## 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.

Initially, it contained 815,597 rows and 18 columns.

## 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

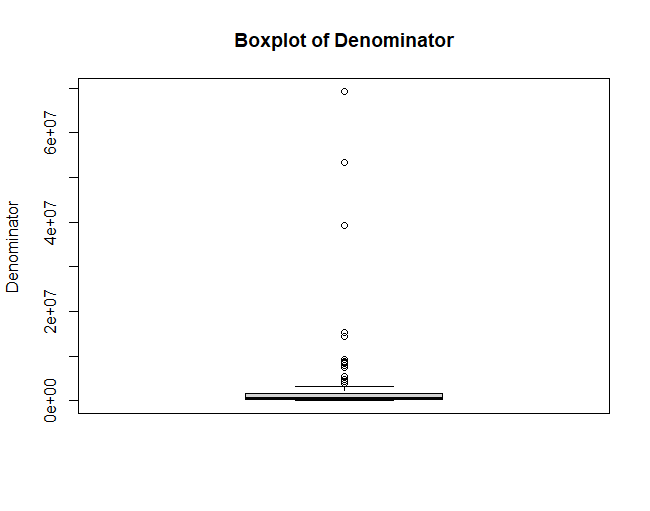
## 

## Outliers

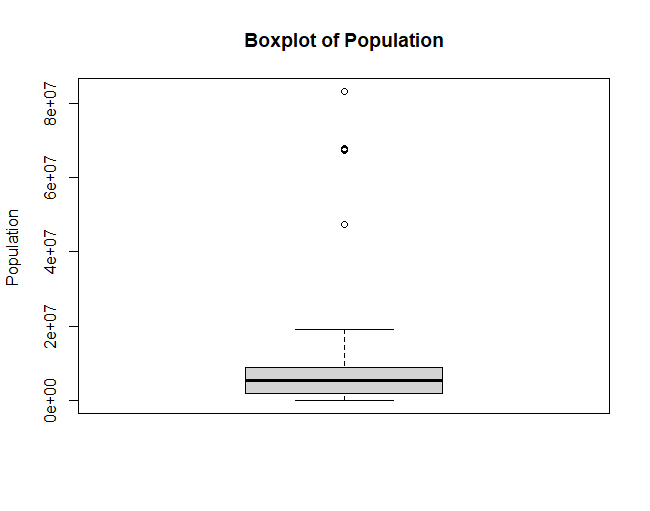
In terms of outliers, we can identify 2 different columns in which outliers could occur, such that it could influence our later exploration. Those columns are Denominator (fig 1) and Population (fig 2).

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*

## 

## Statistical Parameters

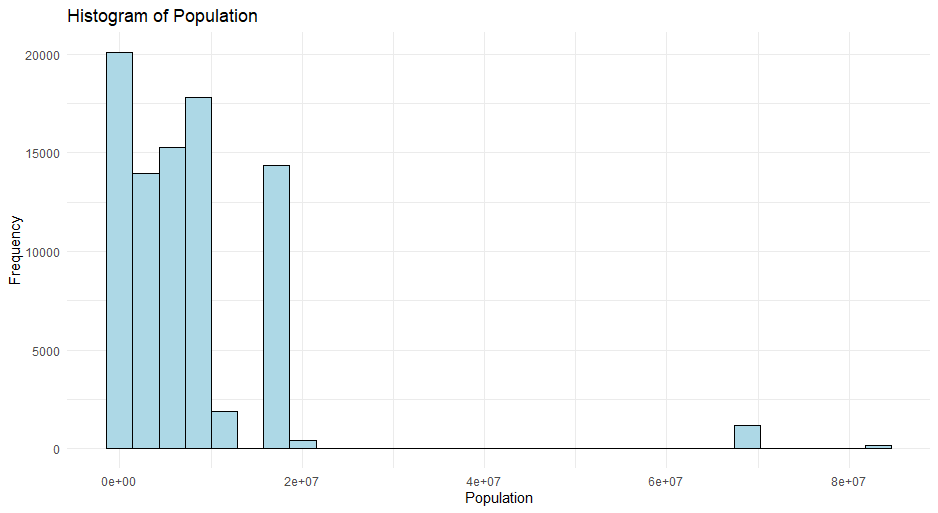
| **Column Name** | **Minimum** | **Median** | **Max** | **Mean** | **Standard Deviation** |
| --- | --- | --- | --- | --- | --- |
| Denominator | 313 | 503217 | 69373865 | 2297011 | 7229396 |
| Number of Doses Received | 0 | 0 | 13092598 | 9498 | 153786.5 |
| Number of Doses Exported | 0 | 0 | 6488820 | 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

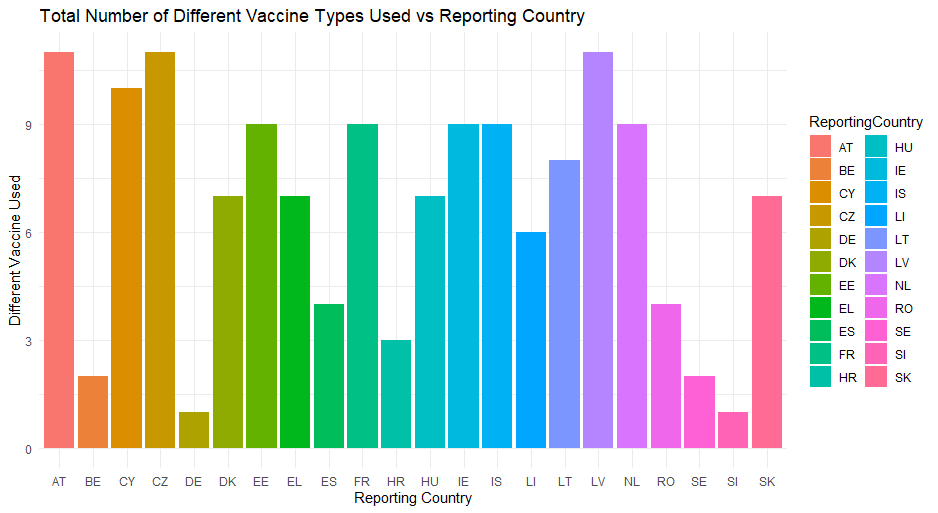
### Variations within dataset

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*

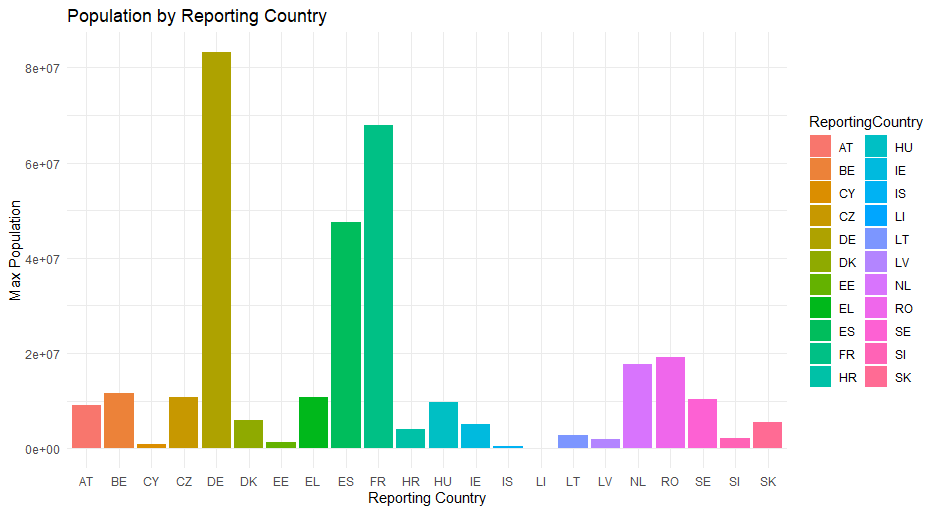
Each country reports every week. And since their population does not change significantly within a short period of time, we get consistent values of population across observations from distinct countries. We can also conclude that the majority of reporting countries' population falls under 20 million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

### Covariation within dataset

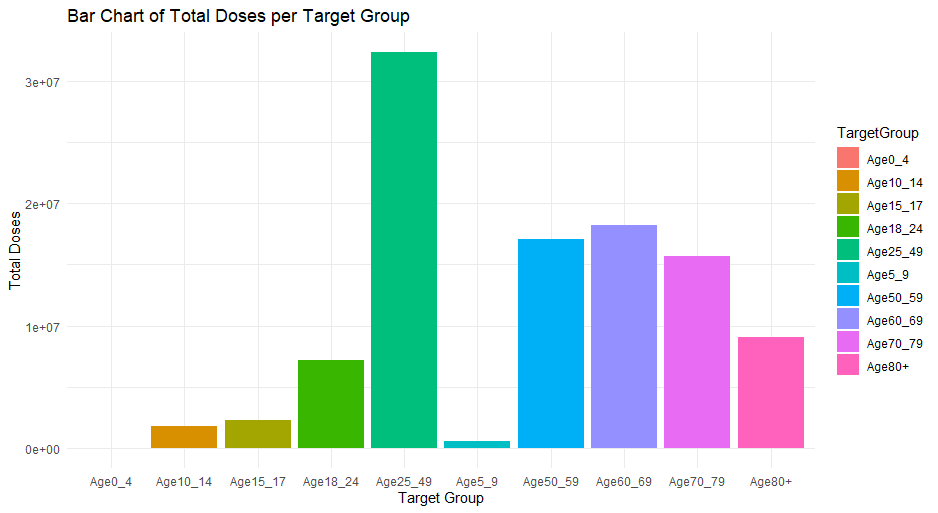
### // to do

### 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 6 - Total Doses Administered by Target Group*