TAT

## Gun Violence in the United States of America

The investigation of the number of murders in the United States due to shootings is alarming. Therefore, the statistical assessment of the prevalence of gun violence should relay meaningful insights concerning the deplorable state and mitigation measures to counter its existence.

library(dslabs)

## Warning: package 'dslabs' was built under R version 4.0.5

The first step is to load the dataset and assign the dataset to the object data1.

data("murders")  
data1 <- murders

# Exploratory Data Analysis

Notably, the data frame consists of 51 observations and five variables(features), where chr stands for character type, i.e., state and abb and the region is a factor hence it’s a categorical data, population and total are numeric features

str(data1)

## 'data.frame': 51 obs. of 5 variables:  
## $ state : chr "Alabama" "Alaska" "Arizona" "Arkansas" ...  
## $ abb : chr "AL" "AK" "AZ" "AR" ...  
## $ region : Factor w/ 4 levels "Northeast","South",..: 2 4 4 2 4 4 1 2 2 2 ...  
## $ population: num 4779736 710231 6392017 2915918 37253956 ...  
## $ total : num 135 19 232 93 1257 ...

Exploring numeric variables

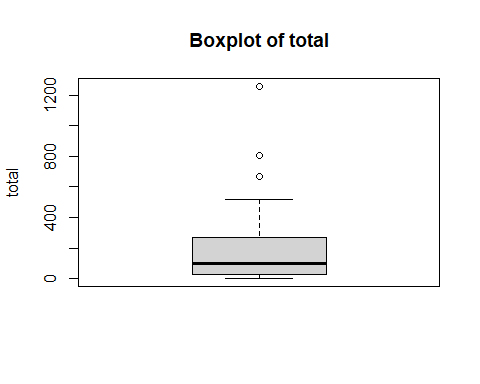
Considering our output, we note that the minimum value of the total is 2 while the average is 184, which is sensitive to outliers. Furthermore, the maximum is 1257, which is way above the mean; hence, it is an outlier. The proof of this numeric information will be visualized using both the histogram and boxplot.

Notably, the total histogram is also skewed towards the right, and it has a heavy tail. To check the measure of central tendency and the spread of our data, one considers the variance and standard deviation. Notably, the outliers in the total variable have outliers from 600.

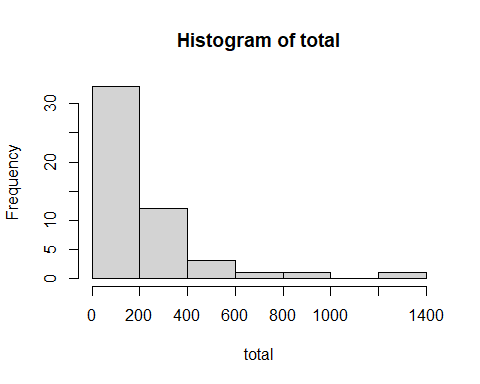
summary(data1[c('population','total')])

## population total   
## Min. : 563626 Min. : 2.0   
## 1st Qu.: 1696962 1st Qu.: 24.5   
## Median : 4339367 Median : 97.0   
## Mean : 6075769 Mean : 184.4   
## 3rd Qu.: 6636084 3rd Qu.: 268.0   
## Max. :37253956 Max. :1257.0

boxplot(data1$total, main="Boxplot of total",ylab="total")



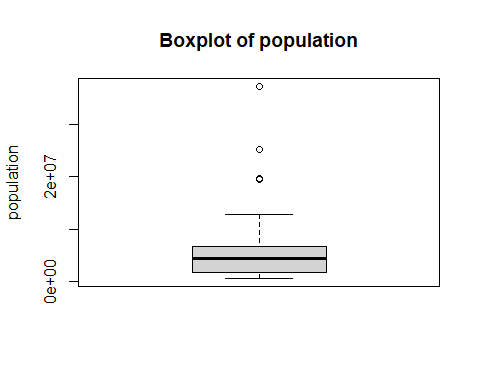
hist(data1$total, main = "Histogram of total",xlab = "total")



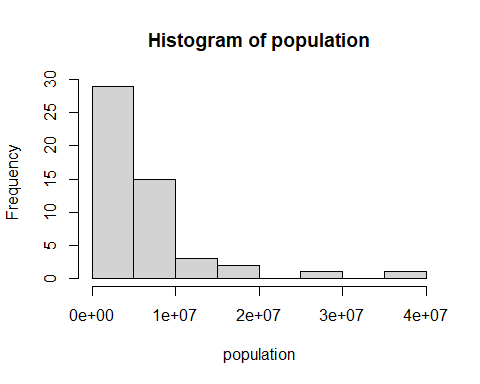
**Visualizing Numeric Variables using Histograms**

In the population variable, the histogram will portray the numerical description of the variable. We note that the standard distribution feature is skewed towards the right with a heavy tail considering the histogram. This shows the data is more spread toward the right

boxplot(data1$population, main="Boxplot of population",ylab="population")



hist(data1$population, main = "Histogram of population",xlab = "population")



We note why the the population feature is far more spread towards the right this is because of the high deviation in population of 6860669.

var(data1$population)

## [1] 4.706878e+13

sd(data1$population)

## [1] 6860669

var(data1$total)

## [1] 55755.56

sd(data1$total)

## [1] 236.1261

**EXPLORING CATEGORICAL DATA**

We note that the mode of the state features is one throughout the column. We note that the southern region has the highest mode while the Northeast has the least mode.Northeast(9),South(17),North Central(12),West(13).

str(data1)

## 'data.frame': 51 obs. of 5 variables:  
## $ state : chr "Alabama" "Alaska" "Arizona" "Arkansas" ...  
## $ abb : chr "AL" "AK" "AZ" "AR" ...  
## $ region : Factor w/ 4 levels "Northeast","South",..: 2 4 4 2 4 4 1 2 2 2 ...  
## $ population: num 4779736 710231 6392017 2915918 37253956 ...  
## $ total : num 135 19 232 93 1257 ...

table(data1$state)

##   
## Alabama Alaska Arizona   
## 1 1 1   
## Arkansas California Colorado   
## 1 1 1   
## Connecticut Delaware District of Columbia   
## 1 1 1   
## Florida Georgia Hawaii   
## 1 1 1   
## Idaho Illinois Indiana   
## 1 1 1   
## Iowa Kansas Kentucky   
## 1 1 1   
## Louisiana Maine Maryland   
## 1 1 1   
## Massachusetts Michigan Minnesota   
## 1 1 1   
## Mississippi Missouri Montana   
## 1 1 1   
## Nebraska Nevada New Hampshire   
## 1 1 1   
## New Jersey New Mexico New York   
## 1 1 1   
## North Carolina North Dakota Ohio   
## 1 1 1   
## Oklahoma Oregon Pennsylvania   
## 1 1 1   
## Rhode Island South Carolina South Dakota   
## 1 1 1   
## Tennessee Texas Utah   
## 1 1 1   
## Vermont Virginia Washington   
## 1 1 1   
## West Virginia Wisconsin Wyoming   
## 1 1 1

table(data1$abb)

##   
## AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS   
## 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   
## MT NC ND NE NH NJ NM NV NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV WY   
## 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

table(data1$region)

##   
## Northeast South North Central West   
## 9 17 12 13

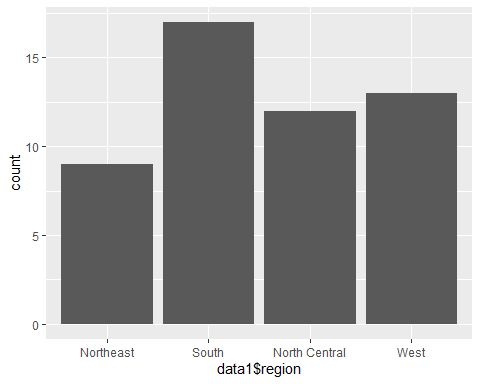
After visualizing a region feature we note that the south region has the highest bar count.

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.0.5

ggplot(data1, aes(x=data1$region)) +  
 geom\_bar()

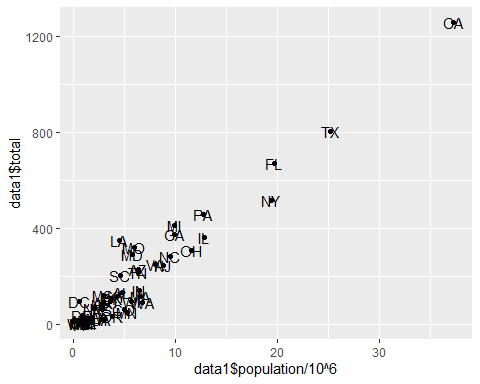
## Warning: Use of `data1$region` is discouraged. Use `region` instead.



**BIVARIATE ANALYSIS**

To obtain the correlation of numerical variables between population and total

ggplot(data = data1)+ geom\_point(aes(data1$population/10^6, data1$total)) +  
 geom\_text(aes(data1$population/10^6, data1$total, label = data1$abb))

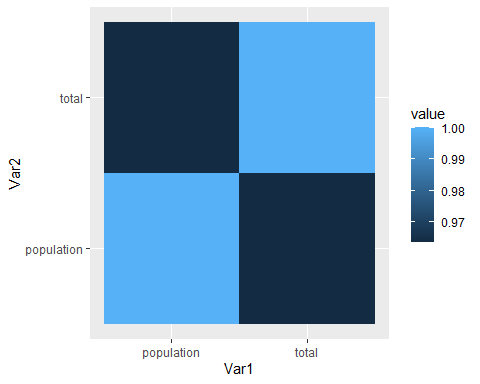
 #heatmap

library(reshape2)

data2 <- cor(data1[sapply(data1,is.numeric)])  
  
data3 <- melt(data2)  
head(data3)

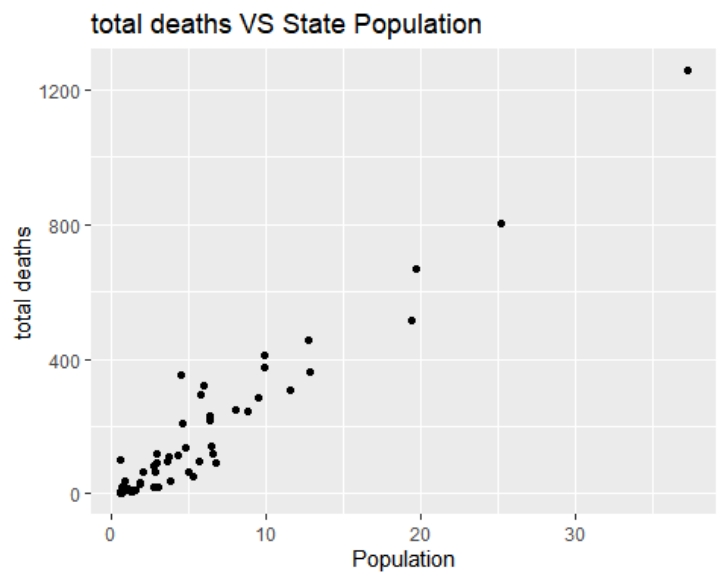
## Var1 Var2 value  
## 1 population population 1.0000000  
## 2 total population 0.9635956  
## 3 population total 0.9635956  
## 4 total total 1.0000000

ggplot(data3, aes(x = Var1,  
 y = Var2,  
 fill = value))+geom\_tile()



exploration report Examining the data before fitting models by creating a scatter plot ,Scatter plots can help visualize any linear relationships between the dependent (response) variable and independent (predictor) variables.

ggplot(data1, aes(population, total))+  
 geom\_point()+  
 labs(title = "total deaths VS State Population",  
 x = "Population",  
 y = "total deaths")

 # Examining the data by also finding the correlation coefficient Correlation is a statistical measure that suggests the level of linear dependence between two variables.Correlation can take values between -1 to +1. If we observe for every instance where population increases,the total number of deaths for each state also increases along with it, then there is a high positive correlation between them and therefore the correlation between them will be closer to 1. The opposite is true for an inverse relationship, in which case, the correlation between the variables will be close to -1.

cor(data1$population/10^6, data1$total)

## [1] 0.9635956

The linear model function lm, used below will create the relationship model between the predictor and the response variable. data1C:\Users\canne\AppData\Local\Temp\ksohtml16656\wps2.jpgpopulation/10^6 presenting the relation between x and y and murders the vector on which the formula will be applied.

simple\_lm <- lm(total~population, data1)  
simple\_lm

##   
## Call:  
## lm(formula = total ~ population, data = data1)  
##   
## Coefficients:  
## (Intercept) population   
## -1.713e+01 3.316e-05

anova(simple\_lm)

## Analysis of Variance Table  
##   
## Response: total  
## Df Sum Sq Mean Sq F value Pr(>F)   
## population 1 2588498 2588498 636.47 < 2.2e-16 \*\*\*  
## Residuals 49 199280 4067   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Generating the anova table. This table will contain the sums of squares, degrees of freedom, F statistic, and p value

pred1 <- predict(simple\_lm, data1)  
pred1

## 1 2 3 4 5 6   
## 141.390359 6.427580 194.860725 79.577913 1218.379075 149.663555   
## 7 8 9 10 11 12   
## 101.406039 12.652641 2.828975 635.802800 311.864240 27.986775   
## 13 14 15 16 17 18   
## 34.861130 408.393666 197.904722 83.903781 77.495187 126.785776   
## 19 20 21 22 23 24   
## 133.219839 26.927504 174.349691 200.021507 310.658382 158.774784   
## 25 26 27 28 29 30   
## 81.281868 181.492478 15.686556 43.442722 72.435391 26.533146   
## 31 32 33 34 35 36   
## 274.451257 51.164660 625.536720 299.111956 5.179271 365.474660   
## 37 38 39 40 41 42   
## 107.284565 109.928532 404.140230 17.780955 136.270701 9.874988   
## 43 44 45 46 47 48   
## 193.338080 816.811161 74.535827 3.625518 248.222511 205.888658   
## 49 50 51   
## 44.326654 171.478780 1.565510

Predicting the response variables

**Data exploration report**

library(DataExplorer)

DataExplorer::create\_report(data1)

##   
##

# Data Profiling Report

* [Basic Statistics](file:///C:\Users\canne\OneDrive\Desktop\interswitch\Data%20Profiling%20Report.docx#basic-statistics)
  + [Raw Counts](file:///C:\Users\canne\OneDrive\Desktop\interswitch\Data%20Profiling%20Report.docx#raw-counts)
  + [Percentages](file:///C:\Users\canne\OneDrive\Desktop\interswitch\Data%20Profiling%20Report.docx#percentages)
* [Data Structure](file:///C:\Users\canne\OneDrive\Desktop\interswitch\Data%20Profiling%20Report.docx#data-structure)
* [Missing Data Profile](file:///C:\Users\canne\OneDrive\Desktop\interswitch\Data%20Profiling%20Report.docx#missing-data-profile)
* [Univariate Distribution](file:///C:\Users\canne\OneDrive\Desktop\interswitch\Data%20Profiling%20Report.docx#univariate-distribution)
  + [Histogram](file:///C:\Users\canne\OneDrive\Desktop\interswitch\Data%20Profiling%20Report.docx#histogram)
  + [Bar Chart (with frequency)](file:///C:\Users\canne\OneDrive\Desktop\interswitch\Data%20Profiling%20Report.docx#bar-chart-with-frequency)
  + [QQ Plot](file:///C:\Users\canne\OneDrive\Desktop\interswitch\Data%20Profiling%20Report.docx#qq-plot)
* [Correlation Analysis](file:///C:\Users\canne\OneDrive\Desktop\interswitch\Data%20Profiling%20Report.docx#correlation-analysis)
* [Principal Component Analysis](file:///C:\Users\canne\OneDrive\Desktop\interswitch\Data%20Profiling%20Report.docx#principal-component-analysis)

### Basic Statistics

#### Raw Counts

|  |  |
| --- | --- |
| **Name** | **Value** |
| Rows | 51 |
| Columns | 5 |
| Discrete columns | 3 |
| Continuous columns | 2 |
| All missing columns | 0 |
| Missing observations | 0 |
| Complete Rows | 51 |
| Total observations | 255 |
| Memory allocation | 10.2 Kb |

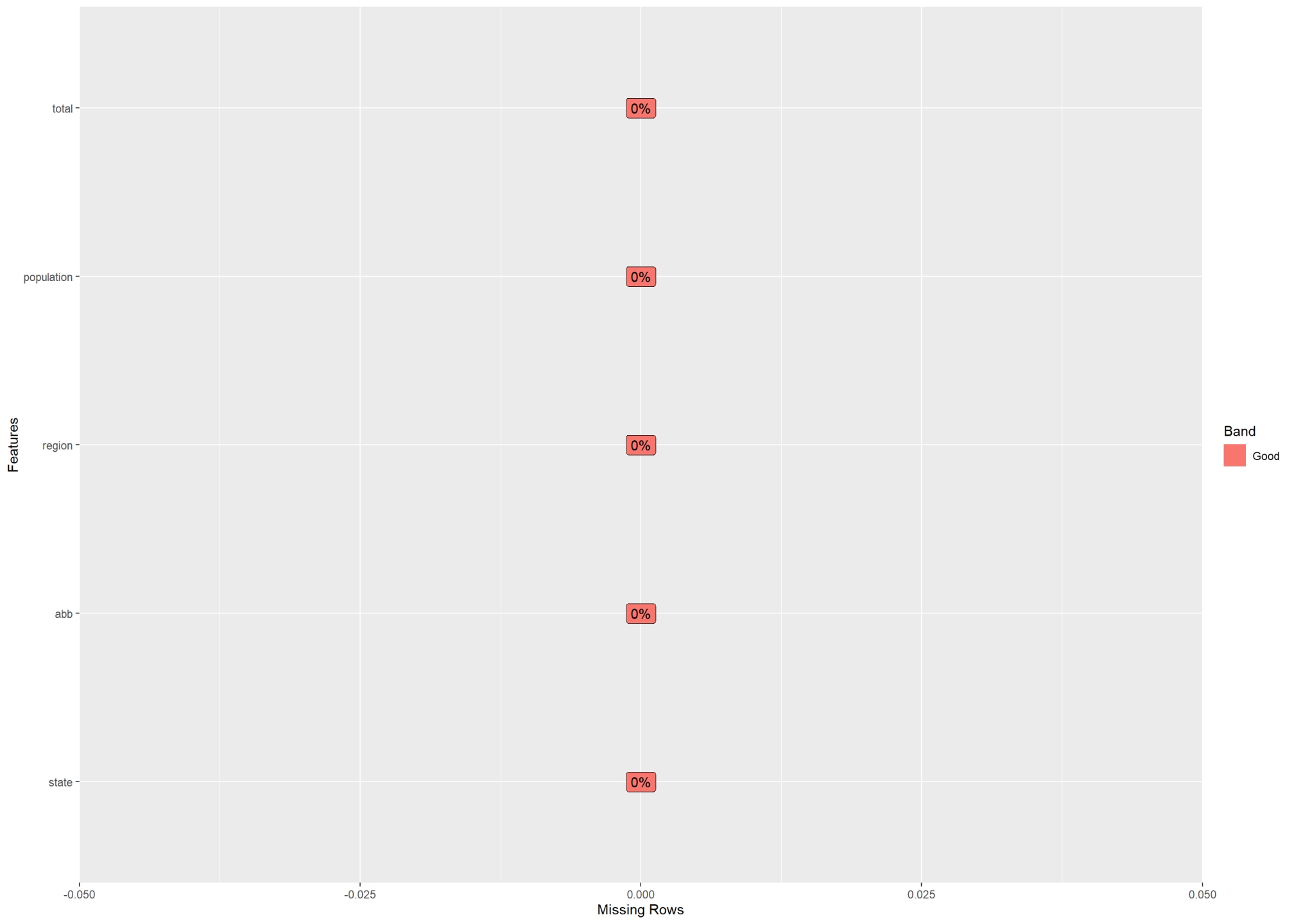
#### Percentages

#### C:\Users\canne\AppData\Local\Temp\ksohtml12976\wps1.jpg

### Data Structure

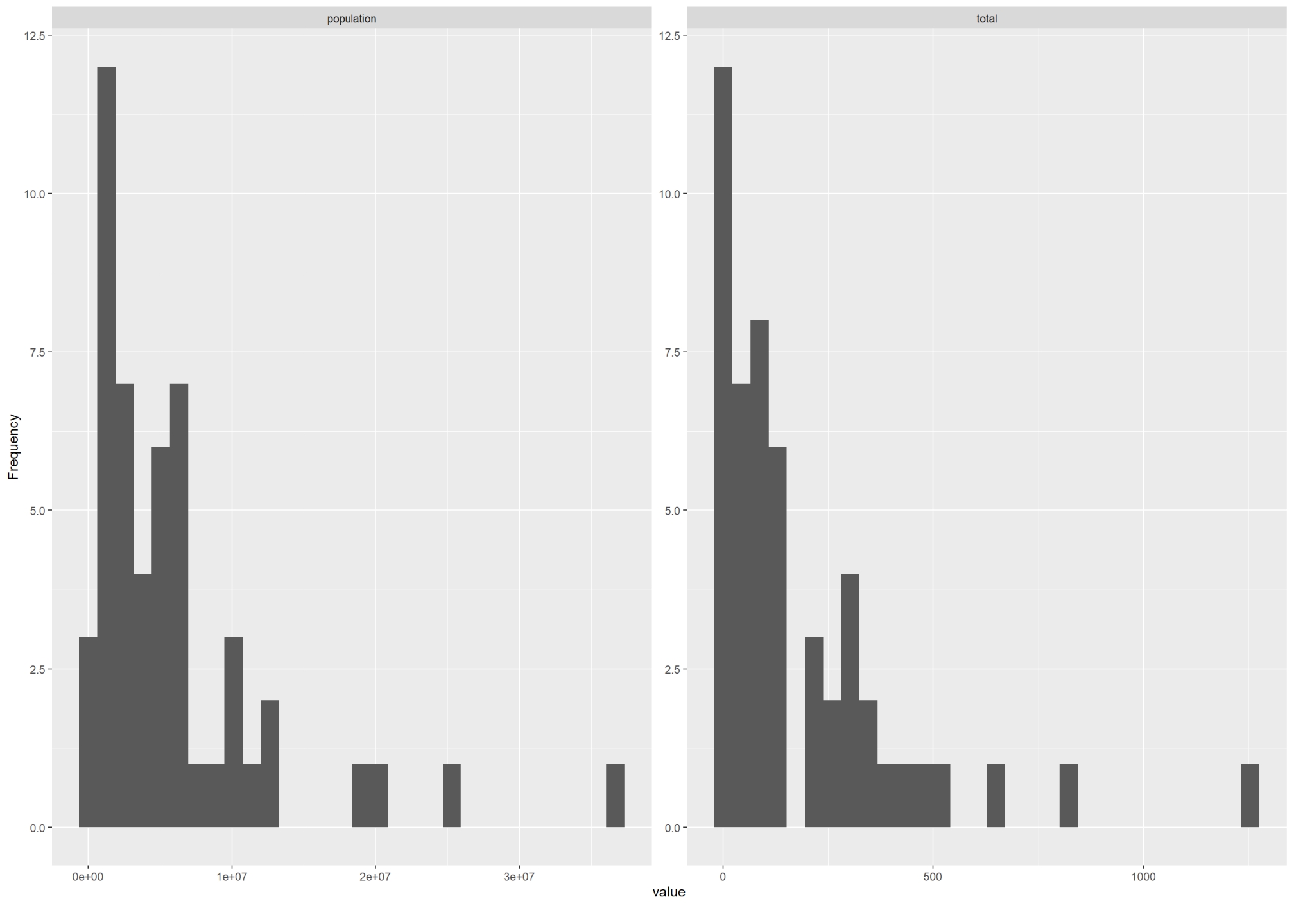
root (Classes 'data.table' and 'data.frame': 51 obs. of 5 variables:)state (chr)abb (chr)region (Factor w/ 4 levels "Northeast","South",)population (num)total (num)

### Missing Data Profile



### Univariate Distribution

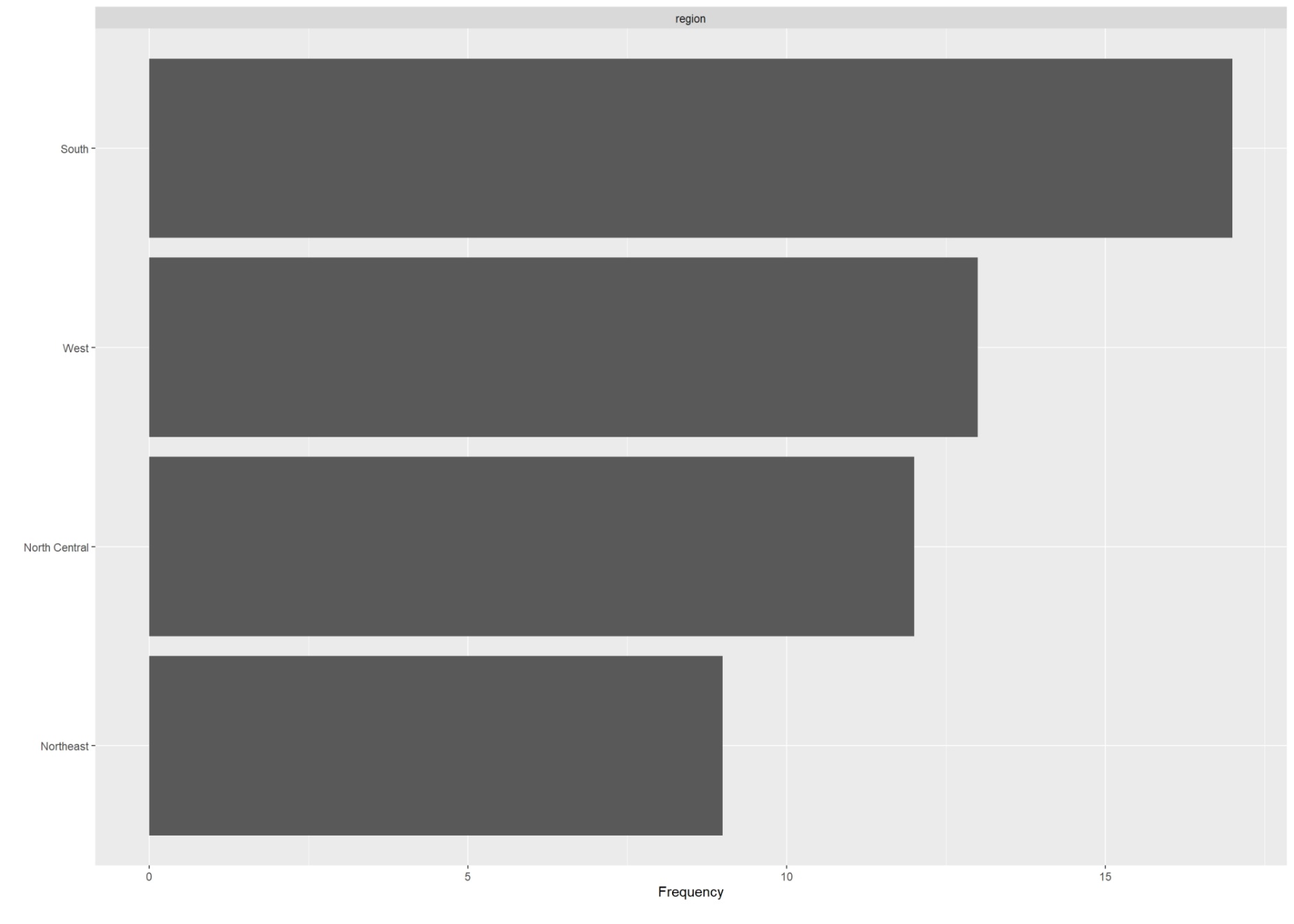
#### Histogram



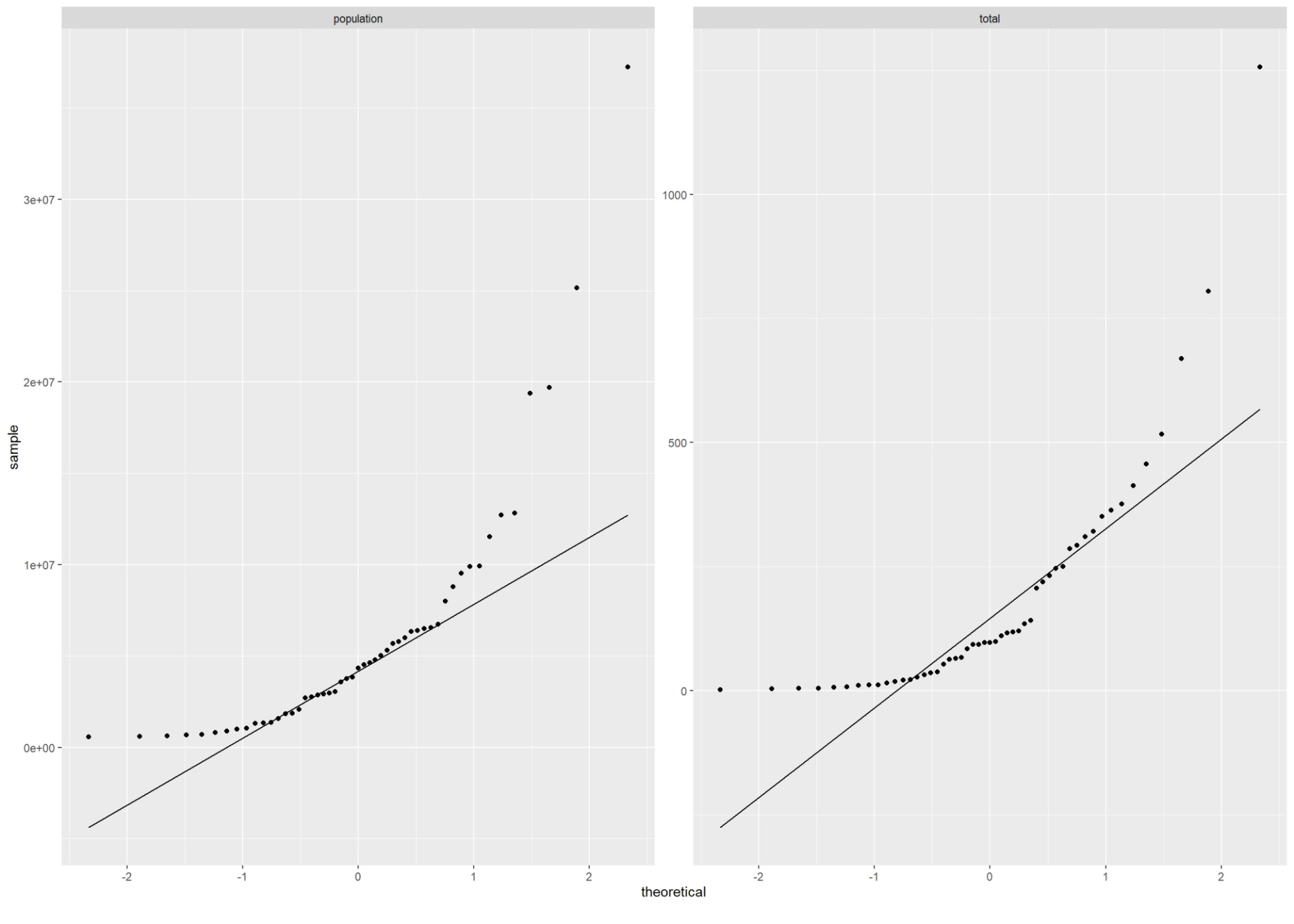
#### Bar Chart (with frequency)

## 2 columns ignored with more than 50 categories.

## state: 51 categories

## abb: 51 categories

#### QQ Plot

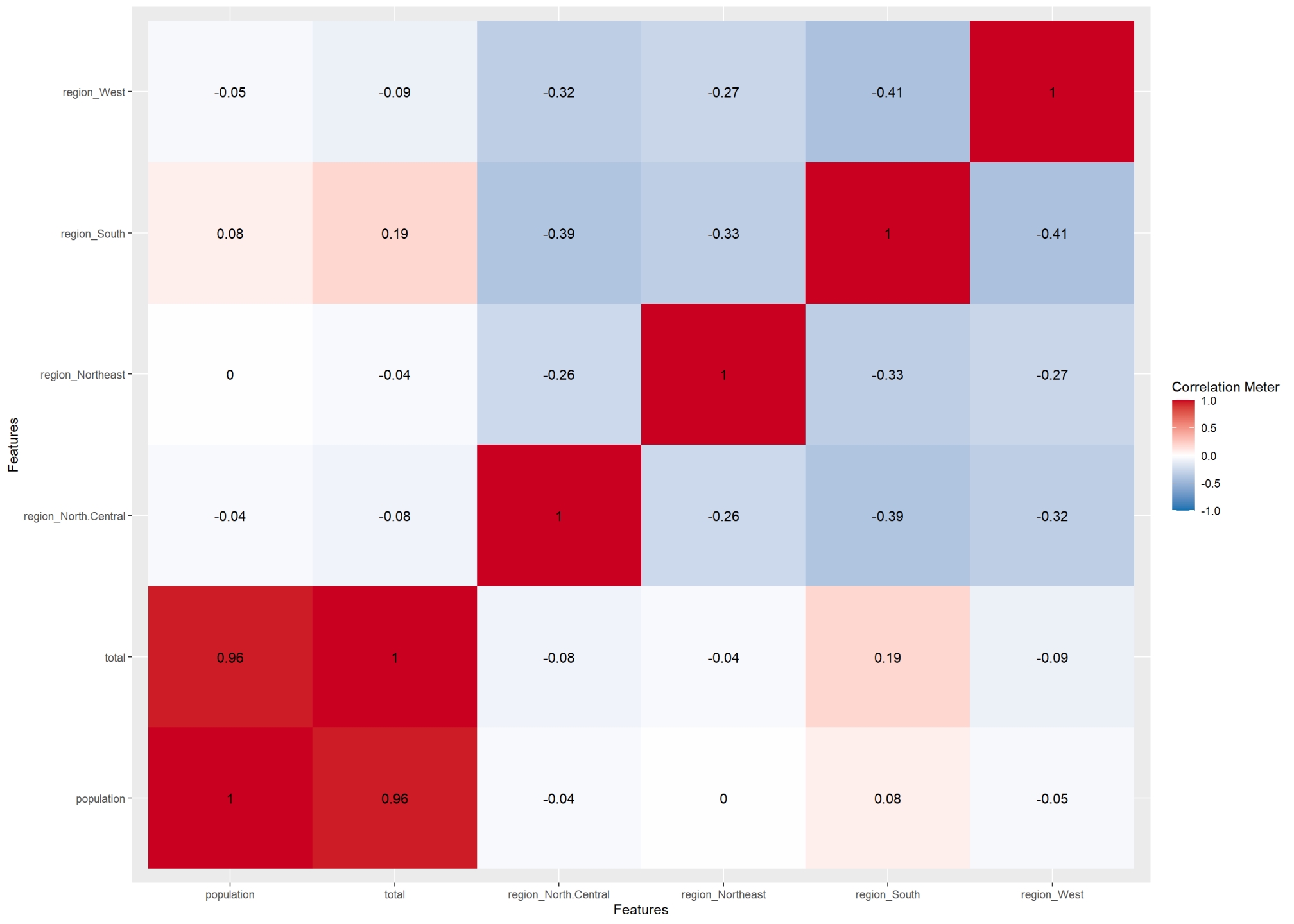


### Correlation Analysis

## 2 features with more than 20 categories ignored!

## state: 51 categories

## abb: 51 categories



### Principal Component Analysis

## 2 features with more than 50 categories ignored!

## state: 51 categories

## abb: 51 categories

