Corruption and Parking Violations

C. Akkineni, K. Hanna, A. Thorp

September 26, 2016

```
setwd("C:/Users/kevin/OneDrive/School/MIDS/W203 - Statistics for Data Science/Lab 1/W203_lab1_corruptio
#library(car)
#library(grid)
library(ggplot2)
library(knitr)
library(kableExtra)

load("Corrupt.Rdata")

## Correct Data Problems
#Fix majoritymuslim where value = -1, should be 0
FMcorrupt[FMcorrupt$majoritymuslim == -1 & ! is.na(FMcorrupt$majoritymuslim), "majoritymuslim"] = 0


# Add missing counties.  Reference: https://www.worldatlas.com/aatlas/ctycodes.htm
FMcorrupt[FMcorrupt$wbcode == "ARE", "country"] = "UNITED ARAB EMIRATES"
FMcorrupt[FMcorrupt$wbcode == "CAF", "country"] = "CENTRAL AFRICAN REPUBLIC"
FMcorrupt[FMcorrupt$wbcode == "CAN", "country"] = "CANADA"
FMcorrupt[FMcorrupt$wbcode == "COL", "country"] = "COLUMBIA"
FMcorrupt[FMcorrupt$wbcode == "ECU", "country"] = "ECUADOR"
FMcorrupt[FMcorrupt$wbcode == "JAM", "country"] = "JAMAICA"
FMcorrupt[FMcorrupt$wbcode == "LVA", "country"] = "LATVIA"
FMcorrupt[FMcorrupt$wbcode == "NOR", "country"] = "NORWAY"
FMcorrupt[FMcorrupt$wbcode == "PAN", "country"] = "PANAMA"
FMcorrupt[FMcorrupt$wbcode == "SWE", "country"] = "SWEDEN"
FMcorrupt[FMcorrupt$wbcode == "TUR", "country"] = "TURKEY"

# Mexico is part of NORTH America, not SOUTH America
FMcorrupt[FMcorrupt$wbcode == "MEX", "region"] = 1


# Create named regions variable using region
FMcorrupt$region_name = NA
FMcorrupt[FMcorrupt$region == 1 & ! is.na(FMcorrupt$region), "region_name"] = "North America"
FMcorrupt[FMcorrupt$region == 2 & ! is.na(FMcorrupt$region), "region_name"] = "South America"
FMcorrupt[FMcorrupt$region == 3 & ! is.na(FMcorrupt$region), "region_name"] = "Europe"
FMcorrupt[FMcorrupt$region == 4 & ! is.na(FMcorrupt$region), "region_name"] = "Asia"
FMcorrupt[FMcorrupt$region == 5 & ! is.na(FMcorrupt$region), "region_name"] = "Oceania"
FMcorrupt[FMcorrupt$region == 6 & ! is.na(FMcorrupt$region), "region_name"] = "Africa"
FMcorrupt[FMcorrupt$region == 7 & ! is.na(FMcorrupt$region), "region_name"] = "Middle East"

FMcorrupt$region_name = factor(FMcorrupt$region_name)

# Remove 66 rows that do not have relevant data to the key analyses
corrupt = subset(FMcorrupt, !is.na(violations) & !is.na(mission) & !is.na(staff) )
```

```r
# split data in to pre and post, before and after enforcement changes
cor_pre = subset(corrupt, prepost == "pre")
cor_pos = subset(corrupt, prepost == "pos")

# Merge both the above to one line with pre and pos appeneded to variable names (prepos removed)
cor_oneline = merge(cor_pre, cor_pos, by = "wbcode", suffixes = c(".pre", ".pos"))

# Grab only the variables that are needed.
cor_oneline = cor_oneline[, c("wbcode", "violations.pre",  "violations.pos", "fines.pre", "fines.pos",
                              "mission.pre", "staff.pre", "spouse.pre", "gov_wage_gdp.pre", "pctmuslim.p
                              "majoritymuslim.pre", "trade.pre",
                              "cars_total.pre", "cars_mission.pre", "pop1998.pre", "gdppcus1998.pre", "e
                              "milaid.pre", "corruption.pre", "totaid.pre",
                              "r_africa.pre", "r_middleeast.pre", "r_europe.pre", "r_southamerica.pre",
                              "country.pre", "distUNplz.pre",
                              "region.pre", "region_name.pre"
                              )]

# Remove suffix where not needed.
colnames(cor_oneline) =  c("wbcode", "violations.pre",  "violations.pos", "fines.pre", "fines.pos",
                              "mission", "staff", "spouse", "gov_wage_gdp", "pctmuslim", "majoritymuslin
                              "cars_total", "cars_mission", "pop1998", "gdppcus1998", "ecaid", "milaid"
                              "r_africa", "r_middleeast", "r_europe", "r_southamerica", "r_asia",
                              "country", "distUNplz",
                              "region", "region_name"
                              )

# Rename FMcorrupt to ensure we don't use it accidentally
cor_nas = FMcorrupt
remove(FMcorrupt)

# Variable alterations and transformations for analysis
correlation_matrix_input = cor_oneline[, c("corruption", "violations.pre", "fines.pre", "violations.pos
                                           "staff", "spouse", "majoritymuslim", "pctmuslim", "trade",
                                           "cars_total", "cars_mission", "totaid", "gdppcus1998"
                                                       )]
#add log transformations
#(with the addition of 1 to the log transform to circumvent issues with the values between 0 and 1)
correlation_matrix_input$violations.pos.log = log(correlation_matrix_input$violations.pos+1)
correlation_matrix_input$violations.pre.log = log(correlation_matrix_input$violations.pre+1)
# add violations treated with total cars
correlation_matrix_input$violations_weighted.cars_total.pre =
  correlation_matrix_input$violations.pre/correlation_matrix_input$cars_total
correlation_matrix_input$violations_weighted.cars_total.pos =
  correlation_matrix_input$violations.pos/correlation_matrix_input$cars_total

#add log transformations
#The addition of 1 to the normalized number of violations would not be acceptable for predictive or cau
# will help us circumvent the 0 - 1 log transformation issues while maintaining the sequence of observa

correlation_matrix_input$violations_weighted.cars_total.pre.log =
  log(correlation_matrix_input$violations_weighted.cars_total.pre+1)
```

```r
correlation_matrix_input$violations_weighted.cars_total.pos.log =
  log(correlation_matrix_input$violations_weighted.cars_total.pos+1)

# add violations treated with staff
correlation_matrix_input$violations_weighted.staff.pre =
  correlation_matrix_input$violations.pre/correlation_matrix_input$staff

correlation_matrix_input$violations_weighted.staff.pos =
  correlation_matrix_input$violations.pos/correlation_matrix_input$staff

# add violations treated with total people (staff + spouse)
correlation_matrix_input$total_people =
  correlation_matrix_input$staff + correlation_matrix_input$spouse

correlation_matrix_input$violations_weighted.total_people.pre =
  correlation_matrix_input$violations.pre/correlation_matrix_input$total_people

correlation_matrix_input$violations_weighted.total_people.pos =
  correlation_matrix_input$violations.pos/correlation_matrix_input$total_people

# add violations treated with cars_mission
correlation_matrix_input$violations_weighted.cars_mission.pre =
  correlation_matrix_input$violations.pre/correlation_matrix_input$cars_mission
correlation_matrix_input$violations_weighted.cars_mission.pos =
  correlation_matrix_input$violations.pos/correlation_matrix_input$cars_mission


#add log transformations
#The addition of 1 to the normalized number of violations would not be acceptable for predictive or cau
# will help us circumvent the 0 - 1 log transformation issues while maintaining the sequence of observa

correlation_matrix_input$violations_weighted.staff.pre.log =
  log(correlation_matrix_input$violations_weighted.staff.pre+1)
correlation_matrix_input$violations_weighted.staff.pos.log =
  log(correlation_matrix_input$violations_weighted.staff.pos+1)
correlation_matrix_input$violations_weighted.total_people.pre.log =
  log(correlation_matrix_input$violations_weighted.total_people.pre+1)
correlation_matrix_input$violations_weighted.total_people.pos.log =
  log(correlation_matrix_input$violations_weighted.total_people.pos+1)
correlation_matrix_input$violations_weighted.cars_mission.pre.log =
  log(correlation_matrix_input$violations_weighted.cars_mission.pre+1)
correlation_matrix_input$violations_weighted.cars_mission.pos.log =
  log(correlation_matrix_input$violations_weighted.cars_mission.pos+1)

correlation_matrix_input$totaid.log = log(correlation_matrix_input$totaid + 1)

# Utility functions
#plot the relationship
ggplotRegression <- function (fit, title, x, y) {

  require(ggplot2)

  ggplot(fit$model, aes_string(x = names(fit$model)[2], y = names(fit$model)[1])) +
```

```
    geom_point() +
    stat_smooth(method = "lm", col = "red") +
    labs(title = title, subtitle = paste("Adj R2 = ",signif(summary(fit)$adj.r.squared, 3),
                        "Intercept =",signif(fit$coef[[1]], 3),
                        " Slope =",signif(fit$coef[[2]], 3),
                        " P =",signif(summary(fit)$coef[2,4], 3)),
                         x = x,
                         y = y)
}

round_df <- function(x, digits) {
    # round all numeric variables
    # x: data frame
    # digits: number of digits to round
    numeric_columns <- sapply(x, mode) == 'numeric'
    x[numeric_columns] <-  round(x[numeric_columns], digits)
    x
}
```

Introduction

Research Question

Prior to 2002 diplomats at UN missions were exempt from parking violations and fines in New York City, by virtue of their diplomatic immunity. There was wide variation in diplomats' willingness to adhere to local parking laws. This analysis attempts to understand whether the variation in adherance to local parking law was related to cultural norms around corruption in the diplomats' home countries. For the period prior to 2002, we examine the relationship between perceptions of corruption in that country and that country's diplomats' willingness to incur parking violations. In 2002 NYC parking enforcement acquired the right to confiscate license plates from vehicles belonging to foreign diplomats if they had accumulated unpaid parking violations, thereby making payment a function of both cultral norms and legal enforcment. This had a notable compressing effect on parking law adherence, with violations incurred dropping dramatically across the sample.

Question: Does an index of perceived corruption in the diplomats' home country have explanatory power for a given diplomatic mission's compliance with local parking regulations?

Description of Dataset

Our dataset has a total of 364 observations. Of the 364, 66 observations contain only economic data leaving NA for our dependent variable, violations. Considering the countries that are among these 66 and the variables for which they have valid data, we suspect these rows result from a merge of economic data with the violations data. As such, we believe these 66 countries represent a data artefact from that data merge. As these observations do not contain valid values for key variables, we remove them from our dataset. Of note, those 66 observations appear to contain many which do not even have a mission or staff in New York City, and as such are not relevant for this study on diplomatic parking violations in New York City. Given these considerations, we feel comfortable that we are not biasing the results of the study by removing these observations. With those 66 rows removed, we're left with 298 observations (two observations for each of 149 countries where corruption data exists.) Each country has one observation from prior to the 2002 regulation change and one observation from after. Only the 'violations' and 'fines' variables differ between the two observations for a given country, while other variables remain constant. As such, we recombine the data to

result in 149 observations with pre and post columns for the violations and fines variables (the only ones that differ).

Univariate Analysis of Key Variables

We start our exploratory analysis with a high level look at all our variables to better understand what our variables represent and to find and correct quality problems.

High Level Univariate Analysis

```
# Use CSV version of Google Sheet 'Variable Description for Introduction': https://docs.google.com/spre
variable_description = read.csv("Lab 1 - Variable Descriptions for Introduction.csv",
                                header = TRUE, sep = ",", quote = "\"", allowEscapes = TRUE)
#summary(variable_description)

kable(variable_description, "latex", longtable = TRUE, booktabs = TRUE, caption = "Dataset Variables") %
  kable_styling(full_width = TRUE, latex_options = c("HOLD_position", "striped", "repeat_header"), row_
```

Table 1: Dataset Variables

Variable	Description	Observations	Alterations
wbcode	Three Letter Country code		
country	country name	Some missing values, though wbcode are available.	Missing values were filled using wbcode variable.
corruption	Country corruption index	Min. = -2.58299 (Least corrupt), Max. = 1.58281 (Most corrupt)	
violations.pre	The number of violations accumulated before enforcement changes.	Values have 6 decimal places, we don't quite understand why, however the instructions we were given with the data set suggest these values should be integer, so we will treat it as the sum.	
violations.pos	The number of violations accumulated after enforcement changes.	Values have 6 decimal places, we don't quite understand why, however the instructions we were given with the data set suggest these values should be integer, so we will treat it as the sum.	
fines.pre	Summed cost of violations.pre adjusted for inflation.		

Table 1: Dataset Variables *(continued)*

Variable	Description	Observations	Alterations
fines.pos	Summed cost of violations.pos adjusted for inflation.		
staff	# of mission employees		
spouse	# of mission employee spouses		
cars_personal	Cars owned by staff of mission	Total of 740, 10 NA's	
cars_mission	Cars owned by mission	Total of 715, 10 NA's	
cars_total	The number of cars owned by both mission and the employees.	Total of 1455, 10 NA's	
gov_wage_gdp	Percentage of country's GDP paid to all government employees.		
gdppcus1998	GDP Per Capita in USD 1998		
ecaid	US Economic Aid to Country		
milaid	Military Aid		
totaid	milaid + ecaid		
pctmuslim	Percent of country that identifies as Muslim	2 countries have NA's: Bosnia-Herzegovina and Zaire	
trade	Total trade with the US in 1998 USD	2 countries have NA's: Bosnia-Herzegovina and Zaire, 1 country shows 0 Libia	
region	Countries grouped in to goegraphic regions	Mexico was included in South America instead of North America	Moved Mexico to North America (Welcome back Mexico)
region_name	7 Geographic regions		We created this variable using region variable above
pop1998	Country's population in 1998		
distUNplz	Distance from country's embassy to UN Plaza in Miles		
majoritymuslim	Is country majority Muslim	This variable had some values of -1 which occured if and only if pctmuslim was 0.	We replaced all -1's with -0's.
mission	If country has mission in NYC	All values are true.	

Both the corruption and violations variables are essential to our examination. There are several variables which reflect the size of the mission (staff, spouse, cars) These are explored in the analysis of key relationships. Total aid provided by the US and the country's GDP per capita are also of interest and we'll explore those more in the Analysis of Secondary Effects.

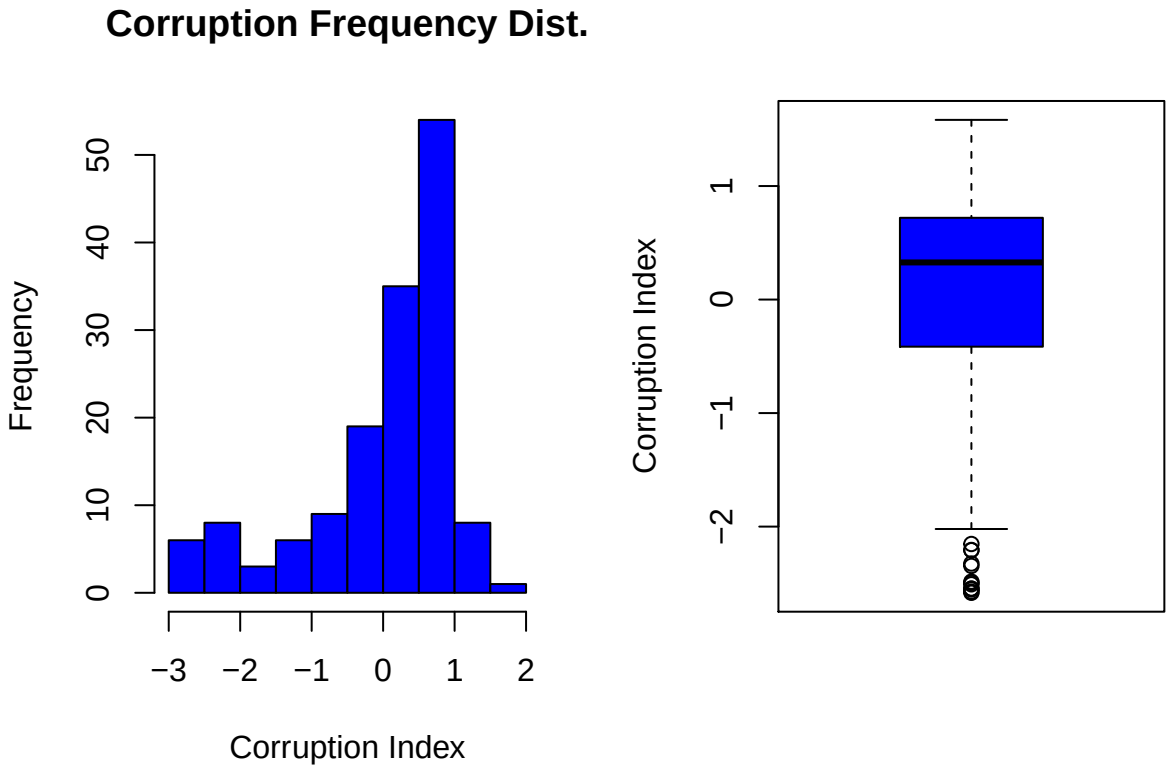
Further Analysis of Key Variables

Key Variable Summary Data

	Mean	Sum	Standard Deviation
Country Violations Before Regulation Change	198.07	29,512.54	405.28
Country Violations After Regulation Change	3.69	549.46	6.01
Country Violations/Total Cars Before Regulation Change	22.42	3,116.38	31.71
Country Violations/Total Cars After Regulation Change	0.62	86.11	1.53
Total Cars for Mission	10.47	1,455.00	13.98
Trade with US (1998 Billions $US)	10.25	1,506.59	35.65
Total Economic and Military Aid Provided by US (1998 $US)	82.32	12,101.00	385.72
GDP Per Capita (1998 $US)	5,044.09	751,569.00	7,971.67

Corruption

```
par(mfrow = c(1, 2))
hist(correlation_matrix_input$corruption, col="blue", main = "Corruption Frequency Dist.", xlab = "Corru
boxplot(correlation_matrix_input$corruption, col = "blue", ylab = "Corruption Index")
```

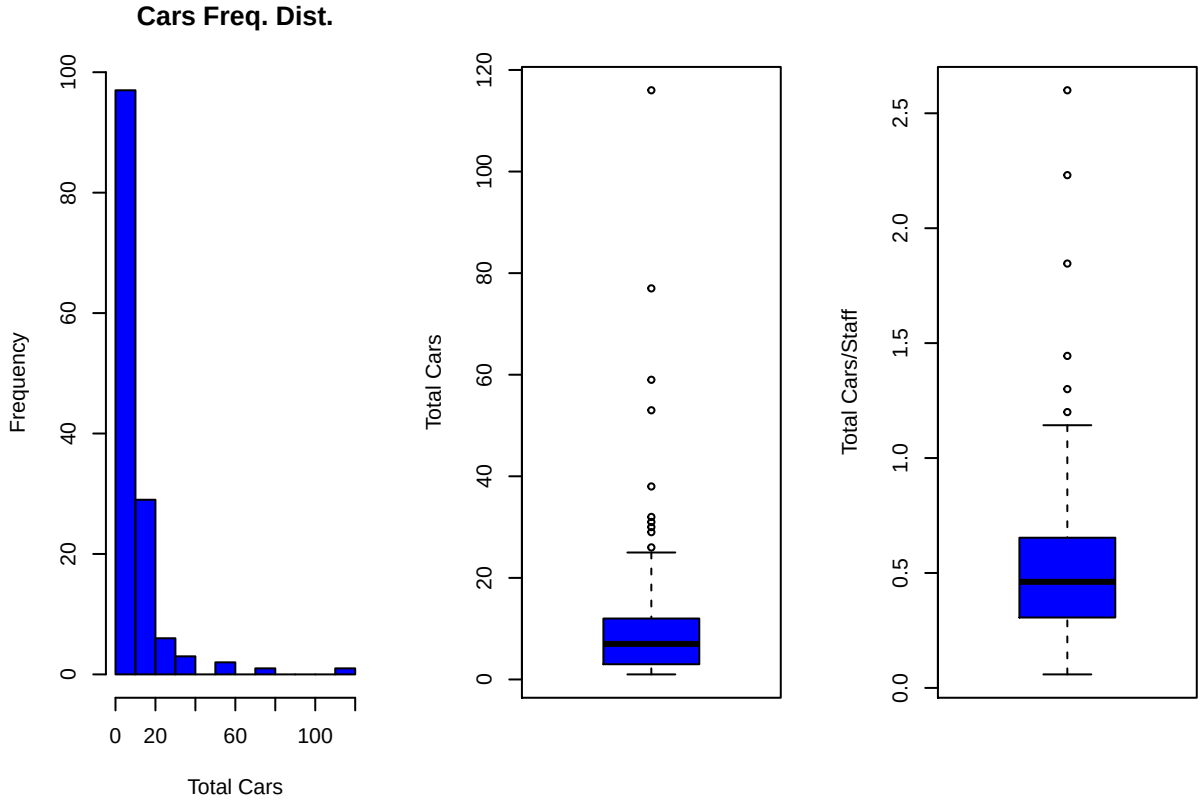


We don't fully understand the corruption variable other than to say it appears to be an index. Neither the dataset we were provided nor the problem description provided a description of the corruption variable's origin. The frequency distribution skews negative, however the median (0.33) is pretty close to 0 which

suggests the values might be deliberately scaled to balance around 0. We do not apply any transformations to corruption at this stage. The outliers in the boxplot are expected with this skew.

Total Cars for Each Mission

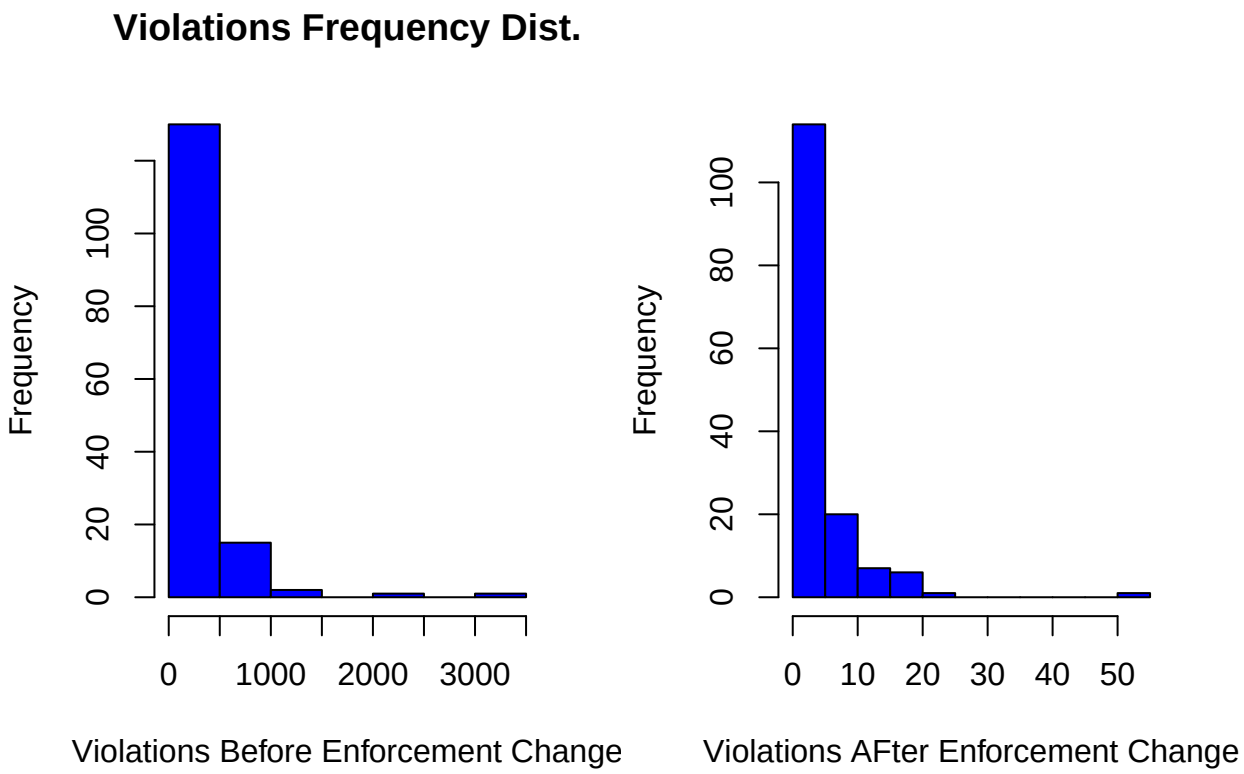
```
par(mfrow = c(1, 3))
hist(correlation_matrix_input$cars_total, col="blue", xlab = "Total Cars", main = "Cars Freq. Dist.")
boxplot(correlation_matrix_input$cars_total, col = "blue", ylab = "Total Cars")
boxplot(correlation_matrix_input$cars_total/(correlation_matrix_input$staff + correlation_matrix_input$s
```



Looking at the histogram on the left we see the vast majority of missions have 0 to 10 cars, which is unsuprising as this variable is reflective of mission size, which is in turn reflective of country population and wealth - both characteristics subject to power laws with notable rightward skew. The two charts on the right show a normalization - examining the total number of cars by the the number of staff and spouses. The maximum is 2.6, which could be explained by staff having more than one car or the mission having different cars for different purposes.

Violations

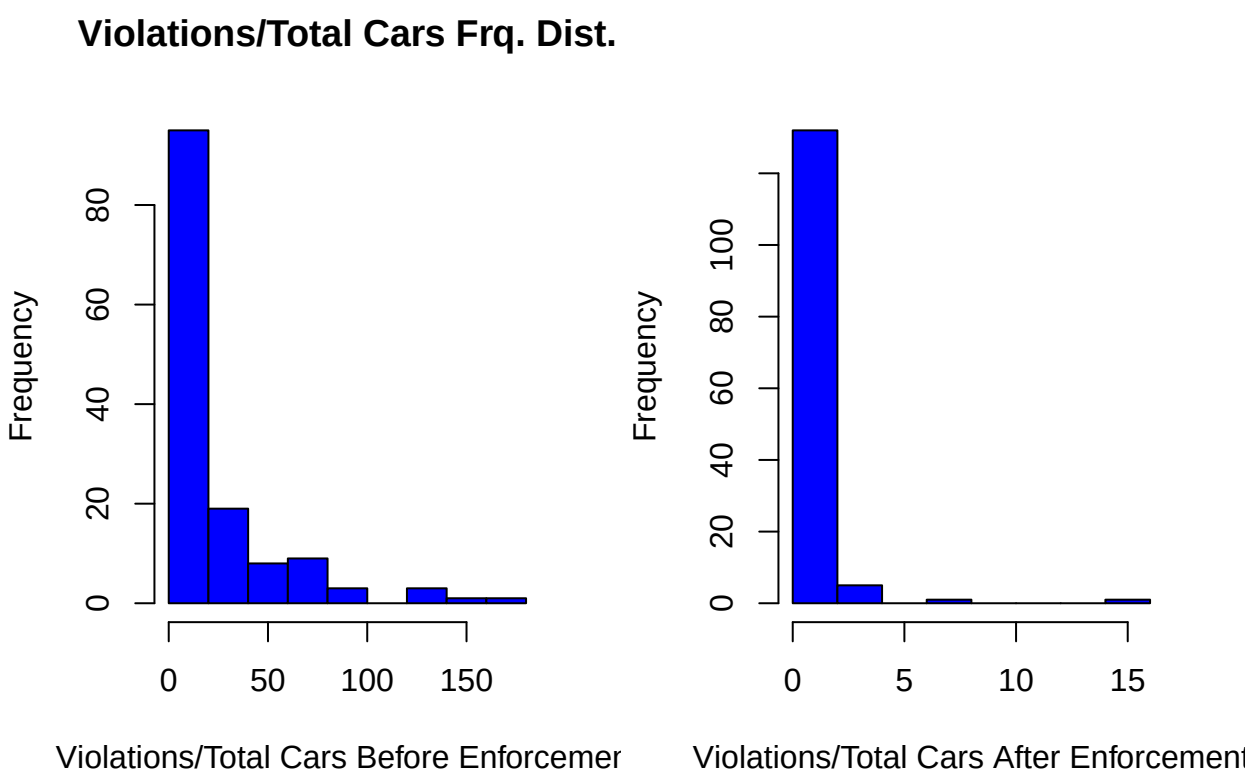
```
par(mfrow = c(1, 2))
hist(correlation_matrix_input$violations.pre, col="blue", xlab = "Violations Before Enforcement Change"
hist(correlation_matrix_input$violations.pos, col="blue", xlab = "Violations AFter Enforcement Change",
```



By examining the x-axis of both charts, it is apparent that the scale of the bins for the violations variable shows a substantial change in behavior after the enforcement change. Though, to make account for the power-law nature of country size on the violations variable itself, we need to create a varable that accounts for the impact of country size on violations.

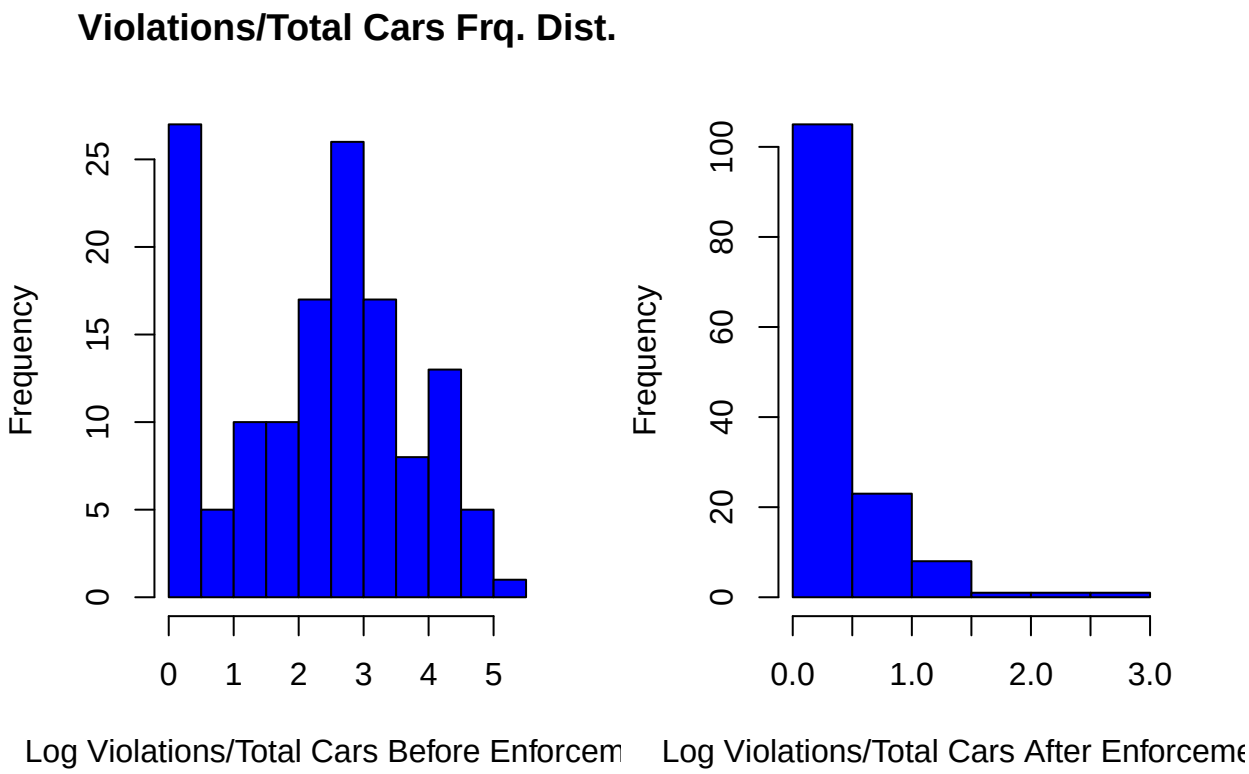
Violations/Total Cars Before and After Parking Violation Enforcement Changes

```
par(mfrow = c(1, 2))
hist(correlation_matrix_input$violations_weighted.cars_total.pre, col="blue", xlab="Violations/Total Ca
hist(correlation_matrix_input$violations_weighted.cars_total.pos, col="blue", xlab="Violations/Total Ca
```



The frequency distribution of the transformation *violations/totalcars* above is skewed positively giving showing that even after compensating for the effect of mission size on violations, certain missions incurred far more violations than others. This is the phenomenon we aim to explore more fully below.

```
par(mfrow = c(1, 2))
hist(correlation_matrix_input$violations_weighted.cars_total.pre.log, col="blue", xlab="Log Violations/
hist(correlation_matrix_input$violations_weighted.cars_total.pos.log, col="blue", xlab="Log Violations/
```



Given the notable right-ward skew and concentration of observations around zero, we perform a log transformation on violations per car. This improves our ability to perform analyses using this variable, with error terms more normally distributed. Though, there is still a large spike in the first bin of both frequency distributions due to a lot of missions having few or no delinquent parking fines.

```
# Reset default back to 1x1
par(mfrow = c(1, 1))
```

```
remove(variable_description)
```

Below is a sample of the key variables being analyzed for 20 countries sorted by the number of violations per car before the 2002 regulation change.

```
summary_table_output = cor_oneline[, c("country", "cars_total", "violations.pre", "violations.pos", "co
summary_table_output$mean_violations_per_car.pre = summary_table_output$violations.pre/summary_table_ou
summary_table_output$mean_violations_per_car.pos = summary_table_output$violations.pos/summary_table_ou

tmp_rounded = round_df(summary_table_output[, c("country", "mean_violations_per_car.pre", "mean_violati
tmp_rounded = tmp_rounded[order(tmp_rounded$mean_violations_per_car.pre, decreasing = TRUE), ]

kable(tmp_rounded[1:20, ],
      "latex", longtable = TRUE, booktabs = TRUE,
      caption = "Top 20 Countries by Parking Violations (Key Variables)",
      col.names = c("Country", "Mean Violations per Car Before 2002 Change", "Mean Violations per Car A
  kable_styling(full_width = TRUE, latex_options = c("HOLD_position", "striped", "repeat_header"), row_
```

Table 3: Top 20 Countries by Parking Violations (Key Variables)

	Country	Mean Violations per Car Before 2002 Change	Mean Violations per Car After 2002 Change	Corruption Index
55	GUINEA	176.22	2.94	0.57
41	EGYPT	141.37	0.33	0.25
77	KUWAIT	132.02	0.08	-1.07
120	SENEGAL	126.05	0.33	0.45
129	CHAD	125.89	0.00	0.84
19	BRAZIL	99.86	0.75	-0.10
119	SUDAN	84.40	0.26	0.75
146	SOUTH AFRICA	81.88	1.19	-0.42
107	PAKISTAN	76.14	1.31	0.76
149	ZIMBABWE	71.79	1.34	0.13
14	BULGARIA	71.42	0.98	0.50
128	SYRIA	71.10	1.82	0.58
93	MOZAMBIQUE	70.08	0.04	0.77
13	BANGLADESH	66.79	0.57	0.40
32	COSTA RICA	64.68	0.44	-0.71
86	MOROCCO	64.56	0.43	0.10
2	ALBANIA	64.16	1.39	0.92
9	BURUNDI	57.32	0.16	0.80
45	ETHIOPIA	54.95	0.56	0.25
11	BENIN	50.41	6.50	0.76

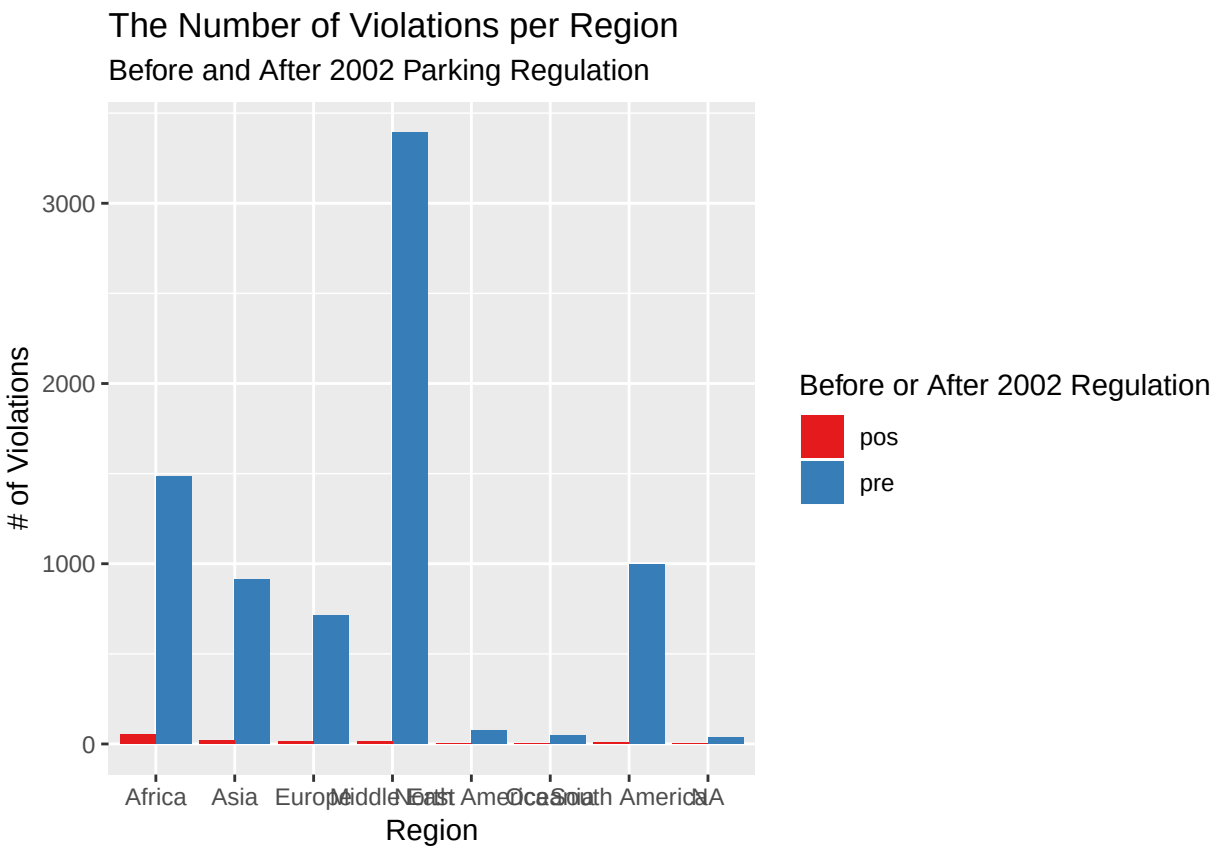
```
remove(summary_table_output, tmp_rounded)
```

Analysis of Key Relationships

Here we turn to analysis of key relationships among the variables. For the exploratory phase, we are interested in determining which of the variables are correlated, and if so, how strongly. For this analysis, where applicable, and based on the discussion in the Univariate section above, we use the log transformation of variables, as opposed to the untransformed version of the variable(s). This helps improve our understanding of relationships as positive skew that is present in some of the variables is compressed helping us to normalize error terms in the linear models employed below.

1) What is the relationship between violations before and after the 2002 introduction of the new parking regulation?

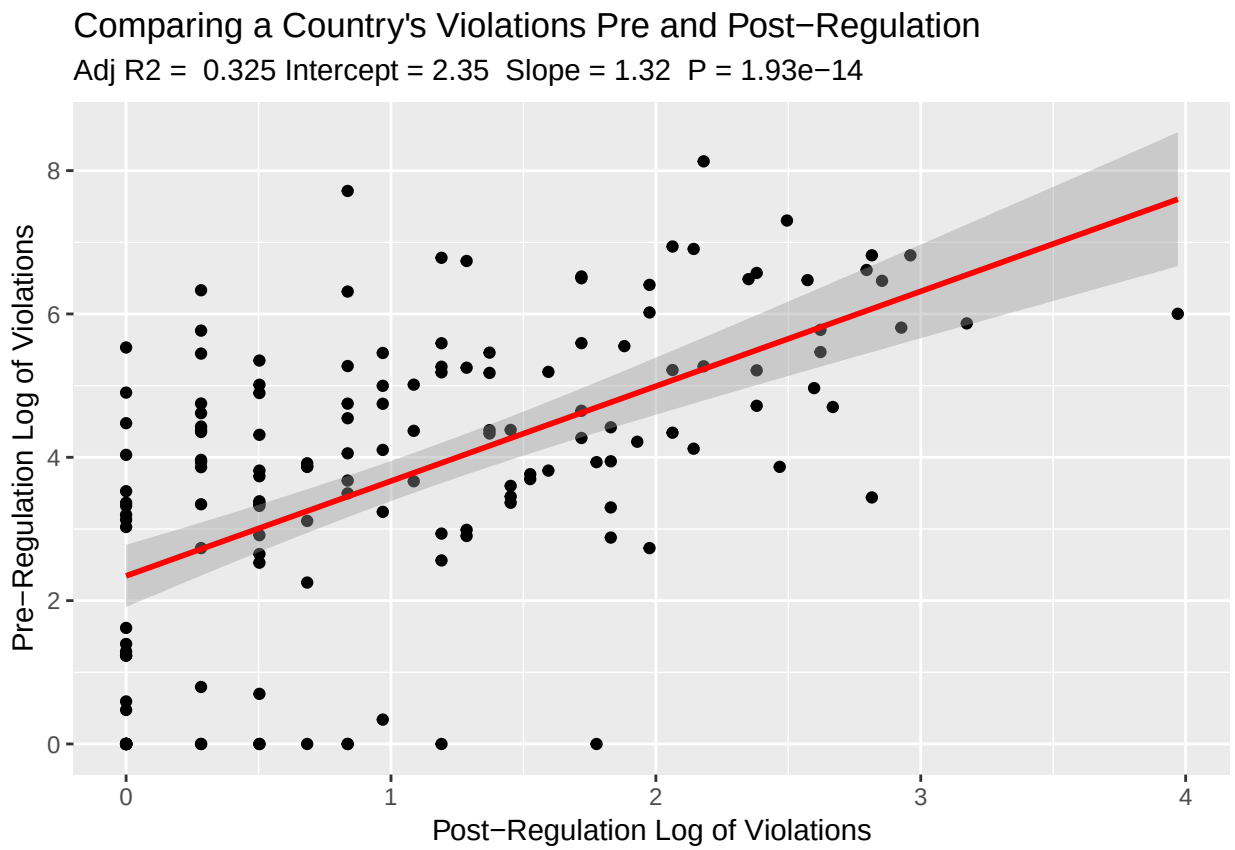
```
p <- ggplot(corrupt, aes(factor(region_name), violations, fill = factor(prepost))) +
  geom_bar(stat="identity", position = "dodge") +
  scale_fill_brewer(palette = "Set1")
p + labs(title = 'The Number of Violations per Region', subtitle = 'Before and After 2002 Parking Regula
```



This shows us that: There is a dramatic difference between the number of violations before and after the 2002 regulation for all regions. (Regions are used here as an expedient bucketing method across the observations to reduce complexity. This is useful for explanation, though not necessarily useful for model-building.)

However, it is difficult to see the rate-of-change relationship between the two variables given their different scales. This relationship can be explored by taking a log transformation of both variables.

```
lm1 <- lm(violations.pre.log ~ violations.pos.log, data=correlation_matrix_input)
ggplotRegression(lm1, 'Comparing a Country\'s Violations Pre and Post-Regulation', 'Post-Regulation Log
```

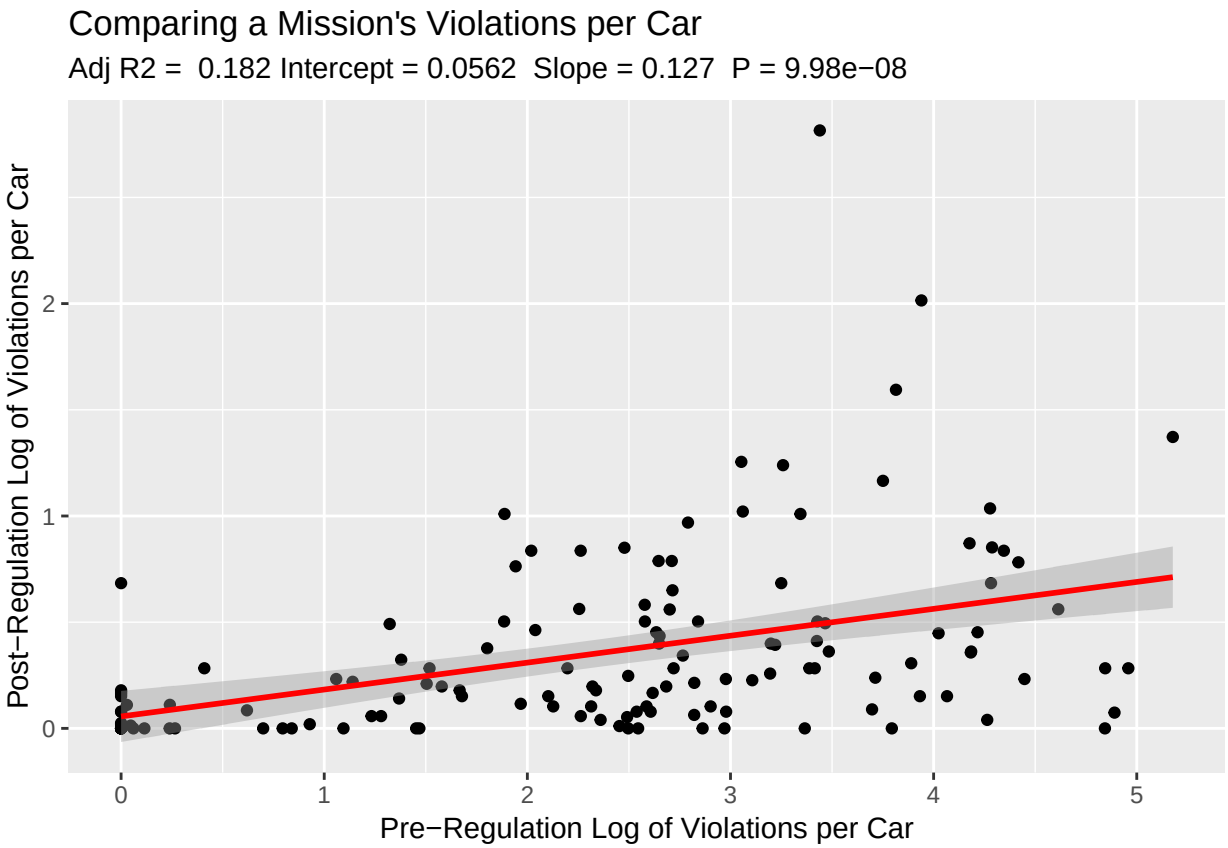


This plot roughly shows that for a 1 percent increase in the number of violations before the regulation, the data reveals a roughly 0.24 percent increase in the number of violations seen after the regulation. At this exploratory stage, there are two important facets to takeaway: 1) The correlation between these two series at 0.33 is notable, 2) The relationship between the two variables is clearly positive.

The interpretation of this is that in general countries with higher amounts of violations before the regulation are likely to also have relatively higher amounts of violations after.

However, there is likely much more going on in this data generation process. As discussed earlier, we speculate that violations is likely related to the size of a mission. This facet was examined through the car-based and the people-based variables. We found that the strongest relationship (considering both correlation and slope) between pre- and post-regulation percent changes in violations was for the number of violations normalized for the number of cars. As such, we will be using this during the remainder of our exploration. Should we be beyond the exploration phase into a phase where we were trying to determine causation or to predict violations, this would be insufficient, but given that we are still in exploration phase, it will suffice. Alternate formulations that were explored to control for size (staff members, total people, staff cars) are included in the Appendix.

```
lm5 <- lm(violations_weighted.cars_total.pos.log~violations_weighted.cars_total.pre.log, data=correlati
ggplotRegression(lm5,'Comparing a Mission\'s Violations per Car','Pre-Regulation Log of Violations per
```

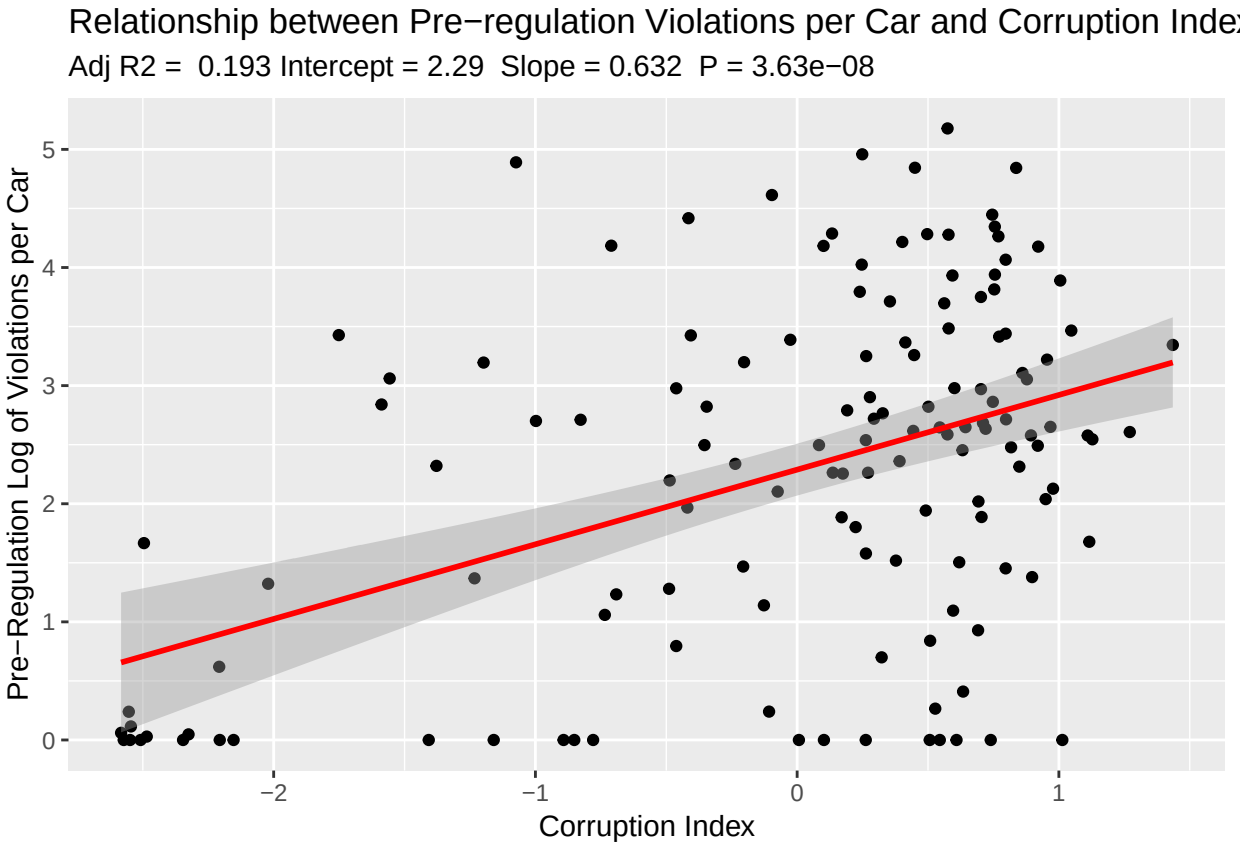


From this chart, we can again take away a sense that those countries which had higher violations prior to the regulations were also likely to have relatively higher violations after the regulation, even after controlling for the conflating of size impacts by normalizing violations with the most appropriate proxy for the mission size-type variables. And, importantly, this formulation of size normalization appears to offer the best preservation of signal on country differences in violations even after controlling for size. This selection also makes intuitive sense as only cars can incur parking fines - people without cars do not.

Now we turn to understanding how other variables in our dataset might help explain the differences in violations per car. We will begin with an examination of the relationship between the provided corruption index and violations per car, with other variables addressed in subsequent sections.

We begin by examining the relationship between pre-regulation # of violations per car and the provided corruption index.

```
lm6 <- lm(violations_weighted.cars_total.pre.log~corruption, data=correlation_matrix_input)
ggplotRegression(lm6,'Relationship between Pre-regulation Violations per Car and Corruption Index','Cor
```

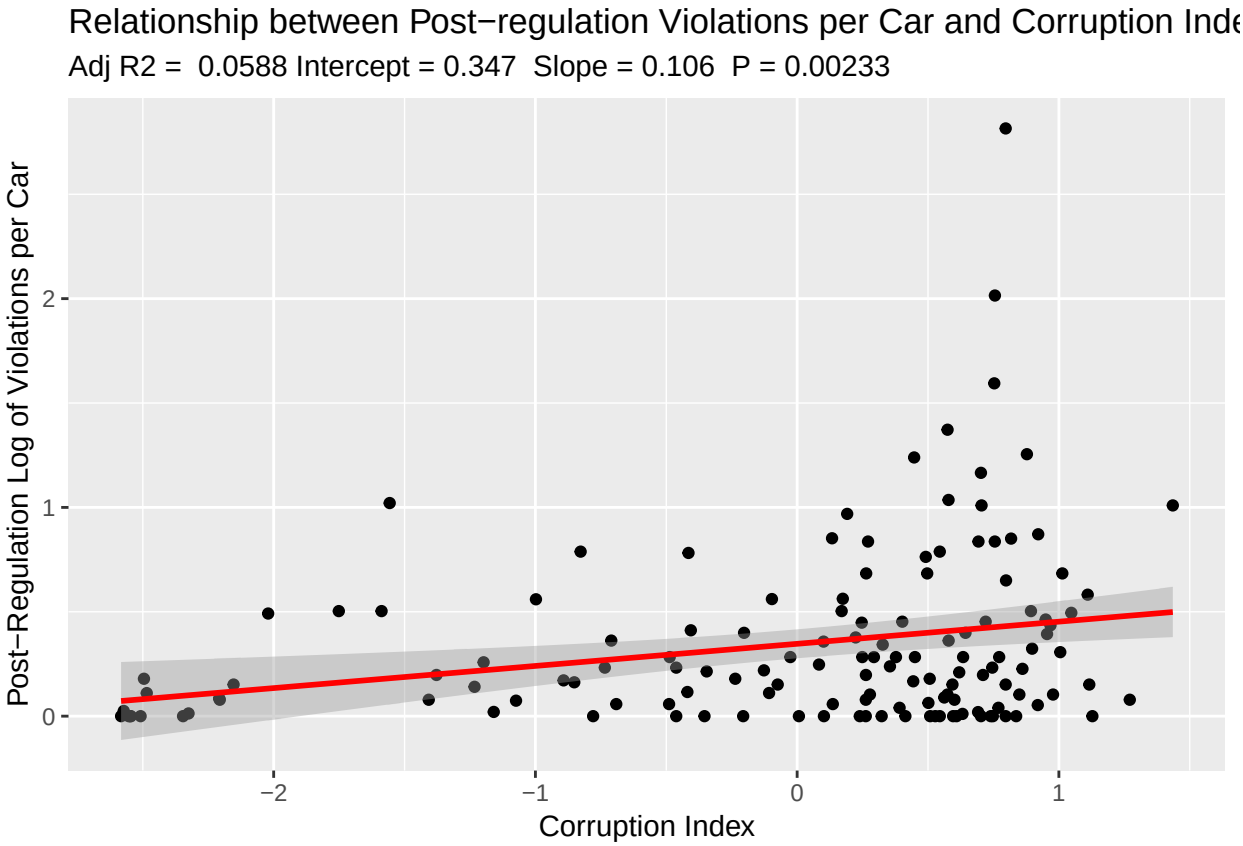


From this chart, we can see a strong positive relationship between these two variables appears clear, although there is much more going on than is characterized here as the corelationship is notable but far from linear. This chart can be interpreted as - for a 1 unit increase in the country's ranking on this corruption index, we expect to see a 63 percent increase in that country's number of parking violations per car.

This seems to confirm our initial inclination that corruption may be a good indicator of a country's willingness to incur parking violations.

Was this relationship changed by the introduction of new parking regulations?

```
lm7 <- lm(violations_weighted.cars_total.pos.log~corruption, data=correlation_matrix_input)
ggplotRegression(lm7,'Relationship between Post-regulation Violations per Car and Corruption Index','Co
```



This chart shows that after the regulation, the positive relationship between a country's corruption index and its willingness to incur parking violations still exists, however it is much weaker and the magnitude much smaller. After the introduction of the parking regulation, we expect a one unit increase in a country's corruption index to result in only an 11 percent increase in parking violations per car. Also of note, the correlation between the two variables decreased markedly, with the relationship seen here clearly positive but non-linear.

It is also important to note that, while our analysis reveals a positive relationship between a country's corruption index and their willingness to incur parking violations, both before and after the regulation, there are numerous countries with a high corruption index that also incurred 0 parking violations. So, while this relationship may help describe the population, individual countries within the population can and do deviate markedly from the statistical relationship seen above.

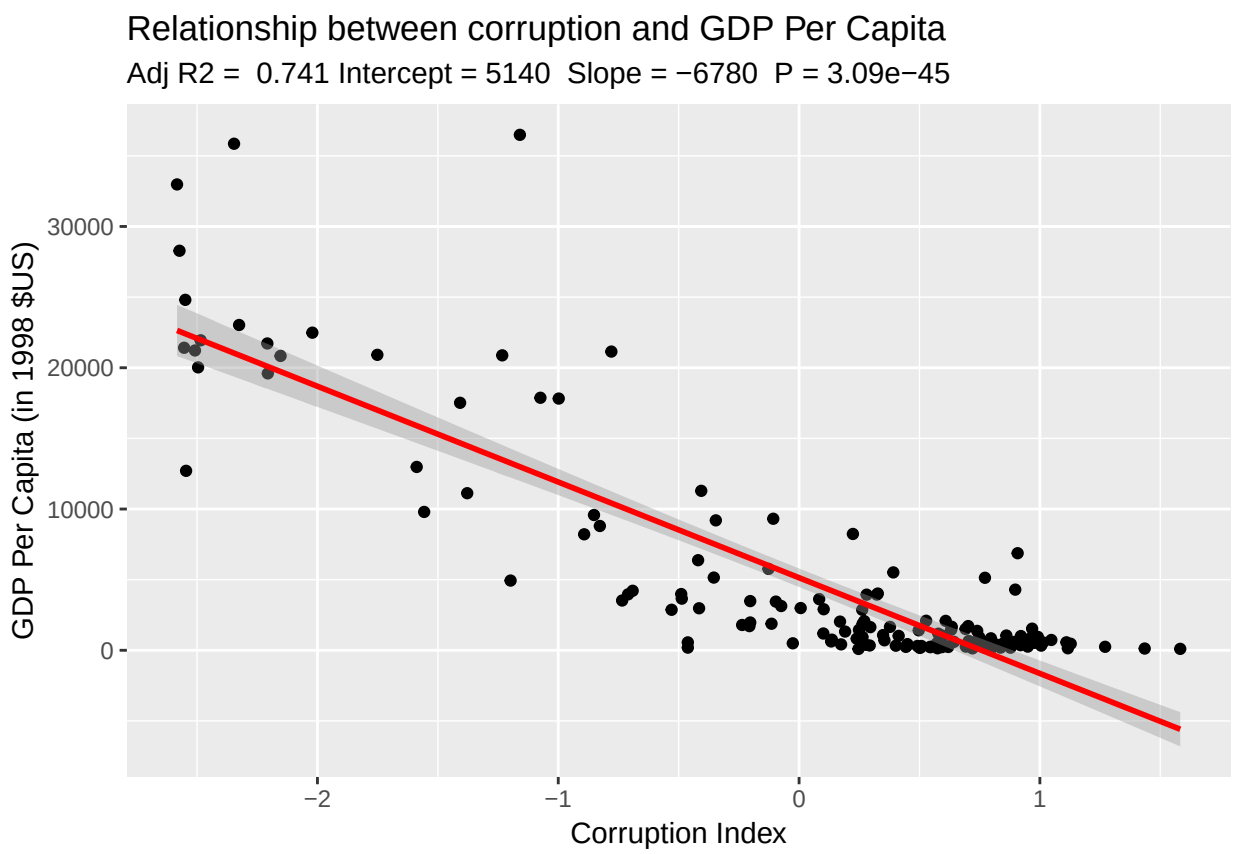
Analysis of Secondary Effects

During our exploratory data analysis, we revealed some relationships to corruption that may confound our key relationship analysis.

Relationship Between Corruption Index and That Country's Gross Domestic Product Per Capita

There are several factors that could be complicating our understanding of the relationship between corruption and parking violations. For example wealth could be seen as having an impact on willingness to incur and pay fines. We examine these relationships below.

```
lm_cor_gdppc <- lm(gdppcus1998~corruption, data=correlation_matrix_input)
ggplotRegression(lm_cor_gdppc,'Relationship between corruption and GDP Per Capita','Corruption Index','
```

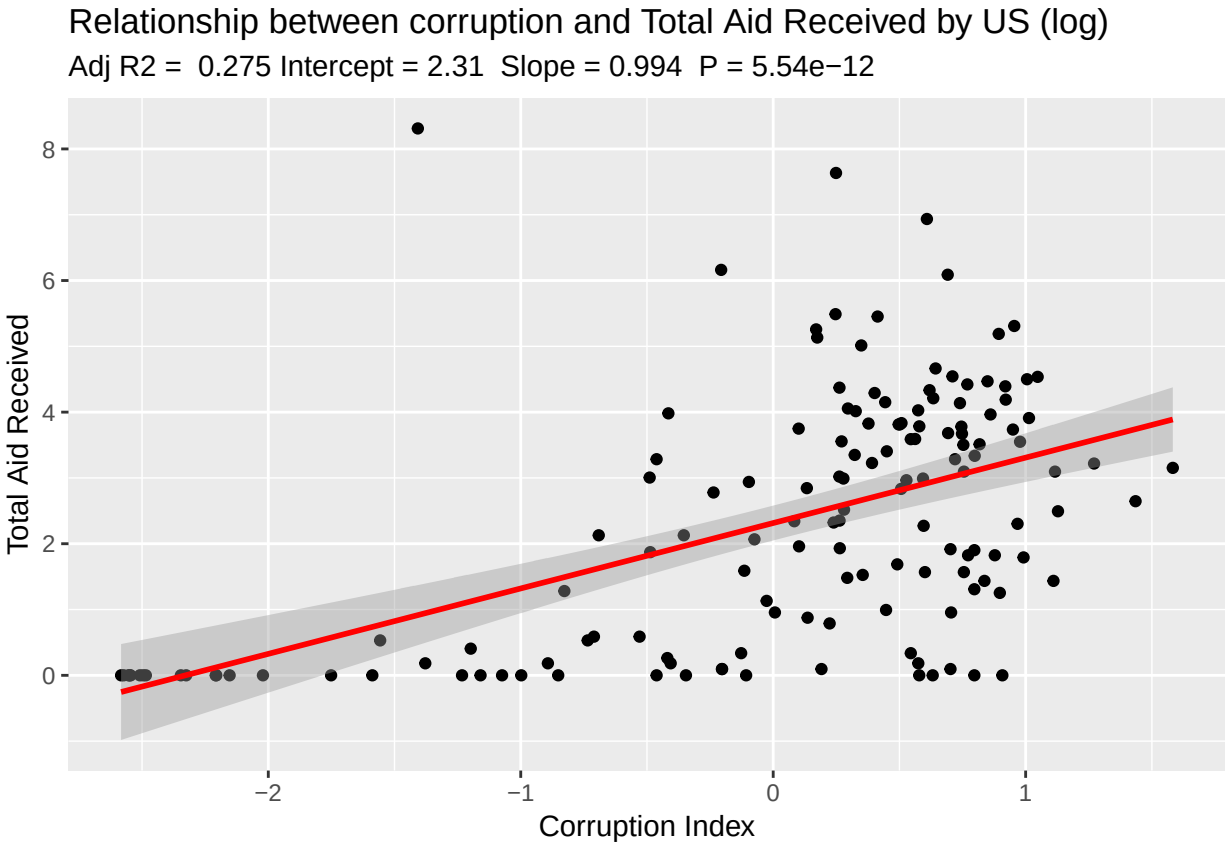


The plot above shows a steep negative relationship of GDP per capita and corrution. This is relevant to the key relationship between violations/total cars and corruption as GPD per capita might positively corelate to the personal wealth of mission employees and wealth might also contribute to a person's decision to pay parking tickets.

It is worth noting that the variable gov_wage_gdp which is the percentage of a country's GDP paid to government employees was analyzed by multiplying it to GDP per capita. The correlation was weaker, however 57 of the 149 countries being analyzed are missing values for gov_wage_gdp, and we have no insights that would provide confidence that wages of mission employees changes in any predictable manner with gov_wage_gdp. We also have little insight to whether gov_wage_gdp can be interpreted in the way we speculate. Little information was provided about what the variable represents or its source. Due to these factors, we do not further examine that variable, despite the positive potential for a well-constructed government employee wage variable.

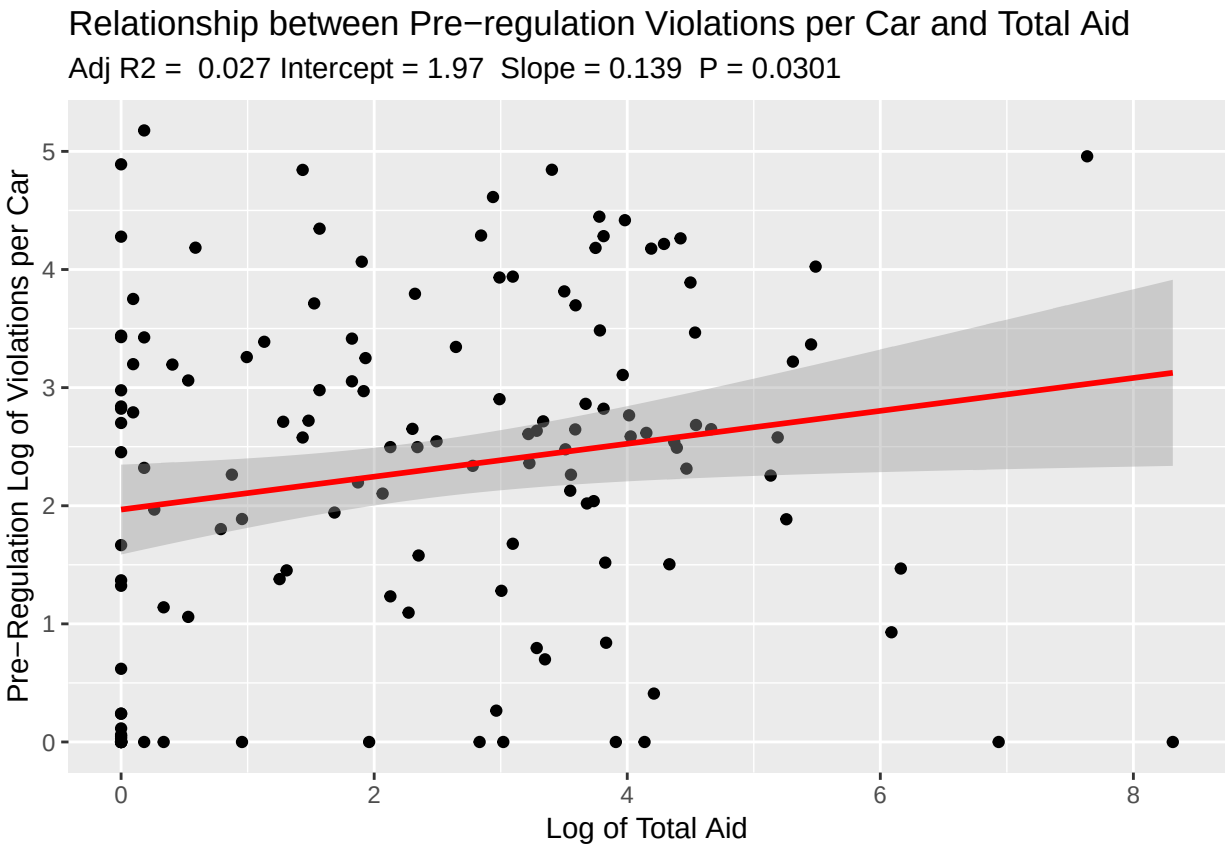
Relationship Between Corruption Index and The Amount of Aid Recieved by the US

```
lm_cor_totaid <- lm(totaid.log~corruption, data=correlation_matrix_input)
ggplotRegression(lm_cor_totaid,'Relationship between corruption and Total Aid Received by US (log)','Co
```



Total military and economic aid provided to the country by the U.S. is also correlated to corruption and violations. This is intuitive as foreign aid tends to go to less developed countries and less developed countries also tend to be more corrupt. As such, we are not convinced that foreign aid really provides any new information in this context.

```
lm9 <- lm(violations_weighted.cars_total.pre.log~totaid.log,data=correlation_matrix_input)
ggplotRegression(lm9,'Relationship between Pre-regulation Violations per Car and Total Aid','Log of Tota
```



We also examined the relationship between violations per car and total aid, taking the log of both variables to compress the distribution. This plot shows that relationship to be largely random - not one worth further exploration. The post-regulation violations per car chart showed a similar lack of relationship.

Percentage of Country Who Identify as Muslim

Our dataset included the percentage of a country's population that identified as Muslim, and that happens to correlate with the number of violations per car. However, the percentage of Muslims also correlates with GDP per capita, and GDP per capita has a much stronger correlation to the number of violations per car. We feel that this variable really contains information similar to other economic data. Additionally, we don't feel any relationships to a single religion is valuable in absense of other major religions having different correlations with independent variables.

Trade Between Countries and the US.

Trade is positively correlated with the number of mission staff and spouses. It seems the more trade there is between a country and the US, the more staff (and cars) there are at the mission in NYC. This makes sense as larger countries naturally trade more in dollar terms. The number of cars at the mission it turns out acts as a proxy variable for trade, and it is more relevant to our analysis as it can be used to find the mean violations. Therefore, we have not included analysis on trade.

Conclusion

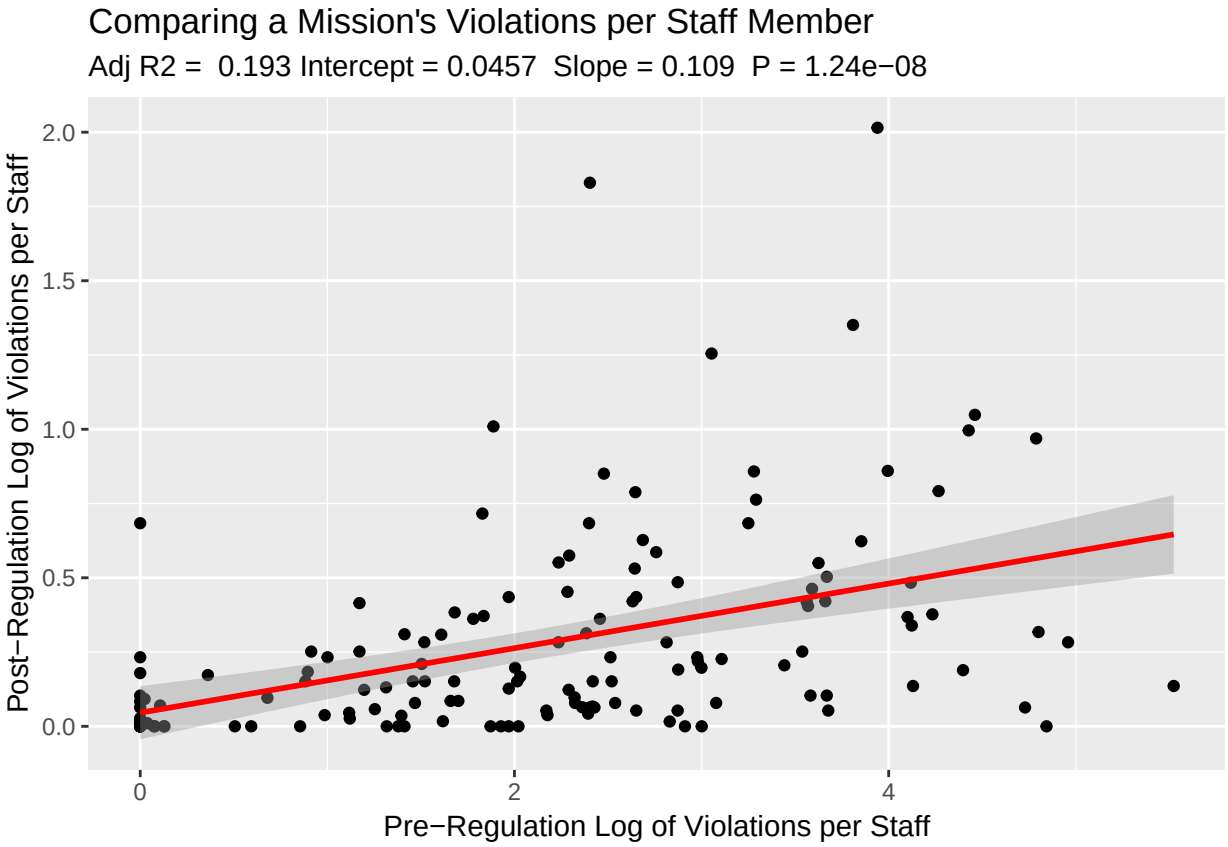
In conclusion, we find a strong positive relationship between the provided corruption index and violations per car. We also note that the total number of parking tickets dropped notably after the regulation change - a sign that the regulation had its intended effect. However, the countries that had the most tickets before also tended to have the most tickets after, though even then they accrued notably fewer.

We also note that data issues dampen the confidence in our conclusions as data origins and data quality hampered our exploratory analysis throughout. Should we be publishing this analysis publically, we would have engaged in much more extensive research around this data quality issue.

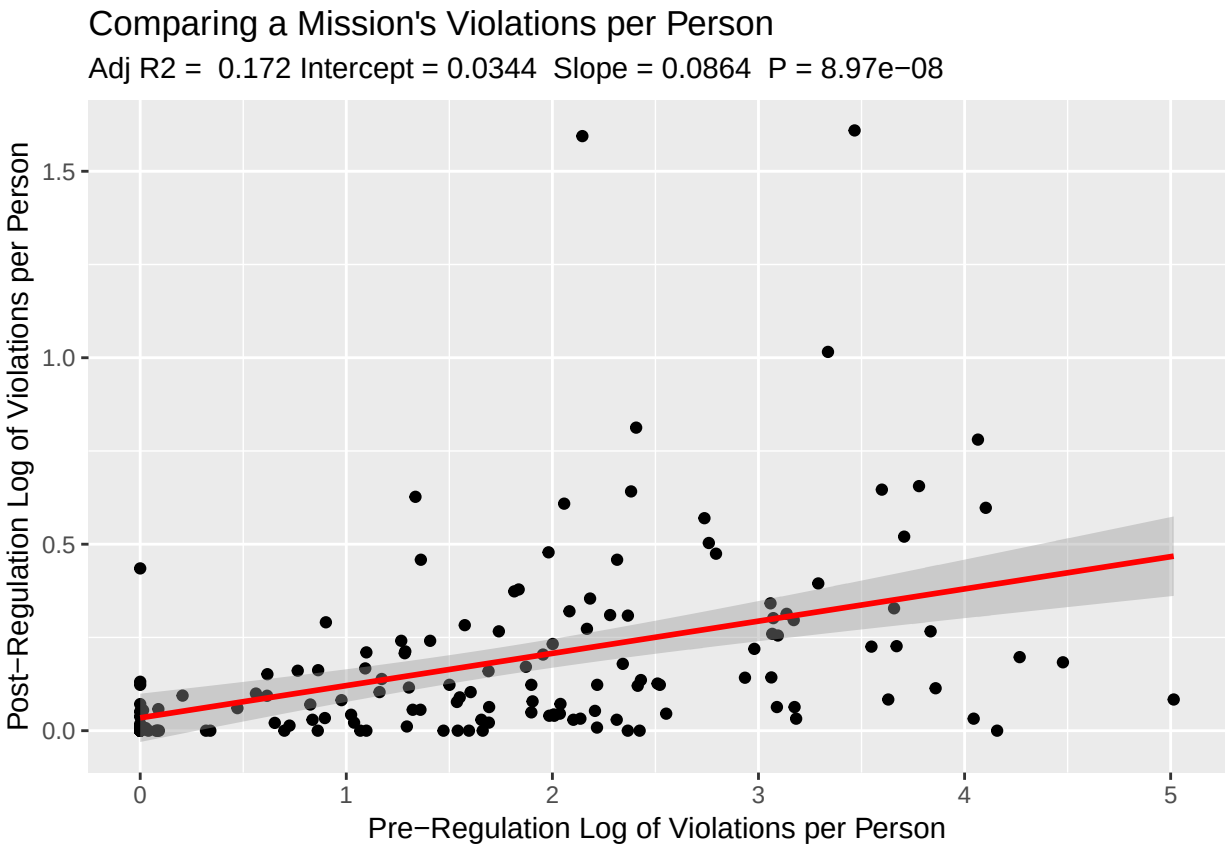
Appendix

From the Analysis of Key Relationships section. We also explored alternate means of controlling for the impact of mission size on the relationship between violations pre and post regulation. These formulations are documented here for the curious reader.

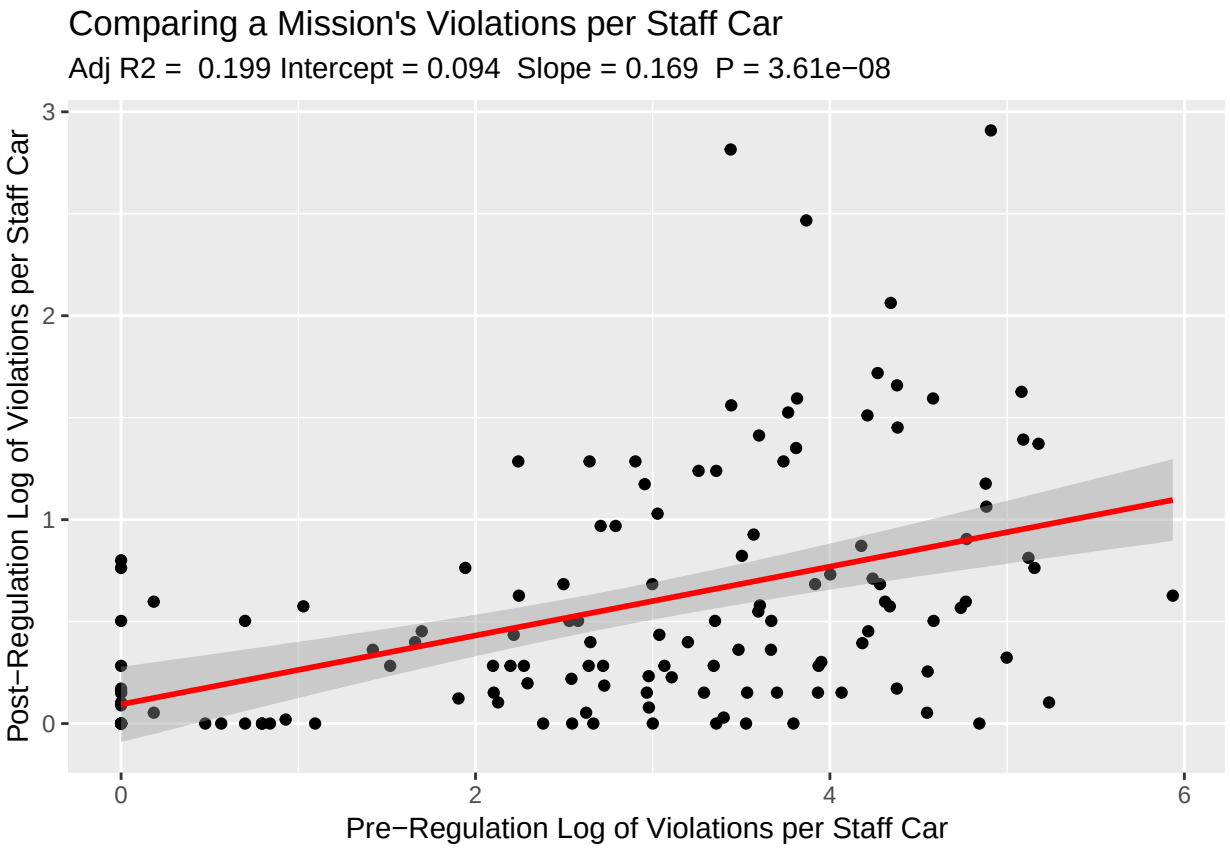
```
lm2 <- lm(violations_weighted.staff.pos.log~violations_weighted.staff.pre.log, data=correlation_matrix_
ggplotRegression(lm2,'Comparing a Mission\'s Violations per Staff Member','Pre-Regulation Log of Violat
```




```
lm3 <- lm(violations_weighted.total_people.pos.log~violations_weighted.total_people.pre.log, data=correl
ggplotRegression(lm3,'Comparing a Mission\'s Violations per Person','Pre-Regulation Log of Violations p
```




```
lm4 <- lm(violations_weighted.cars_mission.pos.log~violations_weighted.cars_mission.pre.log, data=correl
ggplotRegression(lm4,'Comparing a Mission\'s Violations per Staff Car','Pre-Regulation Log of Violations
```



It makes sense that the set of normalizations which best preserved signal from violations normalized by size-type variables was the total cars variable. The three other formulations above are potential substitutes, but appear to preserve less of the signal than the total cars formulation. As mentioned above, it makes intuitive sense for the cars variables to carry more signal than the number of staff variables as only cars can incur parking violations. Also, we considered using the number of violations per staff car variable as it also appear to preserve signal well, however the normalization method used here (violations divided by staff car) is compromised where staff cars == 0, as is the case for four observations.