An Exploratory Analysis of Corruption and Parking Violations

Kenneth Chen, Shiraz, Praba Santhanakrishnan

May 28, 2018

Introduction

Background Imagine that you have been hired by the World Bank to study the eect of cultural norms and legal enforcement in controlling corruption by analyzing the parking behavior of United Nations officials in Manhattan. Until 2002, diplomatic immunity protected UN diplomats from parking enforcement actions, so diplomats actions were constrained by cultural norms alone. In 2002, enforcement authorities acquired the right to confiscate diplomatic license plates of violators, after which diplomatic behavior was constrained by both cultural norms and the legal penalties of unpaid tickets.

Data You are given a dataset for a selection of UN diplomatic missions, Corrupt.R. The dependent (or target) variable in this data is named violations. The labels of some of the variables are listed below; the rest of the variables should be self-explanatory.

corruption: Country corruption index, 1998
violations: Unpaid New York City parking violations
trade: total trade with the United States (1998 US$)

Objective The World Bank would like to know what if any relationship there is between corruption and parking violations both pre and post 2002 and if there are any other relevant explanatory variables.

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

(b) How does the number of diplomats contribute to the frequency of violations?

(c) Did World Trade Center attack on September 11, 2001 have any impact on the parking violations in Manhattan NY?

Data Exploration

Our team, Kenneth Chen, Shiraz, Praba Santhanakrishnan from W203, will address this question using exploratory data analysis (EDA) techniques. Our data is composed of the number of parking violations, number of diplomats from countries all over the world, individual country corruption index so on and so forth. We are interested in investigating the correlation between parking violations and the corruption of the diplomat countries. How is corruption relevant to the frequency of the parking violations happened in

Manhattan NY before and after the year 2002. The year '2002' was intuitively chosen to reflect the World Trade Center attack on September 11, 2001.

Setup

First, we load the car library, which gives us a convenient scatterplotMatrix function.

```
library(car)

## Loading required package: carData

# Load the data
load("Corrupt.Rdata")
```

Data Selection

We observe that we have 364 observations and 28 variables.

```
nrow(FMcorrupt)

## [1] 364

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 ...
##  $ 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 ...
```

We looked at the total number of violations and found that the violations could be as low as 0 and could also go as frequent as 3392.96. This shows a wide discrepancy in violations, from which we could gather some insightful information regarding other factors such as corruption index and the number of diplomats visits to the US.

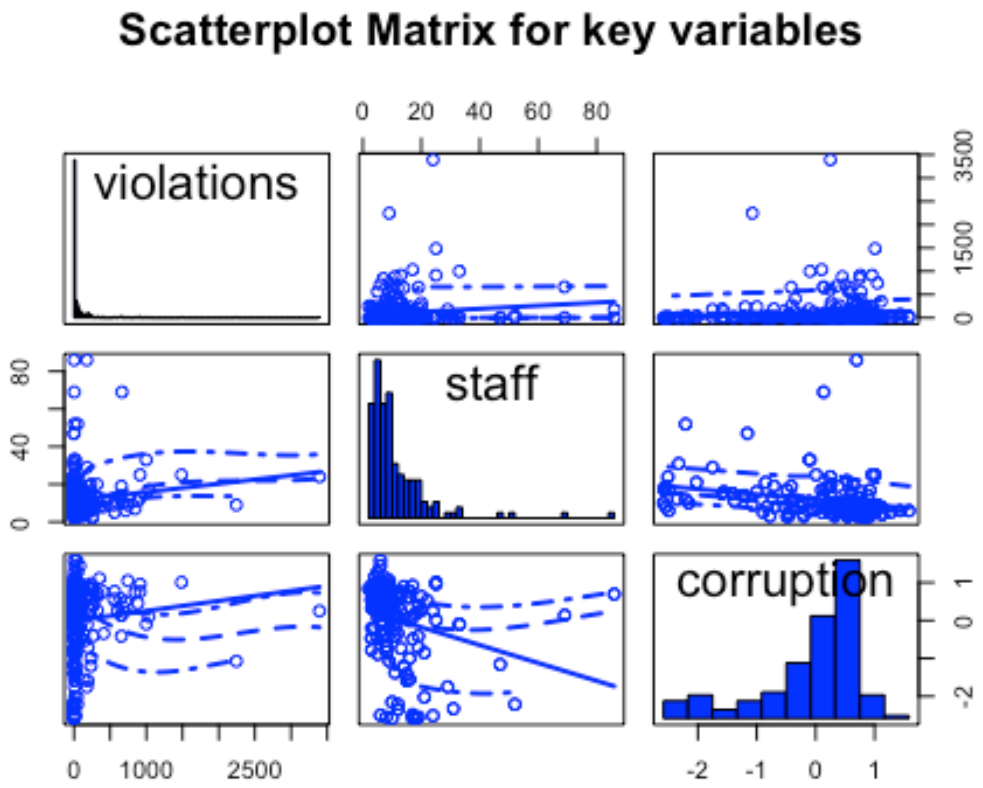
Looking at the diplomat variable, i.e., staff, we notice that diplomat numbers stay between 0 and 86

```
summary(FMcorrupt$violations)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##    0.000    0.654    5.724  100.879   51.915 3392.961       66

summary(FMcorrupt$staff)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    5.00    9.00   11.65   14.00   86.00      62
```

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

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

```
scatterplotMatrix(~ violations + staff + corruption, data=FMcorrupt, diagonal
= list(method='histogram'), main = "Scatterplot Matrix for key variables")
```

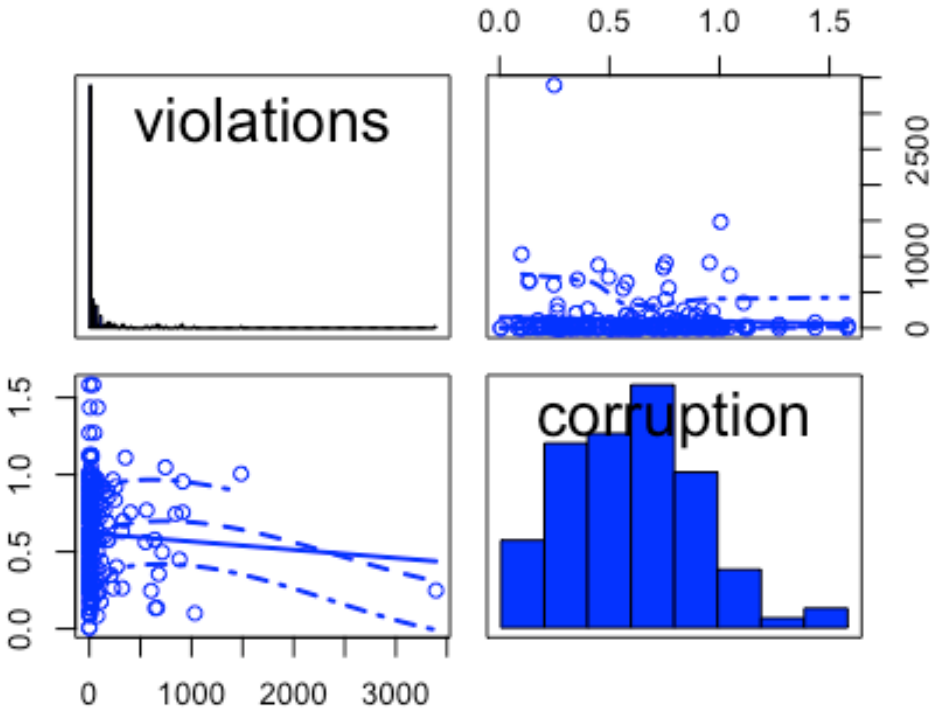


Our first step is to subset the corruption index data to further zoom in to the most corrupted countries. We created subcases with below and above zero.

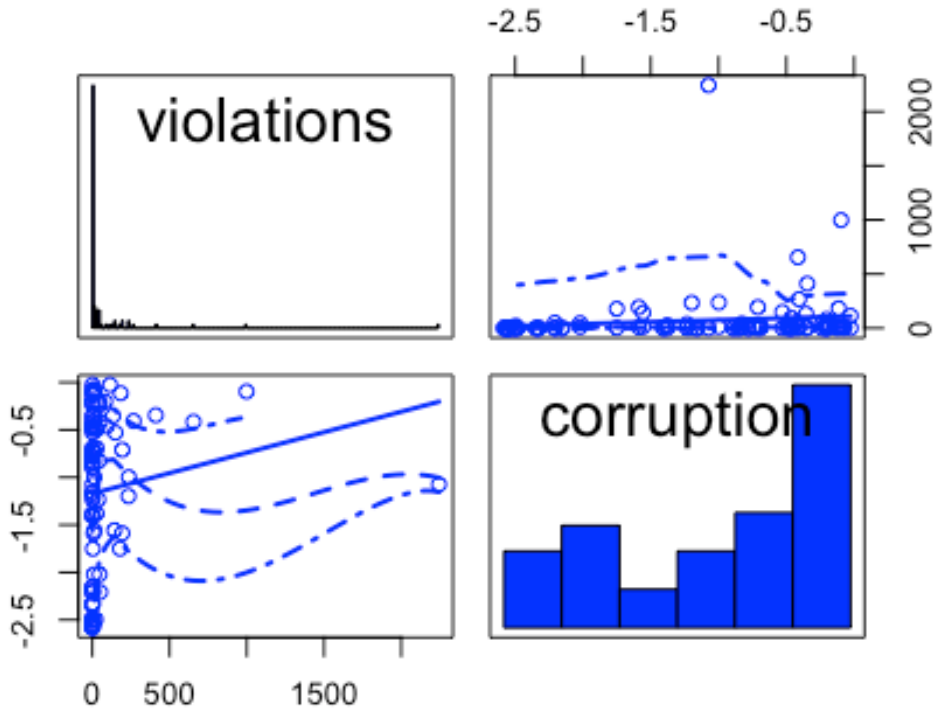
```
subcases_above_zero = 0 <= FMcorrupt$corruption &
!is.na(FMcorrupt$corruption)

subcases_below_zero = 0 >= FMcorrupt$corruption &
!is.na(FMcorrupt$corruption)

FM_subcases_above_zero = FMcorrupt[subcases_above_zero, ]
nrow(FM_subcases_above_zero)

## [1] 196

FM_subcases_below_zero = FMcorrupt[subcases_below_zero, ]
nrow(FM_subcases_below_zero)

## [1] 107
```

We also removed any corruption observations where the event is “NA”. Using the logical vector to pull out from the original data, we found that the total number of observation above corruption index 0 is 196 and observation below corruption index 0 is 107 .

```
scatterplotMatrix(~ violations + corruption, data=FM_subcases_above_zero,
diagonal=list(method="histogram"))
```




```
scatterplotMatrix(~ violations + corruption, data=FM_subcases_below_zero,
diagonal=list(method="histogram"))
```




```
cor(FMcorrupt$corruption, FMcorrupt$violations, use="complete.obs")

## [1] 0.07884143

cor_below = cor(FM_subcases_below_zero$corruption,
FM_subcases_below_zero$violations, use ='complete.obs')
cor_below

## [1] 0.1242881

cor_above = cor(FM_subcases_above_zero$corruption,
FM_subcases_above_zero$violations, use ='complete.obs')
cor_above

## [1] -0.05543683
```

Results

Upon checking the violations Vs corruption based on corruption index centered at '0', we observed that corruption is relevant in predicting the parking violation when the index is below 0 as indicated by our correlation value at 0 . However observation above the corruption index of "1"", we do not observe a strong relationship between the corruption and the parking violations as indicated by the negative value -0.06 . This somehow

indicates that we need to further fine tune our data analysis with more variables in investigation of corruption index and parking violations.

(b) How does the number of diplomats contribute to the frequency of violations?

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

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

```
##     wbcode             prepost            violations
##  Length:298         Length:298         Min.   :   0.000
##  Class :character   Class :character   1st Qu.:   0.654
##  Mode  :character   Mode  :character   Median :   5.724
##                                        Mean   : 100.879
##                                        3rd Qu.:  51.915
##                                        Max.   :3392.961
##
##      fines               mission      staff           spouse
##  Min.   :     0.00   Min.   :1   Min.   : 2.00   Min.   : 0.000
##  1st Qu.:    65.41   1st Qu.:1   1st Qu.: 6.00   1st Qu.: 3.000
##  Median :   579.72   Median :1   Median : 9.00   Median : 6.000
##  Mean   :  5579.60   Mean   :1   Mean   :11.81   Mean   : 7.758
##  3rd Qu.:  2999.05   3rd Qu.:1   3rd Qu.:14.00   3rd Qu.:10.000
##  Max.   :186163.17   Max.   :1   Max.   :86.00   Max.   :81.000
##
##   gov_wage_gdp      pctmuslim        majoritymuslim        trade
##  Min.   : 0.100   Min.   :0.000000   Min.   :-1.0000   Min.   :0.000e+00
##  1st Qu.: 1.300   1st Qu.:0.006375   1st Qu.: 0.0000   1st Qu.:8.911e+07
##  Median : 1.900   Median :0.050000   Median : 0.0000   Median :5.194e+08
##  Mean   : 2.828   Mean   :0.280317   Mean   : 0.2517   Mean   :1.025e+10
##  3rd Qu.: 3.625   3rd Qu.:0.547500   3rd Qu.: 1.0000   3rd Qu.:4.796e+09
##  Max.   :11.800   Max.   :0.999000   Max.   : 1.0000   Max.   :3.290e+11
##  NA's   :114      NA's   :4          NA's   :4         NA's   :4
##    cars_total     cars_personal     cars_mission        pop1998
##  Min.   :  1.00   Min.   : 0.000   Min.   :  0.000   Min.   :5.308e+05
##  1st Qu.:  3.00   1st Qu.: 1.000   1st Qu.:  2.000   1st Qu.:3.815e+06
##  Median :  7.00   Median : 2.000   Median :  3.000   Median :8.852e+06
##  Mean   : 10.47   Mean   : 5.324   Mean   :  5.144   Mean   :3.655e+07
##  3rd Qu.: 12.00   3rd Qu.: 6.000   3rd Qu.:  6.000   3rd Qu.:2.341e+07
##  Max.   :116.00   Max.   :64.000   Max.   :116.000   Max.   :1.242e+09
```

```
##  NA's   :20       NA's   :20       NA's   :20
##   gdppcus1998           ecaid             milaid            region
##  Min.   :   95.45   Min.   :   0.00   Min.   :   0.000   Min.   :1.000
##  1st Qu.:  412.07   1st Qu.:   0.00   1st Qu.:   0.000   1st Qu.:3.000
##  Median : 1374.88   Median :   8.70   Median :   0.200   Median :4.000
##  Mean   : 5044.09   Mean   :  49.27   Mean   :  33.048   Mean   :4.372
##  3rd Qu.: 4936.62   3rd Qu.:  40.30   3rd Qu.:   0.775   3rd Qu.:6.000
##  Max.   :36485.64   Max.   :1026.10   Max.   :3120.000   Max.   :7.000
##                     NA's   :4         NA's   :4          NA's   :2
##    corruption           totaid           r_africa       r_middleeast
##  Min.   :-2.58299   Min.   :   0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:-0.41515   1st Qu.:   0.325   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 0.32696   Median :   9.000   Median :0.0000   Median :0.0000
##  Mean   : 0.01364   Mean   :  82.320   Mean   :0.3087   Mean   :0.1007
##  3rd Qu.: 0.72025   3rd Qu.:  42.950   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   : 1.58281   Max.   :4069.100   Max.   :1.0000   Max.   :1.0000
##                     NA's   :4
##     r_europe      r_southamerica       r_asia         country
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Length:298
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   Class :character
##  Median :0.0000   Median :0.0000   Median :0.0000   Mode  :character
##  Mean   :0.2349   Mean   :0.1208   Mean   :0.1678
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##    distUNplz            vpd
##  Min.   : 0.0000   Min.   :  0.00000
##  1st Qu.: 0.2219   1st Qu.:  0.07722
##  Median : 0.2956   Median :  0.60506
##  Mean   : 0.5493   Mean   :  9.86292
##  3rd Qu.: 0.4608   3rd Qu.:  7.80324
##  Max.   :15.0552   Max.   :249.36491
##  NA's   :6
```

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

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

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

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

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

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

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

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

Results

The number of staff does not correlate with the frequency of parking violations in New York Manhattan. Investigation of the average number of violations per diplomats clarified our previous findings that the number of diplomats did not matter. Some countries diplomat committed parking violations as high as 249.36, which rather suggested other underlying causes for such a high frequency per diplomat.