W203 Lab 1: Exploratory Data Analysis of Corruption and Parking Violations

Chi Iong Ansjory, Tsung-Chin Han, Marcelo Queiroz

5/25/2018

Overview

The goal of this lab is to provide you with an opportunity to use R and gain experience performing exploratory data analysis (EDA). In this lab, you will be asked to find new insight into a data set by assessing the underlying structure, evaluating the variables, detecting outliers and anomalies, and so on.

This is a group lab. Each team has been assigned a different dataset to work on. In your assigned folder, you will find a file containing background information on your data, along with a research objective and any instructions that are specific to your team.

Please note that you will be working with real data, but it may have been modified by your instructors to test your abilities.

Although your assigned topic may be the focus of an existing literature, we recommend that you do not spend your time researching what others have done, or gaining significant domain expertise. The purpose of the lab is to see how well you can apply exploratory techniques. Moreover, the background we have provided in your assignment should be sufficient to guide your analysis.[1]

Assignment

Generate an exploratory analysis to address the goals found in your assigned folder.

Be sure to follow the guidelines we covered in class. Remember that you are to use descriptive tools (no inference), but note any features you find that you think would be relevant to statistical modeling.

Your analysis should be thorough, but limit your report to a maximum of 25 pages. This means that you will have to make choices about what variables and relationships to focus on (and justify those choices).

To assist with evaluation, we are providing the following outline for your report. As you work, you may fill in each section with your analysis.

Introduction (20 pts)

State the research question that motivates your analysis.

How is the amount of parking violations received by a UN diplomat related to the country corruption index?

Load your data set into R.

```r
load("Corrupt.RData")
```

[1] We also do not want you to be led astray by the bad advice that is common on the internet.

Describe your data set. What types of variables does it contain? How many observations are there?

We note that we have 364 observations and 28 variables.

```r
nrow(FMcorrupt)
```

```
## [1] 364
```

```r
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 ...
```

Evaluate the data quality. Are there any issues with the data? Explain how you handled these potential issues.

```r
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
```

```r
summary(FMcorrupt$corruption)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
## -2.58299 -0.46186  0.32292 -0.00932  0.71516  1.58281       61
```

Explain whether any data processing or preparation is required for your data set.

```
subcase_nona = !is.na(FMcorrupt$corruption)
FMcorrupt_nona = FMcorrupt[subcase_nona, ]
nrow(FMcorrupt_nona)
```

```
## [1] 303
```

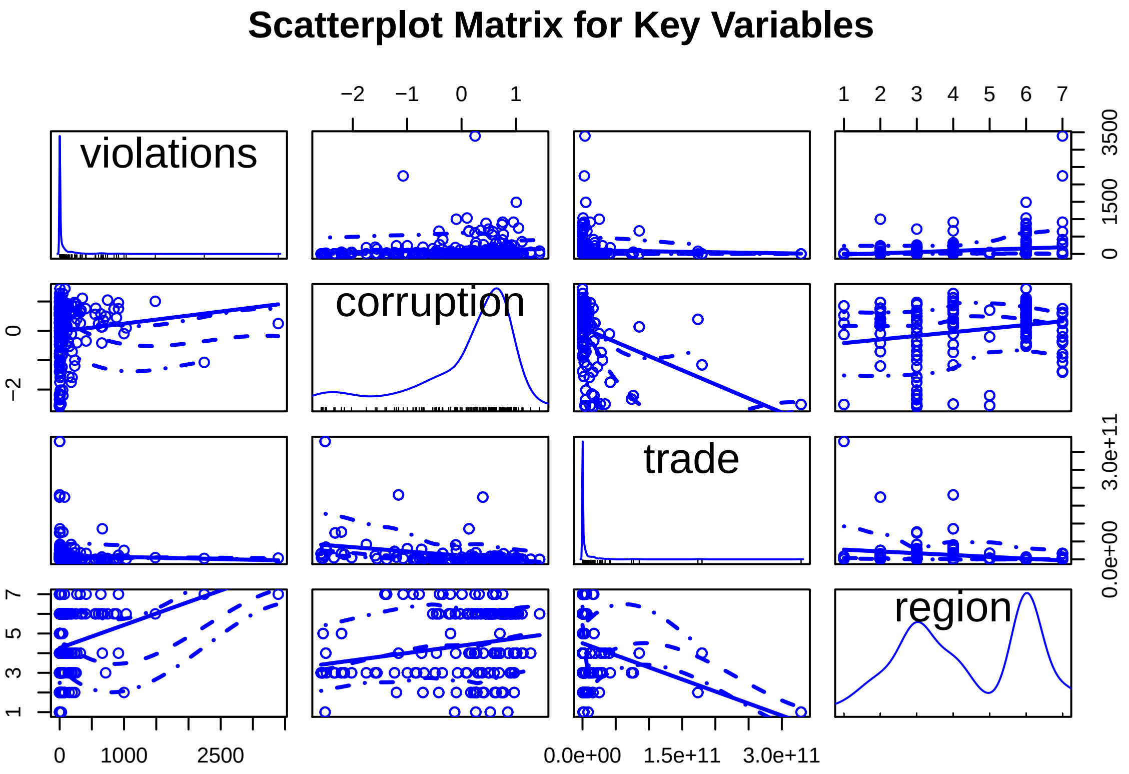
Univariate Analysis of Key Variables (20 pts)

Use visualizations and descriptive statistics to perform a univariate analysis of each key variable. Be sure to describe any anomalies, coding issues, or potentially erroneous values. Explain how you respond to each issue you identify. Note any features that appear relevant to statistical analysis. Discuss what transformations may be appropriate for each variable.

```
library(car)
```

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

```
scatterplotMatrix(~ violations + corruption + trade + region,
                  data = FMcorrupt_nona,
                  main = "Scatterplot Matrix for Key Variables")
```

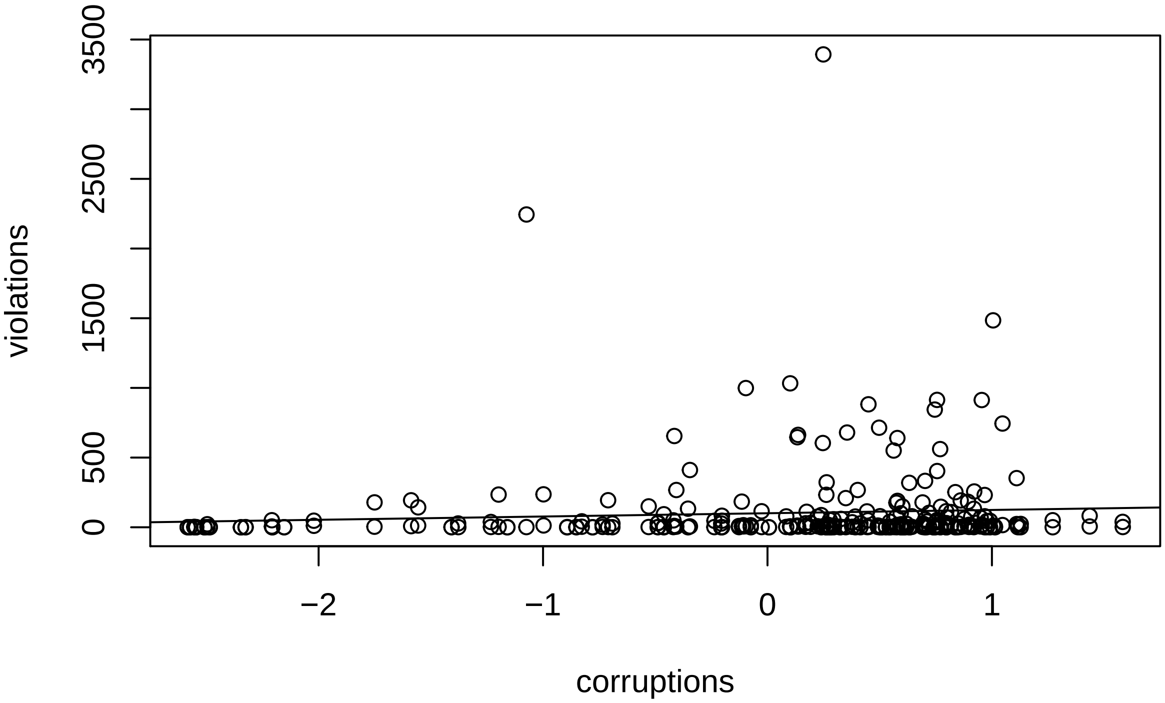


Analysis of Key Relationships (30 pts)

Explore how your outcome variable is related to the other variables in your dataset. Make sure to use visualizations to understand the nature of each bivariate relationship.

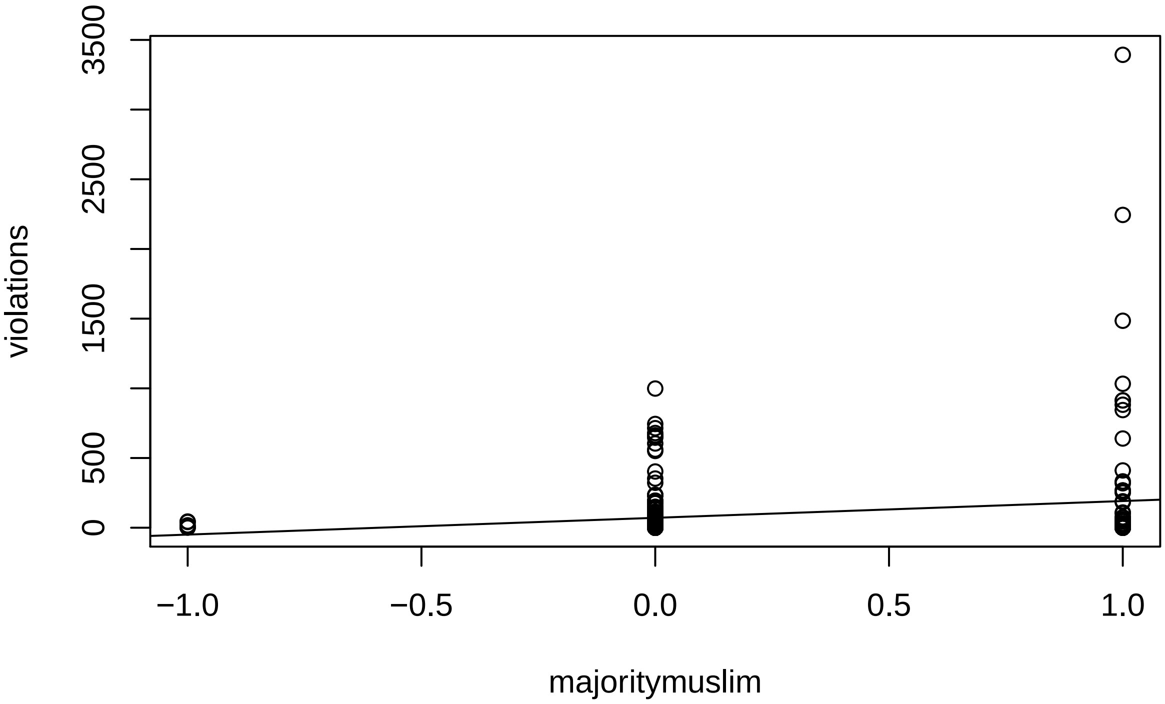
First, use plot to figure out the relationship between violations and corruptions. From the plot, the majority of density of violations take place in the corruptions between 0 and 1.

```
plot(jitter(FMcorrupt_nona$corruption, factor=2), jitter(FMcorrupt_nona$violations, factor=2), xlab = "
abline(lm(FMcorrupt_nona$violations ~ FMcorrupt_nona$corruption))
```



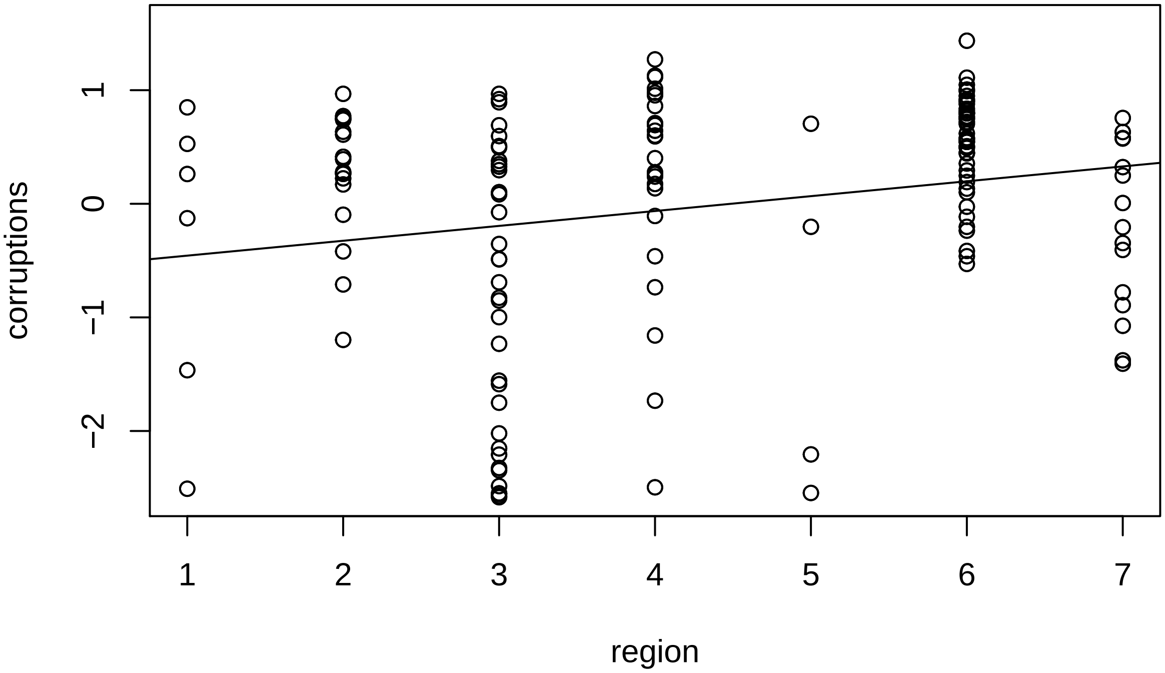
Then, inspect the relationship between violations and majoritymuslim. There is no clear correlation between them.

```
plot((FMcorrupt_nona$majoritymuslim),(FMcorrupt_nona$violations), xlab = "majoritymuslim", ylab = "viola
abline(lm(FMcorrupt_nona$violations ~ FMcorrupt_nona$majoritymuslim))
```



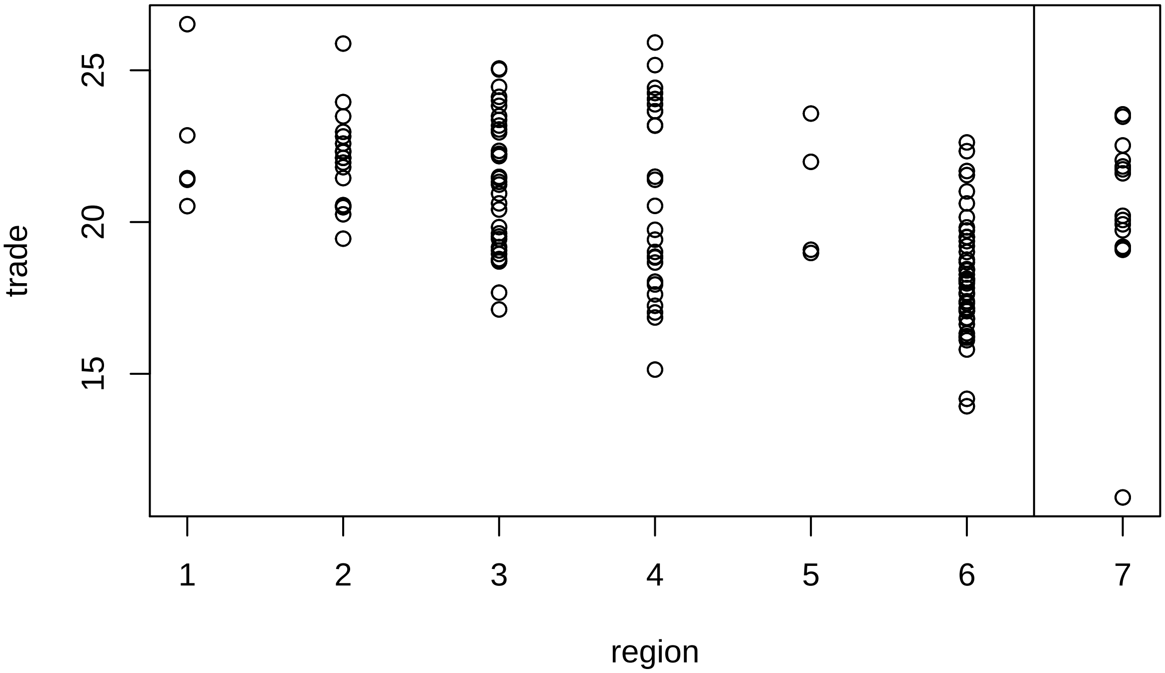
Now, trying to plot the corruptions against each region. Region 6 has the most distribution of corruptions between 0 and 1.

```
plot((FMcorrupt_nona$region), (FMcorrupt_nona$corruption), xlab = "region", ylab = "corruptions")
abline(lm(FMcorrupt_nona$corruption ~ FMcorrupt_nona$region))
```



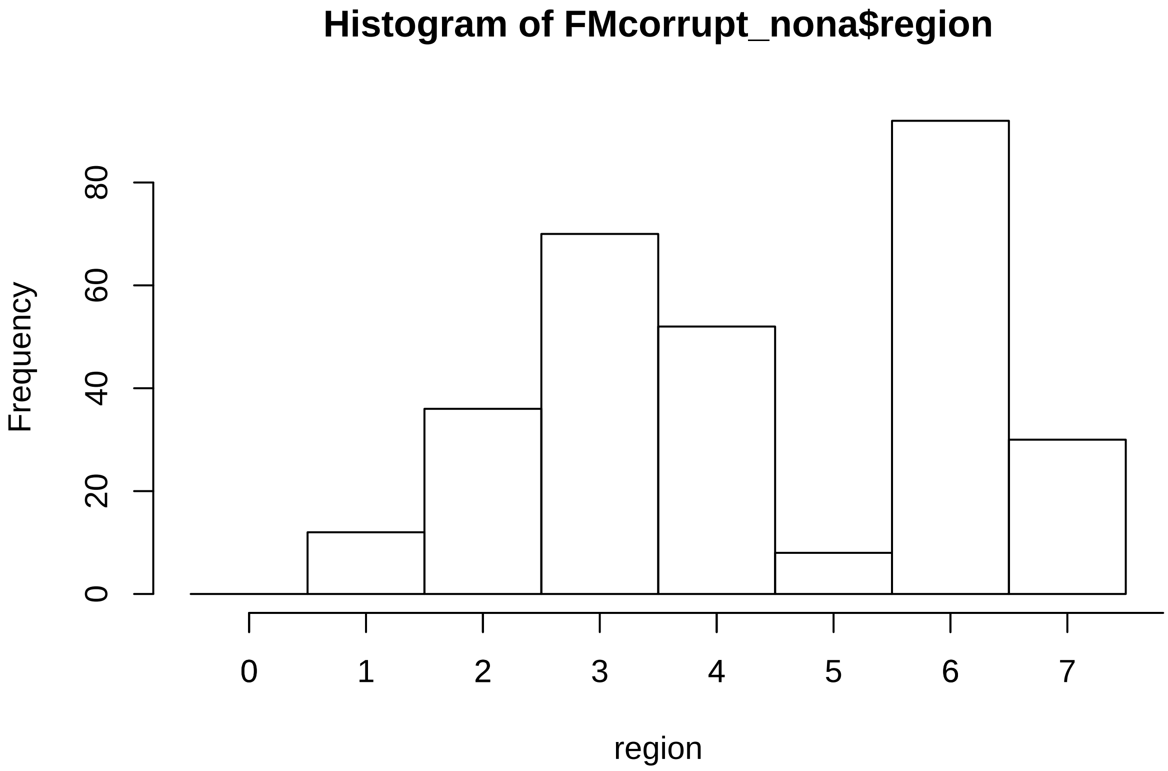
Region 6 tends to have lower total trade with United States.

```
plot(FMcorrupt_nona$region, log(FMcorrupt_nona$trade), xlab = "region", ylab = "trade")
abline(lm(FMcorrupt_nona$trade ~ FMcorrupt_nona$region))
```



Region 6 has the most UN diplomats presence.

```
hist(FMcorrupt_nona$region, breaks = 0:8 - 0.5, xlab = "region")
axis(1, at = 0:8)
```



What tranformations can you apply to clarify the relationships you see in the data? Be sure to justify each transformation you use.

Now, use transformation to find relationship between region number and the region name:

Region 6 = Africa

```
subcase_africa = FMcorrupt_nona$r_africa == 1 & !is.na(FMcorrupt_nona$r_africa)
FMcorrupt_africa = FMcorrupt_nona[subcase_africa, ]
summary(FMcorrupt_africa$region)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       6       6       6       6       6       6
```

Region 7 = Middle East

```
subcase_middleeast = FMcorrupt_nona$r_middleeast == 1 & !is.na(FMcorrupt_nona$r_middleeast)
FMcorrupt_middleeast = FMcorrupt_nona[subcase_middleeast, ]
summary(FMcorrupt_middleeast$region)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       7       7       7       7       7       7
```

Region 3 = Europe

```
subcase_europe = FMcorrupt_nona$r_europe == 1 & !is.na(FMcorrupt_nona$r_europe)
FMcorrupt_europe = FMcorrupt_nona[subcase_europe, ]
summary(FMcorrupt_europe$region)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       3       3       3       3       3       3 
```

Region 2 = South America

```
subcase_southamerica = FMcorrupt_nona$r_southamerica == 1 & !is.na(FMcorrupt_nona$r_southamerica)
FMcorrupt_southamerica = FMcorrupt_nona[subcase_southamerica, ]
summary(FMcorrupt_southamerica$region)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       2       2       2       2       2       2 
```

Region 4 = Asia

```
subcase_asia = FMcorrupt_nona$r_asia == 1 & !is.na(FMcorrupt_nona$r_asia)
FMcorrupt_asia = FMcorrupt_nona[subcase_asia, ]
summary(FMcorrupt_asia$region)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       4       4       4       4       4       4 
```

Analysis of Secondary Effects (10 pts)

What secondary variables might have confounding effects on the relationships you have identified? Explain how these variables affect your understanding of the data.

The variable prepost can further segment the data set to see if there is any effect of legal penalities against the violations after 2002.

```
subcase_pre = FMcorrupt_nona$prepost == "pre"
FMcorrupt_pre = FMcorrupt_nona[subcase_pre, ]
nrow(FMcorrupt_pre)
```

```
## [1] 151
```

```
summary(FMcorrupt_pre)
```

```
##     wbcode             prepost            violations     
##  Length:151         Length:151         Min.   :   0.00  
##  Class :character   Class :character   1st Qu.:  17.22  
##  Mode  :character   Mode  :character   Median :  51.65  
##                                        Mean   : 198.07  
##                                        3rd Qu.: 189.59  
##                                        Max.   :3392.96  
##                                        NA's   :2        
##      fines             mission             staff             spouse      
##  Min.   :     0.0   Min.   :0.0000   Min.   : 0.00   Min.   : 0.000  
##  1st Qu.:   930.7   1st Qu.:1.0000   1st Qu.: 5.00   1st Qu.: 3.000  
##  Median :  2838.8   Median :1.0000   Median : 9.00   Median : 5.000  
##  Mean   : 10806.3   Mean   :0.9868   Mean   :11.65   Mean   : 7.656  
##  3rd Qu.: 10362.6   3rd Qu.:1.0000   3rd Qu.:14.00   3rd Qu.:10.000  
##  Max.   :186163.2   Max.   :1.0000   Max.   :86.00   Max.   :81.000  
##  NA's   :2                                                           
##   gov_wage_gdp      pctmuslim        majoritymuslim        trade          
##  Min.   : 0.100   Min.   :0.0000   Min.   :-1.0000   Min.   :0.000e+00  
##  1st Qu.: 1.300   1st Qu.:0.0060   1st Qu.: 0.0000   1st Qu.:9.532e+07  
##  Median : 1.900   Median :0.0500   Median : 0.0000   Median :5.443e+08  
##  Mean   : 2.828   Mean   :0.2766   Mean   : 0.2416   Mean   :1.034e+10  
##  3rd Qu.: 3.625   3rd Qu.:0.5400   3rd Qu.: 1.0000   3rd Qu.:4.904e+09
##  Max.   :11.800   Max.   :0.9990   Max.   : 1.0000   Max.   :3.290e+11
##  NA's   :59       NA's   :2        NA's   :2         NA's   :3
##    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.:3.788e+06
##  Median :  7.00   Median : 2.000   Median :  3.000   Median :8.257e+06
##  Mean   : 10.47   Mean   : 5.324   Mean   :  5.144   Mean   :3.613e+07
##  3rd Qu.: 12.00   3rd Qu.: 6.000   3rd Qu.:  6.000   3rd Qu.:2.296e+07
##  Max.   :116.00   Max.   :64.000   Max.   :116.000   Max.   :1.242e+09
##  NA's   :12       NA's   :12       NA's   :12
##   gdppcus1998           ecaid            milaid            region
##  Min.   :   95.45   Min.   :   0.00   Min.   :   0.00   Min.   :1.000
##  1st Qu.:  415.14   1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:3.000
##  Median : 1416.04   Median :   8.70   Median :   0.20   Median :4.000
##  Mean   : 5223.74   Mean   :  49.27   Mean   :  33.05   Mean   :4.347
##  3rd Qu.: 5139.20   3rd Qu.:  39.90   3rd Qu.:   0.75   3rd Qu.:6.000
##  Max.   :36485.64   Max.   :1026.10   Max.   :3120.00   Max.   :7.000
##                     NA's   :4         NA's   :4         NA's   :1
##    corruption            totaid           r_africa       r_middleeast
##  Min.   :-2.582988   Min.   :   0.00   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:-0.440568   1st Qu.:   0.35   1st Qu.:0.0000   1st Qu.:0.00000
##  Median : 0.322920   Median :   9.00   Median :0.0000   Median :0.00000
##  Mean   :-0.007721   Mean   :  82.32   Mean   :0.3046   Mean   :0.09934
##  3rd Qu.: 0.715164   3rd Qu.:  42.90   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   : 1.582807   Max.   :4069.10   Max.   :1.0000   Max.   :1.00000
##                      NA's   :4
##     r_europe      r_southamerica       r_asia         country
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Length:151
##  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   :0.2318   Mean   :0.1192   Mean   :0.1722
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##    distUNplz
##  Min.   : 0.0000
##  1st Qu.: 0.2219
##  Median : 0.2956
##  Mean   : 0.5532
##  3rd Qu.: 0.4609
##  Max.   :15.0552
##  NA's   :5
```

```r
subcase_post = FMcorrupt_nona$prepost == "pos"
FMcorrupt_post = FMcorrupt_nona[subcase_post, ]
nrow(FMcorrupt_post)
```

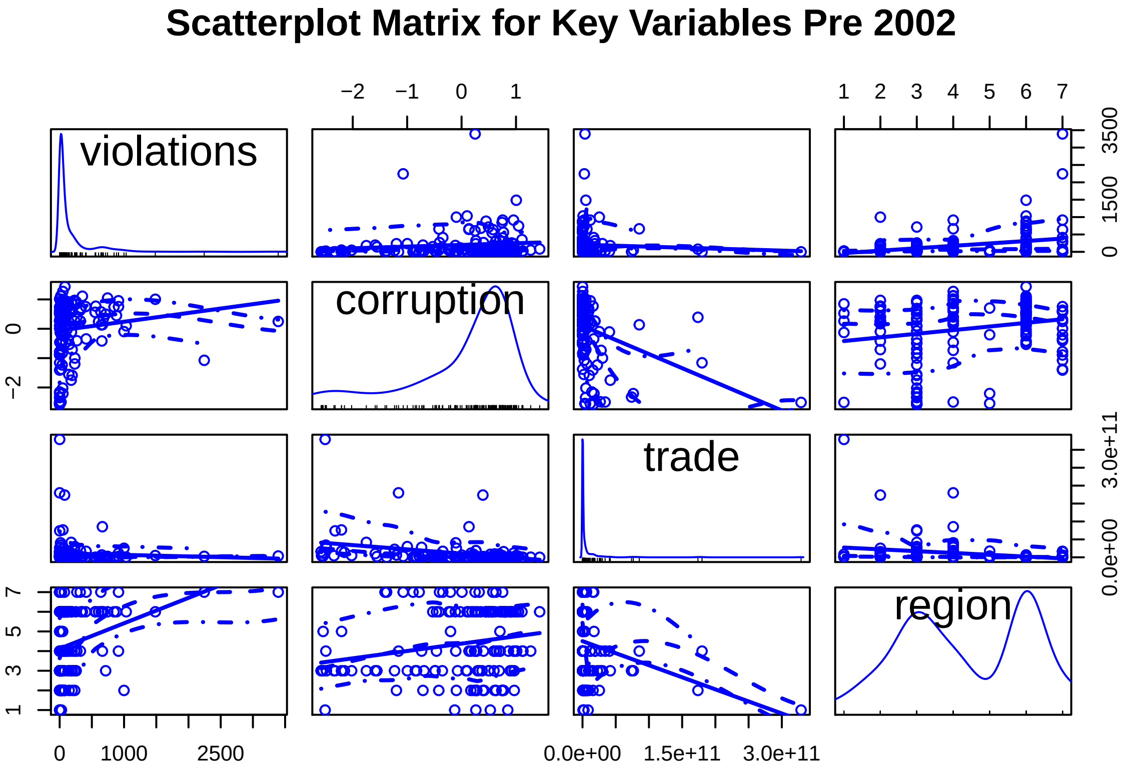
```
## [1] 151
```

```r
summary(FMcorrupt_post)
```

```
##     wbcode            prepost            violations          fines
##  Length:151         Length:151         Min.   : 0.0000   Min.   :   0.00
##  Class :character   Class :character   1st Qu.: 0.3271   1st Qu.:  37.61
##  Mode  :character   Mode  :character   Median : 1.3082   Median : 143.91
##                                        Mean   : 3.6877   Mean   : 352.92
##                                        3rd Qu.: 4.5789   3rd Qu.: 470.97
##                                        Max.   :52.0027   Max.   :5100.52
##                                        NA's   :2         NA's   :2
##     mission           staff          spouse        gov_wage_gdp
##  Min.   :0.0000   Min.   : 0.00   Min.   : 0.000   Min.   : 0.100
##  1st Qu.:1.0000   1st Qu.: 5.00   1st Qu.: 3.000   1st Qu.: 1.300
##  Median :1.0000   Median : 9.00   Median : 5.000   Median : 1.900
##  Mean   :0.9868   Mean   :11.65   Mean   : 7.656   Mean   : 2.828
##  3rd Qu.:1.0000   3rd Qu.:14.00   3rd Qu.:10.000   3rd Qu.: 3.625
##  Max.   :1.0000   Max.   :86.00   Max.   :81.000   Max.   :11.800
##                                                    NA's   :59
##    pctmuslim      majoritymuslim        trade           cars_total
##  Min.   :0.0000   Min.   :-1.0000   Min.   :0.000e+00   Min.   :  1.00
##  1st Qu.:0.0060   1st Qu.: 0.0000   1st Qu.:9.532e+07   1st Qu.:  3.00
##  Median :0.0500   Median : 0.0000   Median :5.443e+08   Median :  7.00
##  Mean   :0.2766   Mean   : 0.2416   Mean   :1.034e+10   Mean   : 10.47
##  3rd Qu.:0.5400   3rd Qu.: 1.0000   3rd Qu.:4.904e+09   3rd Qu.: 12.00
##  Max.   :0.9990   Max.   : 1.0000   Max.   :3.290e+11   Max.   :116.00
##  NA's   :2        NA's   :2         NA's   :3           NA's   :12
##  cars_personal     cars_mission       pop1998           gdppcus1998
##  Min.   : 0.000   Min.   :  0.000   Min.   :5.308e+05   Min.   :   95.45
##  1st Qu.: 1.000   1st Qu.:  2.000   1st Qu.:3.788e+06   1st Qu.:  415.14
##  Median : 2.000   Median :  3.000   Median :8.257e+06   Median : 1416.04
##  Mean   : 5.324   Mean   :  5.144   Mean   :3.613e+07   Mean   : 5223.74
##  3rd Qu.: 6.000   3rd Qu.:  6.000   3rd Qu.:2.296e+07   3rd Qu.: 5139.20
##  Max.   :64.000   Max.   :116.000   Max.   :1.242e+09   Max.   :36485.64
##  NA's   :12       NA's   :12
##      ecaid             milaid            region       corruption
##  Min.   :   0.00   Min.   :   0.00   Min.   :1.000   Min.   :-2.582988
##  1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:3.000   1st Qu.:-0.440568
##  Median :   8.70   Median :   0.20   Median :4.000   Median : 0.322920
##  Mean   :  49.27   Mean   :  33.05   Mean   :4.347   Mean   :-0.007721
##  3rd Qu.:  39.90   3rd Qu.:   0.75   3rd Qu.:6.000   3rd Qu.: 0.715164
##  Max.   :1026.10   Max.   :3120.00   Max.   :7.000   Max.   : 1.582807
##  NA's   :4         NA's   :4         NA's   :1
##      totaid           r_africa       r_middleeast        r_europe
##  Min.   :   0.00   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:   0.35   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :   9.00   Median :0.0000   Median :0.00000   Median :0.0000
##  Mean   :  82.32   Mean   :0.3046   Mean   :0.09934   Mean   :0.2318
##  3rd Qu.:  42.90   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :4069.10   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##  NA's   :4
##  r_southamerica       r_asia         country            distUNplz
##  Min.   :0.0000   Min.   :0.0000   Length:151         Min.   : 0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   Class :character   1st Qu.: 0.2219
##  Median :0.0000   Median :0.0000   Mode  :character   Median : 0.2956
##  Mean   :0.1192   Mean   :0.1722                      Mean   : 0.5455
##  3rd Qu.:0.0000   3rd Qu.:0.0000                      3rd Qu.: 0.4578
##  Max.   :1.0000   Max.   :1.0000                      Max.   :15.0552
##                                                       NA's   :5
```

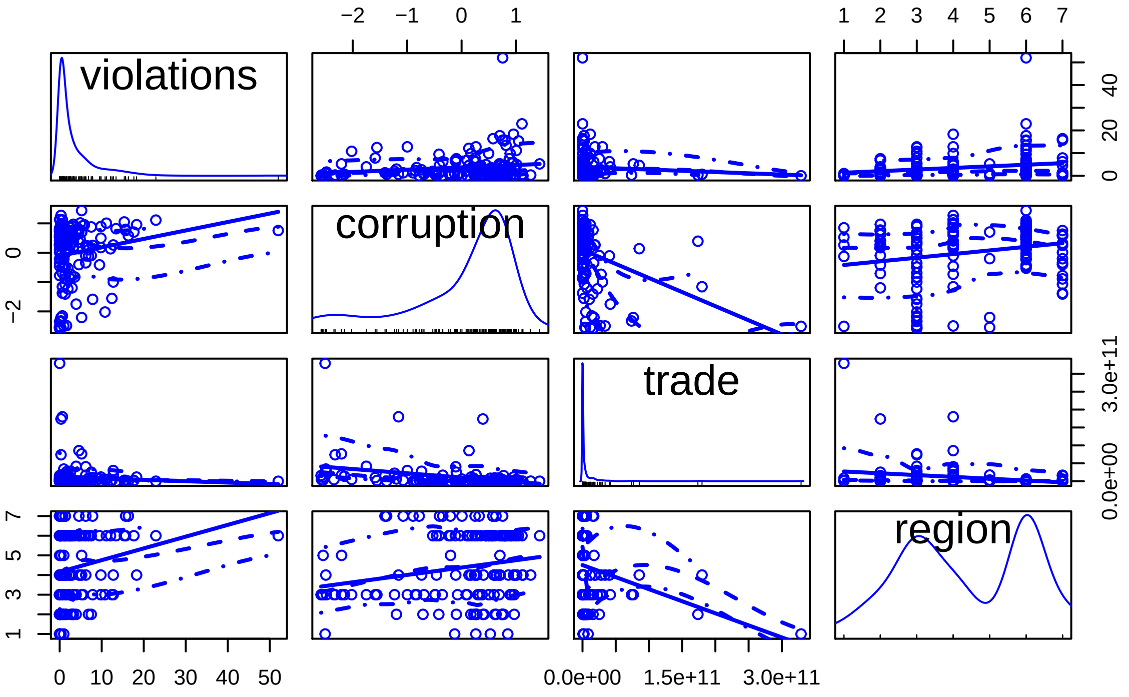
The overall trends between Pre 2002 and Post 2002 are basically the same. However, the overall violations are comparatively lower Post 2002 with same trend as Pre 2002.

```
scatterplotMatrix(~ violations + corruption + trade + region,
                  data = FMcorrupt_pre,
                  main = "Scatterplot Matrix for Key Variables Pre 2002")
```



```
scatterplotMatrix(~ violations + corruption + trade + region,
                  data = FMcorrupt_post,
                  main = "Scatterplot Matrix for Key Variables Post 2002")
```

Scatterplot Matrix for Key Variables Post 2002



Conclusion (20 pts)

Summarize your exploratory analysis. What can you conclude based on your analysis?

The major contribution to diplomatic behavior is cultural norm and legal penalties don't alter the behavior but only suppress the violations. The diplomats from Africa have the significant contribtions to the violations indicated by the Country corruption index between 0 and 1, and lower total trade with the United States. The amount of African diplomats is the most among diplomats from other regions.

Evaluation

We will evaluate your report for technical correctness, but also clarity and overall effectiveness. A point distribution is provided with the above outline. In addition to these point totals, we will impose penalties for output dumps, unclear language, and other errors.

Submission

Only one student in the team needs to submit via the ISVC. Make sure that you include the names of all group members in your report.

You must turn in

1. Your pdf report. In this report, do not suppress the R code that generates your output.
2. The source file you use to generate your report (i.e. your Rmd file)

Use the following naming convention for your files:

- lastname1_lastname2_lab1.pdf
- lastname1_lastname2_lab1.Rmd

Due Date

This lab is due 24 hours before the week 4 live session.

Presentation

During your week 4 live session, your team should present your analysis to the class. Please limit your presentation to 15 minutes (10 minutes plus 5 minutes for questions). You should use this presentation to highlight the process you followed in your EDA, as well as any aspects of your data that find particularly interesting.