CA1 Project

Lecturer: Dr. Muhammad Iqbal

CCT COLLEGE

STUDENT NAME: MARCO DOS SANTOS
STUDENT NUMBER: 2020333

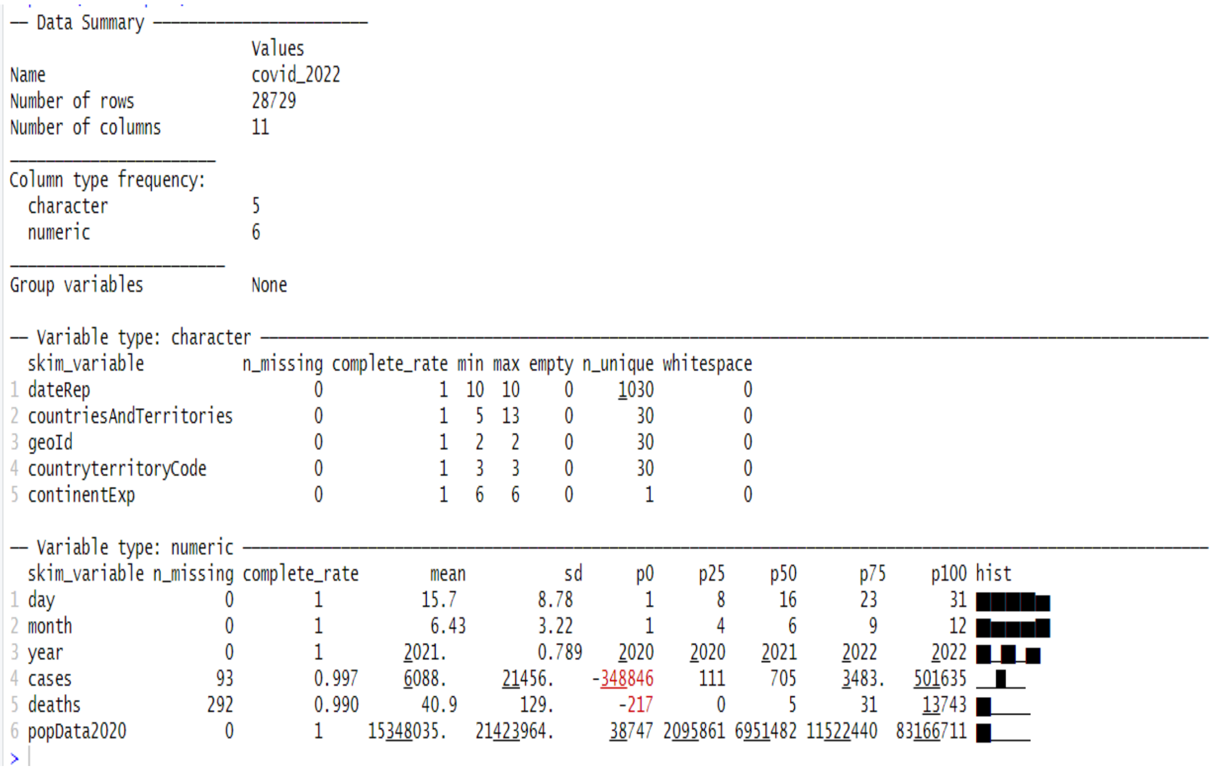
Summary

Introduction

This report, based on the R language provided by Professor Dr. Muhammad Iqbal, leverages Exploratory Data Analysis (EDA) on a dataset obtained from the European Centre for Disease Prevention and Control. Encompassing categorical, discrete, and continuous variables, the analysis uses diverse graphical representations, statistical analysis and dimensionality reduction techniques to reveal intricate patterns and characteristics inherent in the data set.

A- Identify which variables are categorical, discrete and continuous.

Skimr in R language streamlines data analysis with a concise presentation: Data Overview describes the structure; Column types highlight variable types; Character summaries and numeric variables provide important statistics aiding quick interpretation and insights into data distribution. Below you will see the figure that represents all the information provided by the Covid_2022.csv dataset:

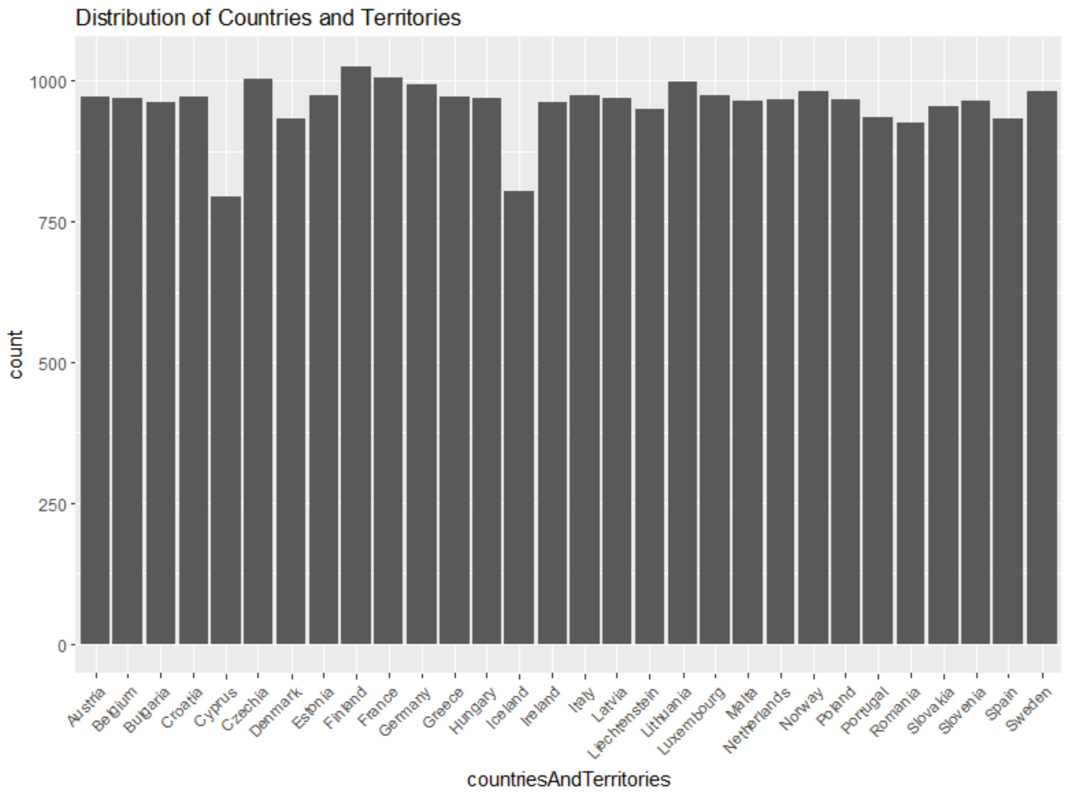


```
— Data Summary ————————————————————————
                           Values
Name                       covid_2022
Number of rows             28729
Number of columns          11
_______________________
Column type frequency:
  character                5
  numeric                  6
________________________
Group variables            None

— Variable type: character ——————————————————————————————————————————————————————————————————————————————————————
  skim_variable          n_missing complete_rate min max empty n_unique whitespace
1 dateRep                        0             1  10  10     0     1030          0
2 countriesAndTerritories        0             1   5  13     0       30          0
3 geoId                          0             1   2   2     0       30          0
4 countryterritoryCode           0             1   3   3     0       30          0
5 continentExp                   0             1   6   6     0        1          0

— Variable type: numeric ————————————————————————————————————————————————————————————————————————————————————————
  skim_variable n_missing complete_rate       mean         sd      p0     p25     p50      p75     p100 hist
1 day                   0         1           15.7       8.78       1       8      16       23       31 ▇▇▇▇▆
2 month                 0         1           6.43       3.22       1       4       6        9       12 ▇▆▆▆▇
3 year                  0         1          2021.      0.789    2020    2020    2021     2022     2022 ▇▁▇▁▇
4 cases                93         0.997      6088.     21456.  -348846     111     705    3483.   501635 ▁▇▁▁▁
5 deaths              292         0.990       40.9       129.    -217       0       5       31    13743 ▇▁▁▁▁
6 popData2020           0         1     15348035.  21423964.    38747 2095861 6951482 11522440 83166711 ▇▁▁▁▁
>
```

Categorical Variables

Categorical variables are often used to classify data into groups or classes based on some qualitative property or attribute. According to research (Frost,n.d.)“*A categorical variable has values that you can put into a countable number of distinct groups based on a characteristic. For a categorical variable, you can assign categories but the categories have no natural order.”*

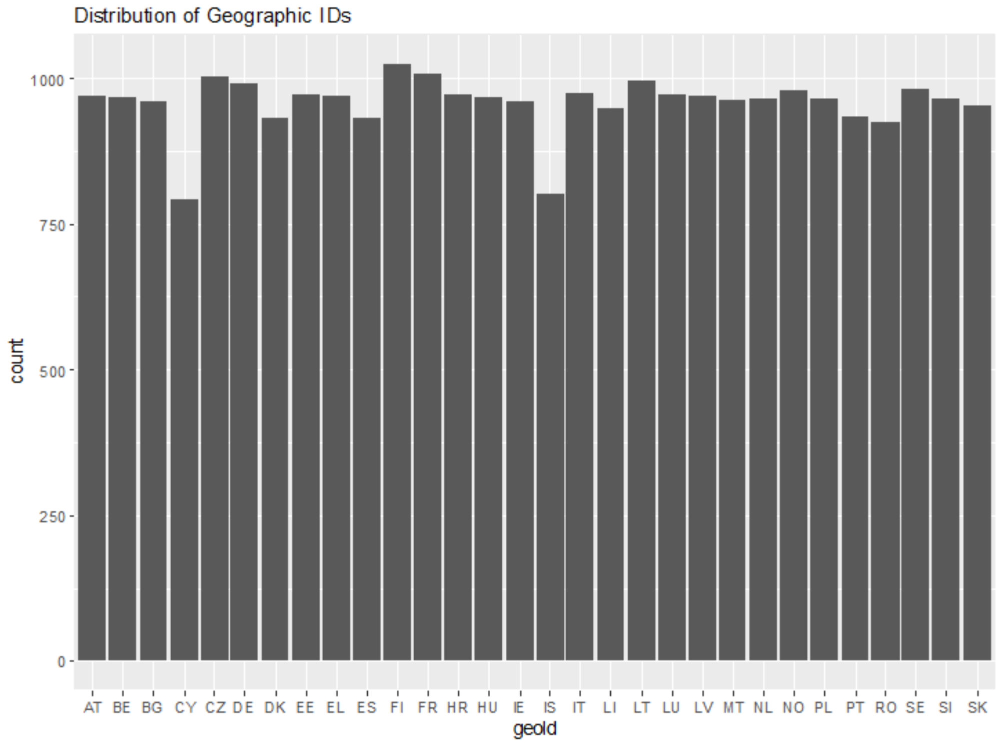
Based on the previous statement, some graphs were built using ggplot in R language where it is possible to show some categorical variables found in the covid_2022 dataset. below you can see the Distribution of Countries and Territories graph:



In the previous graph where the Distribution of Countries and Territories is represented, the sample value varies in size, ranging from 0 to approximately 1000.

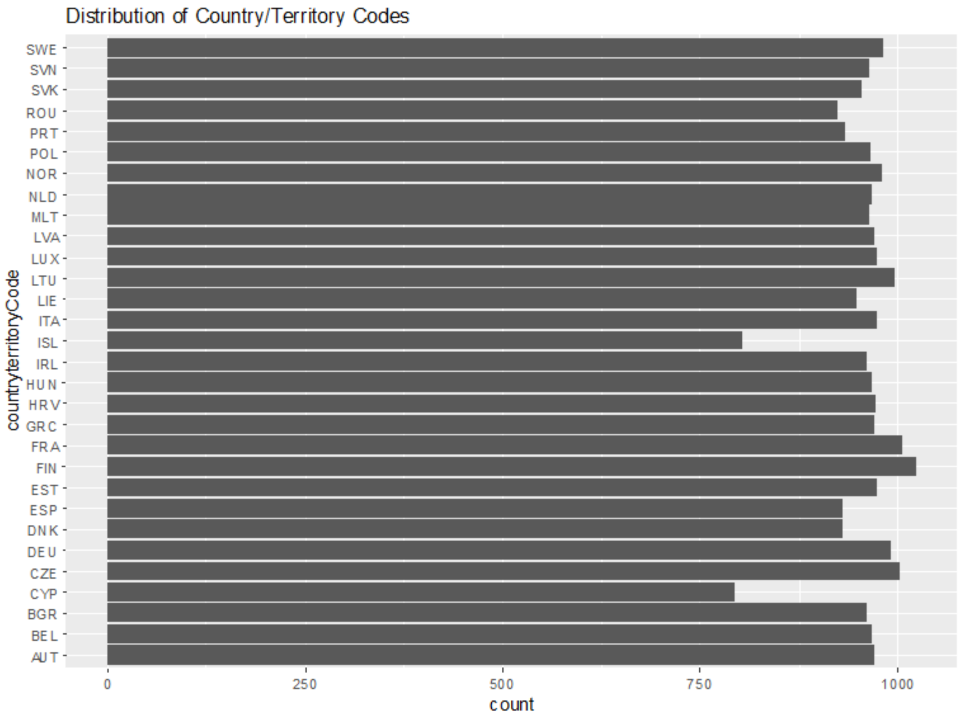
In the graph, there is a significant difference in size between countries and territories. The taller bar represents the largest country or territory, while the shorter bar represents the smallest.

Next graph Distribution of geographic IDs follows the same premises of previous graph, where the sample value varies in size, ranging from 0 to approximately 1000. Below you can see the Distribution of Geographic IDs graph:

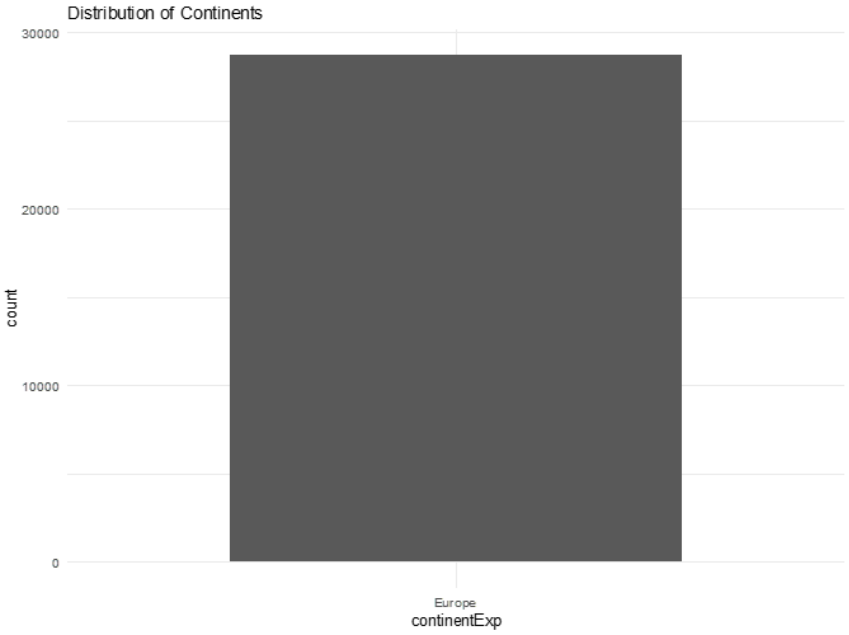


The graph displays a Distribution of geographic locations, with each location represented by a bar. it also provides a visual representation of the distribution of geographic locations, allowing for quick comparison between regions.

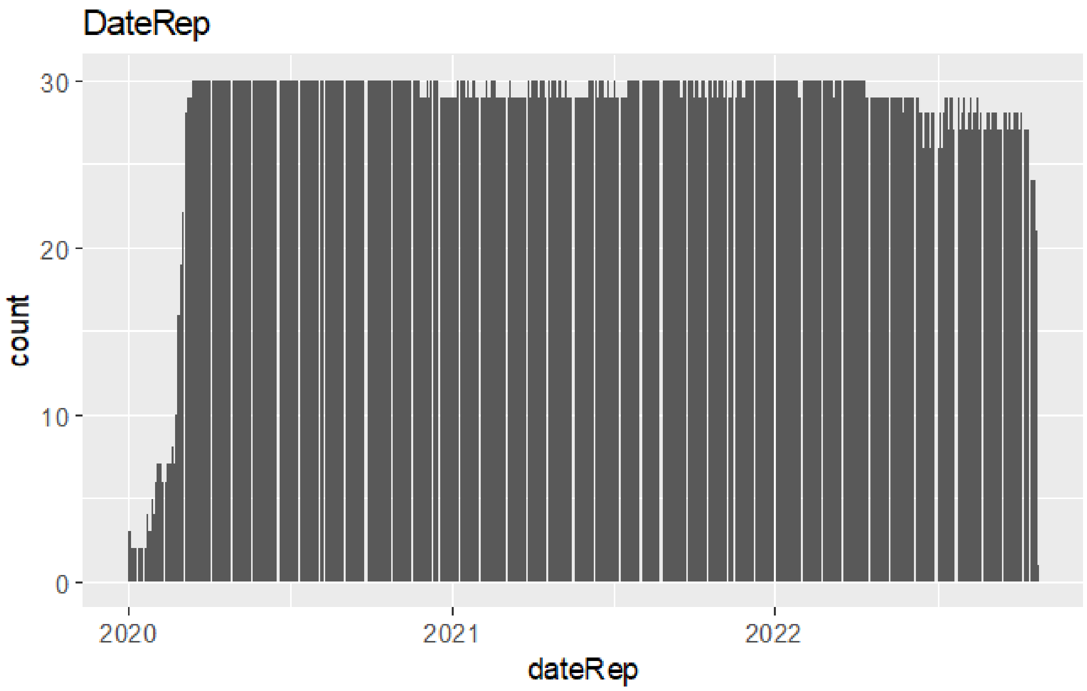
Next graph displays Distribution of country/territory Codes, For better visualization The bars are arranged in a vertical line, with the height of each bar corresponding to the number of territories in the country. Also the sample value varies in size, ranging from 0 to approximately 1000. Below you can see the Distribution of Country/territory Code:



The following graph shows the distribution of continents, as all countries are located in Europe, the best way to represent them was by the histogram graph. Also the sample value varies in size, ranging from 0 to approximately 3000. below you can see the Distribution of Continents graph:



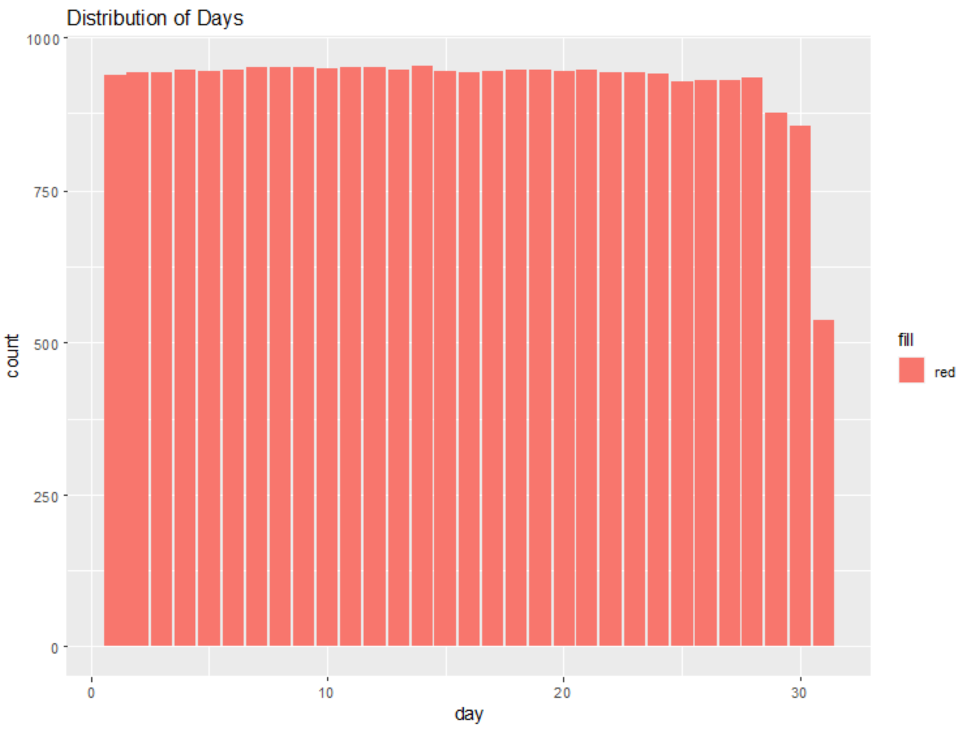
The following DateRep chart shows the distribution of data between 2020 and 2022. Furthermore, the sample value varies in size, ranging from 0 to approximately 30 . Below you can see the graphical representation:



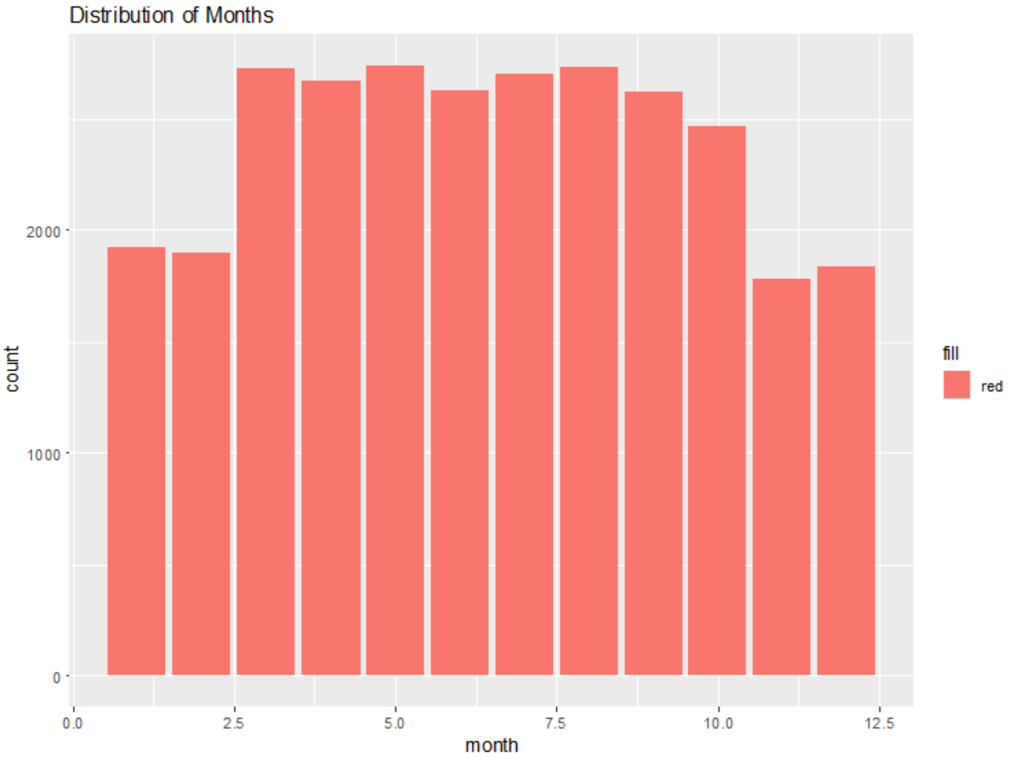
Discrete variables

According to Muhmmad (2022) "*In other words, the variable can only take on certain specific values and not any values within a range*."These values are typically countable and finite, or countably infinite. Discrete variables differ from continuous variables, which can take on any value within a range.

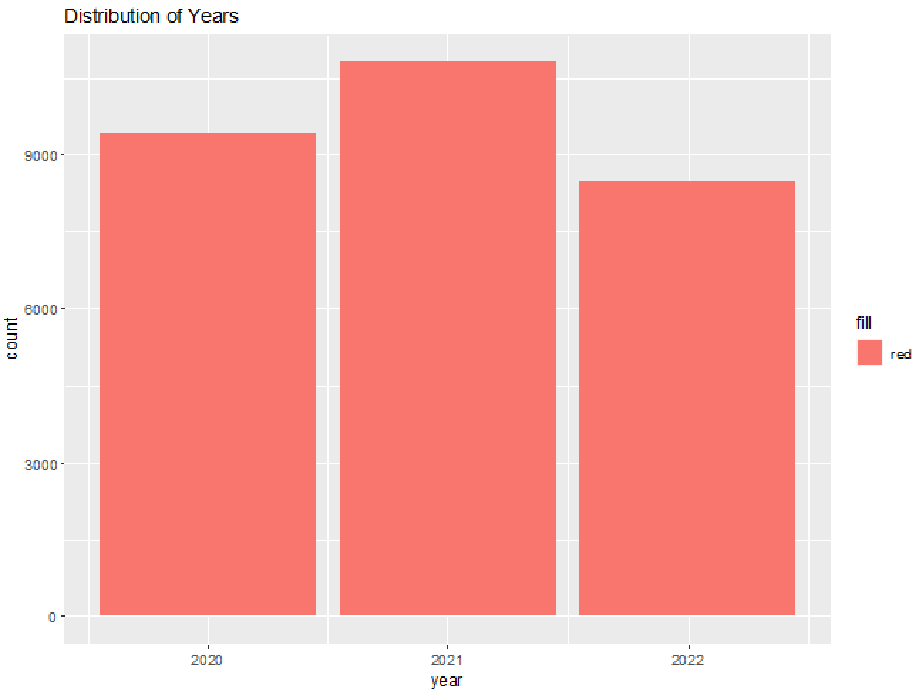
The following graph show the Distribution of Days with total the 30 days and the sample value varies in size, ranging from 0 to approximately 30. What can be observed is that most days have a high average value. however, towards the end of the 30 days, the values drop a little as you can see below:



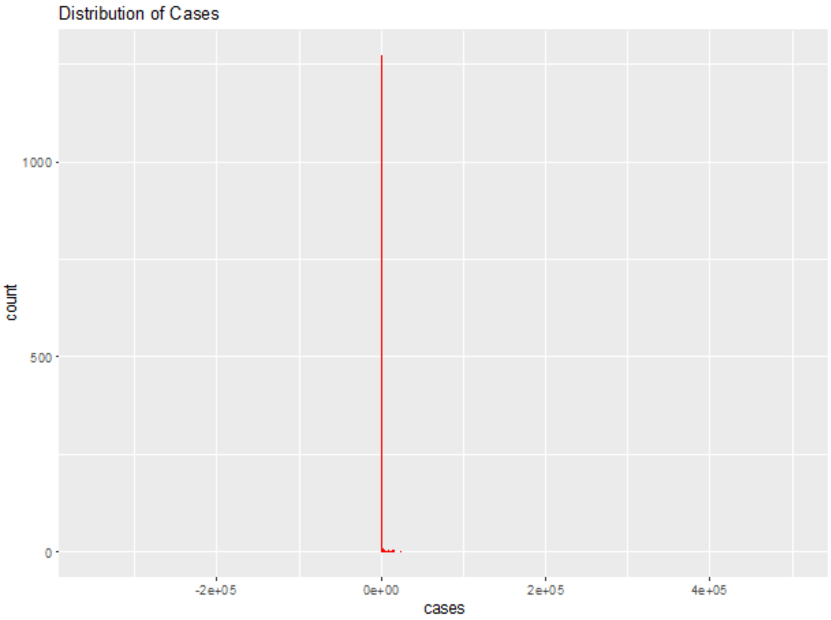
In the next Distribution of Months graph, it is possible to observe that between 2.5 and 10 the highest values occurred. Below you can see the graph:



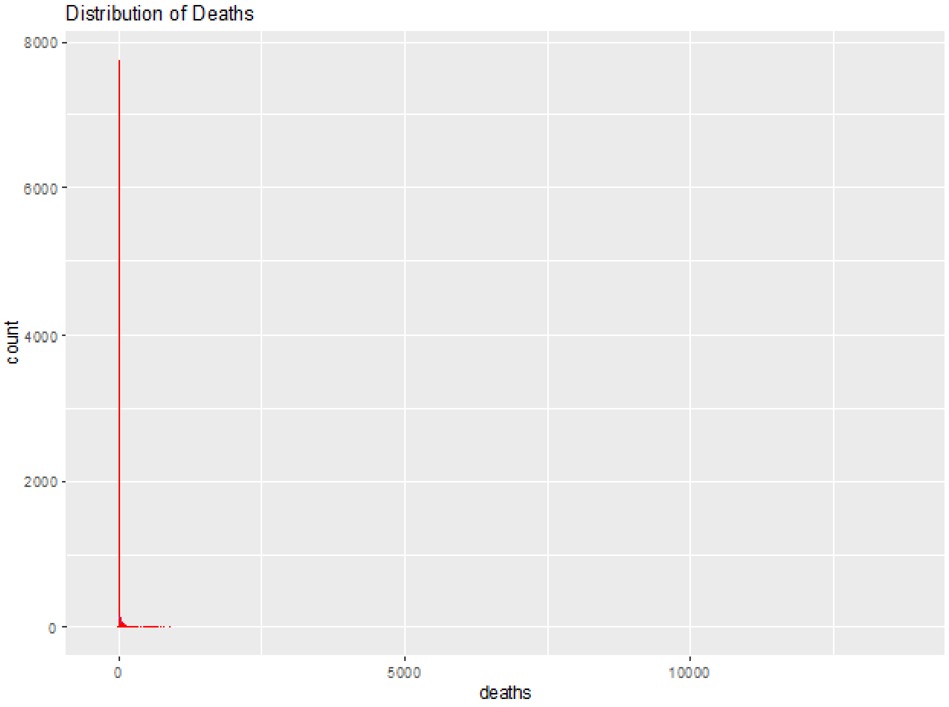
Next we have the Distribution graph of the years with the sample value varying in size, ranging from 0 to approximately 900. It is possible to observe that 2021 had the highest value:



Next graph the Distribution of Cases years with the sample value varying in size, ranging from 0 to approximately 100. Below you will find the graph:



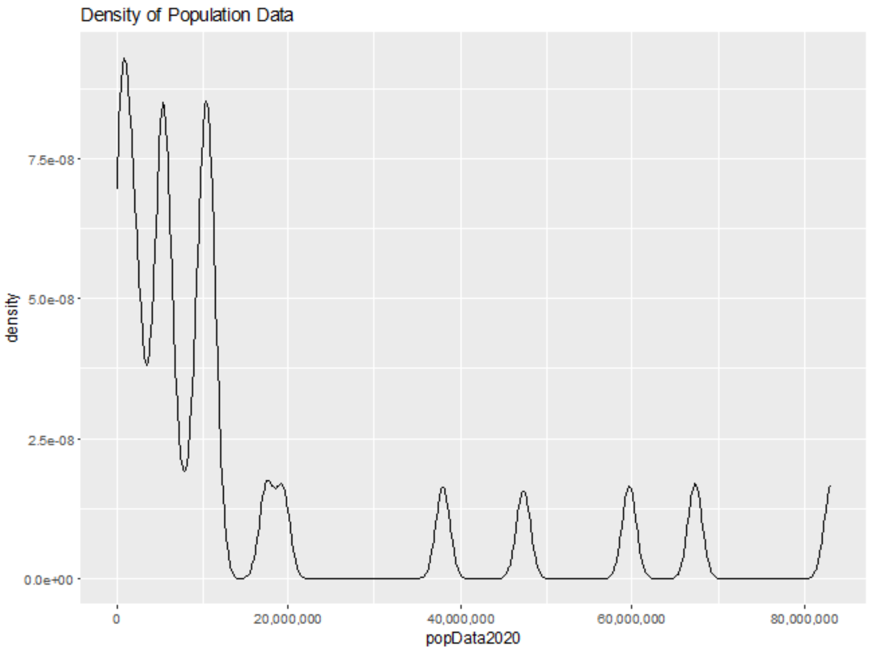
The following graph shows the Distribution of Deaths with the sample value varying in size, ranging from 0 to approximately 800. Can you observe the highest value is nearly the top in the range:



Continuous variables

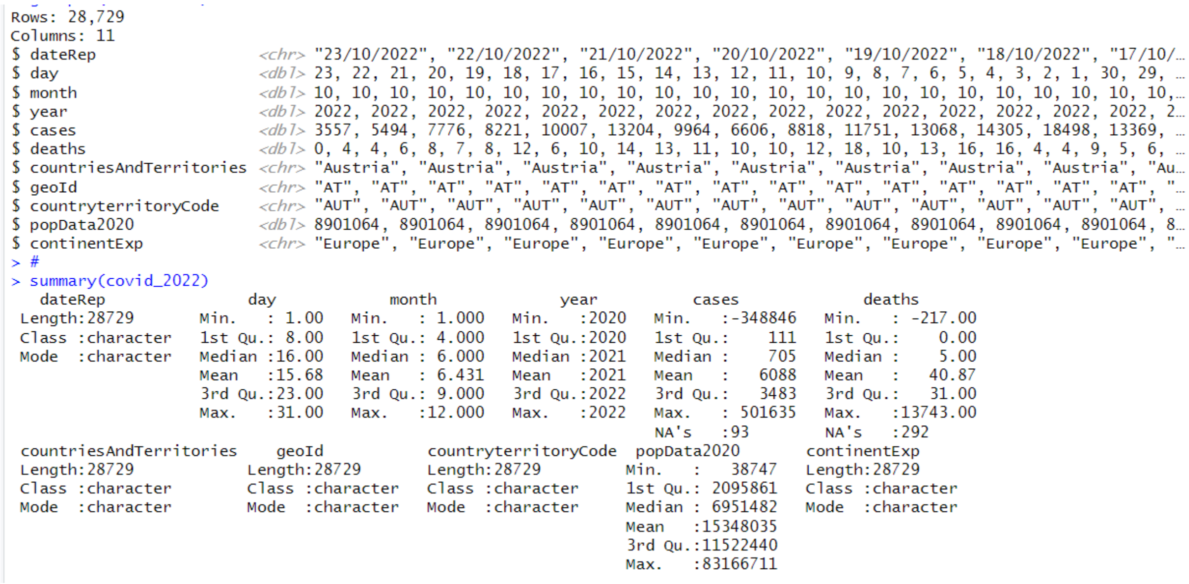
Unlike discrete variables, which can only take on distinct, separate values, continuous variables have an infinite number of possible values within a given range. according to Boddie (2022) "*Continuous variables can take on values that are not restricted to just integers and instead also include decimal values and fractional values*."

The following graph show the Density of population data, it is possible to have an observation where we have the population density peak between 0 and 20,000,000. Below you can see this in the graph:



B- mean, median, minimum, maximum, and standard deviation.

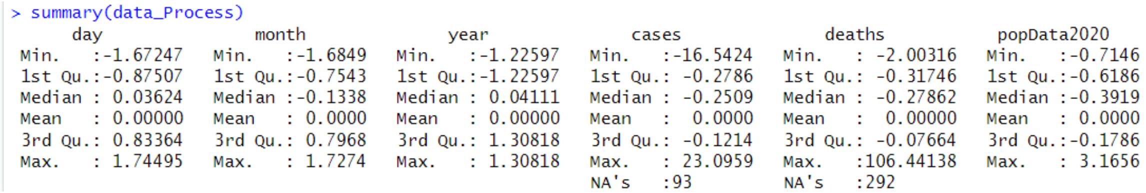
To provide a concise and informative overview of a data frame, we used a glimpse of a compact display that fits a lot of information in a small space, making it easy to get a quick overview of your data, in a way it works like print() methods. Also summary is a quick and useful way to get an initial sense of the data's distribution and characteristics, then we summary to try to show in details all information about the numerical values as mean, median, minimum, maximum, and standard deviation). As you can see in the next image:



```
Rows: 28,729
Columns: 11
$ dateRep                 <chr> "23/10/2022", "22/10/2022", "21/10/2022", "20/10/2022", "19/10/2022", "18/10/2022", "17/10/…
$ day                     <dbl> 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 30, 29, …
$ month                   <dbl> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,…
$ year                    <dbl> 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2…
$ cases                   <dbl> 3557, 5494, 7776, 8221, 10007, 13204, 9964, 6606, 8818, 11751, 13068, 14305, 18498, 13369, …
$ deaths                  <dbl> 0, 4, 4, 6, 8, 7, 8, 12, 6, 10, 14, 13, 11, 10, 10, 12, 18, 10, 13, 16, 16, 4, 4, 9, 5, 6, …
$ countriesAndTerritories <chr> "Austria", "Austria", "Austria", "Austria", "Austria", "Austria", "Austria", "Austria", "Au…
$ geoId                   <chr> "AT", "AT", "AT", "AT", "AT", "AT", "AT", "AT", "AT", "AT", "AT", "AT", "AT", "AT", "AT", "…
$ countryterritoryCode    <chr> "AUT", "AUT", "AUT", "AUT", "AUT", "AUT", "AUT", "AUT", "AUT", "AUT", "AUT", "AUT", "AUT", …
$ popData2020             <dbl> 8901064, 8901064, 8901064, 8901064, 8901064, 8901064, 8901064, 8901064, 8901064, 8901064, 8…
$ continentExp            <chr> "Europe", "Europe", "Europe", "Europe", "Europe", "Europe", "Europe", "Europe", "Europe", "…
> #
> summary(covid_2022)
   dateRep               day            month             year          cases             deaths        
 Length:28729       Min.   : 1.00   Min.   : 1.000   Min.   :2020   Min.   :-348846   Min.   : -217.00  
 Class :character   1st Qu.: 8.00   1st Qu.: 4.000   1st Qu.:2020   1st Qu.:    111   1st Qu.:    0.00  
 Mode  :character   Median :16.00   Median : 6.000   Median :2021   Median :    705   Median :    5.00  
                    Mean   :15.68   Mean   : 6.431   Mean   :2021   Mean   :   6088   Mean   :   40.87  
                    3rd Qu.:23.00   3rd Qu.: 9.000   3rd Qu.:2022   3rd Qu.:   3483   3rd Qu.:   31.00  
                    Max.   :31.00   Max.   :12.000   Max.   :2022   Max.   : 501635   Max.   :13743.00  
                                                                    NA's   :93        NA's   :292       
 countriesAndTerritories    geoId           countryterritoryCode  popData2020        continentExp      
 Length:28729            Length:28729       Length:28729         Min.   :   38747   Length:28729      
 Class :character        Class :character   Class :character     1st Qu.: 2095861   Class :character  
 Mode  :character        Mode  :character   Mode  :character     Median : 6951482   Mode  :character  
                                                                 Mean   :15348035                     
                                                                 3rd Qu.:11522440                     
                                                                 Max.   :83166711                     
```

C- Min-Max Normalization, Z-score Standardization and Robust scalar

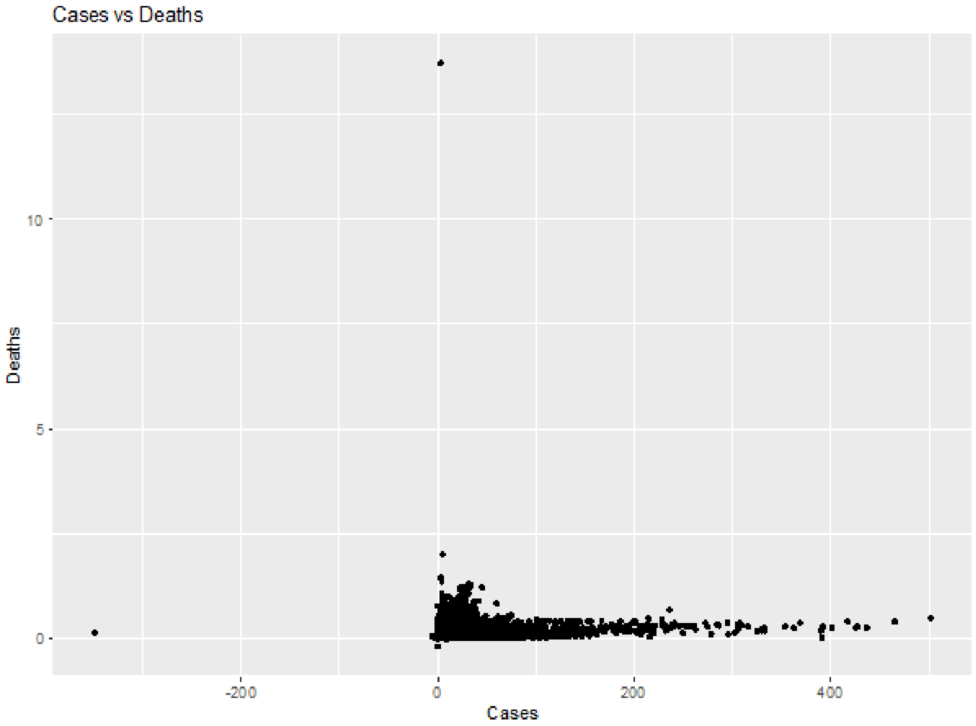
The data going through Min-Max normalization, scaling values between 0 and 1. Z score standardization centres the values around an average of 0, indicating deviations. Summary was used to show all the implementation as you can see below:



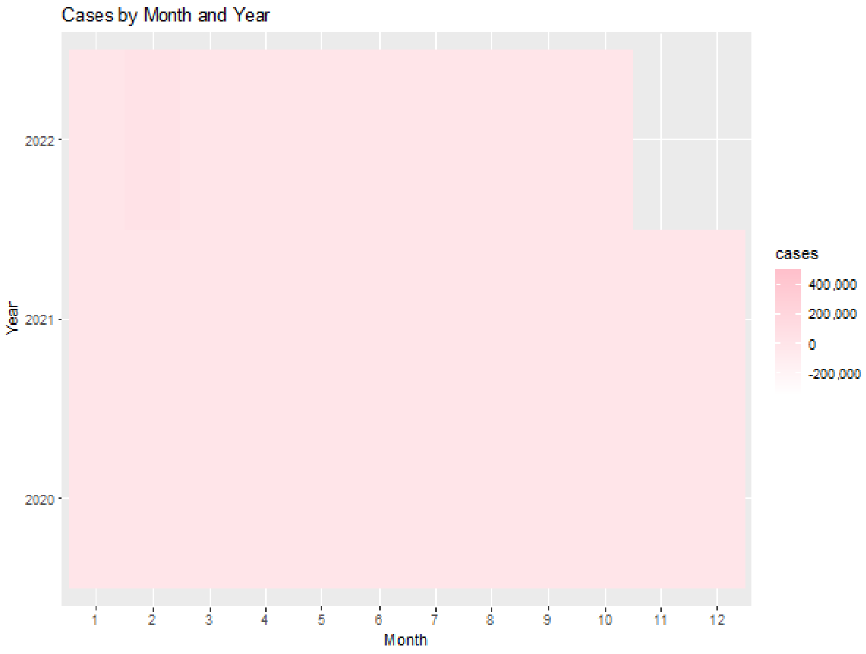
```
> summary(data_Process)
      day                month               year               cases              deaths          popData2020
 Min.   :-1.67247   Min.   :-1.6849   Min.   :-1.22597   Min.   :-16.5424   Min.   : -2.00316   Min.   :-0.7146
 1st Qu.:-0.87507   1st Qu.:-0.7543   1st Qu.:-1.22597   1st Qu.: -0.2786   1st Qu.: -0.31746   1st Qu.:-0.6186
 Median : 0.03624   Median :-0.1338   Median : 0.04111   Median : -0.2509   Median : -0.27862   Median :-0.3919
 Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000   Mean   :  0.0000   Mean   :  0.00000   Mean   : 0.0000
 3rd Qu.: 0.83364   3rd Qu.: 0.7968   3rd Qu.: 1.30818   3rd Qu.: -0.1214   3rd Qu.: -0.07664   3rd Qu.:-0.1786
 Max.   : 1.74495   Max.   : 1.7274   Max.   : 1.30818   Max.   : 23.0959   Max.   :106.44138   Max.   : 3.1656
                                                          NA's   :93        NA's   :292
```

D- Line, Scatter and Heatmaps

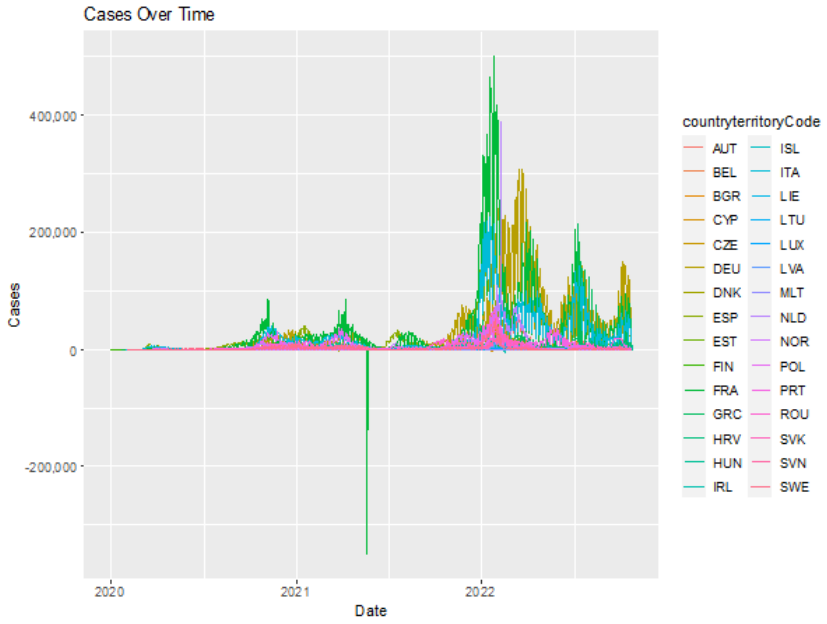
The graph shows a line with a series of dots, indicating the number of cases and deaths over time. As you can see in the scatter plot Cases vs Deaths:



In This graph when can see some in the number of cases, there was a peak pattern during the first 10 months throughout the year 2020 to 2022. As you can see in heatmaps graph below:



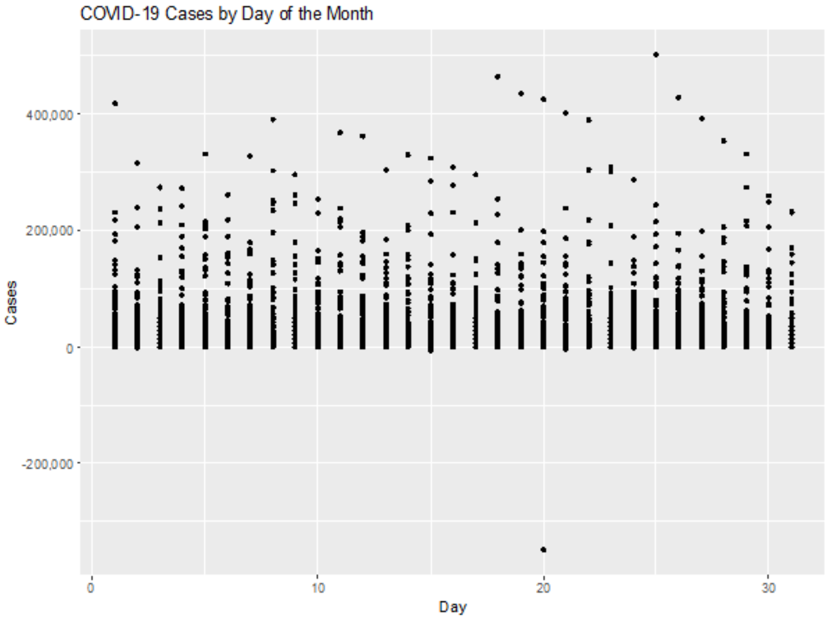
The follow graph provides a visual representation of the Cases Over Time in Europe and the different levels of cases in different countries. Below you can see the line graph :



E- Data Exploratory analysis (EDA).

The main goal of EDA is to understand the structure of data, identify patterns, relationships, and potential outliers, and generate hypotheses that can inform further analysis. According to IBM(2020) “*Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods*.”

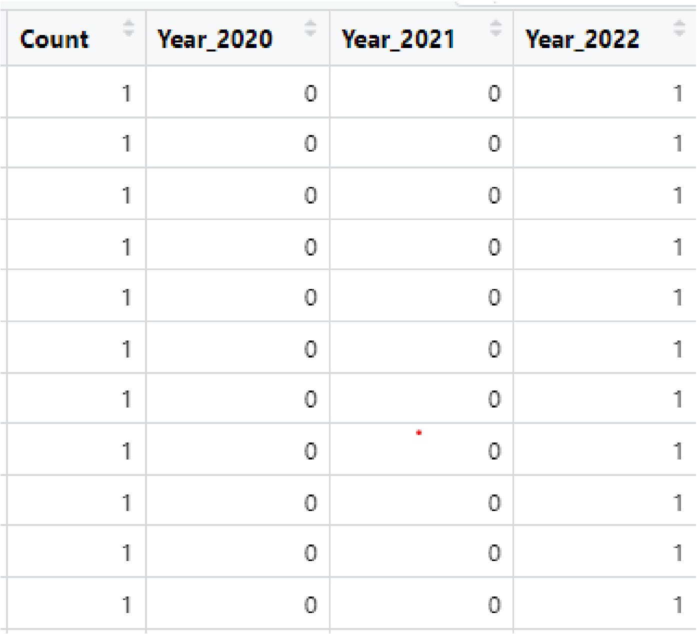
A subgroup of features we can explore are Covid-19 cases by day of the month, as you can see in the Scatter plot image:



We say this scatterplot that visualizes daily COVID-19 cases over the month, this scatterplot is a subgroup of graphical heatmaps of cases by months and year as shown earlier. The x-axis represents days, the y-axis represents the number of cases, and each point on the graph corresponds to the case count for a specific day.

F- dummy encoding to categorical variables

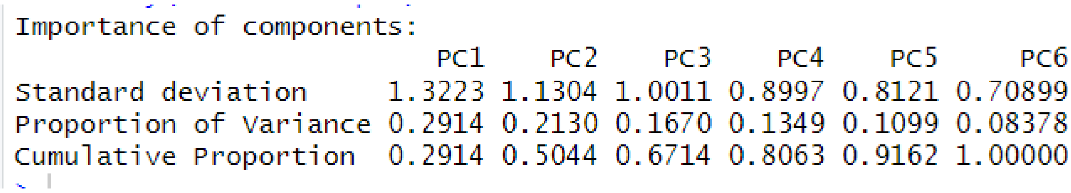
Dummy coding is a technique used to convert categorical variables into a binary format (0 or 1) to make them suitable for machine learning algorithms, for example. According to Pramoditha (2021) "*The dummy variable trap occurs when we use one-hot encoding to encode categorical variables*." From the dataset, dummy variables were created for the variable 'year' in the 'covid_2022' dataset. Below you can see the image:



Count	Year_2020	Year_2021	Year_2022
1	0	0	1
1	0	0	1
1	0	0	1
1	0	0	1
1	0	0	1
1	0	0	1
1	0	0	1
1	0	0	1
1	0	0	1
1	0	0	1
1	0	0	1

G- Apply PCA with your chosen number of components.

The image below provided represents the results of Principal Component Analysis (PCA), where components are labelled PC1 to PC6. Each component has associated standard deviations, variance proportions and cumulative proportions.



```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6
Standard deviation     1.3223 1.1304 1.0011 0.8997 0.8121 0.70899
Proportion of Variance 0.2914 0.2130 0.1670 0.1349 0.1099 0.08378
Cumulative Proportion  0.2914 0.5044 0.6714 0.8063 0.9162 1.00000
```

The first components typically capture the most significant information, allowing for simplified analysis and visualization. According to Jaadi (2019) "*Principal component analysis, or PCA, is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.*" PCA is useful for dimensionality reduction and identifying patterns in high-dimensional data.

Individual resource loads on each component can be examined to understand which variables contribute the most to each major component. In this case, the first six components collectively explain 100% of the variation in the data. Therefore, we can say that the cumulative proportion provides a measure of how much variability in the original data is captured by the chosen components.

H- dimensionality reduction for data analysis.

Dimensionality reduction is valuable in situations where data analysts need to seek to improve visualization, manage computational complexity, improve model performance, avoid overfitting, and gain a deeper understanding of the underlying structure of the data. According to Habs (2022). "*Dimensionality reduction is just like its name, which simply implies that we reduce the dataset's dimensionality.*"

Therefore, dimensionality reduction offers several benefits in different situations as:

Improve model performance: Dimensionality reduction can lead to better generalization and performance of the model, for example Machine learning algorithms suffer from the "curse" of dimensionality, where performance decreases as the number of features increases.

Improve visualization: Dimensionality reduction projects data into lower-dimensional spaces, allowing for easier visualization and interpretation of patterns and relationships.

Computational complexity: Dimensionality reduction mitigates this challenge by reducing the number of features, making computations more accurate. For example, performing computations on high-dimensional datasets often requires more computational resources and time for analysis.

We can conclude that dimensionality reduction is a versatile tool that addresses challenges associated with high-dimensional datasets across multiple domains.

Conclusion

In conclusion, while recognizing that a more comprehensive report could be produced with additional time and expertise, this exploration, facilitated by the professor's code provided by Dr. Muhammad Iqbal, serves as a significant step in understanding the European Centre for Prevention dataset. of Diseases and control. As a newcomer to the field, the simplicity in dataset selection allowed for a basic understanding of exploratory data analysis (EDA). Gratitude extends to the accessibility of the code, allowing for the interpretation of intricate patterns in the data. This experience lays the foundation for further learning and refinement of analytical skills in future endeavours. For a detailed look at all the code, visit my student GitHub account via the following link: https://github.com/Marco2020333/CA1_ProjetDEP.git

Reference:

Boddie, K. (2022). *How to Identify Continuous Variables*. [online] Study.com. Available at: https://study.com/skill/learn/how-to-identify-continuous-variables-explanation.html [Accessed 6 Dec. 2023].

European Centre for Disease Prevention and Control. (2021). *Data on the daily number of new reported COVID-19 cases and deaths by EU/EEA country*. [online] Available at: https://www.ecdc.europa.eu/en/publications-data/data-daily-new-cases-covid-19-eueea-country [Accessed 3 Dec. 2023].

GeeksforGeeks. (2021). *Create Heatmap in R Using ggplot2*. [online] Available at: https://www.geeksforgeeks.org/create-heatmap-in-r-using-ggplot2/ [Accessed Dec. 2023].

Frost, J. (n.d.). *Categorical variables*. [online] Statistics By Jim. Available at: https://statisticsbyjim.com/glossary/categorical-variables/ [Accessed Dec. 2023].

Jaadi, Z. (2019). *A Step by Step Explanation of Principal Component Analysis*. [online] Built In. Available at: https://builtin.com/data-science/step-step-explanation-principal-component-analysis [Accessed 7 Dec. 2023].

Habs (2022). *Principal Components Analysis (PCA) using R programming.* [online] Medium. Available at: https://medium.com/@hablo/principal-components-analysis-pca-using-r-programming-1c0b59190f12 [Accessed 7 Dec. 2023].

IBM (2020). *What Is Exploratory Data Analysis? | IBM*. [online] www.ibm.com. Available at: https://www.ibm.com/topics/exploratory-data-analysis [Accessed 7 Dec. 2023].

Hassan, M. (2022). *Discrete Variable - Definition, Examples, Types*. [online] Research Method. Available at: https://researchmethod.net/discrete-variable/ [Accessed 6 Dec. 2023].

Pither, J. (n.d.). *4.6 Example question / answer | Tutorials for BIOL202: Introduction to Biostatistics*. [online] *ubco-biology.github.io*. Available at: https://ubco-biology.github.io/BIOL202/example_answer.html [Accessed 3 Dec. 2023].

Pramoditha, R. (2021). *What is the Dummy Variable Trap and How to Avoid it?* [online] Data Science 365. Available at: https://medium.com/data-science-365/what-is-the-dummy-variable-trap-and-how-to-avoid-it-aeb227c2cd92 [Accessed 8 Dec. 2023].