CCT College Dublin

Data Exploration and Preparation

CA1

Robert Szlufik

2020358

GitHub repository:
https://github.com/2020358/data-exploration-and-communication

Dataset 2
Challenges 3
Data Preparation 3
N.A values and repetitive columns 3
Identification of variable types and description 3
Outliers 4
EDA 7
Initial Observations 7
Correlation between numerical variables 9
Fig 6 - Heat map between numerical variables in dataset 9
Additional Observations 9
Dummy and One-hot Encoding 11
PCA 11
Conclusions 13
Appendix 14
Code Explanation 14
References 14

Dataset

Dataset chosen for this assignment is called “Data on COVID-19 vaccination in EU/EEA (European Centre for Disease Prevention and Control, 2021), and was obtained from the European Centre for Disease Prevention and Control website.

Dataset contains information about vaccination between week 34 of 2021 and week 30 2022. Vaccination information is divided into eight columns, first dose, second dose, additional dose 1 - 5 and unknown dose. Each entry in the dataset contains information code of the country and region that reports this information. Additionally, it contains information such as population and vaccine type. This data set was obtained from European governing bodies and includes information about the European countries.

Initially, the dataset contained 815,597 rows and 18 columns.

Challenges

In my opinion the biggest challenge was to find a suitable dataset. Data analysis depends on the quality of collected data, for example a data set which contains a lot of empty values is more difficult to analyse and draw conclusions from.
There are many dataset online, however, for this project, the dataset had to be in one of three domains. I decided to use this dataset as it seemed to me as the best choice for my purpose and the way I would like to analyse data.

The biggest issue I found is the fact that the data set is very large, it contains over 80,000 rows, that is 10 times the number of requirements. With a dataset that big, for example, it would be very hard to draw conclusions for one country. That's why I decided to analyse this dataset in a more statistical fashion.

Data Preparation

N.A values and repetitive columns

In the first step of cleaning data, I identified that one of the columns, called "FirstDoseRefused" contained only "N.A" values. We have to remove this column in order to proceed with any further exploration.

Further filtration with na.omit() function has not change dimensions of the dataset

Upon further investigation of the original dataset we can conclude that two of the columns have very similar values. Those columns are Reporting Country and Region. We can identify that there are 22 reporting countries and 24 regions, and from the PDF documentation included with the dataset, we can find that regions are specific to the country. Therefor, we can remove region column from the dataset

Dataset dimensions after cleaning of NA values:

- Number of rows: 85140
- Number of columns: 16

Identification of variable types and description

Categorical:

- YearWeekISO - year and week reported
- ReportingCountry - country code

- Target Group - age group
- Vaccine - vaccine code

Discrete Variables:

- Denominator
- Number of Doses Received
- Number of Doses Exported
- First Dose
- Second Dose
- Dose Additional 1
- Dose Additional 2
- Dose Additional 3
- Dose Additional 4
- Dose Additional 5
- Unknown Dose

Continuous Variables:

- Population

Outliers

In terms of outliers, we can identify 2 different columns in which outliers could occur, such that it could influence our later exploration. Those columns are Denominator (fig 1) and Population (fig 2).
We can further read from the PDF document obtained along with a dataset from the website (European Centre for Disease Prevention and Control, 2021) that Denominator is *"Population denominators for target groups (total population and age- specific population obtained from Eurostat/UN). Denominators reported by countries for TargetGroup = "HCW" and TargetGroup = "LTCF"."* and Population is "*Age-specific population for the country*".

In terms of those 2 variables, outliers would indicate that the population in a particular country is larger than the average, however, this would not change further observations since the number of vaccines received should be corresponding to population number.

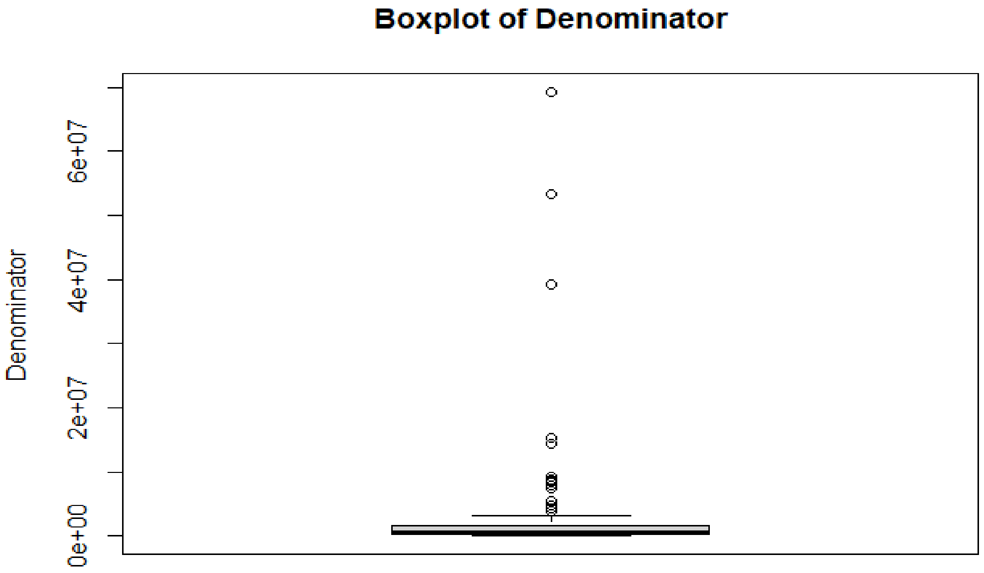

Fig 1 - Box plot of Denominator

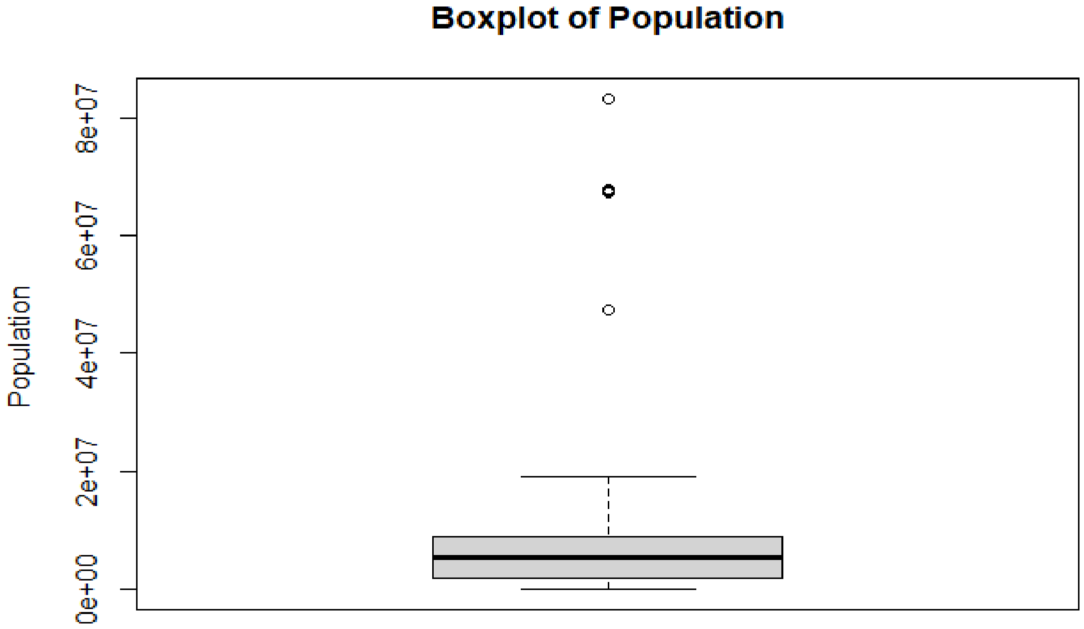

Fig 2 - Box plot of Population

Column Name	**Minimum**	**Median**	**Max**	**Mean**	**Standard Deviation**
Denominator	313	503217	69373865	2297011	7229396
Number of Doses Received	0	0	13092598	9498	153786.5
Number of Doses Exported	0	0	6488820	1052	153786.5
First Dose	0	0	4021725	2278	47468.14
Second Dose	0	0	4149209	2175	46673.87
Dose Additional 1	0	0	6647843	1932	55879.54
Dose Additional 2	0	0	598894.0	528.2	9542.449
Dose Additional 3	0	0	337594.0	122.3	3570.396
Dose Additional 4	0	0	34714.0	7.2	301.7767
Dose Additional 4	0	0	30509.000	2.228	189.5138
Unknown Dose	0	0	22367.000	5.238	204.3932

As we can see that standard deviation is mostly a large number which indicates that data is spread out. Standard deviation is larger than mean, therefore the majority of values lay in the higher range.

EDA

Initial Observations

We can find 22 distinct countries reporting in this dataset. Since this dataset is obtained from the EU, it only includes European countries. We can display the distribution of population across this dataset with histogram where Population is on x axis and frequency of given population occurring on y axis. (Fig

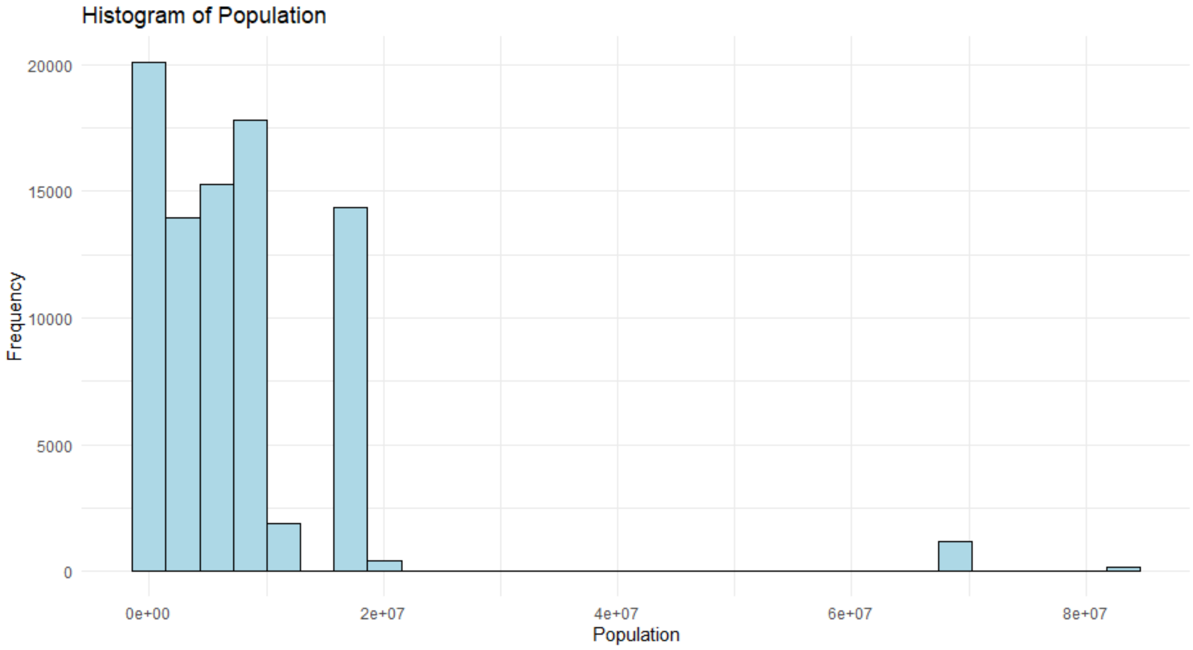

Fig 3 - Histogram of population

Each country reports every week. And since their population does not change significantly within a short period of time, we get consistent values of population across observations from distinct countries. We can also conclude that the majority of reporting countries' population falls under 20 million people.

Another observation that we can make is total a total number of different vaccine types by country reporting (Fig 4)

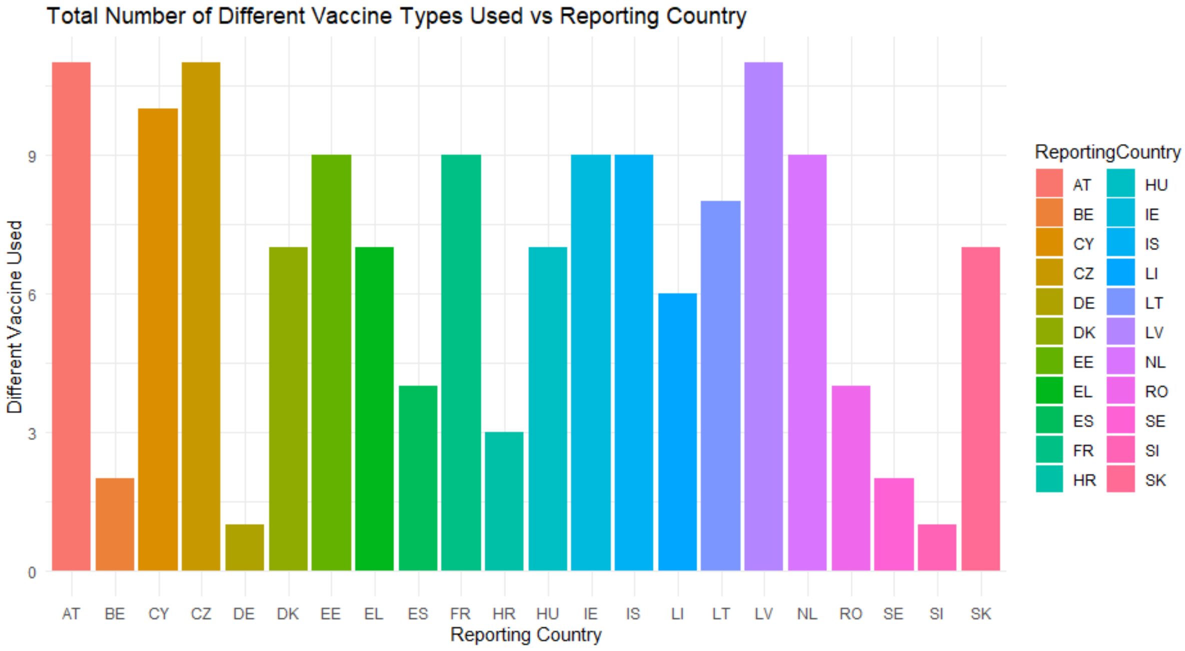

Fig 4 - Number of different vaccine types by country

We can also derive the population of each reporting country (Fig 5).

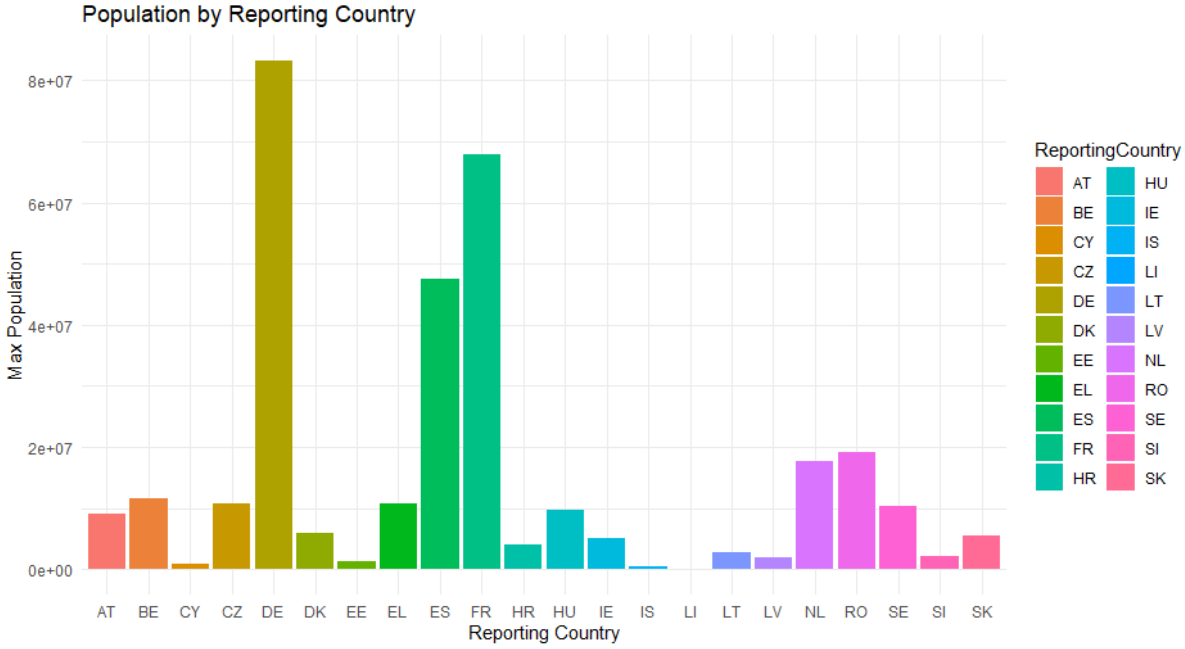

Fig 5 - Population by reporting country

Correlation between numerical variables

We can show correlation between numerical variables in the dataset by creating a heat map (fig 6). We can observe that, there is correlation between first and second dose, however, in general correlation between variables is weak. The correlation between NumberDosesReceived, FirstDose and SecondDose indicates that reporting countries were able to administer more first and second doses as they receiving number increased.

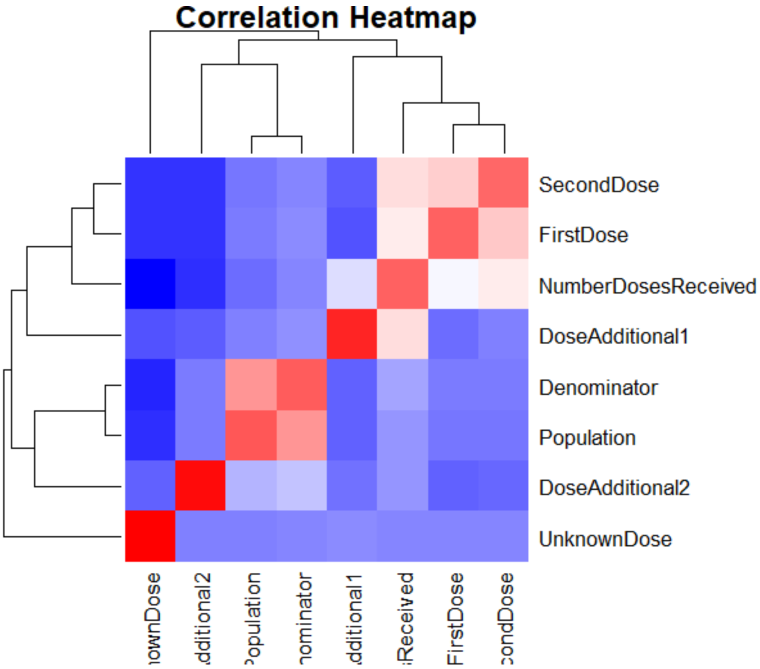

Fig 6 - Heat map between numerical variables in dataset

Additional Observations

We can calculate and observe the total number of doses administered by the target group, which in our case is an age group.

In order to achieve it, we need to create additional columns in each observation, sum up values from columns: FirstDose, SecondDose, AdditionalDose 1- 5, and assign it as Total Dose. This will represent the total dose administered by TargetGroup in a given week. Then, we need to create a table of groups and total doses administered. We need to remove several target groups such as All, AgeUNK, HCQ,LTCF,AGE<18. After cleaning the table, we can create a barchart. (Fig 6)

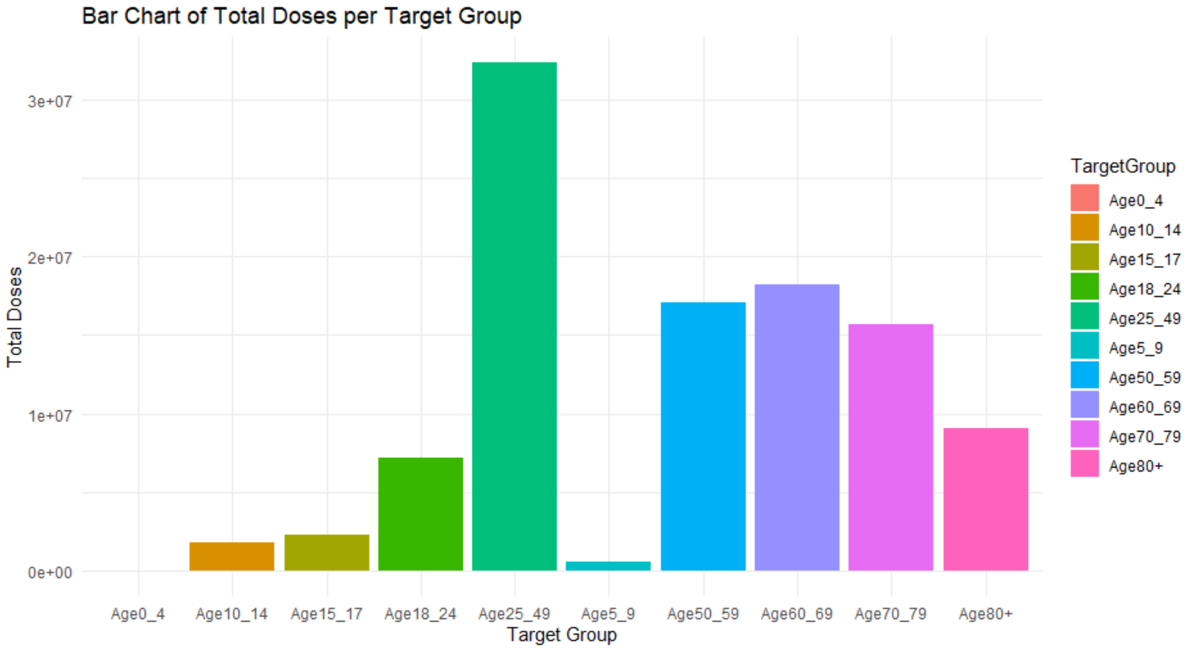

Fig 7 - Total Doses Administered by Target Group

Another interesting observation is the distribution of vaccine types across dataset (Fig 8). We can observe that the biggest 4 vaccines administered to the population were AZ (AstraZeneca), COM (Pfizer/BioNTech), MOD(Moderna), JANSS (Janssen).

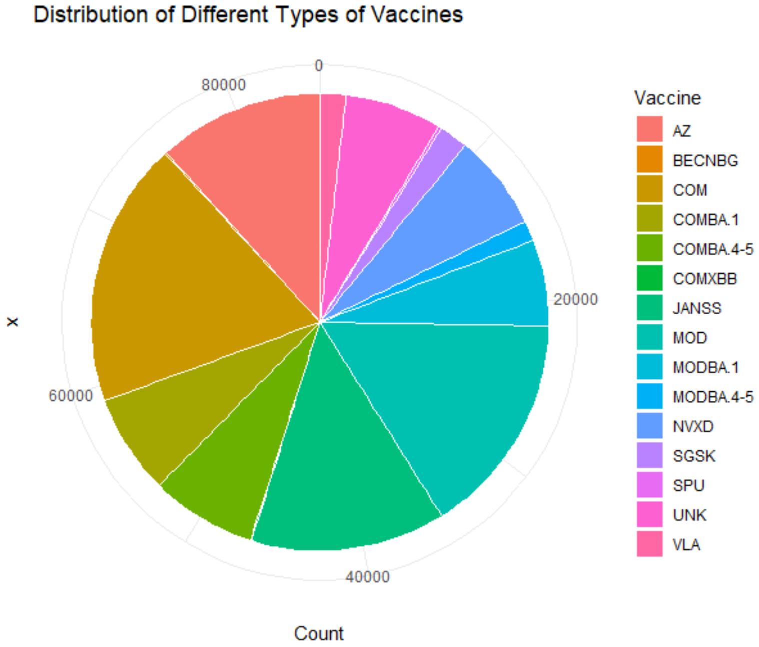

Fig 8 - Distribution of vaccine type

Dummy and One-hot Encoding

In my opinion there are 3 categorical variables that are candidates for one-hot encoding. We can choose Vaccine, Target Group or Country Reporting as a target variable for encoding. I decided to hot-one encode Vaccine. Ideal variable would have only two categories, then we could assign binary encoding, either true (1) or false(0) value. However in the case of this dataset, each of the variables have many different categories, in the case of the Reporting Country is 24, Vaccine is 16 and Target Group is 15.

As a consequence of one-hot encoding, we will add 16 columns to our dataset.

As presented by (Garavaglia and Sharma, n.d.), dummy encoding has a great benefit when it comes to preparing dataset for machine learning models and algorithms. With such encoding, we can provide a model with categorical variables as numeric. This further enhances our ability to carry out analysis and correlation between encoded variables, as opposed to not being able to compute variables of categorical type.

PCA

As we can deduct from the heat map previously presented (Fig 6), there is little correlation between variables in this dataset. The only variables that show correlation are DosesReceivede, FirstDose and SecondDose. I decided to use those three variables as targets for PCA analysis.

For this PCA analysis I decided to use the FactoMineR library.

The result of PCA performed in R are:

- Dim.1 - variance: 2.32, % of variance: 77, cumulative % of variance: 77
- Dim.2 - variance: 0.3, % of variance: 13, cumulative % of variance: 90
- Dim.3 - variance: 0.2, % of variance: 9, cumulative % of variance: 100

We can observe that with Dim.1 and Dim.2 combined together have over 90% of data, and those are to components with the largest relative importance (Fig 9)

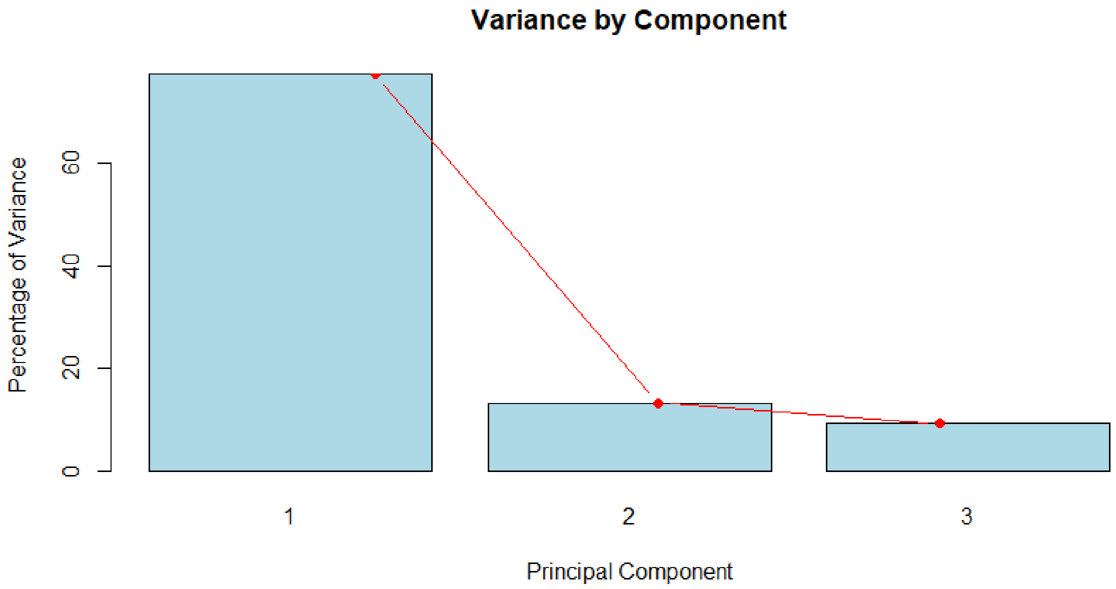

Fig 9 - Relative importance of each component

We can further observe and graph correlation between each variable in those to results (Fig 10).

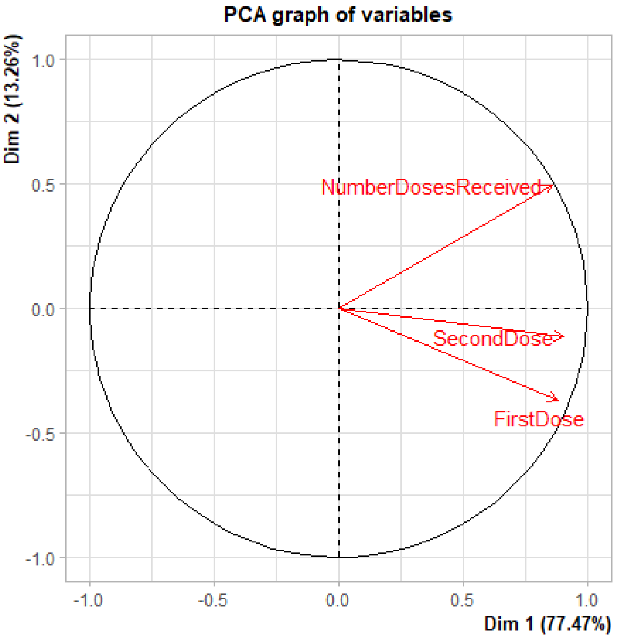

Fig 10 - Correlation between variables in PCA

We can observe that all variables exhibit positive correlation among each other, however, the graph indicates that correlation between First and Second dose is stronger than other correlations. This graph exhibits 90% of data.

Upon further investigation, I found that if we increase the number of columns to include AdditionalDose 1 - 5, the correlation positive among initial 3 variables seems to be even higher. (Fig 11) This might be due to the fact we operate with much more data, and proportion has changed.

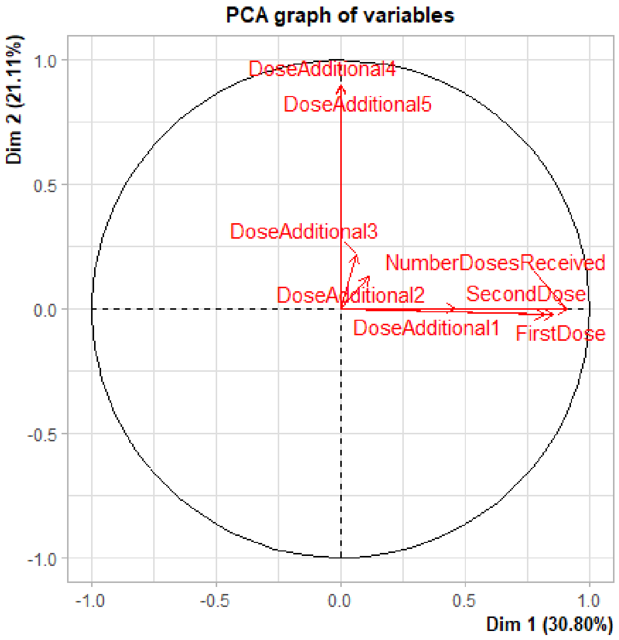

Fig 11 - Correlation with more variables.

Additionally, we can observe strong positive correlation between Dose Additional 1 and Dose Additional 2, Dose Additional 3,4 and 5 do not exhibit strong correlation.

Unfortunately, the factor map graph for this dataset does not provide meaningful visualisation due to the large number of observations in the dataset.

Conclusions

This dataset included a vast amount of information. It includes information on vaccination from 24 European countries. It is ordered chronologically by ISO timestamp, and this resulted in certain variables having values only in the last quadrant. For example, additional doses 4 and 5 would be administered to the population after most of the population was given the first and second dose.

We also observed which vaccines were the most popular and which age group was vaccinated the most.

Finally, we have seen correlation between the number of variables, and concluded that the first and second dose and number of doses received show positive correlation.

Appendix

Code Explanation

- Line 1 - set working directory to current folder
- Lines 7,8 - import data
- Lines 10-40 - cleaning dataset from na values and removing unnecessary columns
- Lines 40-45 - checking for outliers
- Lines 47-108 - statistical analysis (min, max, median, mean, standard deviation) (www.rdocumentation.org, n.d.)
- Lines 108-176 - min-max, z-score and robust scalar normalisation (Kalpana, 2020)
- Lines 178-274 - EDA and supporting graphs (ggplot2.tidyverse.org, n.d.), (Holtz, 2018)
- Lines 276-278 - One-Hot encoding (GeeksforGeeks, 2023)
- Lines 281-330 - PCA and supporting graphs (Iqbal, 2023), (Husson et al., 2020)

All code is commented within the R file attached.

References

European Centre for Disease Prevention and Control. (2021). *Data on COVID-19 vaccination in the EU/EEA*. [online] Available at: https://www.ecdc.europa.eu/en/publications-data/data-covid-19-vaccination-eu-eea [Accessed 20 Nov. 2023].

Garavaglia, S. and Sharma, A. (n.d.). *A SMART GUIDE TO DUMMY VARIABLES: FOUR APPLICATIONS AND A MACRO*. [online] Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=efec2a88f0e74a1668df4b80b571c04af9ebf707 [Accessed 30 Nov. 2023].

GeeksforGeeks. (2023). *Encoding Categorical Data in R*. [online] Available at: https://www.geeksforgeeks.org/encoding-categorical-data-in-r/ [Accessed 28 Nov. 2023].

ggplot2.tidyverse.org. (n.d.). *Histograms and frequency polygons — geom_freqpoly*. [online] Available at: https://ggplot2.tidyverse.org/reference/geom_histogram.html [Accessed 28 Nov. 2023].

Holtz, Y. (2018). *Heatmap | the R Graph Gallery*. [online] r-graph-gallery.com. Available at: https://r-graph-gallery.com/heatmap [Accessed 28 Nov. 2023].

Husson, F., Josse, J., Le, S. and Maintainer, J. (2020). *Package ‘FactoMineR’ Title Multivariate Exploratory Data Analysis and Data Mining*. [online] Available at: https://cran.r-project.org/web/packages/FactoMineR/FactoMineR.pdf.

Iqbal, M. (2023). *Lecture_DE&P_7*. [online] CCT College Dublin . Available at: https://moodle.cct.ie/course/view.php?id=2584.

Kalpana, N.S.S. (2020). *Data Normalisation With R*. [online] The Startup. Available at: https://medium.com/swlh/data-normalisation-with-r-6ef1d1947970 [Accessed 30 Nov. 2023].

The Average and SD in R. (n.d.). Available at: https://www.carlislerainey.com/teaching/pols-209/files/notes-10-average-sd-r.pdf [Accessed 29 Nov. 2023].

www.rdocumentation.org. (n.d.). *summary function | R Documentation*. [online] Available at: https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/summary [Accessed 27 Nov. 2023].