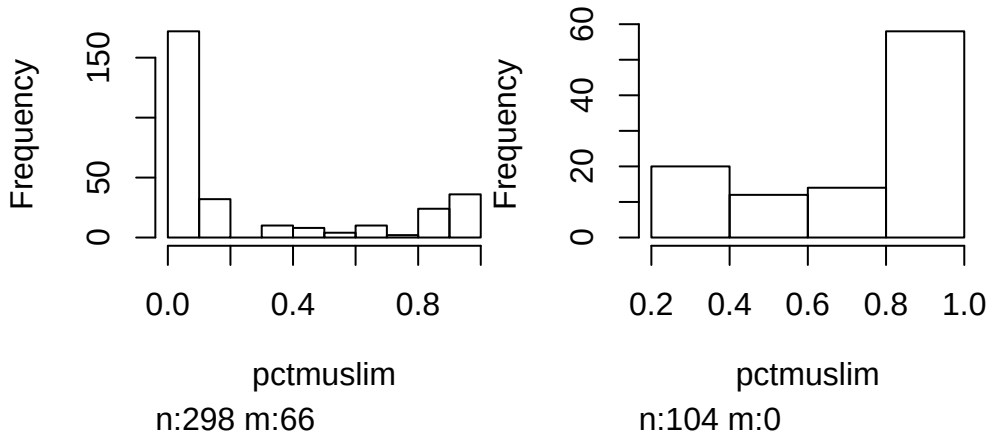
```
# Using ggplot object to plot corruption index distribution

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
```

- 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.

3. Analysis of key relationships

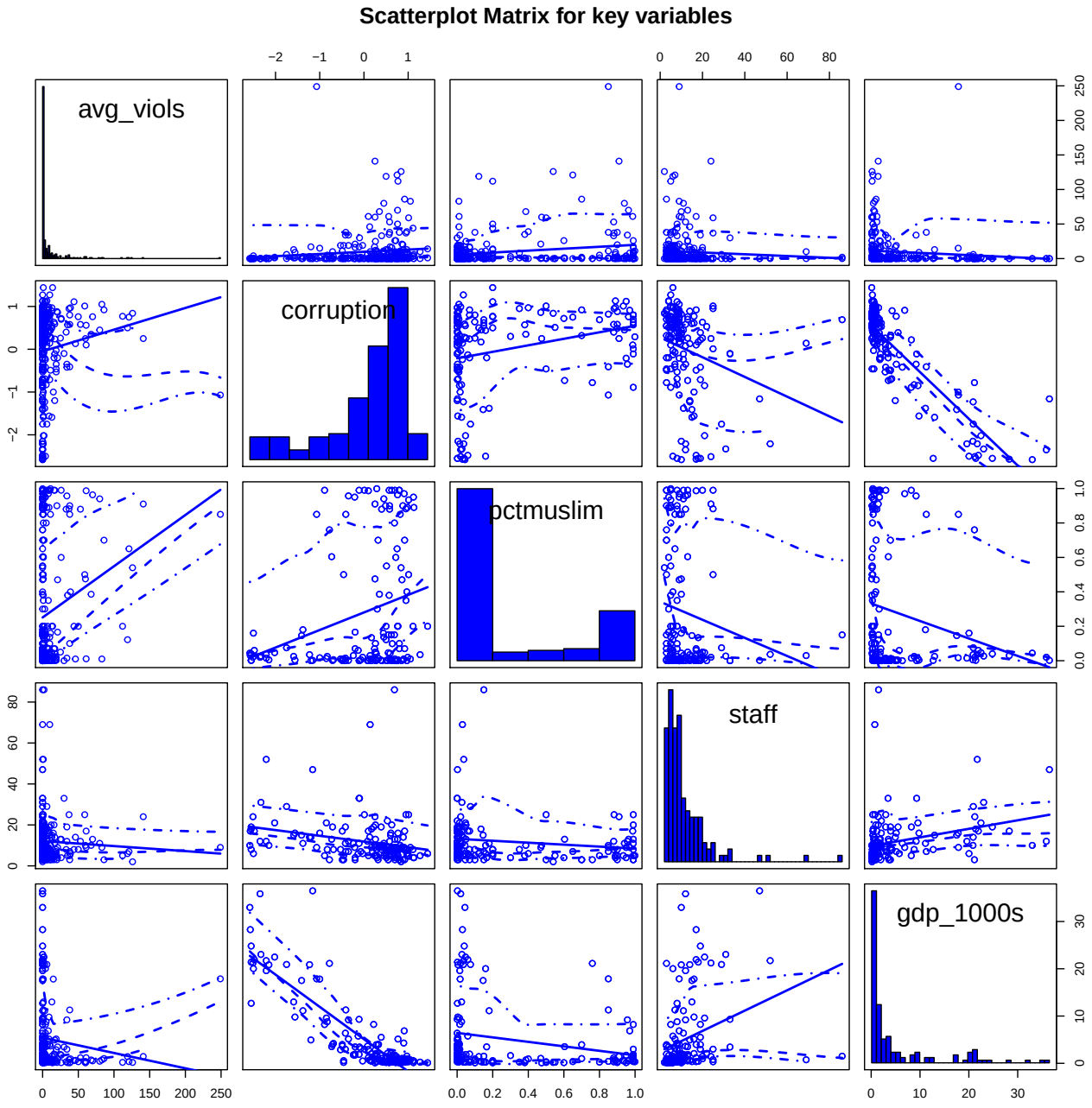
We did a quick scatter plot matrix to identify interesting relationshis we want to explore. We make few key association that interested us as following,

```
tib_computed_pre = dplyr::select(filter(tb_un_revised, prepost ==
    "pre"), avg_viols, corruption, pctmuslim, region, staff,
    gdp_1000s)
kable(corrr::correlate(tib_computed_pre)[1:7, ], caption = "Correlation - Pre 2002")
```

Table 9: Correlation - Pre 2002

rowname	avg_viols	corruption	pctmuslim	region	staff	gdp_1000s
avg_viols	NA	0.1830667	0.3061743	0.3499746	-0.0818401	-0.1518583
corruption	0.1830667	NA	0.2785027	0.2128406	-0.2578231	-0.8619694
pctmuslim	0.3061743	0.2785027	NA	0.5778721	-0.1582079	-0.2194083
region	0.3499746	0.2128406	0.5778721	NA	-0.1976593	-0.2120288
staff	-0.0818401	-0.2578231	-0.1582079	-0.1976593	NA	0.3021291
gdp_1000s	-0.1518583	-0.8619694	-0.2194083	-0.2120288	0.3021291	NA
NA	NA	NA	NA	NA	NA	NA

```
car::scatterplotMatrix(~avg_viols + corruption + pctmuslim +
    staff + gdp_1000s, data = tb_un_revised, diagonal = list(method = "histogram"),
    main = "Scatterplot Matrix for key variables")
```

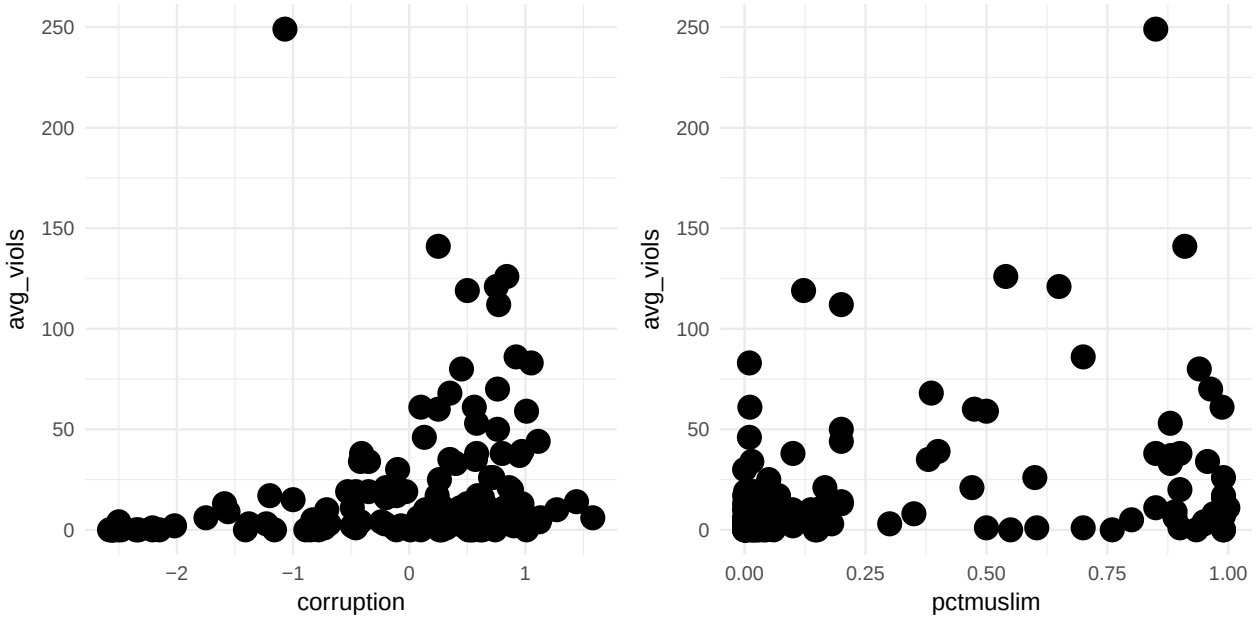


3.1 Key Relationships

i) Violations relationship to Corruption and Percent muslim. Violations do not seem to have a strong correlation to corruption index. More perceptage of muslim population seems to increase the violations. We discuss this more in statistical terms below.

```
plot1 <- ggplot(tib_computed_pre, aes(corruption, avg_viols)) +
    geom_point(shape = 16, size = 5) + theme_minimal()
plot2 <- ggplot(tib_computed_pre, aes(pctmuslim, avg_viols)) +
    geom_point(shape = 16, size = 5) + theme_minimal()

grid.arrange(plot1, plot2, ncol = 2)
```

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



Looking at the two plots we can see that corruption index which seems to be a central data element here does not show a strong relationship to violations.

ii) Corruption

There seems to be a strong relationship with Percent muslim, Staff, GDP.

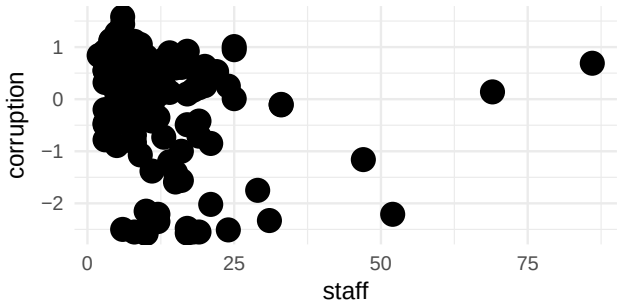
Corruption seems to increase with increased percent muslim, which also contributes to violations. We also see more staffers increasing corruption and countries with higher GDP seem to have significantly lower corruption index values. This leads us to wonder if this has to do with sentiments towards the United states within predominantly muslim nations.

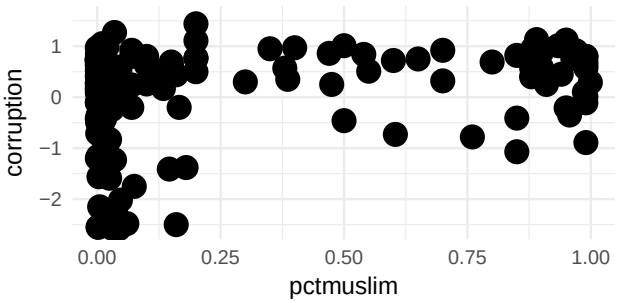
Also missions with more staff seem to have lower corruption, we wonder if this is due to a 'peer effect' where diplomats act more responsibly as a result of a larger team in the US that has a higher risk of exposure if they engage in small acts of unlawful actions and get caught.

GDP also seems to negetively effect corruption, makes us think if perhaps a country with higher GDP perhaps means better governance which may cause a lower corruption index aggregae value.

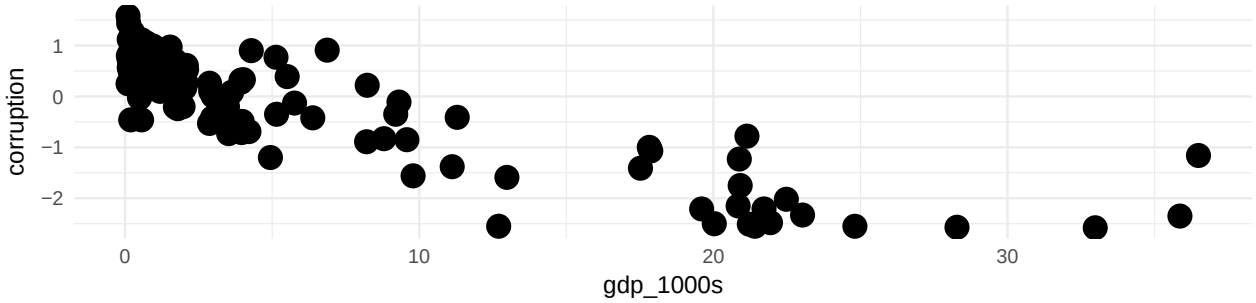
```
plot3 <- ggplot(tib_computed_pre, aes(staff, corruption)) + geom_point(shape = 16,
    size = 5) + theme_minimal()
plot4 <- ggplot(tib_computed_pre, aes(pctmuslim, corruption)) +
    geom_point(shape = 16, size = 5) + theme_minimal()


grid.arrange(plot3, plot4, ncol = 2)
```

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



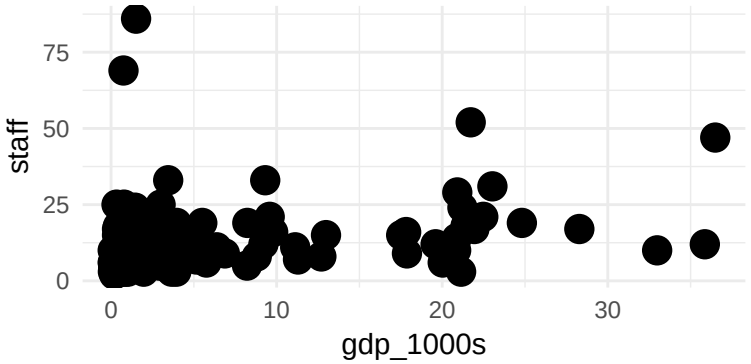



```
ggplot(tib_computed_pre, aes(gdp_1000s, corruption)) + geom_point(shape = 16,
    size = 5) + theme_minimal()
```



iii) GDP relationship to Staff. Higher the GDP, we notice the staff numbers also increase. This simply tells us that nations who have a higher per capita income are able to have a larger mission size at the UN.

```
ggplot(tib_computed_pre, aes(gdp_1000s, staff)) + geom_point(shape = 16,
    size = 5) + theme_minimal()
```



Key relationship recap

1. We see a 0.18 correlation with parking violations. Given that our data is highly summarised and does not represent raw data, we are inclined to consider that only percent of muslims has a strong effect on violations and not the corruption index.
2. We see that percentage of muslim variable has a 0.30 correlation with violations and we see this as a stronger correlation that corruption index.
3. Corruption has a 0.27 correlation with percent of muslim variable, -0.25 with staff and -.86 with GDP. These are our strongest relationships. Telling us that corruption index is really sensitive to GDP and more staffers also effect corruption index to go down.
4. We find that percent of muslim variable correlates to violations at 0.30, but also to corruption index. Leading us to think this could be a result of overall negetive sentiments towards the US or the UN

by these nations and perhaps their diplomats behave with some disregard to US laws due to their antipathy towards the US.

5. We observe that the legislative change in October 2002 dramatically reduced volume of violations. Due to this major event, we decided to only look at 'cultural' paramters like corruption index, percent muslim. . . all before this legislation. We think the pre legislation data really represents the true behavior and post enforcement we find the behaviors to be dramatically different.

4. Analysis of Secondary Effects (10 pts)

We found some confounding things in this dataset that give us pause and reason to further investigate.

1) We list here as exhibit the top 25 missions that had the most parking violations. Kuwait is at the top of the list, but their corruption index is -1.07. Next, Egypt is the second highest on this list but they too have a 0.25 corruption index total. This leads us to our first major realization that corruption index as is does not have a statistically significant relationship to violations.

2) Regions variable is nominal and thus we can't correlate "regions" with "violations". But there are visibly large magnitude of violations in two regions here. We cannot explain why this is the case. Is this simply a by product of more nations in a certain region, or more total staff due to having more nations. These remain unexplainable from the dataset.

EXAMPLE 1

```
computed_full = subset(tb_un_revised, select = c(country, region,
    prepost, corruption, ecaid, milaid, violations, avg_viols,
    trade_mil, pop_mil, gdp_1000s))
com1 = select(mutate(filter(computed_full, prepost == "pre"),
    country = country, pre_2002_violations = avg_viols), country,
    pre_2002_violations, gdp_1000s, corruption)
com2 = select(mutate(filter(computed_full, prepost == "pos"),
    country = country, pos_2002_violations = avg_viols), country,
    pos_2002_violations)
com_final = arrange(filter(merge(com1, com2), country != ""),
    desc(pre_2002_violations))
kable(com_final[1:25, ], caption = "Parking Violation side by side ( pre / post enforcement )")
```

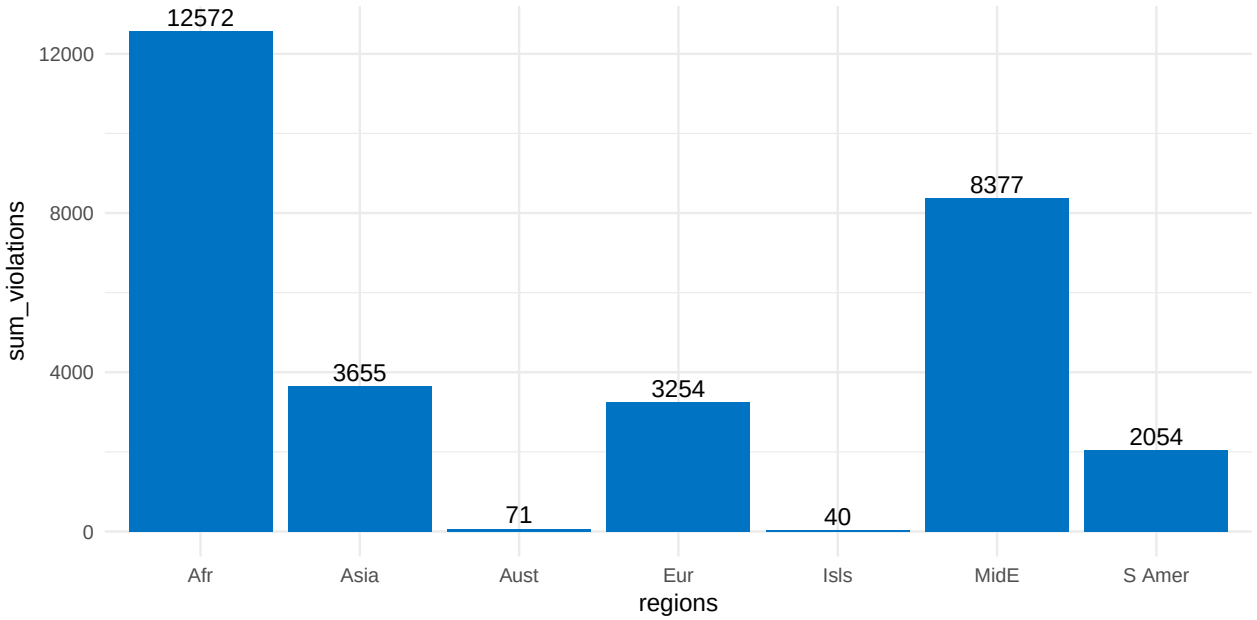
Table 10: Parking Violation side by side (pre / post enforcement)

country	pre_2002_violations	gdp_1000s	corruption	pos_2002_violations
KUWAIT	249	17.87	-1.07	0
EGYPT	141	1.45	0.25	0
CHAD	126	0.19	0.84	0
SUDAN	121	0.36	0.75	0
BULGARIA	119	1.42	0.50	2
MOZAMBIQUE	112	0.20	0.77	0
ALBANIA	86	1.01	0.92	2
ANGOLA	83	0.73	1.05	2
SENEGAL	80	0.44	0.45	0
PAKISTAN	70	0.52	0.76	1
IVORY COAST	68	0.71	0.35	0
MOROCCO	61	1.19	0.10	0
ZAMBIA	61	0.32	0.56	0
ETHIOPIA	60	0.10	0.25	1
NIGERIA	59	0.33	1.01	0

country	pre_2002_violations	gdp_1000s	corruption	pos_2002_violations
SYRIA	53	1.18	0.58	1
BENIN	50	0.34	0.76	7
ZIMBABWE	46	0.63	0.13	1
CAMEROON	44	0.57	1.11	3
MONTENEGRO & SERBIA	39	0.94	0.97	0
BAHRAIN	38	11.29	-0.41	1
BURUNDI	38	0.11	0.80	0
MALI	38	0.21	0.58	1
INDONESIA	37	0.78	0.95	1
BOSNIA-HERZEGOVINA	35	1.08	0.35	0

EXAMPLE 2

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

country_base <- tb_un_clean %>% group_by(region) %>% summarise(sum_violations = round(sum(violations)))
# 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 = sum_violations)) +
    geom_bar(fill = "#0073C2FF", stat = "identity") + geom_text(aes(label = sum_violations),
    vjust = -0.3) + theme_minimal()
```



5. Final Summary

We had started with this construct that every region and country has some form of corruption, a prevailing diplomatic relationship with the UN and by extention, The United States. 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 except pctmuslim (muslim population).

We find however that the corruption index has a strong relationship to GDP, percentage of muslims in nation and this gives us pause, is this an indication of attitutes towards the US and UN based on how foriegn policy effecs these regions of the world.

We further wanted to explore how diplomatic attitudes are tuned into economic development of nation & Level of crime in respective nations. We found the corruption index to have a very strong relationship with GDP and also with number of staff and finally percent of muslims in that nation.

We find a linear relationship between the corruption index and GDP at a correlation of -0.86. We do not find corruption to have a strong linear relathinship to violations from this dataset.

We wanted to understand the effects of the current world events, how they effect attitudes of the diplomats. we found a strong effect as observed in reduction of violations irrespective of the corruption index. We notice an across the chart decline in violations by the top 25 offending countries: Kuwait, Egypt, Chad, Sudan - after the enforcement and in fact 13 months after World Trade Center terrorist attacks. We think this is due to diplomats finally coming outside their umbrella of immunity when it comes to parking violations.

Finally we find our most interesting relationships, Per capita income and crime index. We note here that Kuwait has a -1.07 corruption index, that is low on corruption scale, yet very high violations before enforcement. Kuwait is a oil rich country with a very high GDP, compared to Egypt, Chad and Sudan who have a GDP that is 20 times smaller than Kuwait have a corruption index that is greater than zero, yet much lower violations than Kuwait. This informs us that the index does not effect violations and diplomats from poor nations are not nessasarity displaying a culture of corruption.

In other words, some countries do fit the profile of the conventional thinking that lower GDP nations have a poorer law and order environment. But, we don't see a linear relationship of diplomats from poor nations with a higher corruption index engaging in more violations, i.e. they do not show a higher rate of violations.

for example, We find Nigeria to not fit this 'conventional thinking' profile, as they have a much lower number

of violations despite a GDP 20 times lower than Kuwait and a corruption index more than 1, still they only have 59 violations and are 15th on the top violators list.

Bibliography & Credits

1. Data Import with tidyverse and working with the results as tibbles, and reshape messy data with tidyr. (link : https://github.com/rstudio/cheatsheets/raw/master/data-import.pdf)
2. Data transformation cheatsheet where dplyr was used to change and create tables. (link :https://github.com/rstudio/cheatsheets/raw/master/data-transformation.pdf)
3. R Markdown guidance on how to knit, format, code and create awesome R markdown projects (link: https://github.com/rstudio/cheatsheets/raw/master/rmarkdown-2.0.pdf)
4. Data visualization using ggplot2 allowed us to build some visualizations. (link:https://github.com/rstudio/cheatsheets/raw/master/data-visualization-2.1.pdf)
5. Stackexchange helped us resolve over 50 issues in cleaning, categorizing data. It helped us resolve issues with coding and R markdown pitfalls.