# Corruption and Parking Violations

*Akkineni, Hanna, 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 majoritymulsim 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"

# Create named regions variable using region
FMcorrupt$region_name = NA
FMcorrupt[FMcorrupt$region == 1 & ! is.na(FMcorrupt$region), "region_name"] = "Caribbean"
FMcorrupt[FMcorrupt$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$region), "region_name"] = "Asia" # "South Asia"
FMcorrupt[FMcorrupt$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" # "Western

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) )


# 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)
```

```r
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
                              "cars_total.pre", "cars_mission.pre", "pop1998.pre", "gdppcus1998.pre", "e
                              "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", "majoritymuslim
                              "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)
```

# Introduction

## Research Question

Prior to 2002 there were no measures New York City (NYC) was able to take to enforce payment of parking violations from UN officials who have diplomatic emmunity. Therefore, payment was largely a function of cultural norms. 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, thus making payment an function of both cultral norms and legal enforcment.

In our exploratory data analysis, we will describe how cultural norms affect the payment of parking violations by researching the number of violations diplomats had accumulated both before and after the enforcement changes using the corruption index of the country of the embassy they are accociated with.

## Description of Dataset

Our data set has a total of 364 observations, there are two observation for each country, where only the number of violations and value of the fines changes, one before NYC gained the ability to enforce non-payment, and another after.

# Univariate Analysis of Key Variables

Key Variables:

- violations
- staff

- trade

There are 66 observations that contain only economic data leaving NA for our dependent variable, violations. Considering the countries that are among these 66 we suspect these rows result from a merge of economic data with the violations data ($economic \cup violations$) and the economic data set had countries that did not have embasies in Manhattan. Removing these superfluous observations results in the intersection of the two data sets ($economic \cap violations$), which is what we desire. With those 66 rows removed, we're left with 298 observations (two observations for each of 149 countries where corruption data exists.)

```
# 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, se
#summary(variable_description)

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

Table 1: Data Set Variables

| Variable | Description | Observations | Alterations |
|---|---|---|---|
| wbcode | Three Letter Country code | | |
| 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. | | |
| fines.pos | Summed cost of violations.pos adjusted for inflation. | | |
| mission | If country has mission in NYC | All values are true. | |
| staff | # of employees | | |
| spouse | # of spouses | | |
| gov_wage_gdp | | | |
| pctmuslim | Percent of country that identifies as Muslim | 2 countries have NA's: Bosnia-Herzegovina and Zaire | |

Table 1: Data Set Variables *(continued)*

| Variable | Description | Observations | Alterations |
|---|---|---|---|
| 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. |
| trade | Total trade with the US in 1998 USD | 2 countries have NA's: Bosnia-Herzegovina and Zaire, 1 country shows 0 Libia | |
| cars_total | The number of cars owned by both mission and the employees. | Total of 1455, 10 NA's | |
| cars_personal | Cars owned by staff and spouses of mission | Total of 740, 10 NA's | |
| cars_mission | Cars owned by mission | Total of 715, 10 NA's | |
| pop1998 | Country's population in 1998 | | |
| gdppcus1998 | GDP Percent in USD 1998(?) | | |
| ecaid | US Economic Assistance grants and loans | | |
| milaid | Military Aid | | |
| region | 7 Geographic regions labelled 1-7 | | |
| region_name | 7 Geographic regions named | | We created this variable using region variable above |
| corruption | Country corruption index | Min. = -2.58299 (Least corrupt), Max. = 1.58281 (Most corrupt) | |
| totaid | milaid + ecaid | | |
| r_africa | Boolean if part of continent. Some countries belong to none, and some others have NA's. | | |
| r_middleeast | " | | |
| r_europe | " | | |
| r_southamerica | " | | |
| r_asia | " | | |
| country | country name | Some missing values, though wbcode are available. | Missing values were filled using wbcode variable. |
| distUNplz | | | |

## Alex notes to be incorporated above

“”“Data quality issues:
~~1) We have several extra observations that have NaN values for key variables. These data likely reflect a merging process wherein the desired dataset regarding parking violations and fines was merged with economic~~

~~data. +wbcode is impacted by this. There are 213 unique values for wbcode, whereas there are only 151 countries with pre and post-legislation observations ++++Suggestion: We drop the N/A variables and note the number of countries for which we have no data.~~

2) The violations and fines data are odd. The instructions do not provide adequate background regarding these variables. This is a particular problem because these variable are at the core of the proposed analysis. Oddities: +violations has 7 decimal places. This is odd because we would expect a positive integer (in the case of a single year) or at least a figure that is recognizable as an average of some sort (in the case of multiple years)

3) The mission, staff, and spouse variables also contain oddities. For example, the entries for HKG and PRI are 0. These are the only zeros in those variables. Looking at the values for other variables for those two observations, some other oddities emerge. For example: +HKG's majoritymuslim (which should be a dummy variable of either 1 or 0) is -1 ++++Suggestion: Given that our research question regards the impact of the corruption on willingness to incur parking violations and fines in NYC, we probably do not care about observations without a mission in NYC. As such, taking all of these oddities and the low desirability of the data, I recommend excluding. where mission equals 0 or NaN

4) The variable gov_wage_gdp is not specified in the instructions, and no source for the data is provided. We don't know if this comes from an official sector body or from some potentially less rigorous institution. Furthermore, the data appears somewhat disconnected from our core violations and fines variables, as there are numerous NaNs for gov_wage_gdp where valid values exist for violations and fines. I do think the concept that this variable hints at - the impact of government workers' wages on their willingness to occur fines - would be interesting to examine in relation to the degree of decline from the pre to the post subsets. The thinking being that those government workers with lower average salaries would be less willing to incur fines out of their own pocket. This is potentially problematic though as it is not clear that the workers themselves would definitely be the ones paying for the fines. ++++Suggestion: Given that we don't know the provenance of this dataset nor how it was calculated, I would not treat results using this dataset with a high degree of confidence. Rather, I would be inclined to largely exclude this variable from the analysis At most, it should be used with caution in it's own separate section.

5) The variable pctmuslim is among the more complete variables outside the core violations and fines dataset. We are not told the origin of this variable either, so we have no idea of its veracity other than a common sense check. The thought behind including this variable may be that religion and perhaps, Islam in particular, would have an impact on ethical or ethical (corrupt) behavior. It is not clear to me why Islam would be included and all other religions excluded. A more appropriate variable would be the percent of population which practices religion. It is also not clear to me that the ethics-religion association is necessarily as strong as some might think it is, though a more valid non-biased variable could be used to test that relationship.

“”"

```
remove(variable_description)
```

```
# This is assuming staff is better than cars, staff+spouse, total_cars.  Though I don't know if that's
summary_table_output = cor_oneline[, c("country", "staff", "violations.pre", "violations.pos", "corrupti
summary_table_output$mean_violations_per_staff.pre = summary_table_output$violations.pre/summary_table_
summary_table_output$mean_violations_per_staff.pos = summary_table_output$violations.pos/summary_table_

#TODO Sort, Round

kable(summary_table_output[1:10, c("country", "mean_violations_per_staff.pre", "mean_violations_per_sta
      "latex", longtable = TRUE, booktabs = TRUE,
      caption = "NOT YET SORTED BY VIOLATIONS - Top 10 Countries by Parking Violations",
      col.names = c("Country", "Mean Violations per Staff Before 2002 Change", "Mean Violations per Sta
  kable_styling(full_width = TRUE, latex_options = c("HOLD_position", "striped", "repeat_header"), row_
```

Table 2: NOT YET SORTED BY VIOLATIONS - Top 10 Countries by Parking Violations

| Country | Mean Violations per Staff Before 2002 Change | Mean Violations per Staff After 2002 Change | Corruption Index |
|---|---|---|---|
| ANGOLA | 82.709025 | 1.7079848 | 1.0475056 |
| ALBANIA | 85.544769 | 1.8533452 | 0.9210790 |
| United Arab Emirates | 0.000000 | 0.0000000 | -0.7794677 |
| ARGENTINA | 3.997751 | 0.3614884 | 0.2235667 |
| ARMENIA | 10.228912 | 0.1635305 | 0.7100782 |
| AUSTRALIA | 0.000000 | 0.0272551 | -2.2059753 |
| AUSTRIA | 2.228080 | 0.5139529 | -2.0210860 |
| AZERBAIJAN | 0.000000 | 0.9811828 | 1.0132217 |
| BURUNDI | 38.214945 | 0.1090203 | 0.7967861 |
| BELGIUM | 2.719994 | 0.1401690 | -1.2325672 |

```
remove(summary_table_output)
```

## Analysis of Key Relationships

## Analysis of Secondary Effects

## Conclusion