Corruption & Parking Violations: An Exploratory Data Analysis

Chandra Shekar Bikkanur & Shweta Sen

September 28, 2018

Introduction

Prior to 2002, UN officials carried diplomatic immunity that exempted them from receiving parking violations. After 2002, however, law enforcement was granted the right to confiscate license plates of UN diplomats with parking violations. Thus, UN diplomats lost their immunity and began to face the legal penalties of unpaid parking tickets. By analyzing the parking behavior of UN officials in Manhattan pre- and post-2002, the client (World Bank) is interested in understanding how corruption is affected by cultural norms (i.e., immunity status) and legal enforcement.

1. Research Questions

The objective of our exploratory data analysis is to answer the following questions:

a. How do legal enforcement and cultural norms affect corruption and parking behavior?

b. What is the relationship between corruption and parking violations before and after 2002?

c. What variables, other than corruption, contribute to parking violations?

2. Loading The Data Set

Before beginning the EDA, the CAR package was installed and the data file (Corrupt.RData) was loaded into the R workspace.

```
# install.packages("ggplot2") # For plotting
# install.packages("scales")
# install.packages("gridExtra") # For formatting graph positions
# install.packages("reshape2") # To melt data frames
# install.packages("qwraps2", repo = "http://cran.rstudio.com") # For tables
# options(qwraps2_markup = "markdown")
# install.packages("dplyr")
# install.packages(car, dependencies = TRUE)

library(car)
```

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

```
install.packages("qwraps2", repo = "http://cran.rstudio.com")
```

```
## Installing package into 'C:/Users/chand/Documents/R/win-library/3.5'
## (as 'lib' is unspecified)
```

```
## package 'qwraps2' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\chand\AppData\Local\Temp\RtmpYvy3pd\downloaded_packages
```

```
options(qwraps2_markup = "markdown")
install.packages("gridExtra", repos = "http://cran.univ-lyon1.fr")
```

```
## Installing package into 'C:/Users/chand/Documents/R/win-library/3.5'
## (as 'lib' is unspecified)
```

```
## package 'gridExtra' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\chand\AppData\Local\Temp\RtmpYvy3pd\downloaded_packages
```

```
# Load data
load('Corrupt.Rdata')
```

3. Description of Data Set

Once the data is loaded into the workspace, we observe the objects in the workspace.

```
objects()
```

```
## [1] "FMcorrupt"
```

The data set contains an object named **FMcorrupt**, which is of type **data.frame**, has 364 rows, along with 28 variables. Using the str function, we can see the structure of the data, including variable names and variable types.

```
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 ...
```

Of the 28 variables in the data set, we see that 3 are character variables, 12 are numerical variables, and 13 are discrete integer variables (of which 7 are dummy coded to represent religious status or geographic region). We also observe that several of the variables contain NA values. In terms of geographic spread, the data set

contains observations for diplomats from 166 unique countries across seven different regions.

4. Evaluation of Data Quality

Data quality is shaped by several factors such as completeness, reliability, accuracy, and consistency, among others. Since the data set contains several missing values (NA) that interfere with completeness, a subset was created to extract rows with only complete observations. By doing so, the number of rows in the data set was reduced from 364 to 298. Although we lose 66 rows by omitting missing values, we still retain 298 complete observations to facilitate the analysis.

```
# Omit rows with incomplete observations (NA)
subcase1 = ! is.na(FMcorrupt$corruption) & ! is.na(FMcorrupt$violations)
CorruptData = FMcorrupt[subcase1, ]
```

a. Discrepancies With Corruption Index Values Given that the corruption variable is gathered from the 1998 Country Corruption Index by Transparency International (https://www.transparency.org/research/cpi/cpi_1998/0#background), it was expected that the values of the corruption variable would be consistent with the scaling of the Corruption Index. Per the 1998 Country Corruption Index, corruption is scaled from 0 to 10, with 0 being wholly corrupt and 10 being corruption-free. The corruption scale in our data set, in contrast, goes from -2.58 (min) to 1.58 (max). As a result, the ambiguous re-scaling of the corruption index casts doubt on the reliability and consistency of the data. For example, if a country has a negative corruption index, is it interpreted as being a highly corrupt or corruption-free? To address the issue, we studied the 1998 Corruption Perceptions Index report to gather a frame of reference for relative country corruption values. Looking at the extremes of the distribution, the report found that Denmark and Finland were least corrupt while Paraguay and Cameroon were most corrupt. Translating the findings to our data set, we note that the corruption indices for Denmark and Finland are -2.57 and -2.55 (respectively) while the corruption indices for Paraguay and Cameroon are 0.97 and 1.11 (respectively). As a result, we can confirm that the lower (or more negative) corruption indices are indicative of less corruption (which is opposite to the scaling used by Transparency International).

b. Mismatch Between Pre- & Post-2002 Variables Since a core output of the analysis is understand the relationship between corruption and violations pre- and post-2002, it was expected that certain variables would time-sensitive. From the data set, we observe that the pre-2002 values of several variables (such as gov_wage_gdp, trade, spouse, total_aid, corruption, and pctmuslim) are completely identical to their respective post-2002 values. While it is certainly possible for some variables to remain the same over time, it is noteworthy that traditionally oscillatory variables like GDP, trade, and total aid do not change value over the time period described. Furthermore, the variables for 1998 population and GDP are provided (pop1998 and gdppcus1998, respectively), but data for their post-2002 counterparts are not provided. As a result, the validity of the data set may require further investigation due to surface anomalies regarding timeline of data collection and the unexpectedly static nature of the data before and after 2002.

c. Lack of Temporal Frame of Reference The data set does not provide a temporal frame of reference beyond "pre-2002" and "post-2002". As a result, we are unable to infer the context of the pre- and post-data. For example, does pre-2002 refer to data from December 2001 and post-2002 refer data from to January 2003? Or, does pre-2002 refer to a data average from 1990 to 2002 while post-2002 refers to the data from an arbitrary date in 2005? Without understanding the time period and sampling strategy that constitute the reference frames of pre-2002 and post-2002, any insights from the data set face the potential of being misconstrued (especially given that we have limited background information on the data collection process).

5. Data Processing & Preparation

The dependent (or target) variable in the data set is violations. During the analysis, data subsets were created for several independent variables of interest in order to remove incomplete observations and more easily study key relationships. In addition to creating data subsets and subcases, population segmentation (by region or prepost status) was used. As such, the data provided for the analysis was largely complete and organized, so significant processing was not required. Any data preparation that was needed was completed in the respective sections of the analysis.

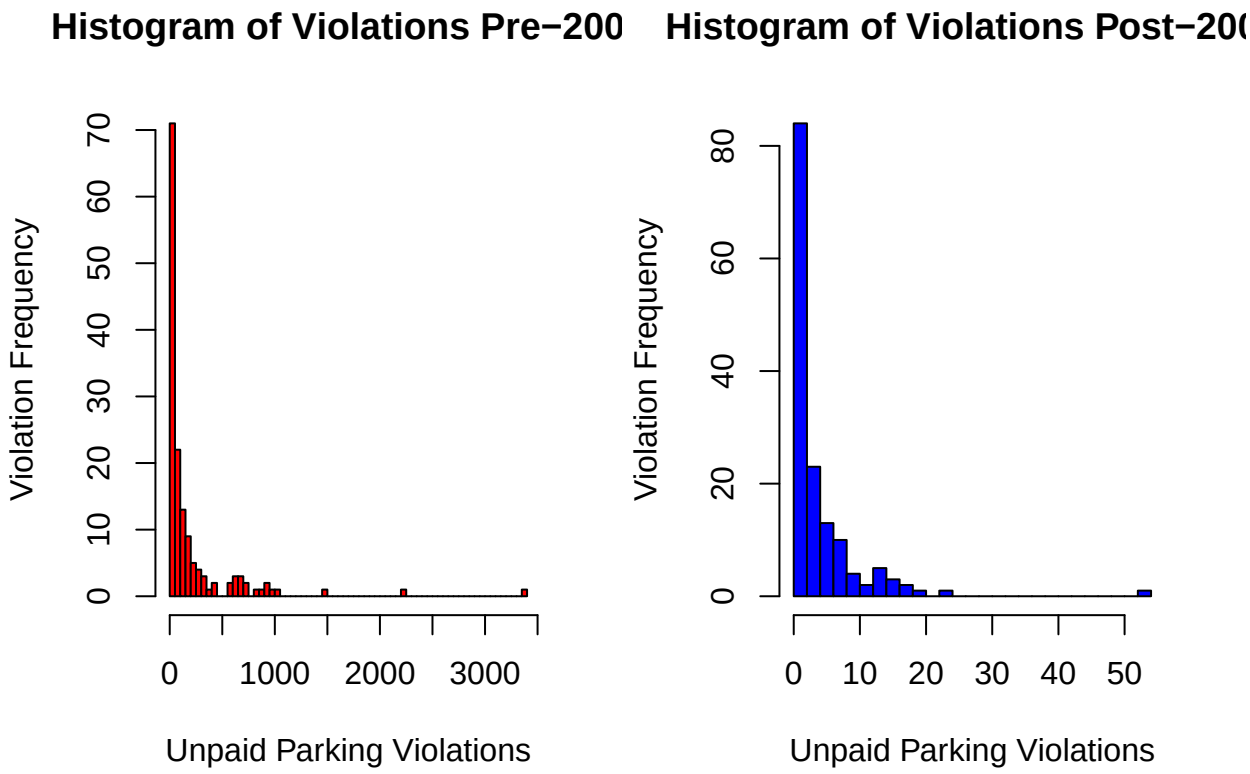
```r
var <- c("prepost", "country", "violations", "fines", "mission", "corruption", "region")
CorruptDataPre2002 <- CorruptData[CorruptData$prepost == 'pre', var ]
CorruptDataPost2002 <- CorruptData[CorruptData$prepost == 'pos', var ]
```

Univariate Analysis of Key Variables

1. Analysis of Violations

The dependent (or target) variable in the data set is violations. Hence, we begin by examining changes in violations pre- and post-2002. From the histogram and descriptive statistics summary, we observe that the number of violations post-2002 are significantly lower than violations pre-2002. In fact, the post-2002 data indicates there is almost a 50-fold reduction in the median and total number of violations. Pre-2002, diplomats from Egypt, Kuwait, Nigeria, and Morocco contributed to the highest violations. After the initiation of legal enforcement, all four countries still remained in the upper quartile of violation distributions. However, post-2002, the rank order changed: diplomats from Benin, Cameroon, Indonesia, and Algeria contributed to the highest violations.

```r
par(mfrow=c(1,2)) # par function used to place graphs side-by-side

# Histograph of violations before 2002
violations_pre <- hist(CorruptDataPre2002$violations, breaks = 'FD', col='red', main = "Histogram of Vi

# Histogram of violations after 2002
violations_post <- hist(CorruptDataPost2002$violations, breaks = 'FD', col='blue', main = "Histogram of
```



```r
library(qwraps2)
```

```
##
## Attaching package: 'qwraps2'
```

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

```r
options(qwraps2_markup = "markdown")

our_summary1 <-
  list("Pre 2002 Viloations" =
       list("min" = ~ min(CorruptData[CorruptData$prepost == 'pre', "violations" ]),
            "max" = ~ max(CorruptData[CorruptData$prepost == 'pre', "violations" ]),
            "mean (sd)" = ~ qwraps2::mean_sd(CorruptData[CorruptData$prepost == 'pre', "violations"]),
            "total" = ~ sum(CorruptData[CorruptData$prepost == 'pre', "violations" ])),
       "Post 2002 Viloations" =
       list("min" = ~ min(CorruptData[CorruptData$prepost == 'pos', "violations" ]),
            "max" = ~ max(CorruptData[CorruptData$prepost == 'pos', "violations" ]),
            "mean (sd)" = ~ qwraps2::mean_sd(CorruptData[CorruptData$prepost == 'pos', "violations" ]),
            "total" = ~ sum(CorruptData[CorruptData$prepost == 'pos', "violations" ])))


summary_table(CorruptData, our_summary1)
```

	CorruptData (N = 298)
Pre 2002 Viloations	
min	0
max	3392.961
mean (sd)	198.07 ± 405.28
total	29512.54
Post 2002 Viloations	
min	0
max	52.00269
mean (sd)	3.69 ± 6.01
total	549.4624

2. Analysis of Corruption

While corruption index may be a logical barometer for violations, it interestingly remains unaffected pre- and post-2002. In particular, the pre- and post- values for corruption index are completely identical, as tested by a correlation index (with 1 as the result, confirming identical values). Given that violations drastically decline after 2002 while corruption index remains constant, it is likely that corruption is not the primary factor governing the frequency and severity of violations.

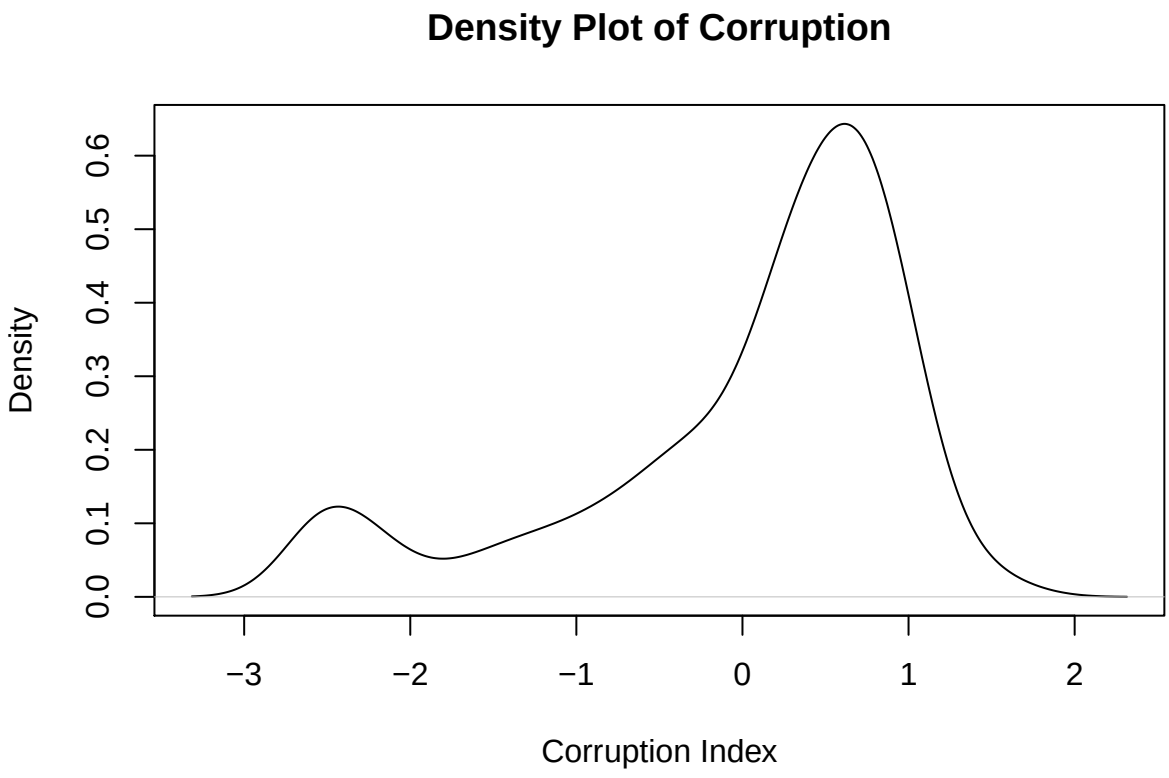
```r
# Correlation of pre-2002 corruption index with post-2002 corruption index
# A correlation value of 1 indicates pre-2002 and post-2002 data is identical.
(cor(CorruptDataPre2002$corruption,CorruptDataPost2002$corruption))
```

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

To better understand the makeup of our corruption data, a density plot was generated. From the plot, it is evident that approximately 60% of the population (i.e., countries) have a corruption index between 0 to 1, which translates to a relatively high level of corruption. A smaller peak forms at a corruption index

between -2 to -3, which indicates that approximately 10% of the population experiences a relatively low level of corruption.

```
# Density plot of corruption
density_corruption <- density(CorruptData$corruption)
plot(density_corruption, xlab = "Corruption Index", main = "Density Plot of Corruption")
```



Analysis of Key Relationships

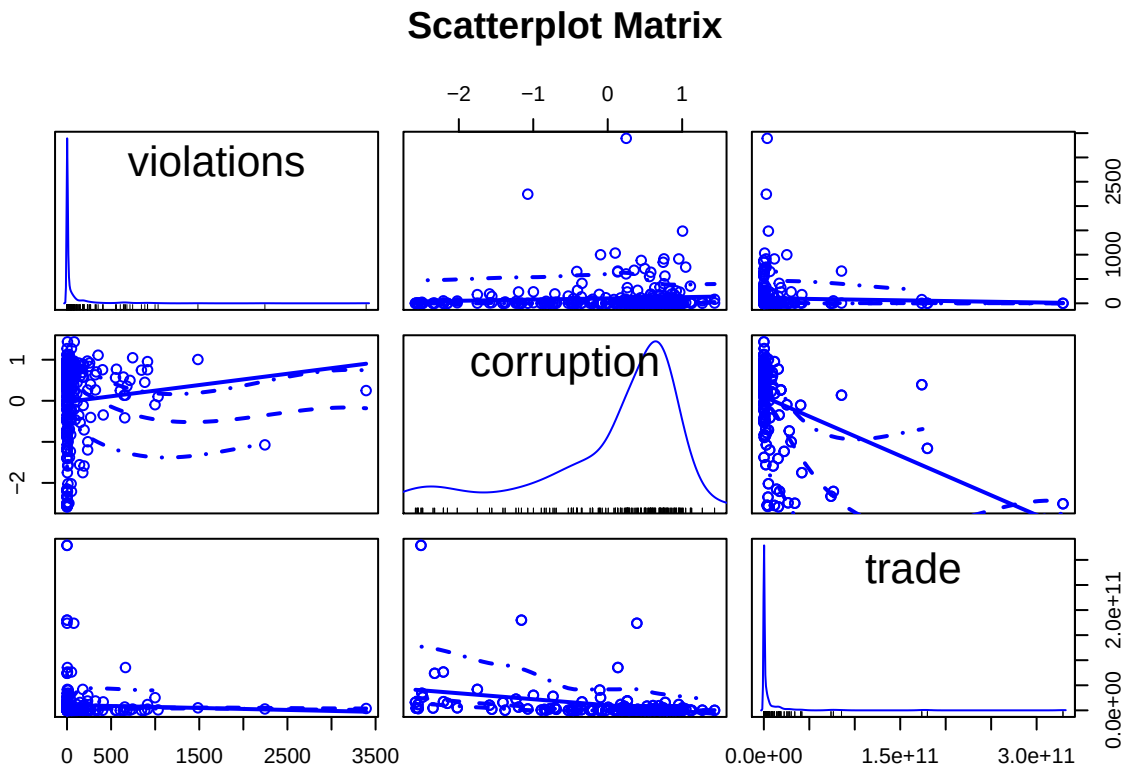
In our analysis, we explore relationships between the following variables:

1. Violations, Corruption, & Trade
2. Violations & Diplomat Region of Origin
3. Violations & GDP Per Capita / Trade Per Capita
4. Economic Aid Received & Percent Contribution to Violations by Region
5. Violations & Fines

1. Starting Analysis of Violations, Corruption, & Trade

To study the relationships between variables (that we hypothesize to be critical or strongly correlated to violations), we first use a scatter plot matrix.

```
library(car)
scatterplotMatrix( ~ violations + corruption + trade, data = FMcorrupt,
                   main = "Scatterplot Matrix")
```



From the scatter plot matrix, the following can be inferred:

1. It is evident that corruption and violations are positively correlated, indicating that parking violations tend to arise more from countries where the corruption index higher. There may be a behavioral or cultural element where diplomats from high-corruption countries may be more accustomed to crime (and hence, tend to ignore it).
2. Corruption and trade have a strong negative correlation, indicating the United States is less likely to engage in trade with countries with high corruption.

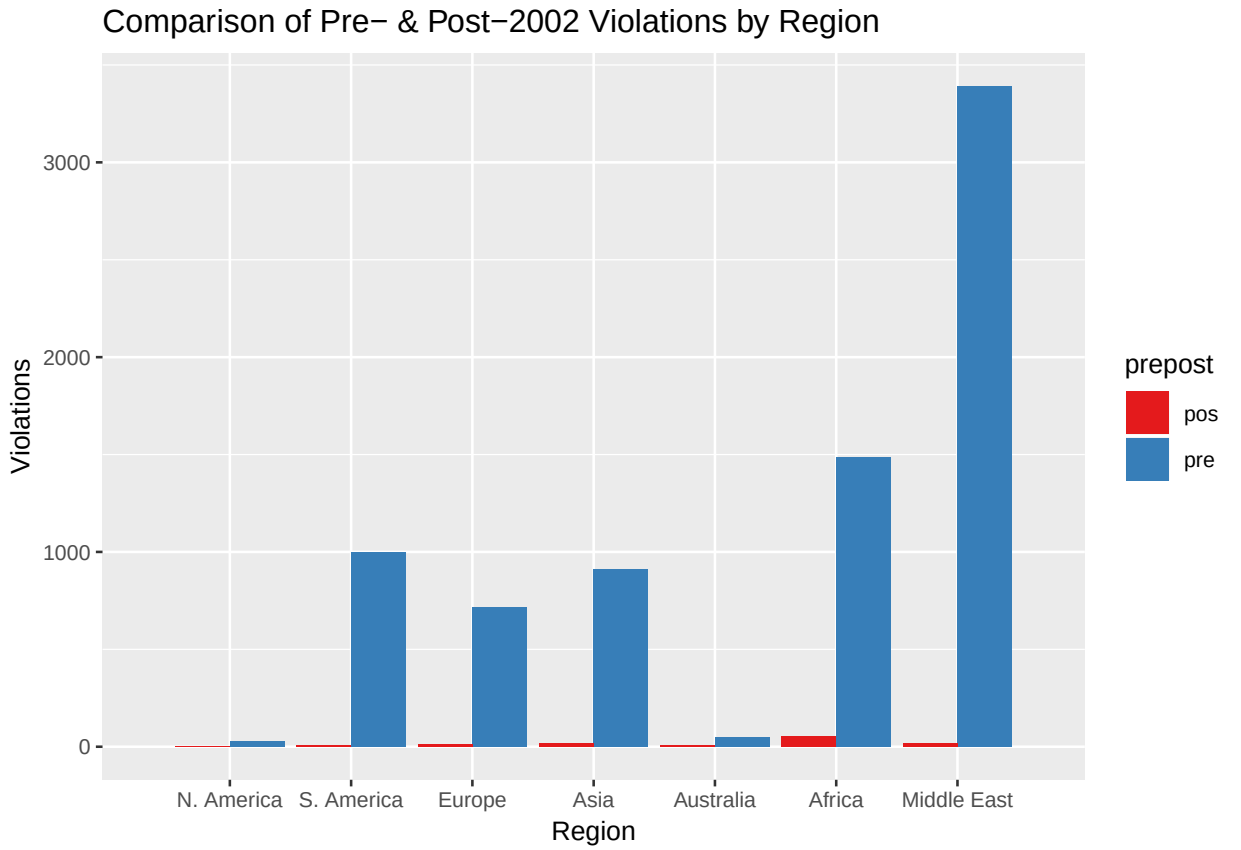
2. Relationship Between Violations & Diplomat Region of Origin

Based on the grouped bar graph studying violations by region pre- and post-2002, it is evident that the Middle East and Africa were the largest contributors to violations pre-2002. Asia and South America had similar violation totals pre-2002, followed by Europe. North America and Australia had significantly lower violation totals pre-2002, when UN diplomats could violate parking rules and still be exempt from punishment or legal action. The observation suggests there may be stark cultural differences in the regional perception of violations and resulting parking behavior.

Post-2002, when UN diplomats were held legally accountable for parking violations, we notice a decrease in violations across all regions. The decrease is especially pronounced for the Middle East, Africa, Asia, and South America: regions that contributed most heavily to violations pre-2002. North America and Australia experienced reductions in violation total as well, however the delta was not nearly as marked as their global counterparts. The post-2002 findings are telling: all regions respond positively to legal enforcement since all regions exhibit a decrease in violations. **For the regions displaying drastic changes pre- and post-2002 (i.e., Middle East, Africa, Asia, and South America), we can deduce that the fear of violations is negligible: diplomats will break parking rules if no penalty is given. For North America and Australia, where violation totals remain low and similar between pre- and post-2002, we can infer that there is a strong cultural norm to abide by the law (with or**

without the absence of penalities).

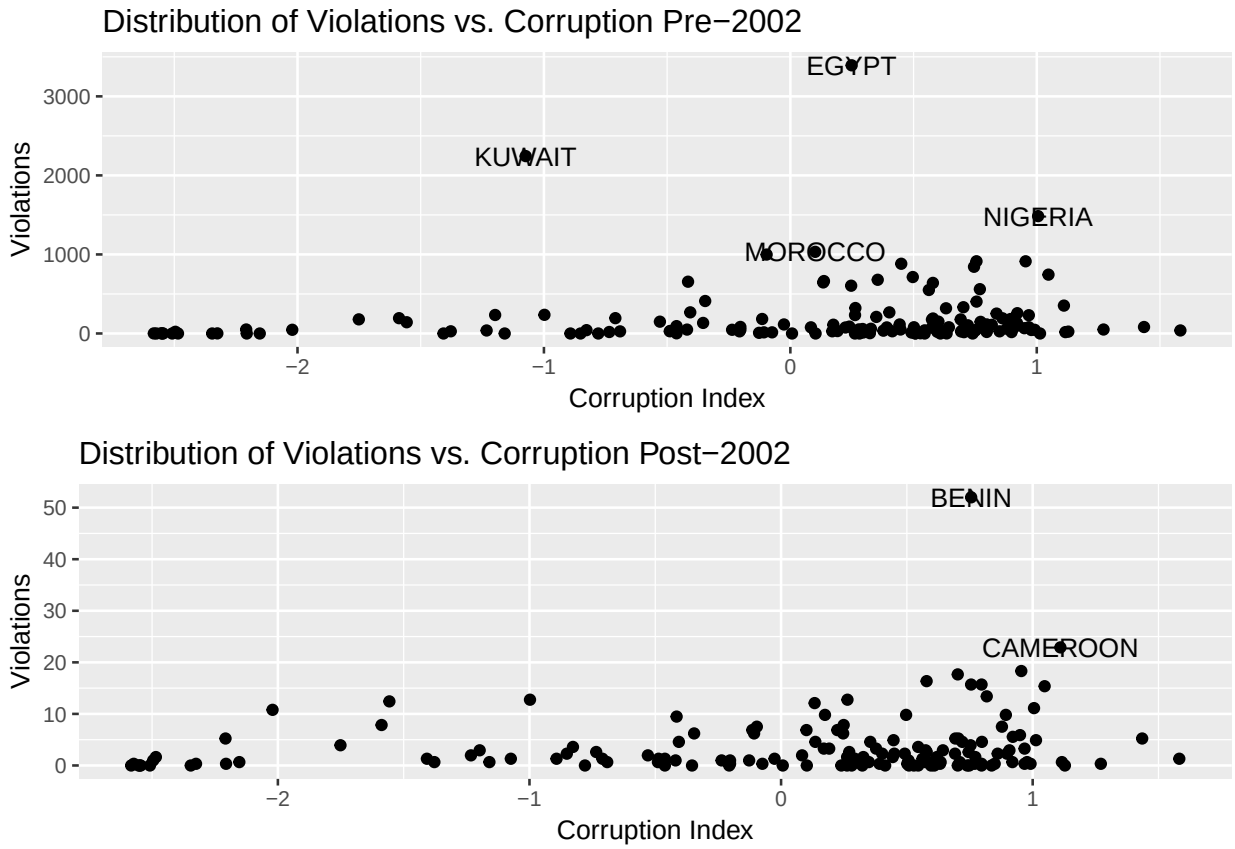
```
# Omit rows with incomplete observations (NA)
subcase2 = ! is.na(FMcorrupt$violations) & ! is.na(FMcorrupt$region) & ! is.na(FMcorrupt$prepost)
RegionAnalysisSubcase = FMcorrupt[subcase2, ]

library(ggplot2)
theme_set(theme_gray(base_size = 10))
p <- ggplot(RegionAnalysisSubcase, aes(region, violations, fill = prepost)) +
  geom_bar(stat="identity", position = "dodge") +
  scale_fill_brewer(palette = "Set1") + scale_x_discrete(name ="Region",
  limits=c("N. America", "S. America","Europe","Asia","Australia","Africa","Middle East"))
p + labs(y = "Violations", title = "Comparison of Pre- & Post-2002 Violations by Region")
```



When we examine the distribution of violations versus corruption, we note that countries with a higher corruption index generally tend to experience higher violations. However, there are several exceptions as we observe countries with high corruption indices yet low violations (particularly pre-2002). While we see a sharp reduction in total violations post-2002, we interestingly also observe that lower corruption countries (between -1 to -2 index) have similar violation totals as higher corruption countries (between 0 to 1 index). The graph suggests that some countries may respond differently to legal enforcement, since some countries have more drastic changes in violation totals than others.

```
library(ggplot2)
theme_set(theme_gray(base_size = 10))
country_pre <- ggplot(CorruptDataPre2002, aes(x = corruption))
country_pre <- country_pre + geom_point(stat = "identity", aes(y = violations))
country_pre <- country_pre + geom_text(data=subset(CorruptDataPre2002, corruption > -2 & violations > 1
```

```
            aes(corruption,violations,label=country), size = 3.5)
country_pre <- country_pre + labs(title = "Distribution of Violations vs. Corruption Pre-2002", x = "Co

country_post <- ggplot(CorruptDataPost2002, aes(x = corruption))
country_post <- country_post + geom_point(stat = "identity", aes(y = violations))
country_post <- country_post + geom_text(data=subset(CorruptDataPost2002, corruption > -2 & violations
            aes(corruption,violations,label=country), size = 3.5)
country_post <- country_post + labs(title = "Distribution of Violations vs. Corruption Post-2002", x =

library("gridExtra")
grid.arrange(country_pre, country_post)
```



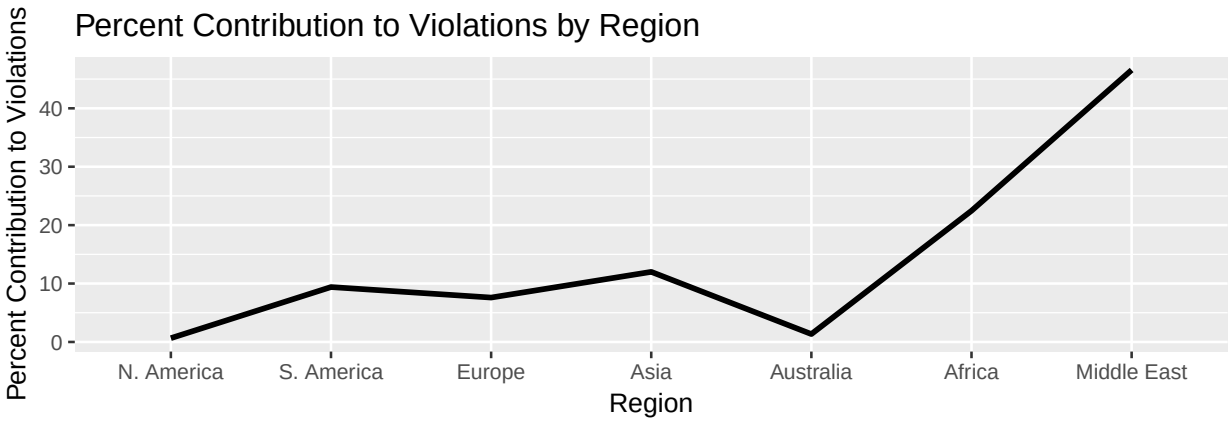
3. Relationship Between GDP Per Capita and Violations

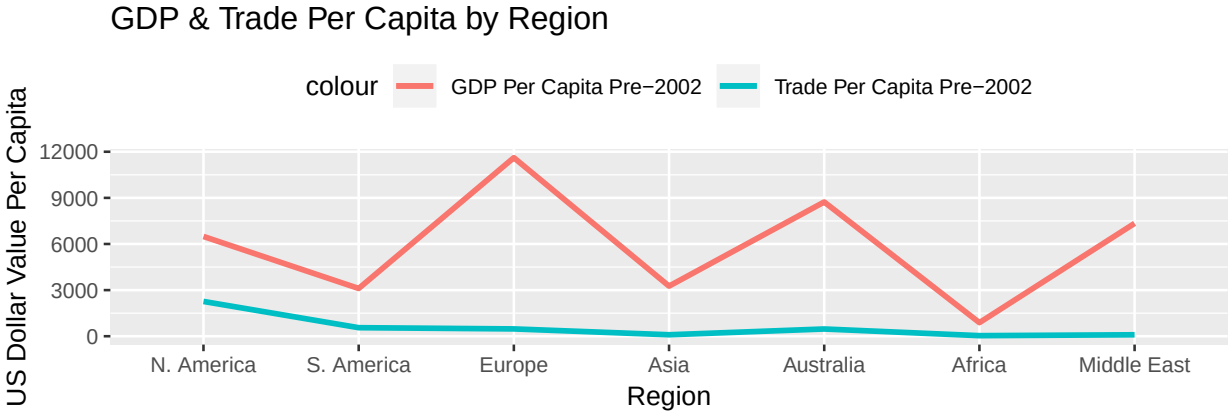
GDP per capita is a strong benchmark of economic strength, poverty rates, and standard of living (https://data.worldbank.org/products/wdi-maps). We expect that countries with a higher gdp per capita to have a lower contribution to violations since countries with a higher GDP per capita are less likely to be marred by corruption and crime (relative to lower income countries). From the visualization, we observe that our hypothesis is not supported: regions with high relative GDP per capita (e.g., North America, Australia, Europe, and the Middle East) experience varying levels of corruption. Nonetheless, the most staggering anomaly is the Middle East, which has one of the highest GDP per capita rates, yet contributes the most to violations (40% contribution). South America and Africa, in contrast, have relatively low GDP per capita rates, but contribute to violations by approximately 10% and 22%, respectively. The findings indicate that GDP per capita is not a strong indicator of region-based propensity to commit a violation.

The impact of trade per capita was also assessed, however the results show that trade per capita values are fairly similar between all the regions. The highest trade per capita is recorded for North and South America,

while the lowest trade per capita is recorded for Africa and the Middle East.

```
library(plyr)
subcase4 = ! is.na(FMcorrupt$trade) & ! is.na(FMcorrupt$violations) & ! is.na(FMcorrupt$pop1998) & ! is
trade_Subset = FMcorrupt[subcase4, ]
# Segmenting and averaging violations by region
violations_mean <- ddply(trade_Subset, .(region), summarize, sort = mean(violations))
# Segmenting and averaging economic aid by region
gdp_mean <- ddply(trade_Subset, .(region), summarize, sort = mean(ecaid))
# Segmenting and averaging 1998 mean population
pop_mean <- ddply(trade_Subset, .(region), summarize, sort = mean(pop1998))
# Segmenting and averaging 1998 GDP
gdppc_mean <- ddply(trade_Subset, .(region), summarize, sort = mean(gdppcus1998))
# Segmenting and averaging trade
tradepc_mean <- ddply(trade_Subset, .(region), summarize, sort = mean(trade))

library(ggplot2)
theme_set(theme_gray(base_size = 10))
gdp_effect <- data.frame(region_matrix = c(1,2,3,4,5,6,7),
                    gdp_avg = c(1.7, 2.69, 1.5, 2.02, 2.25, 5.21, 2.54),
                    violations_avg = c(7.77, 112.20, 90.04, 142.98, 16.10, 267.53, 554.68),
                    percent_violations = c(0.65, 9.4, 7.6, 12, 1.35, 22.46, 46.56),
                    violation_gdp_ratio = c(0.38, 3.49, 5.07, 5.94, 0.57, 4.31, 18.33),
                    trade_region = c(68450148582, 14923721833, 10695630376, 17399073745, 5302756664, 497
                    popmean_98 = c(30247900, 26834152, 22497306, 180547598, 11283000, 13786450, 3746408
                    trade_pc = c(2262.97, 556.15, 475.42, 96.36, 469.98, 36.07, 94.31),
                    gdppc_region = c(6488.9239, 3106.2359, 11613.7570, 3259.8396, 8732.5219, 883.0579, 

# Generating plot for pre-2002 violations by region
plot_violations <- ggplot(gdp_effect, aes(x = region_matrix))
plot_violations <- plot_violations + geom_line(stat = "identity", aes(y = percent_violations), size = 1
plot_violations <- plot_violations + labs(title = "Percent Contribution to Violations by Region", x = "
plot_violations <- plot_violations + scale_x_discrete(name ="Region", limits=c("N. America", "S. Americ

# Defining x-axis as region
plot_gdp <- ggplot(gdp_effect, aes(x = region_matrix, colour = ))
# Adding trade per capita
plot_gdp <- plot_gdp + geom_line(stat = "identity", aes(y = trade_pc, colour = "Trade Per Capita Pre-20
# Adding GDP per capita
plot_gdp <- plot_gdp + geom_line(stat = "identity", aes(y = gdppc_region, colour = "GDP Per Capita Pre-
# Renaming x-axis labels (translating 1, 2, 3,... dummy variables into region names)
plot_gdp <- plot_gdp + scale_x_discrete(name ="Region", limits=c("N. America", "S. America","Europe","A
plot_gdp <- plot_gdp + labs(title = "GDP & Trade Per Capita by Region", x = "Region", y = "US Dollar Va

library("gridExtra")
grid.arrange(plot_violations, plot_gdp)
```





4. Relationship Between Aid Received & Contribution to Violations

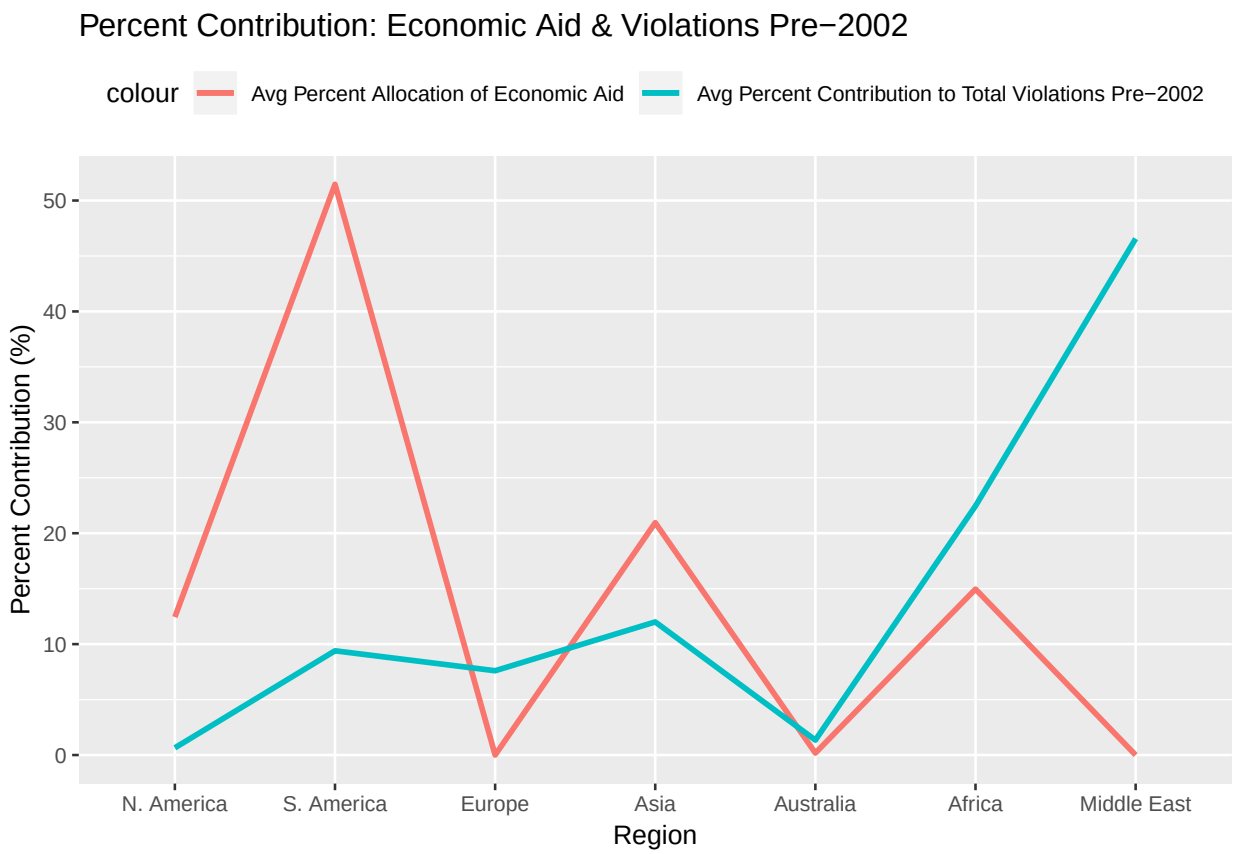
Studying economic aid is one possible approach to assess the strength of global political relationships. A key assumption in the analysis is that more aid is an indicator of a better diplomatic relationship. Given that aid contribution remains the same between pre-2002 and post-2002, we note that the United States provides the most economic aid to South America, followed by Asia, Africa, and itself (i.e., domestic aid). The Middle East, Europe, and Australia receive little to no aid from the United States. Africa is the largest contributor to violations (accounting for almost 46% of all violations) and the poorest recipient of economic aid. In contrast, South America is the highest recipient of economic aid (receiving almost 51% of the aid total) and contributes to 10% of the total violations. Australia and Europe are the only two regions where both aid and contributions to violations are very low. While the visualization does not suggest causation, there does appear to be a relationship between aid and violations at the extremes.

```
library(plyr)
library(dplyr)
```

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

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following object is masked from 'package:gridExtra':
##
##     combine

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

```
##
##     recode

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

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

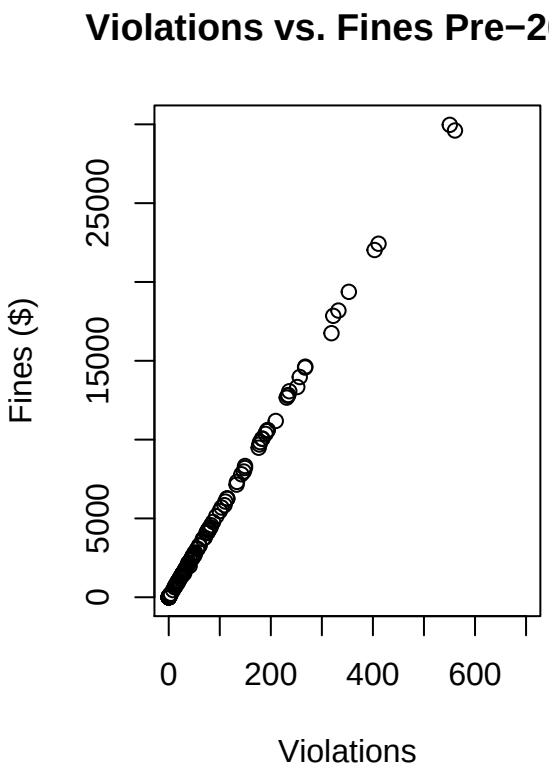
```
subcase5 = ! is.na(FMcorrupt$region) & ! is.na(FMcorrupt$violations) & ! is.na(FMcorrupt$prepost) & FMc
Ecaid_Subset = FMcorrupt[subcase5, ]
# Segmenting and averaging violations by region
violations_mean <- ddply(Ecaid_Subset, .(region), summarize, sort = mean(violations))
# Segmenting and averaging economic aid by region
ecaid_mean <- ddply(Ecaid_Subset, .(region), summarize, sort = mean(ecaid))

library(ggplot2)
theme_set(theme_gray(base_size = 10))
ecaid_effect <- data.frame(region_matrix = c(1,2,3,4,5,6,7),
                    economic_aid_avg = c(23.58, 97.49, 0, 39.7, 0.35, 28.34, 0),
                    percent_ecaid = c(12.44, 51.46, 0, 20.95, 0.18, 14.96, 0),
                    violations_avg = c(7.77, 112.20, 90.04, 142.98, 16.10, 267.53, 554.68),
                    percent_violations = c(0.65, 9.4, 7.6, 12, 1.35, 22.46, 46.56))
# Defining x-axis as region
plot_ecaid <- ggplot(ecaid_effect, aes(x = region_matrix, colour = ))
# Adding economic aid measure: US percent contribution to economic aid by region
plot_ecaid <- plot_ecaid + geom_line(aes(y = percent_ecaid, colour = "Avg Percent Allocation of Economi
# Adding violations measure: percent contribution to violations by region
plot_ecaid <- plot_ecaid + geom_line(aes(y = percent_violations, colour = "Avg Percent Contribution to
# Renaming x-axis labels (translating 1, 2, 3,... dummy variables into region names)
plot_ecaid <- plot_ecaid + scale_x_discrete(name ="Region", limits=c("N. America", "S. America","Europe
plot_ecaid <- plot_ecaid + labs(title = "Percent Contribution: Economic Aid & Violations Pre-2002", x =
plot_ecaid
```

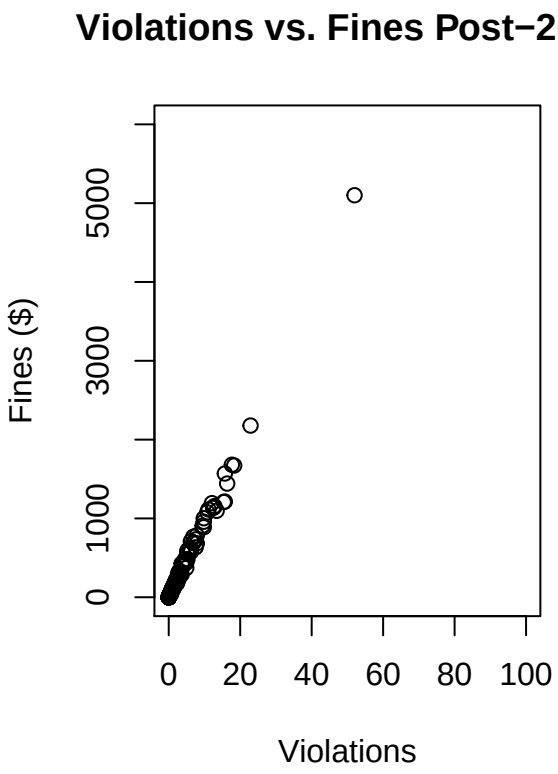


5. Analysis of Violations & Fines

We observe approximately 100% correlation between violations and fines for both pre-2002 and post-2002 data. The result is within expectation because the more the ticketed parking violations, the more the fines. Given there is a highly linear correlation, we can state that violations are a large contributor to fines. However, without further details on the violations (e.g., time issued, reason for violations, degree of repeat infractions), we cannot deduce the effect that fines have on curtailing violations. Moreover, the data set does not provide details on whether a fine was paid or not (or what portion of the fine was paid and by whom). Gaining more information on such subtleties of receiving, handling, and paying a fine can potentially shed light on how fines impact the propensity to violate.

```
par(mfrow=c(1,2)) # par function used to place graphs side-by-side
plot(CorruptDataPre2002$violations,CorruptDataPre2002$fines, xlim = c(0,700),ylim=c(0,30000), main = "V
plot(CorruptDataPost2002$violations,CorruptDataPost2002$fines, xlim = c(0,100),ylim=c(0,6000), main = "
```






```
cor(CorruptDataPre2002$fines,CorruptDataPre2002$violations)
```

```
## [1] 0.9999651
```

```
cor(CorruptDataPost2002$fines,CorruptDataPost2002$violations)
```

```
## [1] 0.9962667
```

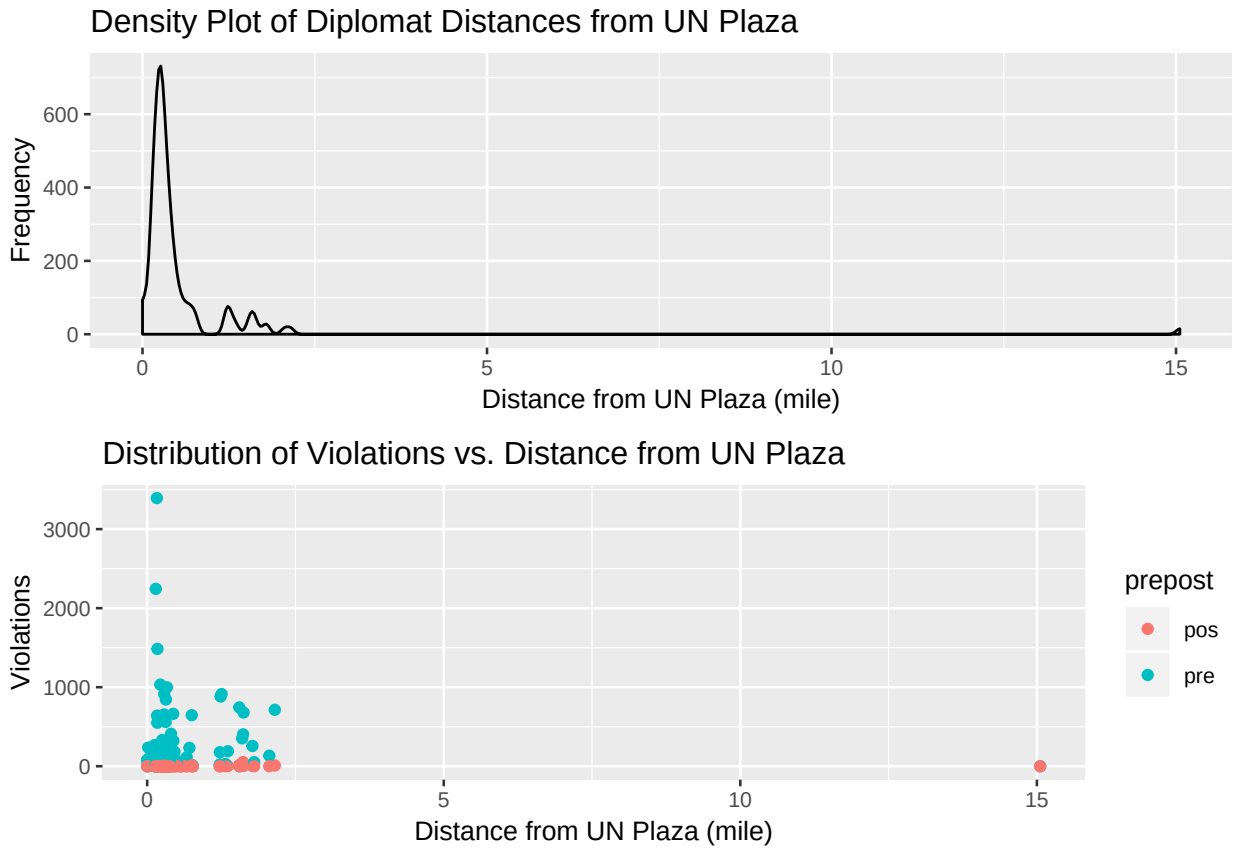
Analysis of Secondary Effects

1. Effect of Distance from the UN Plaza

From the graph, we see that the distance from the UN Plaza should not have significant influence on violations because 87% of the diplomats are within one mile of the UN Plaza. The frequency plot of violations mirrors the distribution of the UN Plaza distances, which counters our prediction because we expected violations pre-2002 to decrease as diplomats became closer to the UN Plaza. Given that a significant portion of violations happened within one mile of the UN Plaza, the graph suggests that there may be a degree of nonchalance to receiving a parking violation pre-2002 (which points to the cultural norms of the time). Post-2002, however, we observe that violations have sharply flat lined near zero irrespective of the distance from UN Plaza, supporting the argument that legal enforcement has strong influence on parking behavior.

```
# Creating a sub-case to extract only complete observations from distUNplz and violations
subcase6 = ! is.na(FMcorrupt$distUNplz) & ! is.na(FMcorrupt$violations) & ! is.na(FMcorrupt$cars_missio
UNdist_Subset = FMcorrupt[subcase6, ]

# Creating frequency plot of distUNplz
UNdist <- ggplot(UNdist_Subset, aes(x = distUNplz))
UNdist <- UNdist + geom_density(aes(y = ..count..))
UNdist <- UNdist + labs(title = "Density Plot of Diplomat Distances from UN Plaza", x = "Distance from
```

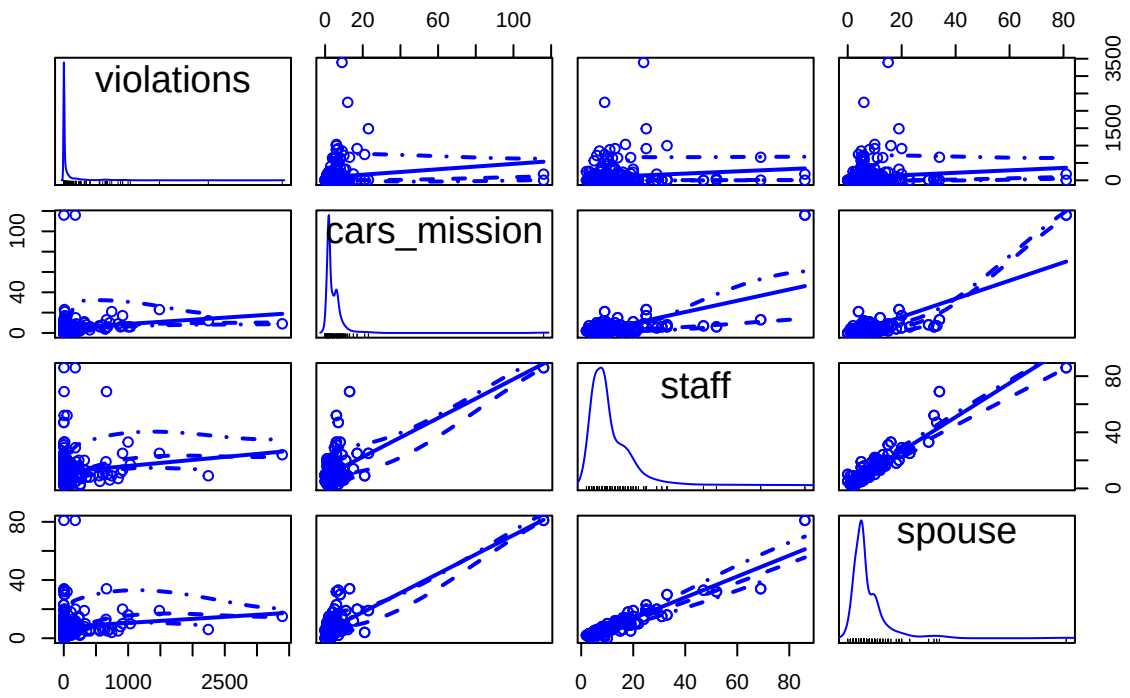
```r
# Adding distribution of violations
violations_dist <- ggplot(UNdist_Subset, aes(x = distUNplz, y = violatons, colour = prepost))
violations_dist <- violations_dist + geom_point(stat = "identity", aes(y = violations, fill = prepost))

# Stacking graphs on top of each other for comparison
library("gridExtra")
grid.arrange(UNdist, violations_dist)
```



2. Effects of Car & Staff Allocation

We observe that violations have a positive but low correlation with the number of mission-designated cars, the number of staff, and the number of spouses for a given country. Further analysis was conducted on the impact of mission car availability and spousal status on violations.

```r
library(car)
scatterplotMatrix( ~ violations + cars_mission + staff + spouse, data = FMcorrupt,
                   main = "Scatterplot Matrix")
```

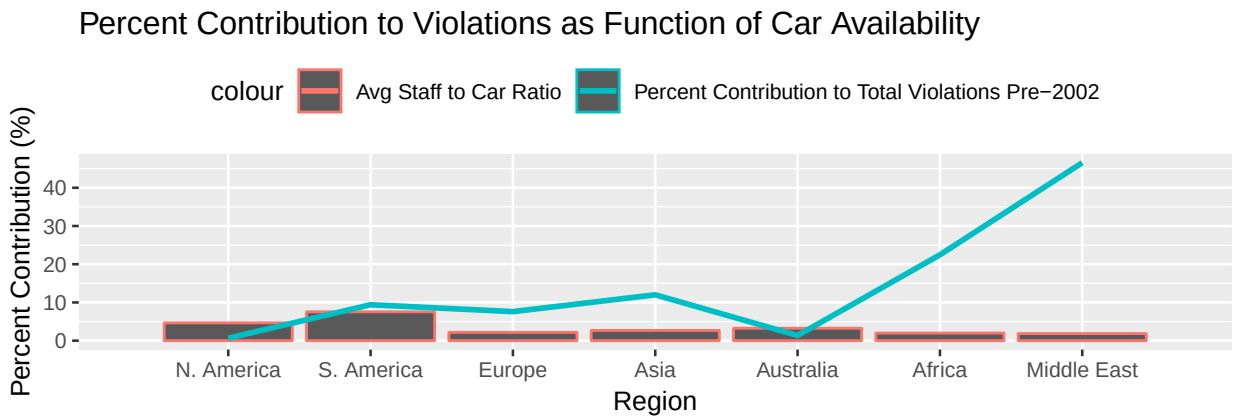


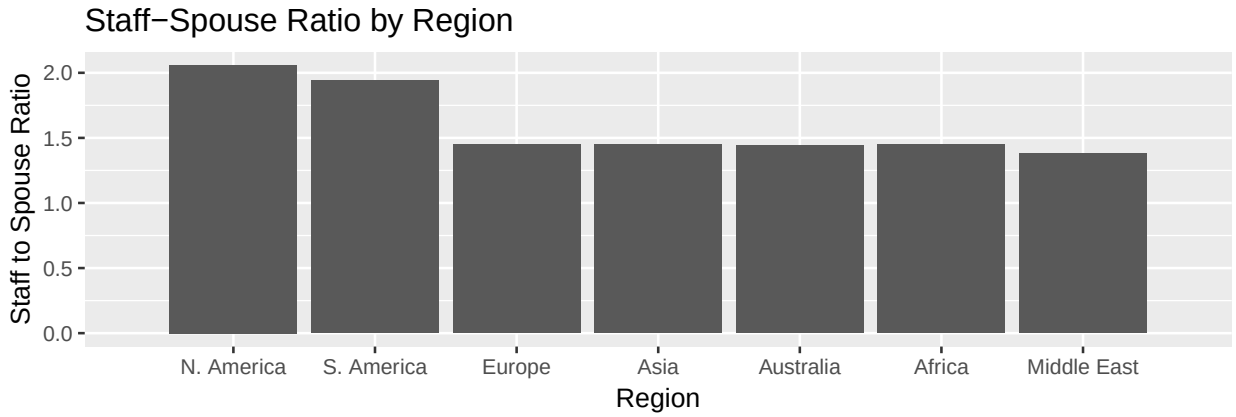
The ratio of staff to mission cars was calculated to gauge how many diplomats share a car. The rationale is that if the staff to cars ratio is high (i.e., lot of diplomats sharing one car), then violations would be high because diplomats would be less likely to feel a personal responsibility towards taking care of the car and parking it in violation-free zones. Similarly, if the staff to car ratio is low, then diplomats may experience a greater sense of ownership with the car and park it in safe zones only. The hypothesis was disputed as Africa and Europe, the two countries with the lowest staff to car ratio, had the highest violations. North America, South America, and Australia had the lowest staff to car ratios along with the lowest relative violation levels.

A secondary analysis was performed on the spouse variable to determine if martial status had any effect on propensity to violate parking rules. North and South America have the most married diplomats, while the remaining regions are similar in their distribution of marital status. The results show that the presence (or absence) of a spouse does not impact violations.

```
# Create new subcase and omit rows with incomplete observations (NA)
subcase8 = ! is.na(FMcorrupt$violations) & ! is.na(FMcorrupt$staff) & ! is.na(FMcorrupt$cars_total)
CarAnalysisSubcase = FMcorrupt[subcase8, ]
# Segmenting and averaging mission car count by region
car_mean <- ddply(CarAnalysisSubcase, .(region), summarize, sort = mean(cars_mission))
# Segmenting and averaging staff numbers by region
staff_mean <- ddply(CarAnalysisSubcase, .(region), summarize, sort = mean(staff))
# Segmenting and averaging spouse count by region
spouse_mean <- ddply(CarAnalysisSubcase, .(region), summarize, sort = mean(spouse))

library(ggplot2)
car_effect <- data.frame(region_matrix = c(1,2,3,4,5,6,7),
                         car_region_avg = c(3, 1.71, 7.53, 5.16, 2, 4.02, 6.33),
                         staff_region_avg = c(14, 12.89, 16.41, 13.84, 6.5, 8.02, 11.73),
```

```
                    staff_car_ratio = c(4.67, 7.54, 2.18, 2.68, 3.25, 1.99, 1.85),
                    violations_avg = c(7.77, 112.20, 90.04, 142.98, 16.10, 267.53, 554.68),
                    percent_violations = c(0.65, 9.4, 7.6, 12, 1.35, 22.46, 46.56),
                    spouse_region_avg = c(6.8, 6.65, 11.28, 9.52, 4.5, 5.54, 8.47),
                    staff_spouse_ratio = c(2.06, 1.94, 1.45, 1.45, 1.44, 1.45, 1.38))

library(ggplot2)
theme_set(theme_gray(base_size = 10))

# Defining x-axis as region
plot_cars <- ggplot(car_effect, aes(x = region_matrix, colour = ))
# Adding staff-to-car calculation
plot_cars <- plot_cars + geom_bar(stat = "identity", aes(y = staff_car_ratio, colour = "Avg Staff to Ca
# Adding violations measure: percent contribution to violations by region
plot_cars <- plot_cars + geom_line(aes(y = percent_violations, colour = "Percent Contribution to Total
# Renaming x-axis labels (translating 1, 2, 3,... dummy variables into region names)
plot_cars <- plot_cars + scale_x_discrete(name ="Region", limits=c("N. America", "S. America","Europe",
plot_cars <- plot_cars + labs(title = "Percent Contribution to Violations as Function of Car Availabili

library(ggplot2)
# Defining x-axis as region
plot_spouse <- ggplot(car_effect, aes(x = region_matrix))
# Adding staff-to-car calculation
plot_spouse <- plot_spouse + geom_bar(stat = "identity", aes(y = staff_spouse_ratio))
# Renaming x-axis labels (translating 1, 2, 3,... dummy variables into region names)
plot_spouse <- plot_spouse + scale_x_discrete(name ="Region", limits=c("N. America", "S. America","Europ
plot_spouse <- plot_spouse + labs(title = "Staff-Spouse Ratio by Region", x = "Region", y = "Staff to Sp

# Stacking graphs on top of each other for comparison
library("gridExtra")
grid.arrange(plot_cars, plot_spouse)
```



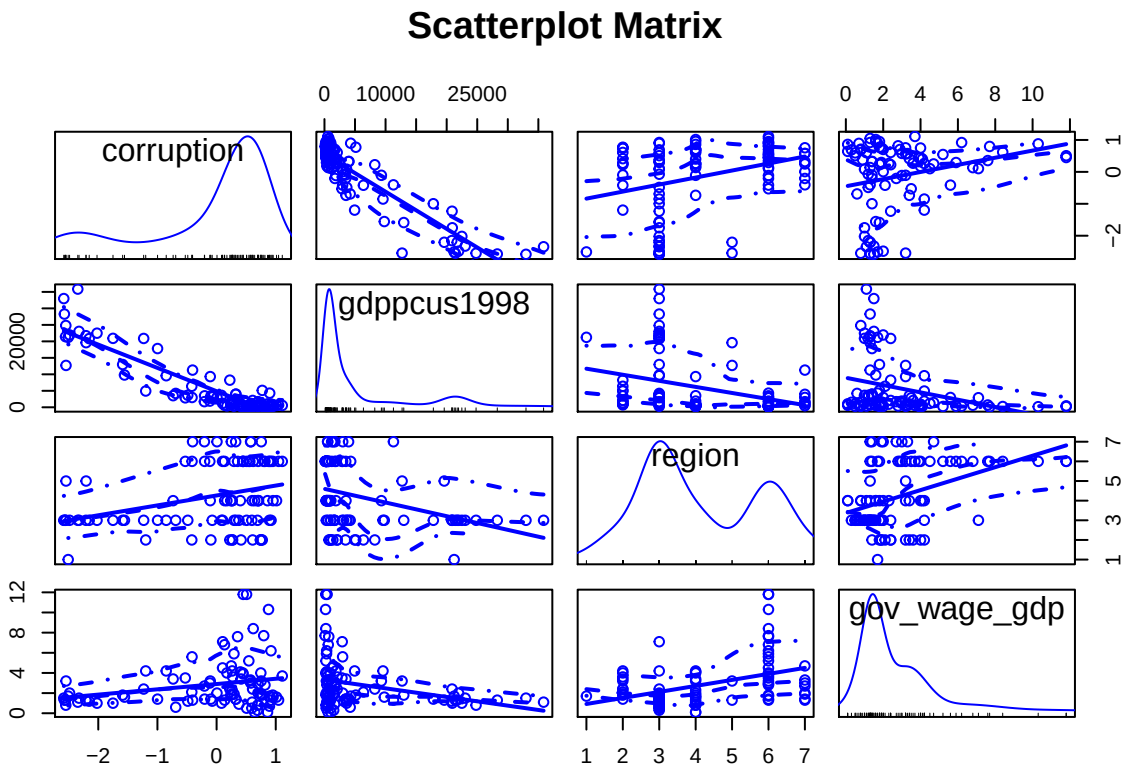


3. Confounding Effects: Corruption

Due to the multivariate nature of the analysis, it is challenging to distinguish between the effects of legal enforcement and social norms on corruption. While we notice that all regions unilaterally respond well to legal enforcement, it is difficult to disentangle the results to attribute certain findings to cultural norms alone. Thus, we experience confounding by the nature of societal function: crime-heavy countries have high corruption, which simultaneously feeds into a weak anti-corruption culture and less legal enforcement. Although violations represent the dependent variable of interest, in order to better understand the factors causing violations, it is necessary to understand the factors causing corruption.

The scatter plot compares corruption to pre-2002 GDP, region, and government wage GDP. Based on region versus corruption comparisons, it is evident that South America, Asia, Africa, and the Middle East (corresponding to 2, 4, 6, and 7 in region dummy variables) have the highest levels of corruption. Throughout the analysis, however, we observed that South America and Asia were one of the lower contributors to violations. We also observe a negative correlation between corruption and GDP per capita, with higher corruption countries having a lower GDP per capita. Nonetheless, our previous analyses have shown that GDP per capita is not a strong indicator for propensity to commit a violation. Hence, it is necessary to gather more contextual details on the variables at hand to more effectively decouple trending observations from causative forces.

```
library(car)
# Create new subcase and omit rows with incomplete observations (NA)
subcase9 = ! is.na(FMcorrupt$region) & ! is.na(FMcorrupt$corruption) & ! is.na(FMcorrupt$gdppcus1998) &
corruption_subset = FMcorrupt[subcase9, ]

scatterplotMatrix( ~ corruption + gdppcus1998 + region + gov_wage_gdp, data = corruption_subset,
                   main = "Scatterplot Matrix")
```



Conclusions

Shedding light on the research questions stated in the Introduction, the conclusions for this exploratory data analysis are three-pronged:

1. Cultural norms and legal inforcement are both important determinants of corruption.

Pre-2002, we observe that diplomats from high-corruption countries accumulated significantly more parking violations than diplomats from low-corruption countries, indicating that cultural norms (i.e., immunity) play a role in parking behavior. After 2002, when authorities acquired the right to confiscate diplomatic license plates, violations dropped sharply, indicating that law enforcement can curtail corruption. Given that corruption indices did not change from pre-2002 to post-2002 despite a decline in violations, we can infer that legal enforcement has a larger effect on parking violations than does the variation in cultural norms across countries.

2. Violations do not impact the corruption index of a country. However, corruption index can impact violations.

When UN diplomats lost immunity, there was a precipitous decline in the total number of violations (by ~50 fold) post-2002. Nonetheless, corruption indices remained identical pre-2002 and post-2002. As a result, we can infer that violations have little to no impact on corruption index. From an inverse perspective, however, we observe that countries with lower corruption indices (e.g., Norway and Denmark) exhibit a relatively low number of violations (even without legal enforcement). In contrast, countries with high corruption indices (e.g., Nigeria, Cameroon) exhibit a relatively high number of violations (even with law enforcement). Hence, corruption can be one predictor of violations and parking behavior, but it is important to note that corruption is not the only variable contributing to violations.

3. Factors that can be predictors of violations (or propensity to commit a violation) include corruption, region of origin, economic aid, legal enforcement, and cultural norms.

From the analysis, we see that countries with a lower corruption index tend to have a lower number of violations. We also observe telling regional differences, with diplomats from Africa, South America, and the Middle East contributing most to violations. Regions that received the least economic aid (e.g., Africa and the Middle East) also had high relative contributions to violations. Factors that do not appear to impact violations include marital status, number of cars, distance from the UN Plaza, and trade per capita. Due to confounding, it is difficult to decouple the effects of legal enforcement and cultural norms on violation reductions. However, it is highly likeley strong cultural norms are at play due to discrepant observations from several of key relationship explored. Legal enforcement is also very effective at curbing violations, as all regions experienced a significant decrease in violations pre- and post-2002. For further analysis, it will be important to identify the root cause(s) of corruption to better understand the propensity of a diplomat to violate parking regulations. Furthermore, it would be helpful to gather data on variables that can paint a better picture of a country's (or region's) cultural norms (e.g., crime rates, happiness indices, average work hours, median household income, poverty rate, rate of immigration/emigration, etc.) In addition, it is necessary to uncover more details on violations (e.g., date/time of violation, reason code, initial dollar value of fine, amount paid towards the fine, the country to which the car was registered, etc.).