CA1 Data Exploration & Preparation

Student:
Charles Franklin Jahn 2020315
Thiago Santos 2020327

Lecturer: Dr Mohammed Iqbal

03.12.2023
BSc (Hons) in Computing in IT - 4nd Year
Module: Data Exploration & Preparation

Word Count: 2992

Introduction

In our first conversation, during the formation of our team composed of Thiago and Charles, we decided to work on the analysis of the criminal dataset in Brazil. Although Brazil is one of the 10 largest world powers, accessing old data is not so simple, as the digitization of public and government information began only after 2005. [1]

The dataset we selected contains data from 2015 to 2022 on 9 different types of crimes that occurred throughout Brazil, having 5 columns (States (UF), Type of Crime, Year, Month, Occurrences) and 23,020 lines.

```
> dataCrimes <- read.csv("C:/Users/charl/Desktop/classes/Data Exploration and Preparation/CA1/Crimes_in_Brazil_2015_2022.csv")
> dim(dataCrimes)
[1] 23020     5
> summary(dataCrimes)
      UF             Tipo.Crime             Ano            Mês             Ocorrências
 Length:23020       Length:23020       Min.   :2015   Length:23020       Min.   :    0.0
 Class :character   Class :character   1st Qu.:2016   Class :character   1st Qu.:    3.0
 Mode  :character   Mode  :character   Median :2018   Mode  :character   Median :   34.0
                                       Mean   :2018                      Mean   :  204.6
                                       3rd Qu.:2020                      3rd Qu.:  169.0
                                       Max.   :2022                      Max.   :10518.0
> # Show 20 first
> head(dataCrimes, 20)
    UF                          Tipo.Crime  Ano      Mês Ocorrências
1  Acre                             Estupro 2022   janeiro          31
2  Acre                     Furto de veículo 2022   janeiro          50
3  Acre                     Homicídio doloso 2022   janeiro          10
4  Acre     Lesão corporal seguida de morte 2022   janeiro           1
5  Acre      Roubo a instituição financeira 2022   janeiro           0
6  Acre                        Roubo de carga 2022   janeiro           0
```

fig.1

However, the dataset was confusing to read and understand. Therefore, we decided to restructure it, converting each crime category into a column and associating each of them with the corresponding values in the "occurrences" column. After this reorganization for a more comprehensive view, all necessary information was translated into English, and is in the CA's R file.

```
> head(dataReorganized, 20)
# A tibble: 20 × 12
   State    Year Month      Rape Vehicle_Theft Homicide Bodily_injury_followed_by_…¹ Robbery_Institution Cargo_Theft
   <chr>   <int> <chr>     <int>         <int>    <int>                        <int>               <int>       <int>
 1 Acre     2022 janeiro      31            50       10                            1                   0           0
 2 Acre     2022 fevereiro    34            55       10                            0                   0           0
 3 Acre     2022 março        57            44       21                            0                   0           0
 4 Acre     2022 abril        28            40       21                           NA                  NA          NA
 5 Acre     2022 maio         45            29       23                           NA                  NA          NA
 6 Acre     2022 junho        45            50       12                           NA                  NA          NA
 7 Acre     2022 julho        62            44       20                           NA                  NA          NA
 8 Acre     2022 agosto       47            34        9                           NA                  NA          NA
 9 Acre     2022 setembro     59            33       16                           NA                  NA          NA
10 Acre     2022 outubro      49            41       14                           NA                  NA          NA
```

fig.2

Dataset

Data Treatment

As illustrated in Fig.2, we observe the presence of NA values. This occurs due to the lack of information on the crime in question in the original dataset, either due to the lack of cases, or the possibility of data corruption during updating and/or uploading. We chose to fill in these NA values using the MEAN of the corresponding state and month, calculated based on the years 2015 to 2022.

```
> # Find the mean of all crimes based in the states and month of a specific crime
> means_crimes <- dataReorganized %>%
+   group_by(State, Month) %>%
+   summarise(across(c(Rape,
+                      Vehicle_Theft, Homicide, Bodily_injury_followed_by_death,
+                      Robbery_Institution, Cargo_Theft,Vehicle_Robbery,
+                      Robbery_Followed_by_Death,
+                      Attempted_Homicide), ~mean(., na.rm = TRUE), .names = "Mean_{.col}"))
`summarise()` has grouped output by 'State'. You can override using the `.groups` argument.
>
> # Replace NA values in the crime columns with the corresponding mean values, keeping only the columns with the results.
> dataReorganized <- dataReorganized %>%
+   left_join(means_crimes, by = c("State", "Month")) %>%
+   mutate(across(c(Rape,
+                   Vehicle_Theft, Homicide, Bodily_injury_followed_by_death,
+                   Robbery_Institution, Cargo_Theft,Vehicle_Robbery,
+                   Robbery_Followed_by_Death,
+                   Attempted_Homicide), ~ifelse(is.na(.x), get(paste0("Mean_", cur_column())), .x)),
+          across(c(Rape,
+                   Vehicle_Theft, Homicide, Bodily_injury_followed_by_death,
+                   Robbery_Institution, Cargo_Theft, Vehicle_Robbery,
+                   Robbery_Followed_by_Death,
+                   Attempted_Homicide), ~as.integer(round(.)))) %>%
+   select(State, Year, Month, Rape, Vehicle_Theft, Homicide, Bodily_injury_followed_by_death, Robbery_Institution, Cargo_Theft,
Vehicle_Robbery, Robbery_Followed_by_Death, Attempted_Homicide)
```

fig.3

After filling in the NA values, a new column called Total_Crimes was also created, representing the sum of crimes for the corresponding state, year and month. For future analysis purposes.

```
> # Add a column  "Total Crimes"
> dataReorganized <- dataReorganized %>%
+   mutate(Total_Crimes = rowSums(select(., -State, -Year, -Month), na.rm = TRUE))
> head(dataReorganized[, c('State', 'Year', 'Month', 'Total_Crimes')], 20)
# A tibble: 20 x 4
   State    Year Month     Total_Crimes
   <chr>   <int> <chr>            <dbl>
 1 Acre     2022 janeiro            186
 2 Acre     2022 fevereiro          167
 3 Acre     2022 março              227
 4 Acre     2022 abril              169
 5 Acre     2022 maio               190
```

fig.4

Statistical Parameters

Calculating statistical parameters is fundamental in the analysis of numerical datasets, it provides essential insights into the distribution and behavior of the data. The Mean calculation shows us a representative value of the set average, helping to understand what is the center. Median is less sensitive to extreme values, providing a robust center point when data is skewed. The Minimum and Maximum values indicate the extent of the data, in other words, they are the lower and upper limits. Finally, the SD informs us about the dispersion of the data in association with the average, allowing us to understand how close or far the values are in relation to the average value.[2]

```r
# Select only numeric columns
numeric_cols <- sapply(dataReorganized_no_year, is.numeric)
numeric_data <- dataReorganized_no_year[, numeric_cols]

# Calculate MEAN MEDIAN MIN MAX SD of all columns that is calculable, Limit to 3 decimal
statistics <- sapply(numeric_data, function(x) {
  c(Mean = round(mean(x, na.rm = TRUE), digits = 3),
    Median = round(median(x, na.rm = TRUE), digits = 3),
    Minimum = round(min(x, na.rm = TRUE), digits = 3),
    Maximum = round(max(x, na.rm = TRUE), digits = 3),
    SD = round(sd(x, na.rm = TRUE), digits = 3))
})
```

Fig.5

For this CA, the dataset used contains crimes recorded throughout Brazil and is represented by nine types of criminal occurrences. By adding the column for a total of all incidents, we can see a broader statistic, comparing each individual category.

```
> statistics
                                  Mean Median Minimum Maximum       SD
Rape                           157.212   80.0       0    1306  197.899
Vehicle_Theft                  714.565  276.0       0   10518 1494.292
Homicide                       136.921   97.0       1     597  116.460
Bodily_injury_followed_by_death  2.417    1.0       0      39    3.561
Robbery_Institution              2.085    1.0       0      21    2.957
Cargo_Theft                     57.272    3.0       0    1329  173.094
Vehicle_Robbery                630.394  279.5       0    7970 1048.058
Robbery_Followed_by_Death        5.516    4.0       0      39    5.673
Attempted_Homicide             111.956   74.0       0     528  101.154
Total_Crimes                  1818.338  850.5       9   19918 2887.298
```

fig.6

Min-Max Normalization

Min-Max normalization is a typical statistical method for rescaling numerical values. These values are transformed to be between 0 and 1, with the minimum value being 0 and the maximum value being 1. [7]

As illustrated in Fig7, we are using the function normalized_MinMax to normalize the data and make sure that the values are on the same scale. After creating the function normalized_MinMax, we applied the normalization to every type of crime in our dataset.

```
# MinMax Normalization function
normalized_MinMax <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}

# MinMax Normalization of crimes
Rape_norm<-normalized_MinMax(dataReorganized$Rape)
Vehicle_Theft_norm<-normalized_MinMax(dataReorganized$Vehicle_Theft)
Homicide_norm<-normalized_MinMax(dataReorganized$Homicide)
Bodily_injury_followed_by_death_norm<-normalized_MinMax(dataReorganized$Bodily_injury_followed_by_death)
Robbery_Institution_norm<-normalized_MinMax(dataReorganized$Robbery_Institution)
Cargo_Theft_norm<-normalized_MinMax(dataReorganized$Cargo_Theft)
Vehicle_Robbery_norm<-normalized_MinMax(dataReorganized$Vehicle_Robbery)
Robbery_Followed_by_Death_norm<-normalized_MinMax(dataReorganized$Robbery_Followed_by_Death)
Attempted_Homicide_norm<-normalized_MinMax(dataReorganized$Attempted_Homicide)
```

fig.7

After applying normalization to all crimes, we normalized the values of homicide and bodily injury followed by death over the years. For that, we have transformed the year into a factor.

With the values normalized over the years, we created a plot using the ggplot2 library to compare the normalized rates of homicide and bodily injury followed by death over the time, in order to identify patterns or trends in the relationship between these two types of crimes that result in death, as can be seen in the figure bellow.

```
# Normalize rates over time (Homicide x Bodily_injury_followed_by_death)
dataReorganized <- dataReorganized %>%
  mutate(Year = as.factor(Year)) %>%
  group_by(Year) %>%
  mutate(
    Homicide_norm = normalized_MinMax(Homicide),
    Bodily_injury_followed_by_death_norm = normalized_MinMax(Bodily_injury_followed_by_death)
  ) %>%
  ungroup()

# Normalized plot - rates over time (Homicide x Bodily_injury_followed_by_death)
ggplot(dataReorganized, aes(x = Year, y = Homicide_norm, color = "Homicide")) +
  geom_point() +
  geom_point(aes(y = Bodily_injury_followed_by_death_norm, color = "Bodily Injury followed by death")) +
  labs(title = "Correlation Over Time",
       x = "Year",
       y = "Normalized Rates",
       color = "Crime") +
  scale_color_manual(values = c("Homicide" = "red", "Bodily Injury followed by death" = "blue"))
```

Z-score Standardization

The Z-Score is a numerical measure that describes the relationship between a value and the mean of a group of values, which can be positive or negative. When the Z-score is positive, the value is above the mean of the set of values, but when the Z-score is negative, the value is below the mean of the set of values. [8]

As shown in figure below, we are using the function standardize_zscore to calculate the Z-Score and we are applying it to every crime in our dataset in dataZscaled.

```
# Calculate Z-score
standardize_zscore <- function(x) {
  return ((x - mean(x, na.rm = TRUE)) / sd(x, na.rm = TRUE))
}

# Z-scores of crimes
dataZscaled <- dataReorganized %>%
  mutate(
    Rape_zscaled = standardize_zscore(Rape),
    Vehicle_Theft_zscaled = standardize_zscore(Vehicle_Theft),
    Homicide_zscaled = standardize_zscore(Homicide),
    Bodily_injury_followed_by_death_zscaled = standardize_zscore(Bodily_injury_followed_by_death),
    Robbery_Institution_zscaled = standardize_zscore(Robbery_Institution),
    Cargo_Theft_zscaled = standardize_zscore(Cargo_Theft),
    Vehicle_Robbery_zscaled = standardize_zscore(Vehicle_Robbery),
    Robbery_Followed_by_Death_zscaled = standardize_zscore(Robbery_Followed_by_Death),
    Attempted_Homicide_zscaled = standardize_zscore(Attempted_Homicide)
  ) %>%
  select(State, Year, Month, Rape_zscaled, Vehicle_Theft_zscaled, Homicide_zscaled, Bodily_injury_followed_by_death_zscaled, Robbery_Institution_zs

View(dataZscaled)
```

fig.9

In the next figure we can see the result of the Z-Score Standardization after using “view(dataZscaled)”.

Filter

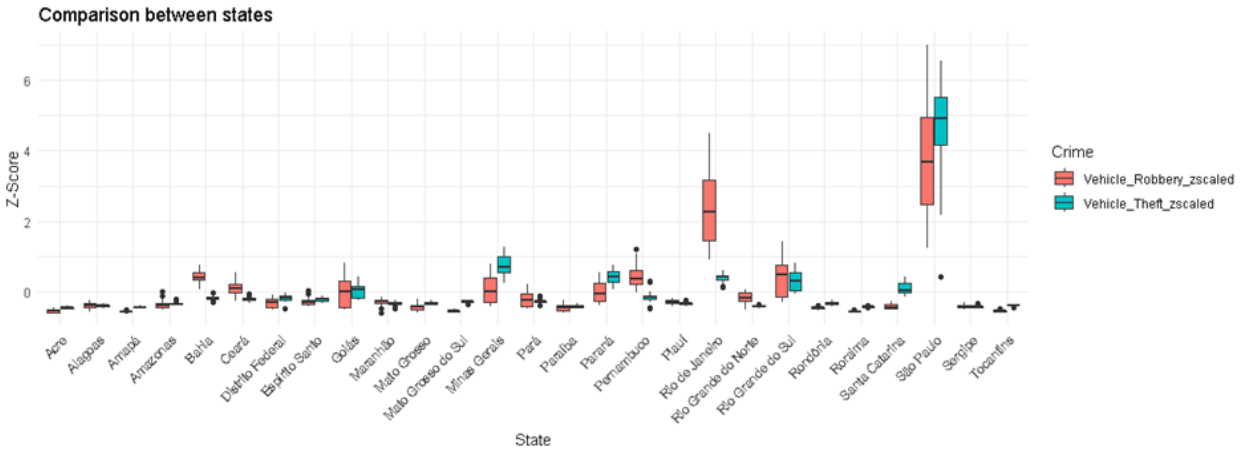
	State	Year	Month	Rape_zscaled	Vehicle_Theft_zscaled	Homicide_zscaled	Bodily_injury_followed_by_death_zscaled
1	Acre	2022	janeiro	-0.637758890	-0.44473571	-1.089825537	-0.3980657
2	Acre	2022	fevereiro	-0.622599638	-0.44138964	-1.089825537	-0.6789003
3	Acre	2022	março	-0.506378699	-0.44875099	-0.995372378	-0.6789003
4	Acre	2022	abril	-0.652918143	-0.45142785	-0.995372378	-0.6789003
5	Acre	2022	maio	-0.567015710	-0.45878919	-0.978199077	-0.6789003
6	Acre	2022	junho	-0.567015710	-0.44473571	-1.072652235	-0.6789003
7	Acre	2022	julho	-0.481113278	-0.44875099	-1.003959029	-0.6789003
8	Acre	2022	agosto	-0.556909542	-0.45544313	-1.098412188	-0.6789003
9	Acre	2022	setembro	-0.496272531	-0.45611234	-1.038305632	-0.6789003
10	Acre	2022	outubro	-0.546803373	-0.45075863	-1.055478934	-0.6789003
11	Acre	2022	novembro	-0.526591036	-0.45075863	-0.926679172	-0.1172311
12	Acre	2022	dezembro	-0.501325615	-0.44339728	-0.986785727	-0.3980657
13	Alagoas	2022	janeiro	-0.506378699	-0.37112223	-0.299853664	-0.6789003

After calculating the Z-Score, we reorganized the data using the dataMelt function to make the process simpler to generate a plot. The resulting data set is longer and simpler for analyzing and, for this example, the crimes vehicle theft and vehicle robbery were analyzed.

```
# Melt data
dataMelt <- dataZscaled %>%
  select(State, Year, Month, Rape_zscaled, Vehicle_Theft_zscaled, Homicide_zscaled, Bodily_injury_followed_by_death_zscaled, Robbery_Institution
  pivot_longer(cols = c(Vehicle_Theft_zscaled, Vehicle_Robbery_zscaled),
               names_to = "Crime", values_to = "Z_Score")

View(dataMelt)

# Plot - comparison between states (Vehicle_Theft_zscaled x Bodily_injury_followed_by_death_zscaled)
ggplot(dataMelt, aes(x = State, y = Z_Score, fill = Crime)) +
  geom_boxplot() +
  labs(title = "Comparison between states",
       x = "State",
       y = "Z-Score",
       fill = "Crime") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



As seen in the plot above, the most significant variations from the average are in the values of vehicle robbery in the states of Rio de Janeiro and São Paulo, as well as in the value of vehicle theft in the state of São Paulo

Robust scalar

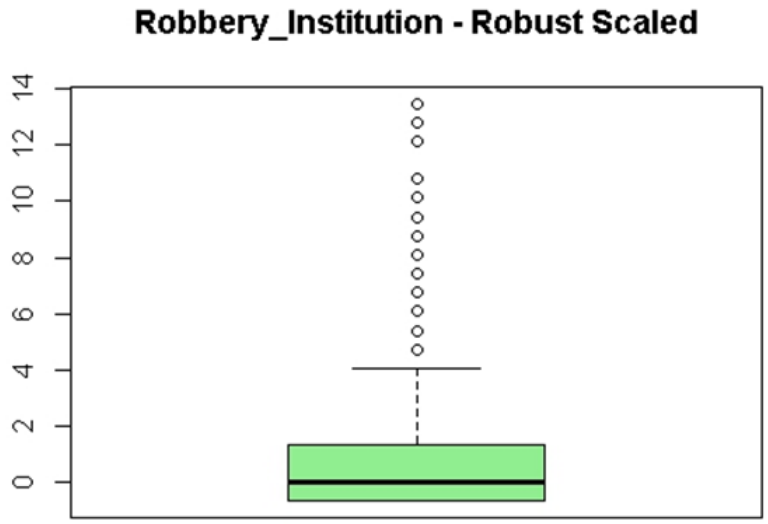
Another data normalization technique is Robust Scaler. It is used to identify outliers in the dataset.

As shown in the figure below, after installing the necessary package, we applied Robust Scaler to every crime in our dataset and created a boxplot plot to check for outliers in the Robbery Institution values.

```r
# Robust Scaler
install.packages("robustbase")
library(robustbase)

# Normalization with Robust Scaler
Rape_robust <- (dataReorganized$Rape - median(dataReorganized$Rape)) / mad(dataReorganized$Rape)
Vehicle_Theft_robust <- (dataReorganized$Vehicle_Theft - median(dataReorganized$Vehicle_Theft)) / mad(dataReorganized$Vehicle_Theft)
Homicide_robust <- (dataReorganized$Homicide - median(dataReorganized$Homicide)) / mad(dataReorganized$Homicide)
Bodily_injury_followed_by_death_robust <- (dataReorganized$Bodily_injury_followed_by_death - median(dataReorganized$Bodily_injury_fol
Robbery_Institution_robust <- (dataReorganized$Robbery_Institution - median(dataReorganized$Robbery_Institution)) / mad(dataReorganiz
Cargo_Theft_robust <- (dataReorganized$Cargo_Theft - median(dataReorganized$Cargo_Theft)) / mad(dataReorganized$Cargo_Theft)
Vehicle_Robbery_robust <- (dataReorganized$Vehicle_Robbery - median(dataReorganized$Vehicle_Robbery)) / mad(dataReorganized$Vehicle_R
Robbery_Followed_by_Death_robust <- (dataReorganized$Robbery_Followed_by_Death - median(dataReorganized$Robbery_Followed_by_Death)) /
Attempted_Homicide_robust <- (dataReorganized$Attempted_Homicide - median(dataReorganized$Attempted_Homicide)) / mad(dataReorganized$

# Boxplot showing outliers
boxplot(Robbery_Institution_robust, main="Robbery_Institution - Robust Scaled", col="lightgreen", border="black")
```



As we can see in the figure above, most of the values for this type of crime are between 0 and 4 monthly, but we have outliers which the values go over that.

Dimensionality Reduction

Dimensionality reduction is a process in which the number of variables or dimensions in a data set is reduced while keeping as much important information as possible. This is done to simplify data analysis and processing, with the aim of:

Simplification and Computational Efficiency:
Reducing Complexity makes it easier to understand a dataset with many variables. With the reduction in size, complexity is simplified, making it more understandable and

interpretable for analysis. Fewer variables also imply fewer computational resources needed for processing, which is advantageous for large-scale analyses.

Redundancy and Noise Removal:

Not all variables contribute equally to understanding the dataset, so removing redundancy is a plausible process. Dimensionality reduction helps eliminate irrelevant variables, reducing noise and focusing on the most important information.

Enhanced View:

Reducing dimensionality makes it much easier to visualize data in two- or three-dimensional graphs, which can help identify patterns and relationships that might not be visible in higher dimensions.

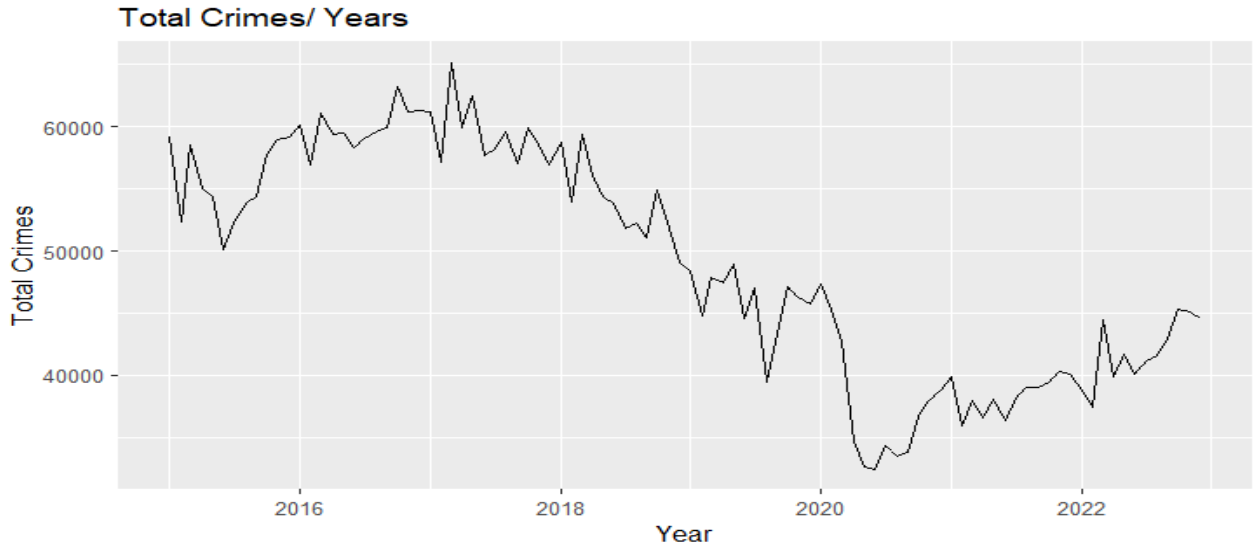
Facilitate Modeling and Understanding:

Reducing the number of variables can also make it easier to build statistical or machine learning models, as the data becomes more understandable and interpretable. Models trained on lower-dimensional data tend to be faster and require less data for training.

Plots Analysis

Line Plot

For an initial presentation, we developed a line graph based on total crimes per year. In a preliminary analysis, we observed a downward trend in the total number of criminal incidents in Brazil from 2015 to 2022.

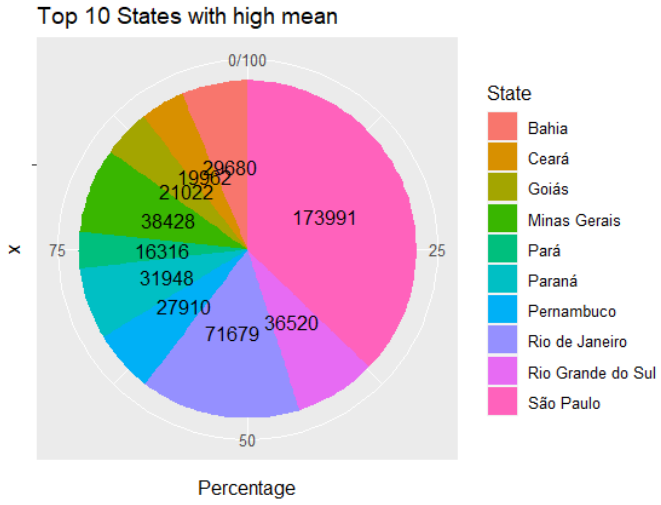


```r
# Aggregating total crimes by date
total_crimes_by_date <- dataReorganized %>%
  group_by(Date) %>%
  summarise(Total_Crimes = sum(Total_Crimes, na.rm = TRUE))

# Graph
ggplot(total_crimes_by_date, aes(x = Date, y = Total_Crimes)) +
  geom_line() +
  labs(title = "Total Crimes vs Year", x = "Date", y = "Total Crimes")
```

Pie chart

For the pie chart, we chose to use the average of total crimes, highlighting the 10 states with the highest average of total crimes. This average is calculated considering the years 2015 to 2022 and the data is separated by states.



This pie chart helps you visualize and present percentages. By separating the top 10 and presenting it in a pie chart, the individual proportion of each state in relation to the entire selected group becomes more visible.

Scatter Plot

We used a scatter plot to examine the relationship between two types of crimes. To accomplish so, we use normalized data to make comparisons between different types of

crime simpler, as the values are transformed to be between 0 and 1.
For this example, we explored the correlation between vehicle theft and vehicle robbery.

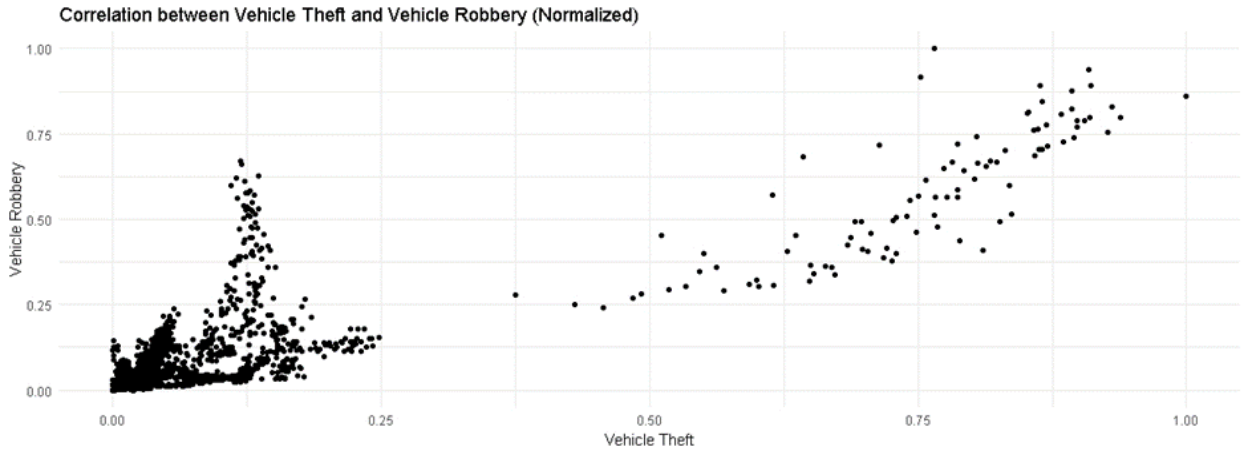
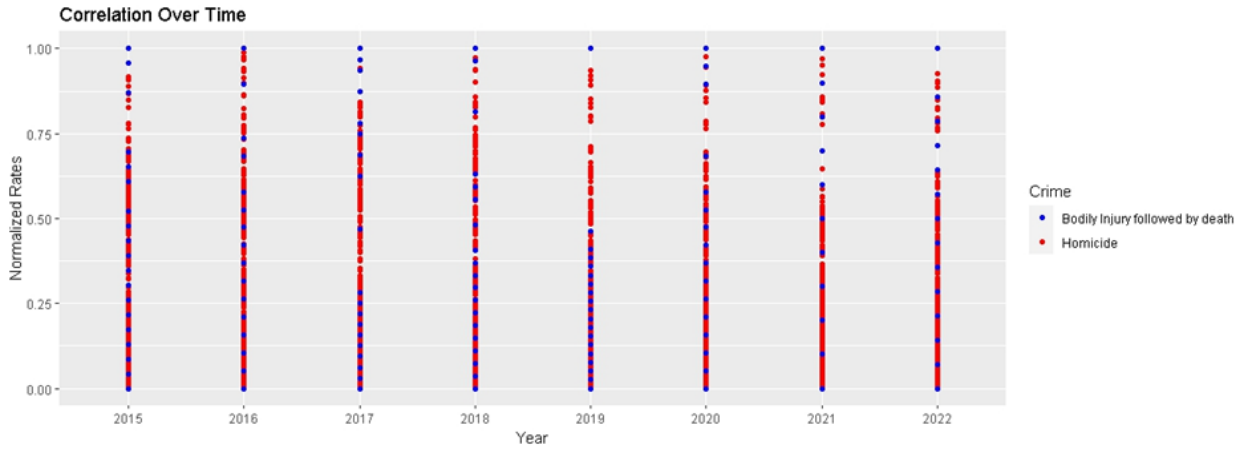


Fig.14

As shown in the plot above, despite having two distinct groups, the two groups have a positive correlation, implying that there is a consistent relationship between the variables in both groups. This means that, within each group, if one variable increases, another tends to increase as well, which could imply that as vehicle thefts increase, the number of vehicle robberies increases as well.

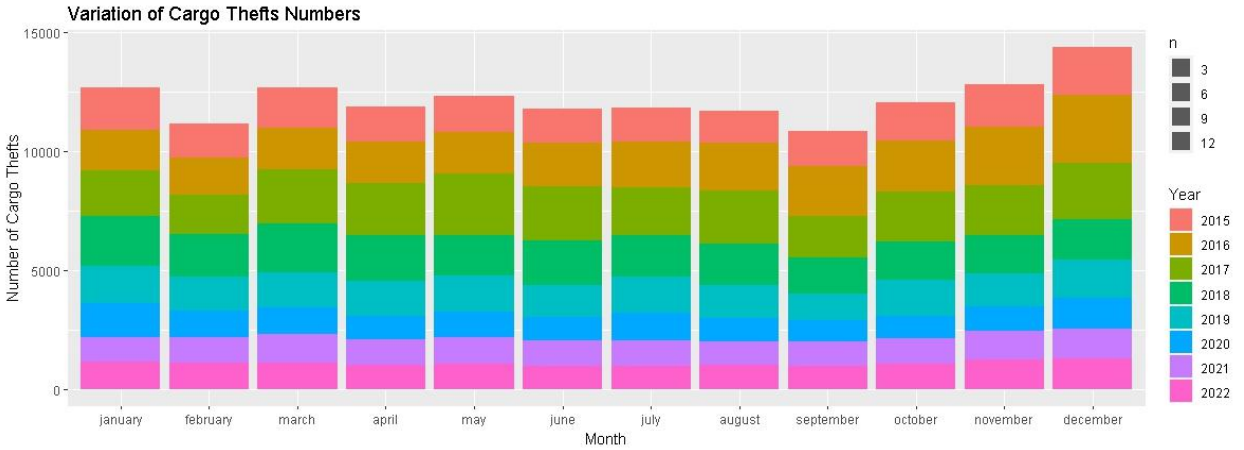
We also used normalized values to create a scatter plot to show the variation in normalized rates over time for two types of crime. For this example, the types of crimes used were bodily injury followed by death and homicide. There are two different colors for the dots in the graphic to distinguish them, a blue one representing bodily injury followed by death and the red one representing homicide.



We can see that the proportion of homicides during the years is much higher in contrast to bodily injury followed by death, with a higher incidence of red dots on the plot. Also, we can see that the rates of bodily injury followed by death are lower in general, being mostly under 0.50, with a peak in 2019, whereas the rates of homicide are more evenly distributed from 0 to 1

Bar Plot

We used a bar plot to analyze the variation in the total number of cargo thefts in each month of the year between 2015 and 2022. Each bar represents a month of the year, and the sections within each bar reflect the proportion of cargo thefts occurring in that particular year.

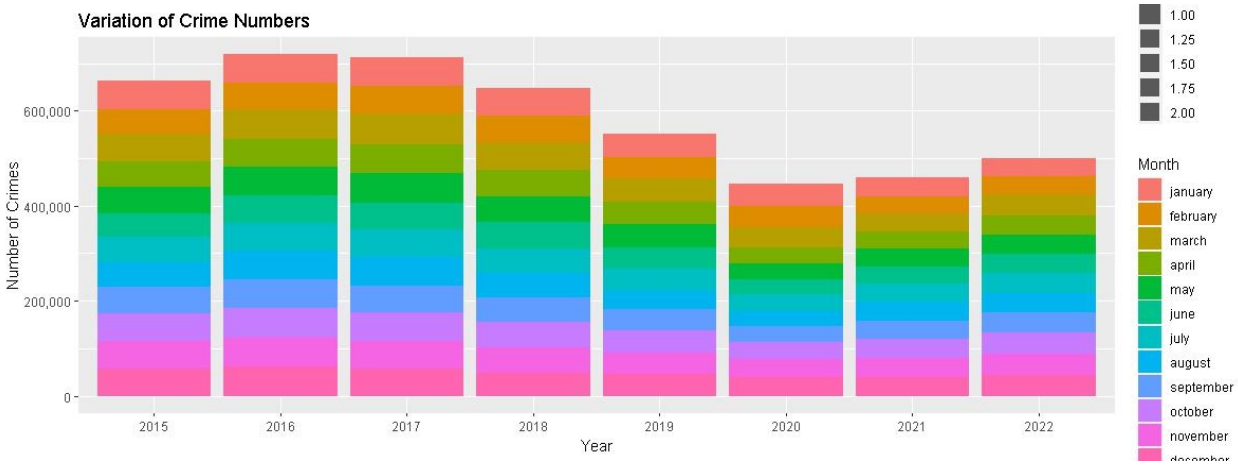


We can see the variation in the number of cargo thefts over time. The month with the most incidents of cargo theft in the analyzed period was December. In addition, each month had a higher number of occurrences than 10000.

We had a higher and more consistent rate of cargo theft in 2019, with the years 2016 and 2017 having the most influence on the total, but we can see a decrease in these numbers beginning in 2020, which may be related to the Covid-19 pandemic.

In the next bar plot, instead of analyzing a specific crime, we will examine the total number of crimes committed over the years. However, the bars now represent the years,

and the sections inside each bar represent the fraction of criminal incidents in a particular month of the year.



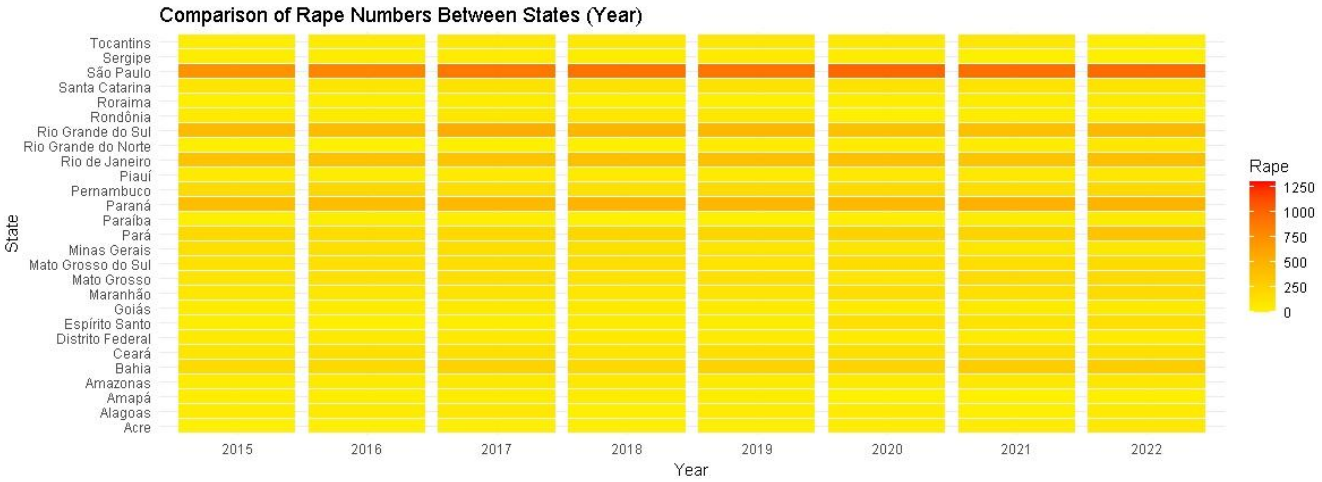
In this example, we can see that the years 2016 and 2017 had the highest number of crimes committed throughout the time frame under study, as well as having the greatest influence on the number of cargo theft incidents.

As of 2020, there will be a noticeable change in the number of crimes committed. However, following the drop in numbers in 2020 and 2021, when the total fell below the 500000 mark for the first time, the number of offenses increased in 2022 and surpassed the 500000 mark once more. This demonstrates that the Covid-19 pandemic had a significant impact on the numbers.

HeatMap

HeatMaps were used to visualize patterns and trends in our dataset. For the next example, we will look at the pattern of the number of rape occurrences by state, between 2015 and 2022.

The plot highlights areas with higher rape rates with darker colors, whereas areas with lower rape rates have lighter colors. This makes it easier to identify variations in the data throughout time and between the states.



The states where the colours are yellow are the ones with lower rape incidents, while the ones which the colour gets closer to red are the ones with more incidents. It is evident that the state of SãoPaulo stands out from the other states. One reason that could explain this difference is that the SãoPaulo state is the most densely urban area in Brazil. We can see that states with a larger rural area, such as Amazonas and Acre, have more yellow colors, indicating a lower incidence of this crime.

We also can see that the colour in SãoPaulo is getting darker during the years, which means that in this case, the Covid-19 pandemic did not have influence to make the numbers of rape crimes go down. They are actually going up. This could point to a conclusion that most rape crimes might happen inside the household enviroment.

The states of Rio Grande do Sul, Rio de Janeiro, Paraná and Bahia, which are the ones with coloured or close to orange, are the ones closer to the middle point.

Dummy Encoding

Dummy encoding is a method of representing categorical variables in a dataset as binary variables. It creates a column with value one when a category event happens and a zero when it does not. [Dummy]

We used Dummy Encoding on the variable "State" in our dataset, as seen in the piece of code below.

```r
# Create Dummy Encoding for State
install.packages('fastDummies')
library('fastDummies')

dataReorganized <- dummy_cols(dataReorganized, select_columns = 'State')

View(dataReorganized)
```

The above piece of code is first installing the fastDummies package. Afterwards, the function dummy_cols is being executed to apply dummy coding in the variable "State" in dataReorganized. Binary columns, one for each state, are being added to the dataset to indicate the presence or absence of each state.

We can view the update dataset with the new columns utilizing View(dataReorganized), and the result is shown in the figure below.

State_Acre	State_Alagoas	State_Amapá	State_Amazonas	State_Bahia	State_Ceará	State_Distrito Federal	State_Espírito Santo	State_Goiás	Stat
1	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	

Each state now has a column starting with “State_” and the name of the State, and the values are 0 or 1. When the State appears the value will be one, as we can see in the column “State_Acre”, which is the first state to appear in the dataset, while the others have the value of 0, until they appear later in the dataset.

Charles' journal

I took on leadership responsibility and began coordinating the start of the AC. Firstly, I took responsibility for creating the repository on GitHub and the document on Google Docs for the report, in addition to searching for the dataset that we would use. After some conversations with Thiago, we came to the conclusion about which dataset to use and how to divide the tasks between us. As Thiago was involved in another project, we decided that I would take on the initial part. I started the file in R, read the CSV and took the first steps to process the files, aiming to apply and analyze the assigned tasks.

I was responsible for reformulating the data, as the original file had main information organized by year, month and state. I chose to transform each type of crime into a column, represented by its occurrences. This made the dataset clearer and less abstract, thus already identifying some data as N.A. values that we would have to solve.

The next step was to translate the information into English, since the dataset referred to Brazil and was in Portuguese. To deal with N.A. values, we discuss the best approach for the work. Initially, I considered N.A. as 0, but Thiago emphasized the need for a more in-depth approach. We came to the conclusion of calculating the averages for each state, year and month, using these averages to replace the N.A.

I was also responsible for working with Statistical Parameters, both in the calculations and in the presentation of line and pie charts. To do this, I found it necessary to include a 'Total_crimes' column, representing the sum of all crimes divided by state, month and year.

One last, but not least, responsibility I took on was reviewing and accepting Push Requests on GitHub. Thiago and I were working on different branches to avoid possible crashes. This is a responsibility I enjoy carrying out.

Thiago's journal

For this Data Exploration & Preparation assignment, we chose to work on a dataset about crime occurrences in Brazil in the period of time between 2015 and 2022. The dataset is called Crimes_in_Brazil_2015_2022.csv. This reflective journal intends to reflect on the process of doing the CA, working as a team, and discussing the obstacles encountered.

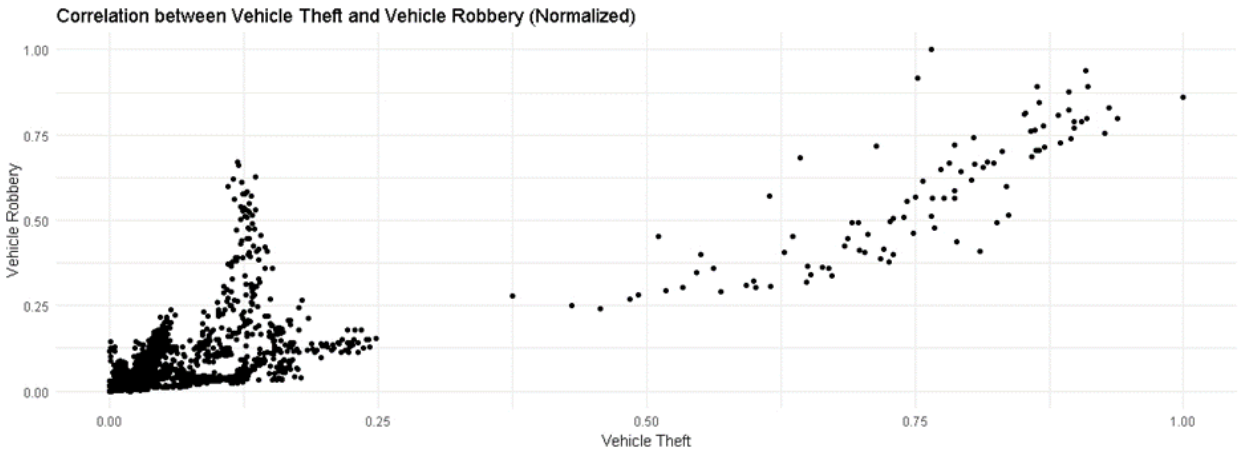
At the beginning of the preparation for this CA, me and Charles met at the library at the College campus to discuss the assignment. We discussed how we would treat the dataset, how we would be communicating while doing the CA and we divided some tasks. It was a good meeting because Charles could help me to understand some parts that I did not understand very well at first. Here I will quickly talk about the tasks that I was responsible for.

```
# MinMax Normalization function
normalized_MinMax <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}

# MinMax Normalization of crimes
Rape_norm<-normalized_MinMax(dataReorganized$Rape)
Vehicle_Theft_norm<-normalized_MinMax(dataReorganized$Vehicle_Theft)
Homicide_norm<-normalized_MinMax(dataReorganized$Homicide)
Bodily_injury_followed_by_death_norm<-normalized_MinMax(dataReorganized$Bodily_injury_followed_by_death)
Robbery_Institution_norm<-normalized_MinMax(dataReorganized$Robbery_Institution)
Cargo_Theft_norm<-normalized_MinMax(dataReorganized$Cargo_Theft)
Vehicle_Robbery_norm<-normalized_MinMax(dataReorganized$Vehicle_Robbery)
Robbery_Followed_by_Death_norm<-normalized_MinMax(dataReorganized$Robbery_Followed_by_Death)
Attempted_Homicide_norm<-normalized_MinMax(dataReorganized$Attempted_Homicide)
```

After Charles reorganized the dataset, because it was confusing to read and understand as we downloaded from the internet, I applied the Min_max Normalization to the numerical values. To achieve that, I wrote the following piece of code:

With normalized values, I could create Scatter Plots to analyze the correlation between different types of crimes in a simpler way to understand, like in the following plot:



I did not find many correlations between the different crimes in the dataset.

After doing the Mn_max Normalization, I applied Z-Score Standardization to the numerical variables in the dataset, as can be seen in the written piece of code bellow:

```
# Calculate Z-score
standardize_zscore <- function(x) {
  return ((x - mean(x, na.rm = TRUE)) / sd(x, na.rm = TRUE))
}

# Z-scores of crimes
dataZscaled <- dataReorganized %>%
  mutate(
    Rape_zscaled = standardize_zscore(Rape),
    Vehicle_Theft_zscaled = standardize_zscore(Vehicle_Theft),
    Homicide_zscaled = standardize_zscore(Homicide),
    Bodily_injury_followed_by_death_zscaled = standardize_zscore(Bodily_injury_followed_by_death),
    Robbery_Institution_zscaled = standardize_zscore(Robbery_Institution),
    Cargo_Theft_zscaled = standardize_zscore(Cargo_Theft),
    Vehicle_Robbery_zscaled = standardize_zscore(Vehicle_Robbery),
    Robbery_Followed_by_Death_zscaled = standardize_zscore(Robbery_Followed_by_Death),
    Attempted_Homicide_zscaled = standardize_zscore(Attempted_Homicide)
  ) %>%
  select(State, Year, Month, Rape_zscaled, Vehicle_Theft_zscaled, Homicide_zscaled, Bodily_injury_followed_by_death_zscaled, Robbery_Institution_zs

View(dataZscaled)
```

After applying Z-Score I had to melt the data to facilitate the creation of a boxplot, to show the states that have more variations from the average when analyzing the numbers of vehicle robbery and vehicle theft.

I also wrote a code for applying Robust Scaler. After applying it to every crime, I generated a boxplot to verify the outliers present in the numbers of the variable Robbery_Institution.

Furthermore, I was responsible for creating some plots and doing some analysis about the data. I wrote the code to generate some bar plots to see the variation on the numbers of crimes between the years or between months.Another

```
# MONTHLY NUMBERS OF A CRIMES OVER THE YEARS (NUMBERS PER MONTH IN DIFFERENT YEARS)
# Specific crime (To view other types of crimes, just change the variable y)
ggplot(dataReorganized, aes(x = Month, y = Cargo_Theft, fill = as.factor(Year))) +
  geom_bar(stat = "sum") +
  labs(title = "Variation of Cargo Thefts Numbers",
       x = "Month",
       y = "Number of Cargo Thefts",
       fill = "Year") +
  scale_x_discrete(labels = translation_dict)
scale_y_continuous(labels = scales::comma_format())

# Total number of crimes
ggplot(dataReorganized, aes(x = Month, y = Total_Crimes, fill = as.factor(Year))) +
  geom_bar(stat = "sum") +
  labs(title = "Variation of Crime Numbers",
       x = "Month",
       y = "Number of Crimes",
       fill = "Year") +
  scale_x_discrete(labels = translation_dict) +
  scale_y_continuous(labels = scales::comma_format())
```

kind of plot I created was the HeatMaps, so I could visualize patterns in the dataset, such as how the rate of crimes is behaving from state to state over the time. Bellow is a part of the code for the HeatMaps:

I also applied Dummy Encoding to a categorical variable in the dataset. I chose the variable “State” to do the Dummy Encoding using the FastDummies package. Additional columns were created in the dataset for each state, now showing just binary values. Here is the code used:

The biggest challenge that I found while doing this CA was when I tried to generate some other types of plots, like line plots, for example. I could not do it in the right way, since sometimes the plots being generated were not making sense to me, so I was probably not coding right.

References:

1. Governo Digital. (n.d.). Do Eletrônico ao Digital. [online] Available at: https://www.gov.br/governodigital/pt-br/estrategia-de-governanca-digital/do-eletronico-ao-digital.
2. Corporate Finance Institute. (n.d.). Parameter. [online] Available at: https://corporatefinanceinstitute.com/resources/data-science/parameter/.
3. Simplilearn (2021). What is Dimensionality Reduction? Overview, and Popular Techniques. [online] Simplilearn.com. Available at: https://www.simplilearn.com/what-is-dimensionality-reduction-article.
4. Pramoditha, R. (2021). 11 Dimensionality reduction techniques you should know in 2021. [online] Medium. Available at: https://towardsdatascience.com/11-dimensionality-reduction-techniques-you-should-know-in-2021-dcb9500d388b.
5. Analytics Vidhya (2018). Comprehensive Guide to 12 Dimensionality Reduction Techniques. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2018/08/dimensionality-reduction-techniques-python/.
6. Bhalla, D. (n.d.). Dimensionality Reduction with R. [online] ListenData. Available at: https://www.listendata.com/2015/06/simplest-dimensionality-reduction-with-r.html [Accessed 3 Dec. 2023].
7. Codecademy. (n.d.). Normalization. [online] Available at: https://www.codecademy.com/article/normalization
8. Mais Retorno. (2022). Z-Score. [online] Available at: https://maisretorno.com/portal/termos/z/z-score [Accessed 3 Dec. 2023].
9. cran.r-project.org. (n.d.). Making dummy variables with dummy_cols(). [online] Available at: https://cran.r-project.org/web/packages/fastDummies/vignettes/making-dummy-variables.html [Accessed 3 Dec. 2023].

GitHub

https://github.com/Charlesjahn/CA1