EDA

Marta Fajlhauer

3 June 2018

Exploratory Data Analysis

We already have some understanding about features that may influence rate of violent crimes in a given area. The knowledge is important as we may try to reduce the level of the negatively correlated variables and increase the level of positively correlted to stay safe. To finalise my work on this dataset I will run Explanatory Data Analysis (EDA) on some of the variables

As from information about dataset we know that: *However, the normalization does not preserve relationships between values between attributes.* This leads us to the conclusion that we cannot compare or visualise 2 or more numerical explanatory variables together. We can still compare the distribution of different attributes across different communities.

In the summary for variables explained below I have included summary statistics for which:

MAD = Mean Absolute Deviation

"In statistics, the median absolute deviation (MAD) is a robust measure of the variability of a univariate sample of quantitative data. For a univariate data set X1, X2, ..., Xn, the MAD is defined as the median of the absolute deviations from the data's median:

$$\textbf{MAD} = median(|X_i - median(X)|)$$

that is, starting with the residuals (deviations) from the data's median, the MAD is the median of their absolute values." [7]

- **skew**

"In probability theory and statistics, skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean." [7]

- **kurtosis**

"Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers." [8]

- **SE.mean**

"The standard error (SE) of a parameter is the standard deviation of its sampling distribution[1] or an estimate of the standard deviation[2]. If the parameter or the statistic is the mean, it is called the standard error of the mean (SEM). The sampling distribution of a population mean is generated by repeated sampling and recording of the means obtained. This forms a distribution of different means, and this distribution has its own mean and variance." [7]

- **CI mean**

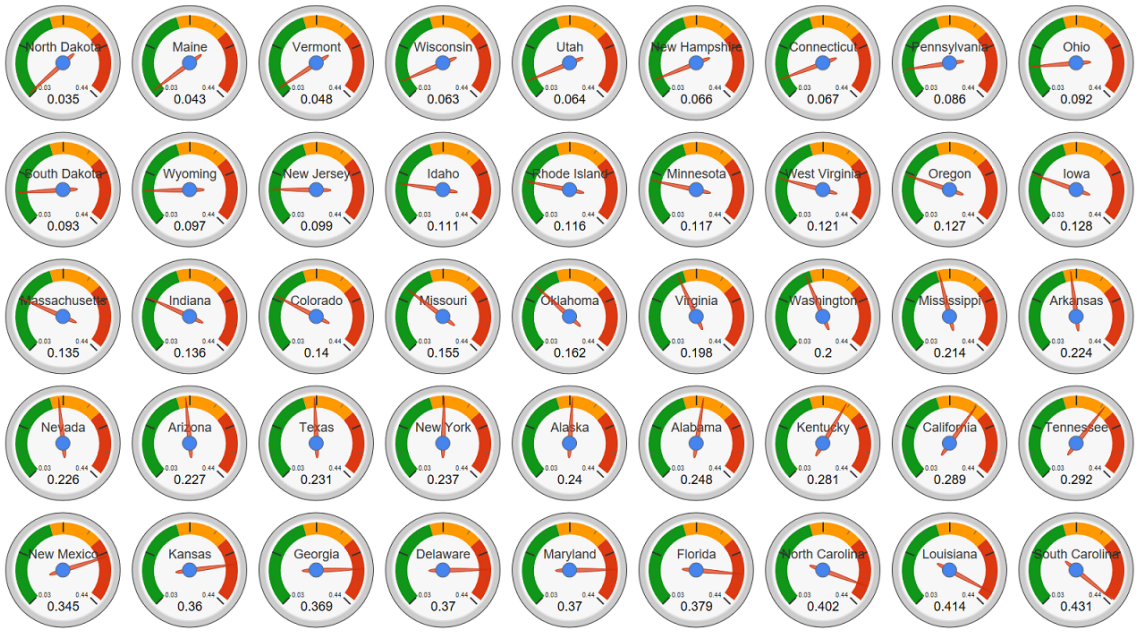
- coef.var variation coefficient = It is the ratio of the standard deviation to the mean.

I also did ANOVA table for the variables described. In the ANOVA table, the mean squares for treatments represent the variance in the given variable per states where mean squares for residuals represents the variance within the particular state. F test is the ratio between the groups and within the group. If there is bigger F-statistics value then there is the bigger difference between different groups.

Violent Crimes per population among different states

The graph shows intensity for a rate of violent crimes per given state. When the average rate for a given state is greater than 0.31 then the arrow inside the circle is pointing to red. From the graph, we see that the most dangerous areas of USA are: South Carolina, Louisiana, North Carolina, Florida, Maryland, Delaware, Georgia, Kansas and New Mexico. If the arrow shows orange then there is a medium rate of intensity for violent crimes. It is taking place in 11 different states including Texas. The rest of the states included in our dataset that is marked as green are the safest areas in the USA with the safest places to be are North Dakota and Maine. I have used the maximum average rate that is considered as safe to be 0.21 so any value that is between 0.21 and 0.31 is considered as orange.

Figure 22: Rates of violent crimes mean in different states



F value from the ANOVA table suggests that there are indeed differences in the number of violent crimes for different states. If the state would have actually an impact on the rate of violent crimes then between the groups would be higher rather than within the group. As F ratio is greater than 1 then this is indeed this situation and we know that state has an impact on the rate of violent crimes rate. Such a small p-value from the test suggests strong evidence that there are differences between the states.

```
   vars    n mean   sd median trimmed  mad  min  max range skew kurtosis
X1    1 1983 0.19 0.17   0.14    0.17 0.12 0.01 0.95  0.94 1.49     2.07
   se
X1  0
```

```
      median         mean      SE.mean CI.mean.0.95          var
 0.140000000  0.194059278  0.003837507  0.007525971  0.029202566
     std.dev     coef.var
 0.170887582  0.880594755
```

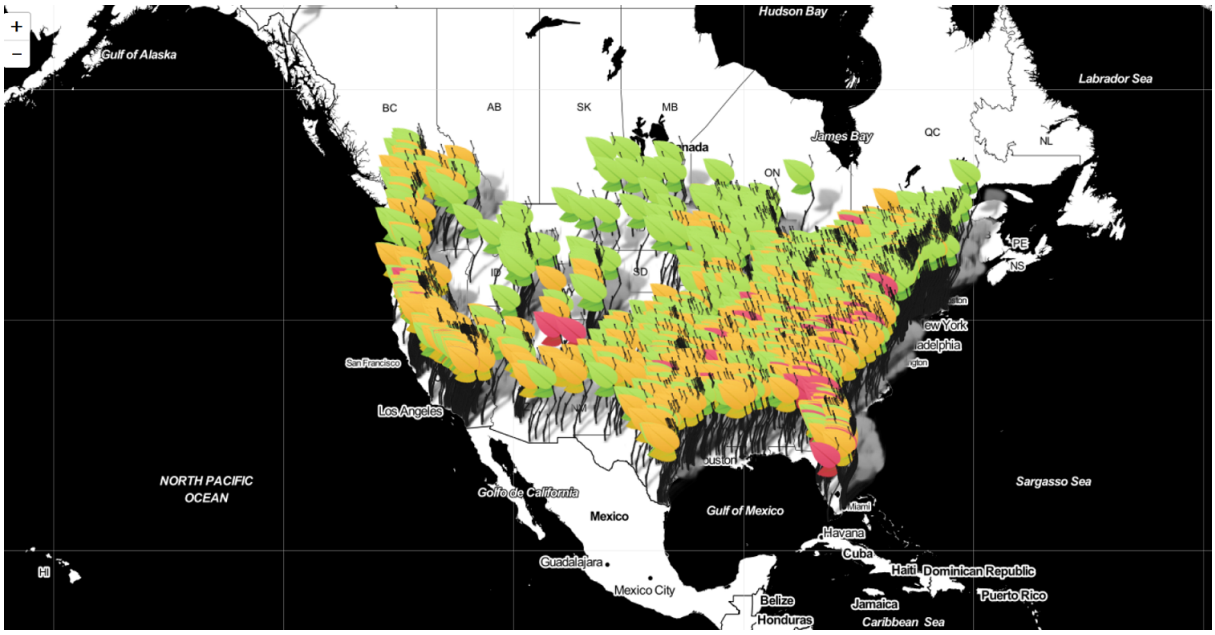
```
              Df Sum Sq Mean Sq F value Pr(>F)
state         44  22.18  0.5042   27.38 <2e-16 ***
Residuals   1938  35.69  0.0184
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

population

We can check not only how violent rate per state looks like but also how it differs across different places within states. Colour of the leaf suggests if there is a big (more than 0.60) rate of violent crime rate represented by

the colour red or low (less than 0.20) which is represented by colour green. Places, where the violent crime is between this range, are represented by the orange colour of the leaf. The map shows me that some areas of the USA with less amount of leaf such as North Dakota, South Dakota, Montana and so on are not described in the data set very well.From dataset description we know that: "The per capita violent crimes variable was calculated using population and the sum of crime variables considered violent crimes in the United States: murder, rape, robbery and assault. There was apparently some controversy in some states concerning the counting of rapes. These resulted in missing values for rape, which resulted in incorrect values for per capita violent crime. These cities are not included in the dataset". That's why we have gaps in the map.

Figure 23: Rates of crimes in a given town, city, community



We can check summary statistics for population variable:

```
   vars    n mean   sd median trimmed  mad  min  max range skew kurtosis
X1    1 1983 0.08 0.12   0.03    0.05 0.03 0.01 0.99  0.98 3.93    18.75
   se
X1  0
```

```
      median         mean      SE.mean CI.mean.0.95          var
 0.030000000  0.075163893  0.002731190  0.005356305  0.014791989
     std.dev     coef.var
 0.121622320  1.618095008
```

```
              Df Sum Sq Mean Sq F value   Pr(>F)
state         44  1.231 0.02798   1.931 0.000262 ***
Residuals   1938 28.087 0.01449
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Smaller F-statistics value rather than in the previous ANOVA table for the violent crimes per population suggests that the variable states have a smaller impact on the population rate than on violent crimes since every single time I am using states as my levels.

householdsize

The smallest range of household sizes is on Alaska and the biggest range in California. In Florida is the smallest value of 0.1 and New Jersey the highest close to 0.8. State Washington has the biggest number of potential outliers.

```
   vars    n mean   sd median trimmed  mad min  max range skew kurtosis se
X1    1 1983 0.44 0.12   0.44    0.44 0.12 0.1 0.77  0.67 0.26    -0.04  0
```

```
      median         mean      SE.mean CI.mean.0.95          var
 0.440000000  0.441387617  0.002730783  0.005355507  0.014787583
     std.dev     coef.var
 0.121604207  0.275504347
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
state        44  4.277 0.09720   7.525 <2e-16 ***
Residuals  1938 25.032 0.01292
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 24

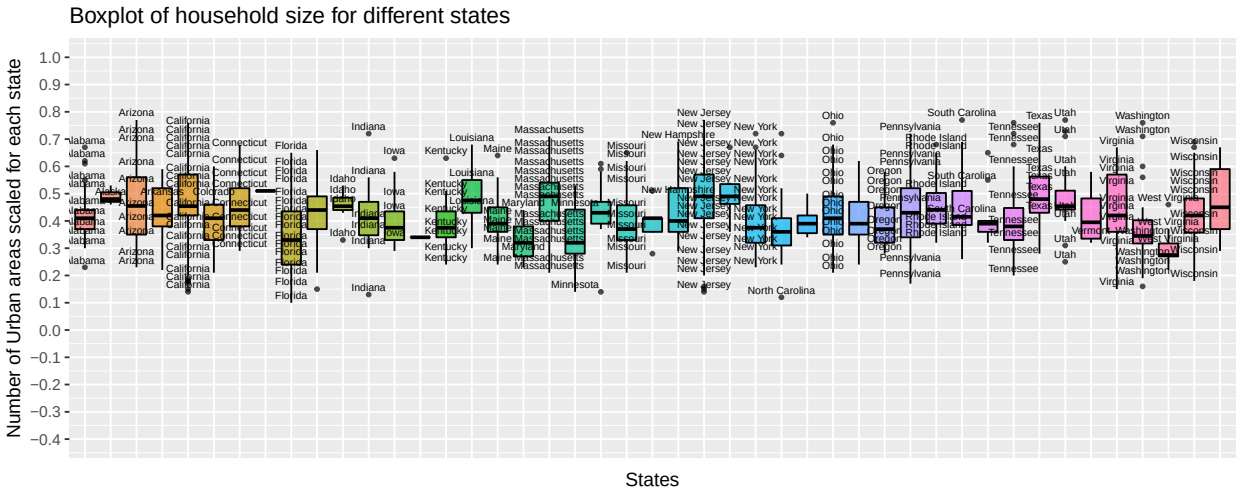
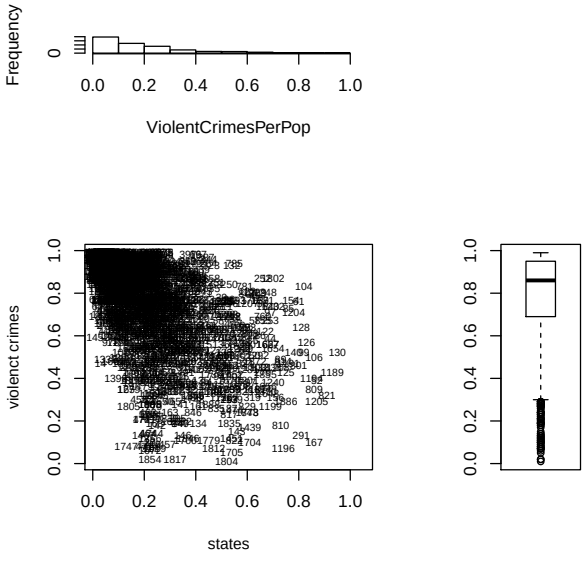


Figure 25: rate of Violent crimes per given states given proportion of Race Percentage White



We can see how Race Percentage White impacts the result of a rate of violent crimes per population rate. Even if those variables are numerical we are expecting them to represent accurate result for the relationship. The scatter plot suggests that the number of violent crimes in the population decreases with increasing percentage of the white race in the population. The histogram shows a distribution of violent crimes per population which suggests that it is right-skewed distribution, most of the time the rate is small. Boxplot shows the distribution of RacePctWhite that suggests that IQR is from 0.7 to 0.9, the minimum value is 0.3 and there are many outliers.

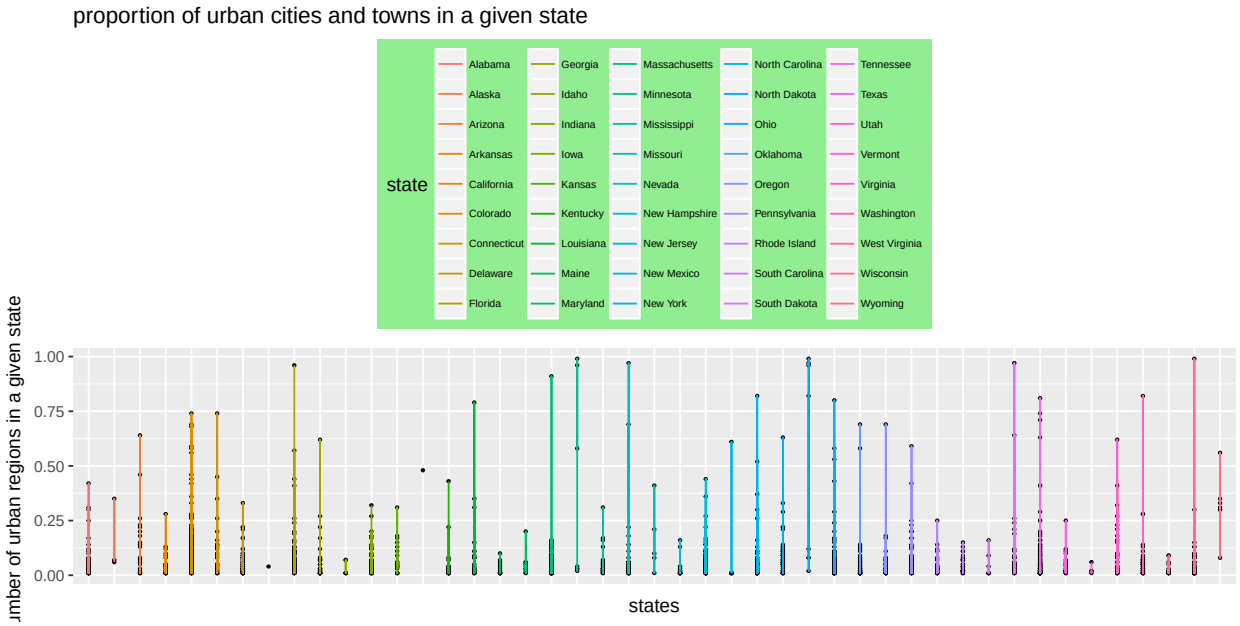
```
   vars    n mean   sd median trimmed  mad  min  max range  skew kurtosis
X1    1 1983 0.78 0.21   0.86    0.82 0.15 0.01 0.99  0.98 -1.34     1.23
   se
X1  0
```

```
      median         mean      SE.mean CI.mean.0.95          var
 0.860000000  0.784526567  0.004678055  0.009174422  0.043396370
     std.dev     coef.var
 0.208317953  0.265533331
```

```
              Df Sum Sq Mean Sq F value Pr(>F)
state         44  41.16  0.9354   40.42 <2e-16 ***
Residuals   1938  44.85  0.0231
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

numbUrban

Figure 26



From this graph, we can see that state Minnesota, North Dakota, Tenessee and Wisconsin are states with the biggest urban areas. States Georgia and Kansas state for which we have the only single record. States Idaho, Vermont, West Virginia and Maine are states with the smallest urban areas.

```
   vars    n mean   sd median trimmed  mad  min  max range skew kurtosis
X1    1 1983 0.08 0.12   0.04    0.05 0.04 0.01 0.99  0.98 4.33    22.74
   se
X1  0
```

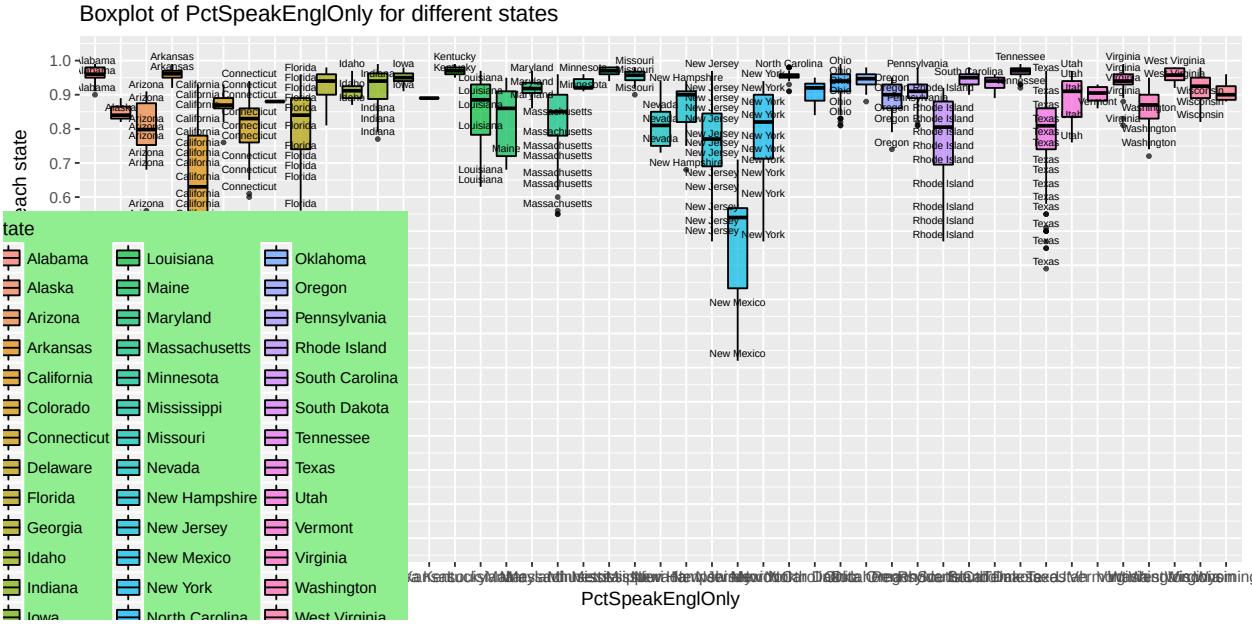
```
      median         mean      SE.mean CI.mean.0.95          var
```

```
 0.040000000  0.076933938  0.002732887  0.005359633  0.014810373
     std.dev     coef.var
 0.121697874  1.581849004
```

```
              Df Sum Sq Mean Sq F value Pr(>F)
state         44  3.893 0.08847   6.734 <2e-16 ***
Residuals   1938 25.461 0.01314
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

PctSpeakEnglishOnly

Figure 27



There is a big difference of proportions for people talking English language only between states. California and New Mexico are states where is the lowest percentage of people talking English language only. Alabama, Arkansas, Kentucky, Mississippi, North Carolina, Tennessee are states with the biggest ratio of people talking English language only. The boxplots suggest that there are no apparent outliers between states.

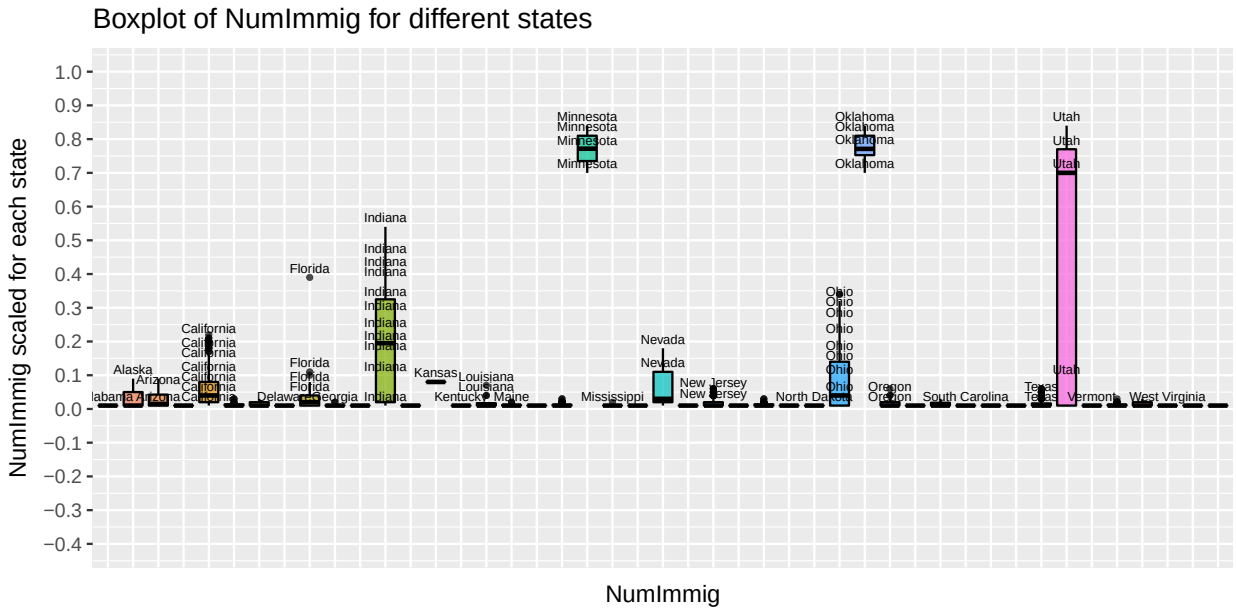
```
   vars    n mean   sd median trimmed mad  min  max range  skew kurtosis
X1    1 1983 0.83 0.16   0.88    0.85 0.1 0.01 0.99  0.98 -1.86     4.08
   se
X1  0
```

```
      median         mean      SE.mean CI.mean.0.95          var
 0.880000000  0.826114064  0.003557158  0.006976162  0.025091643
     std.dev     coef.var
 0.158403418  0.191745214
```

```
              Df Sum Sq Mean Sq F value Pr(>F)
state         44  27.03  0.6143   52.44 <2e-16 ***
Residuals   1938  22.70  0.0117
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

NumImmig

Figure 28



Two states: Minnesota and Oklahoma are states with the biggest proportion of immigrants when Utah has the biggest range.

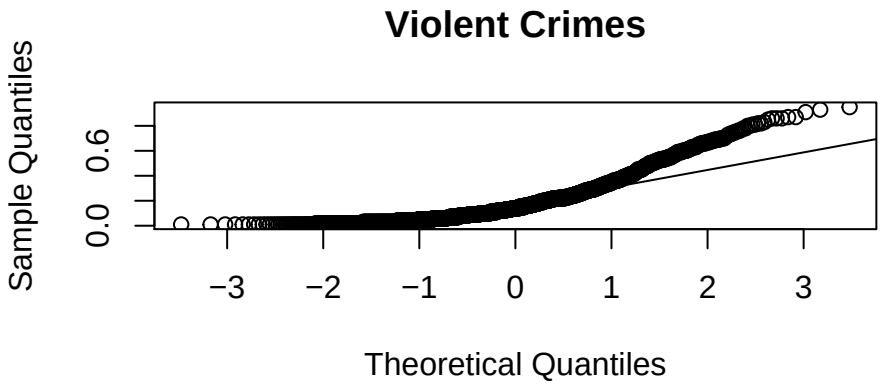
```
   vars    n mean   sd median trimmed mad  min  max range skew kurtosis se
X1    1 1983 0.05 0.13   0.01    0.02   0 0.01 0.84  0.83 4.69    22.16  0

      median         mean      SE.mean CI.mean.0.95          var
 0.010000000  0.050328561  0.002975364  0.005835169  0.017555079
     std.dev     coef.var
 0.132495580  2.632612114

              Df Sum Sq Mean Sq F value Pr(>F)
state         44 28.553  0.6489   201.5 <2e-16 ***
Residuals   1938  6.242  0.0032
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the QQplot below we can see that the normality assumption is not an appropriate assumption, hence the ordinary least squares method is not the best approaching to model the data hence the decision of using advanced regression techniques instead. To final modelling process, I have compared Ridge LASSO and elastic net methods.

Figure 29


Violent Crimes
Sample Quantiles
0.0
0.6
−3
−2
−1
0
1
2
3
Theoretical Quantiles