Università degli Studi dell'Aquila

Dipartimento di Ingegneria e Scienze dell'Informazione e Matematica

Master Degree in Applied Data Science

PROGRAMMING FOR DATA SCIENCE

Professor

Prof. Antinisca Di Marco
Prof. Luca Traini

Students

Francesca Ciccarelli 288804
Fabio D'Andreamatteo 288846
Leonardo Masci 288694

Accademic Year 2022-2023

Contents

Introduction

This work consists in the application of the data mining process to data about the Covid-19 vaccinations in different countries.
In order to reach the goal of computing some statistical analysis to gain useful knowledge, it was necessary to clean the dataset, replacing missing and erroneous entries and dropping useless attributes; to differentiate the analysis were also added information about the belonging of each country to the respective continent.
After the preprocessing phase, the cleaned dataset is ready to be used for the analysis stage.
During this phase, various aspects of the vaccinations have been investigated: a first look was taken to the average number of daily vaccinations per country, in which was obtained a bar chart of the top 10 countries with highest value and a box plot of the top 20 countries in Europe.
Another interesting analysis is related to the study of people that completed the vaccination cycle over people that received only one dose; a similar analysis has been done to the percentage of people fully vaccinated per continent.
Following, through the computation of the correlation between countries, it was possible to highlight the most correlated countries among all. It is also possible to choose a specific country and compute three different analysis regarding the trend of the daily vaccinations.
The last analysis is focused on the type of the vaccines that were administrated in each country.

Chapter 1

Preprocessing Phase

1.1 A closer look to the data

This work starts with the understanding of the dataset *"country_vaccinations"*, which is composed by 65391 rows and 15 columns.
The columns are:

- **country**: it represents the name of the countries;
- **iso_code**: it is the code of the corresponding country;
- **date**: it is the date in which the data about the vaccinations were registered;
- **total_vaccinations**: it represents the number of the total vaccinations per day;
- **people_vaccinated**: it is the total number of the vaccinated people per day;
- **people_fully_vaccinated**: it represents the number of people that have received the total immunization (typically 2 vaccinations, depending on the national rules);
- **daily_vaccinations_raw**: it is the number of vaccinations per day;
- **daily_vaccinations**: it is the number of vaccinations per day;
- **total_vaccinations_per_hundred**: it represents the fraction between the total vaccinations and the number of the population multiplied by 100:

$$\frac{total\ \ vaccinations}{number\ \ of\ \ population} * 100$$

- **people_vaccinated_per_hundred**: it represents the fraction between the vaccinated people and the number of the population multiplied by 100:
$$\frac{people\ \ vaccinated}{number\ \ of\ \ population} * 100$$
- **people_fully_vaccinated_per_hundred**: it represents the fraction between the people fully vaccinated and the number of the population multiplied by 100:
$$\frac{people\ \ fully\ \ vaccinated}{number\ \ of\ \ population} * 100$$
- **daily_vaccination_per_million**: it represents the fraction between the daily vaccinations and the number of the population multiplied by 1000000:
$$\frac{daily\ \ vaccination}{number\ \ of\ \ population} * 1000000$$
- **vaccines**: it is the list of the names of the companies who made the vaccines;
- **source_name**: it is the name of the source of the data (international, national or local organization);
- **source_website**: it is the website of the source of information.

The total number of the countries is 223 and the period of time goes from December 1^{st} 2020 to December 20^{th} 2021, varying for each country.

1.2 Cleaning the Dataset

At first glance, in the dataset there are some missing values, which needs to be dropped or replaced. For the purpose of the analysis, there are some irrelevant columns, such as:

- **iso_code** because it's easier to work with the name of the country instead of the iso_code;
- **daily_vaccinations_raw** because there are a lot of missing values and this column is very similar to *daily_vaccinations*;
- **source_name** because it is irrelevant for the analysis;

- **source_website** because it is irrelevant for the analysis.

The preprocessing phase is carried in three steps:

1. In the dataset there are a lot of missing values and each *NaN* value is replaced with -1. The choice of using this latter instead of 0 is to distinguish the case in which there is no information (the -1 case) from the case in which the value of the entry is exactly 0. It is also needed to convert some float fields in integer type;

2. For simplicity the *date* column's format has been converted from *year-month-day* to the string format *yearmonthday* in order to create different columns to add to the dataset: a column with the new date's format and three new columns named *year*, *month* and *day*;

3. To deepen the analysis a new column, called *continent*, has been added. The labels for each continent are:

   - **EU** for Europe;
   - **AS** for Asia;
   - **NAM** for North America;
   - **SA** for South America;
   - **OC** for Oceania;
   - **AF** for Africa;

   Antartide is not considered because it doesn't have any country.

The new dataset has been saved in a csv file called *clean_country_vaccinations*, used for the following phase.

Chapter 2

Analysis Phase

The analysis phase is composed by different type of statistical analysis with the aim to extract useful insights and knowledge from the data.

2.1 Daily Vaccinations Stats

For the first analysis, the columns that are considered are *country* and *daily_ vaccinations*. After grouping the data by country and computing the average, the countries have been ranked in a descending order and the returned output has been saved in a csv file called *avg_ daily*.

country	daily_vax_avg
China	7192763.89
India	4025984.12
United States	1330219.26
Brazil	961750.09
Indonesia	749900.83
...	...
Tuvalu	54.15
Wallis and Futuna	45.45
Tokelau	17.40
Montserrat	11.02
Pitcairn	0.51

Figure 2.1: Average of daily vaccinations grouped by countries.

From the ranking in figure (2.1) it is possible to plot the first ten countries with highest daily vaccinations average:

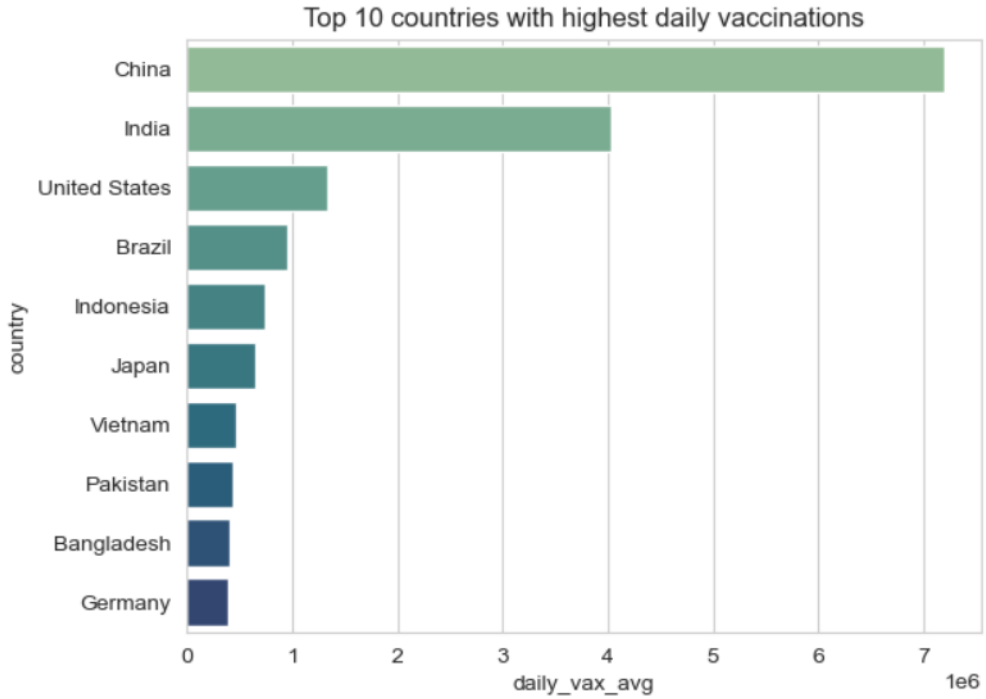


Figure 2.2: Top 10 countries with highest daily vaccinations average.

Prima facie it may seem that China, same as India, are way above the average but, considering the population number, the result is proportionate to the other countries.

2.2 Top 20 Daily Vaccinations in Europe

Using only the columns of *daily_vaccinations* it's possible to reshape the dataset in order to visualize the top 20 countries in Europe with highest average of daily vaccinations.
Through a pivot table, all the countries besides the top 20 have been dropped.

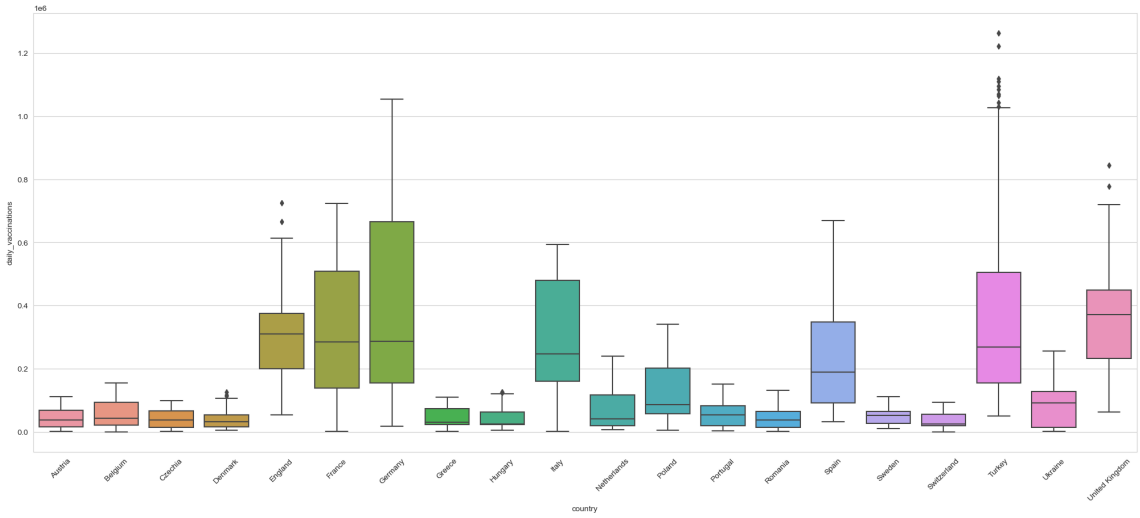


Figure 2.3: Top 20 countries with highest average of daily vaccinations in Europe.

Each box represents the mean distribution of the average of daily vaccinations for the top 20 european countries.
The line in the box refers to the median value; the lower bound of the box is the first quartile while the upper bound is the third quartile. The segments out of the box represents the bounds in which a data is still considered an inlier.
In the figure (2.3) it is possible to notice that there are few countries that have some points, considered outliers, above the upper bound.

2.3 People Vaccinated Stats

The second analysis is the study of the number of people vaccinated (only first dose) compared to the number of people fully vaccinated (people that completed the vaccination cycle).
Considering only the entries different from -1, it is possible to compute different type of statistics, in particular: the average, the standard deviation, the first quartile, the median, the second quartile, the minimum and the maximum.
The results have been stored in a new dataframe called *tot_ stats*.

stat	people_vax	people_fully_vax
avg	2.029387e+07	1.643840e+07
std	1.015797e+08	8.312874e+07
first_quartile	2.708785e+05	2.007805e+05
median	1.366662e+06	1.088111e+06
second_quartile	7.304274e+06	5.734720e+06
min	4.700000e+01	4.700000e+01
max	1.225000e+09	1.076308e+09

Figure 2.4: Dataframe of the statistics about people vaccinated and people fully vaccinated.

Another interesting analysis is the comparison of people vaccinated with only one dose and people that have completed the vaccination cycle, divided by continent. The result is plotted in figure (2.5).

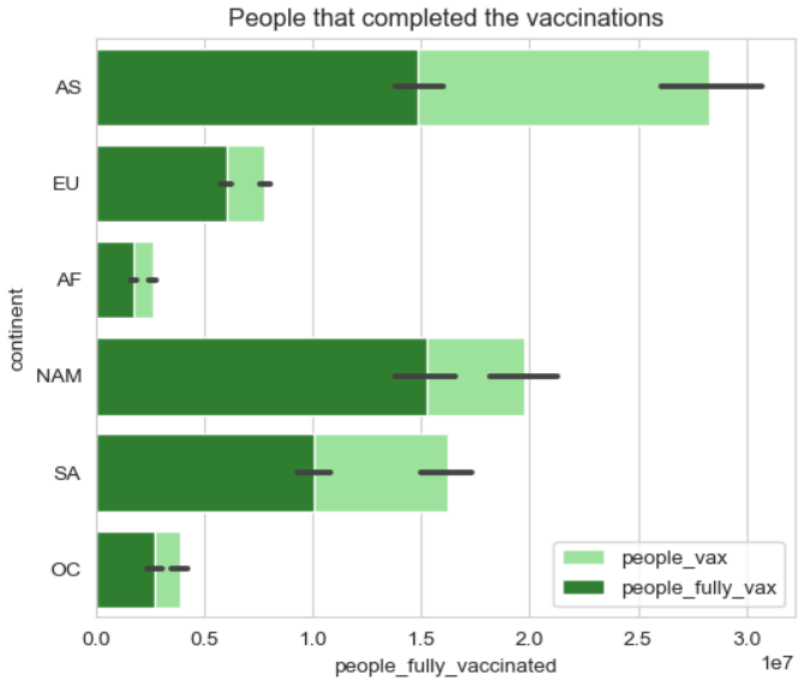


Figure 2.5: Comparison between people vaccinated and people fully vaccinated per continent.

It is noticeable that most of the data were collected in Asia and America. It is evident that the majority of people that received the first dose of the vaccine completed the vaccination cycle. In Europe, North America and South America most of the population completed the cycle, while in Asia only slightly more than half is fully vaccinated. In Africa and Oceania, even though the amount of data is small, the majority of people that received the vaccine is fully vaccinated.
These considerations are confirmed by the following output (saved in a csv file called *perc_stats*):

continent	people_vaccinated	people_fully_vaccinated	percentage
AF	2619672.50	1721879.89	65.73
AS	28293688.70	14852361.47	52.49
EU	7784464.81	5996098.10	77.03
NAM	19735060.19	15230587.91	77.18
OC	3851978.88	2692042.52	69.89
SA	16189087.23	10058359.95	62.13

Figure 2.6: Percentage of people fully vaccinated over people vaccinated.

2.4 People Fully Vaccinated per Continent

Linked to the previous analysis, it is interesting to show the percentage of people fully vaccinated per continent.
Considering the data in the column *people_fully_vaccinated_per_hundred* different from -1, it is possible to compute the average grouping by continent.
The dataframe obtained, in descending order, is the following:

continent	people_fully_vaccinated_per_hundred
EU	62.32
SA	55.57
OC	54.18
NAM	50.89
AS	49.83
AF	15.51

Figure 2.7: Percentage of people fully vaccinated per hundred per continent.

The highest percentage of people fully vaccinated is in Europe, followed by South America and Oceania.
The output of this analysis confirm the results showed in figure (2.5), in which it was conceivable that Europe was at the first place for people fully vaccinated.

2.5 Correlation between Countries

In order to study the correlation between countries, a new dataframe must be created: the date is the index, each column represents a country and each entry is the corresponding value of *daily_vaccinations_per_million*. The new dataframe is used to compute the following correlation matrix:

country	Afghanistan	Albania	Algeria	Andorra	Angola	Anguilla	Antigua and Barbuda	Argentina	Armenia	Aruba	...
country											
Afghanistan	1.000000	0.041743	0.353525	-0.267720	0.606442	-0.209482	-0.093813	0.412950	0.680828	-0.355486	...
Albania	0.041743	1.000000	0.312535	0.157270	-0.169886	-0.002751	-0.489581	0.455372	-0.449591	0.342349	...
Algeria	0.353525	0.312535	1.000000	-0.084236	0.026841	-0.289782	-0.088952	0.455014	-0.005229	-0.123389	...
Andorra	-0.267720	0.157270	-0.084236	1.000000	-0.362763	-0.040558	-0.122533	0.194094	-0.401211	0.152608	...
Angola	0.606442	-0.169886	0.026841	-0.362763	1.000000	-0.352112	0.109830	0.246343	0.879480	-0.456657	...
...	...	...	...	...	...	...	...	...	...	...	...
Wales	-0.153134	-0.176941	-0.399380	0.115983	-0.020217	0.344729	0.025435	-0.333167	-0.031931	0.316142	...
Wallis and Futuna	-0.225305	0.117825	-0.175934	-0.161232	-0.214509	0.331481	-0.052721	-0.595835	-0.264783	0.512193	...
Yemen	0.130221	-0.444998	-0.360270	-0.127078	0.466820	-0.047999	0.533115	-0.389102	0.306126	-0.045955	...
Zambia	0.080095	0.024243	-0.148690	-0.133720	0.239696	-0.138665	-0.231540	0.177071	0.406181	-0.374586	...
Zimbabwe	0.210880	0.381521	0.409048	-0.066829	-0.021725	-0.299268	-0.338445	0.667885	0.017374	-0.331542	...

Figure 2.8: Part of the correlation matrix

Starting from the correlation matrix, it is possible to create a dataframe composed by two columns: on the first one the name of the country, on the second one the values of the maximum index of correlation. Then it is useful to filter all the values that are different from 1 and positive.
The top ten of this analysis is the following:

country	max
United Kingdom	0.996626
England	0.996626
Greece	0.959100
Germany	0.959100
France	0.956021
Italy	0.956021
Switzerland	0.955583
Poland	0.955583
Sudan	0.951228
Congo	0.951228

Figure 2.9: Pairs in the top 10 with highest index of correlation.

The highest index of correlation is between England and United Kingdom: since England is part of United Kingdom it is not meaningful to show this correlation. The following plot represents the correlation between Greece and Germany (the second couple), but a similar result could be obtained by plotting France and Italy (the third couple).

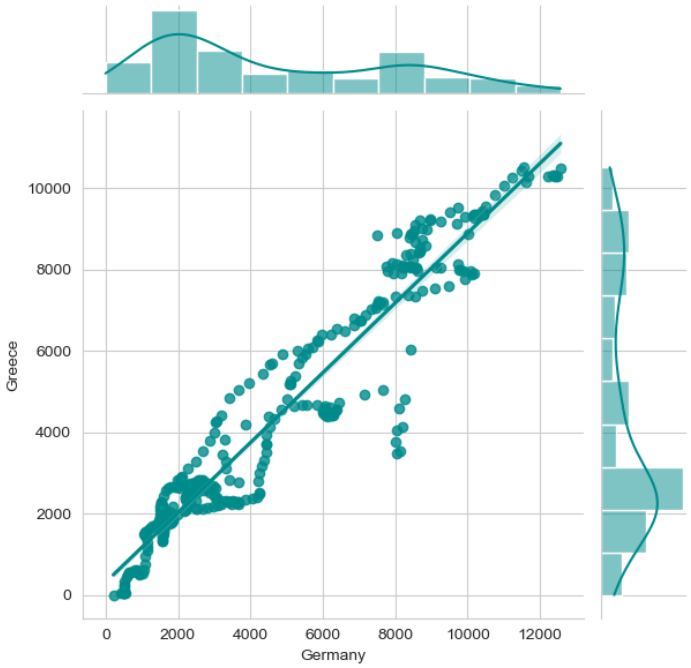


Figure 2.10: Correlation between Germany and Greece.

2.6 Statistics on each countries

The last analysis is about the countries: through a given input it is possible to compute three statistics for the chosen country.
The following results are computed for Italy.

2.6.1 Histogram of daily vaccinations

The first statistics is represented by an histogram in which the mean, the median and the Kernel Density Estimation KDE are highlighted.

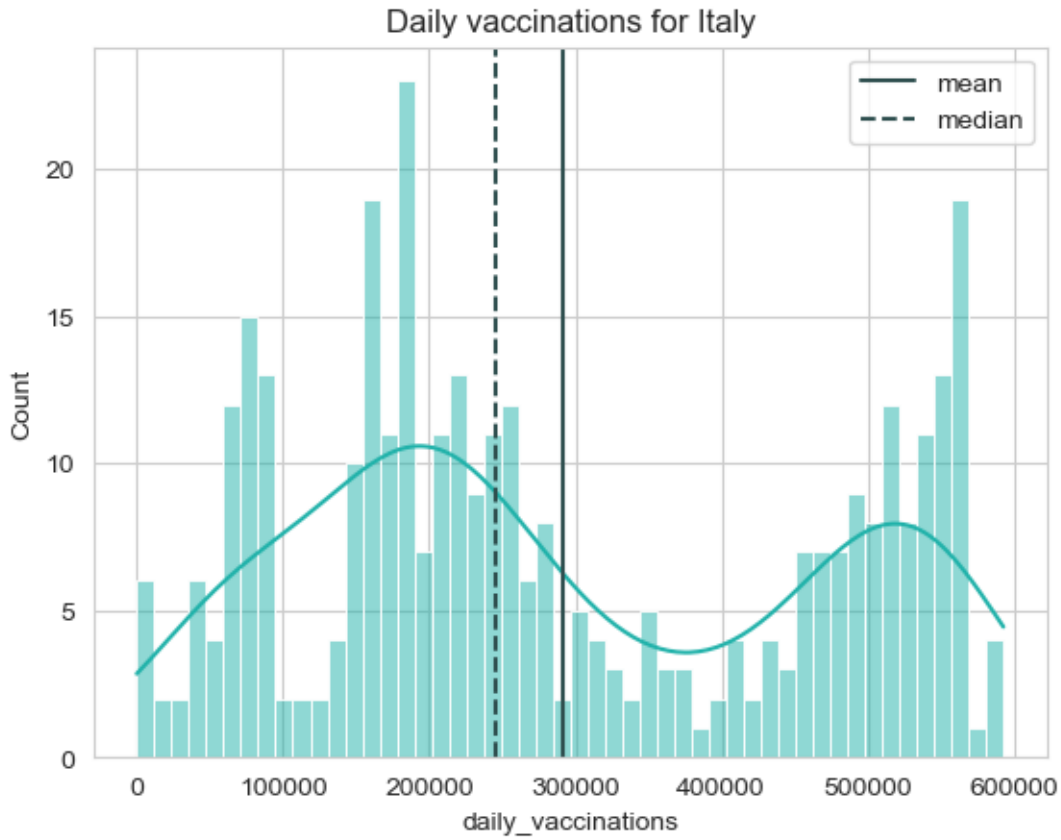


Figure 2.11: Histogram of daily vaccinations in Italy with attention on the mean, the median and the KDE.

2.6.2 KDE of daily vaccinations

The KDE is a type of statistics that allows us to see the distribution of the observations in the dataset: in order to have a better understanding of the plot, it is possible to adjust the smooth of the density curve.

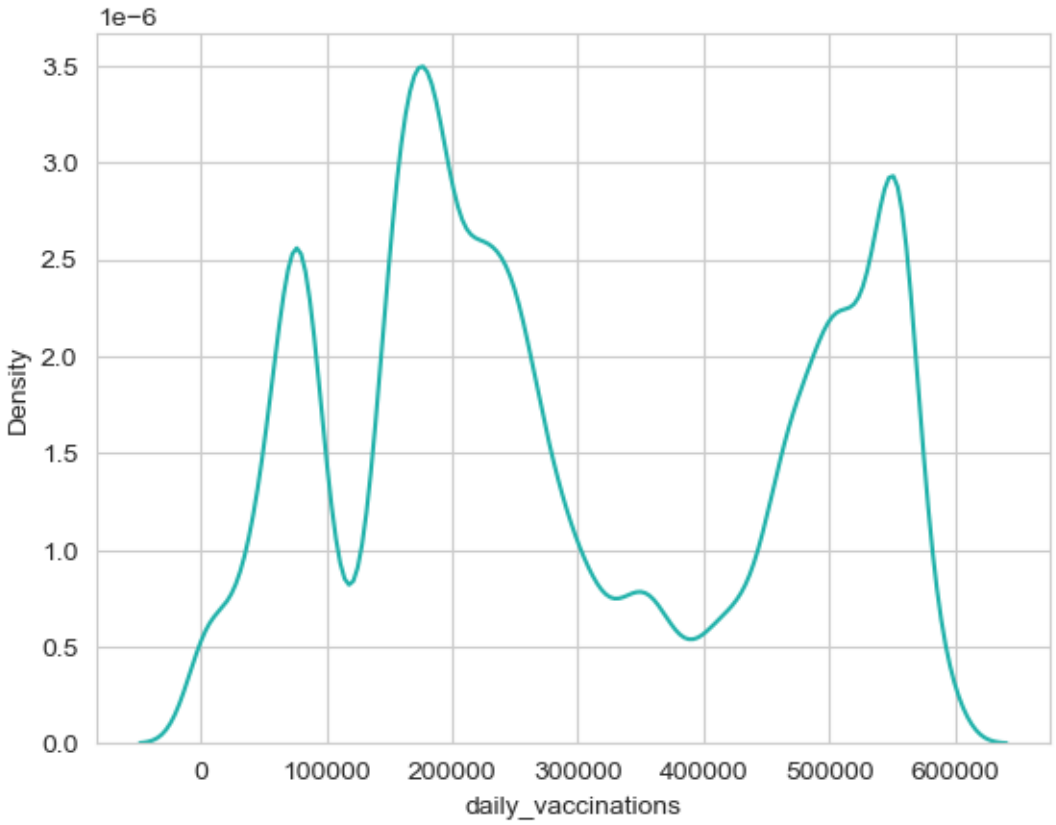


Figure 2.12: KDE of Italy with lower smooth.

A lower value of the smooth creates a more precise representation of the data distribution: in figure (2.12) a smoothing parameter of 0.3 shows the same trend of the histogram plot in figure (2.11).

2.6.3 Barplot of daily vaccinations

Another analysis is obtained by grouping per month the data of daily vaccinations for the chosen country:

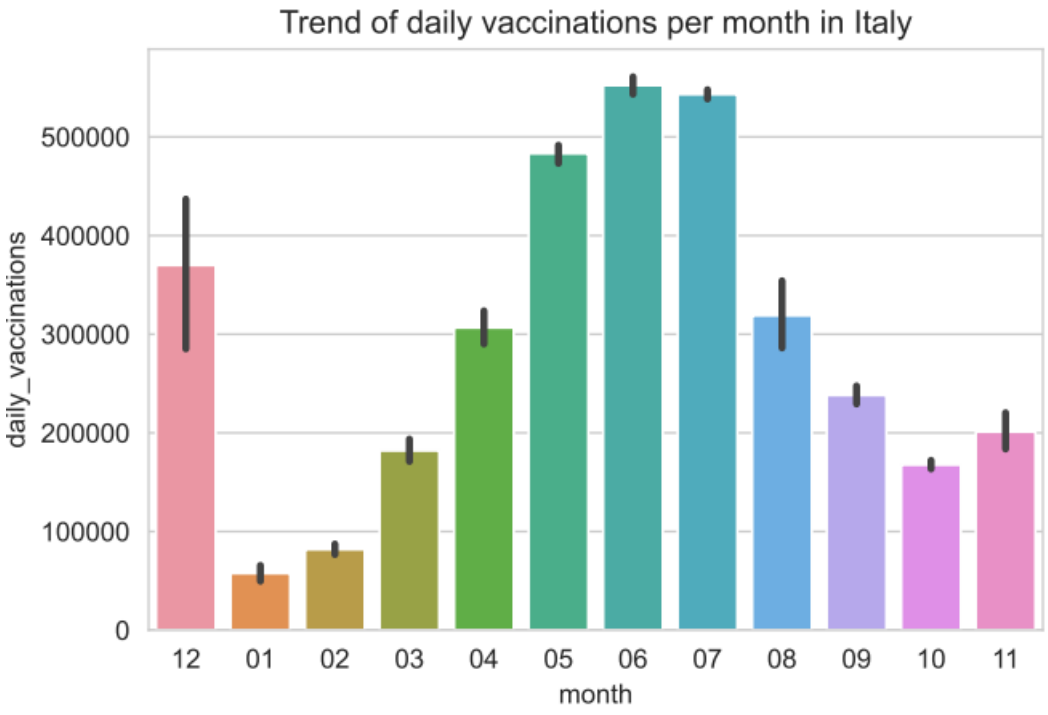
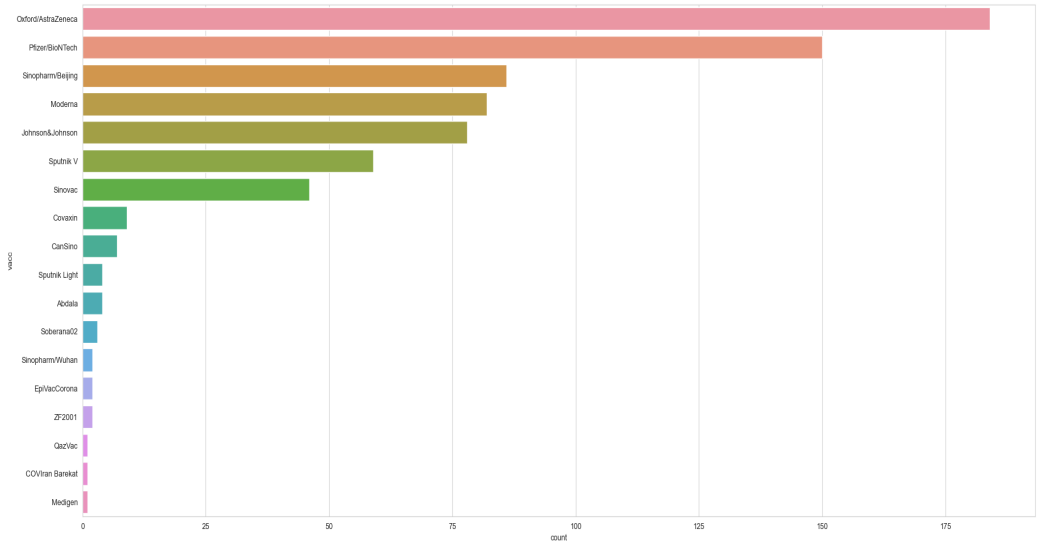


Figure 2.13: Trend per month of daily vaccinations.

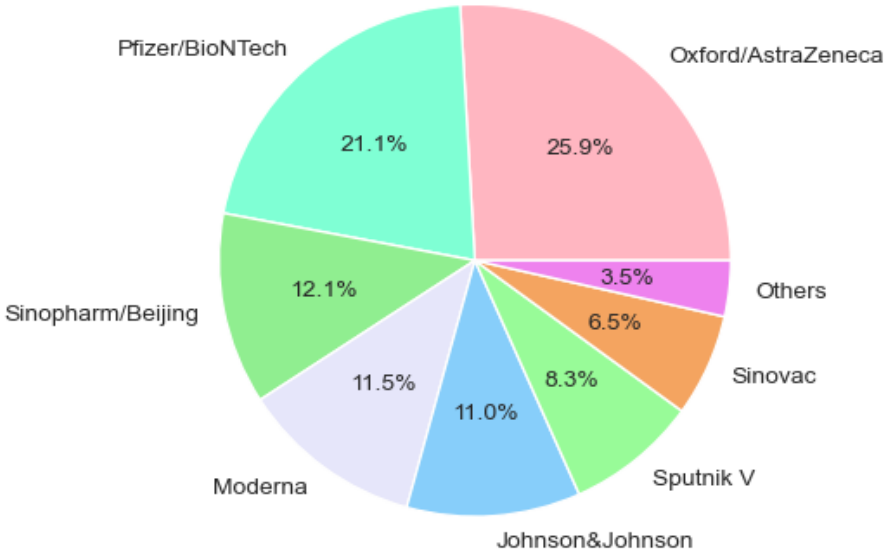
It is possible to see that the highest number of vaccinations in Italy is performed between May and July.
An observation must be done: since the period of time analyzed goes from December 2020 to December 2021, the month of December contains the values of both years.

2.7 Statistics on vaccines

The purpose of this analysis is to find the most used vaccine among countries. The column *vaccines* in the dataset, which contains the brand of vaccines used in each country, has been divided per each different vaccine in order to count in how many countries has been distributed a specific vaccine. The result is showed in the following plot:



It's possible to see that the majority of the countries has dispensed the Oxford/AstraZeneca vaccine, followed by Pfizer/BioNTech, Sinopharm/Beijing, Moderna, Johnson&Johnson, Sputnik V and Sinovac. The other types of vaccines have been dispensed in a more circumscribed way, limited to just few countries.
The same result is illustrated in the following pie chart in which the vaccines that were distributed in less than 10 countries were grouped in the slice *Others*.



It is also relevant to notice that mRNA (Pfizer/BioNTech and Moderna) and Inactive Virus (else) vaccines type have been used approximately in the same number of countries.

Conclusions

The analysis has provided meaningful information about the trend of the first phase of vaccination campaign all over the world.
About the analysis of the daily vaccinations, the results obtained show that China and India are the countries with highest average number of daily vaccinations and this is because of the huge amount of population.
As opposite, the results of the comparison between people that received only one dose and people that completed the vaccination cycle among the continent show that the highest percentage of fully vaccinated people belongs to Europe and North America, where 77% of those who received the first dose completed the cycle, while the lowest percentage belongs to Asia, where the percentage is around the 50%.
Considering the previous results, it is noticeable that the most vaccinated continent is Europe, followed by America, Oceania and Asia, meanwhile Africa has the lowest weight (only 5%).
Another meaningful result is provided by the study of the correlation between countries on the daily vaccination per million people. Analyzing only those countries with the highest correlation, it is noticeable that these couples refers to countries geographically close to each other, such as Germany and Greece, Italy and France and Sudan and Congo.
The same kind of statistical analysis made for continents can be reproduced on a lower scale for a singular country to investigate the trend of daily vaccinations. In the case of Italy, the trend is not linear with peaks of vaccinations in the period between May and July.
A final consideration can be made on the brand of vaccines distributed in each country: Oxford/AstraZeneca and Pfizer/BioNTech are the most distributed ones, reaching almost all countries in the world. This means that mRNA and Inactive Virus technologies have been used on a balanced way to immunized all the population.