National College of Ireland

Project Submission Sheet – 2021/2022

Student Name: Krishnanunni Raju , Nivedita Vishwanath Hiremath , Pramod Ramu , Sai Rajasekhar Reddy Evuri

Student ID: 20232217, 21108471 , 20205759 , 20250151

Programme: MSc. Data Analytics **Year:** 2021

Module: Database and Analytics Programming

Lecturer: Anu Sahni

Submission Due Date: 26-12-2021

Project Title: Data Analysis and Visualization of different trends in EU/EAA Covid-19 Data

Word Count: 3566

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Krishnanunni Raju , Nivedita Vishwanath Hiremath , Pramod Ramu ,Sai Rajasekhar Reddy Evuri

Date: 26-12-2021

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**

5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Data Analysis and Visualization of different trends in EU/EAA Covid-19 Data

Krishnanunni Raju
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x20232217@student.ncirl.ie

Nivedita Vishwanath Hiremath
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x21108471@student.ncirl.ie

Pramod Ramu
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x20205759@student.ncirl.ie

Sai Rajasekhar Reddy Evuri
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x20250151@student.ncirl.ie

Abstract—**In this project, EU/EAA Covid data analysis is performed and provided insights from it. The Covid-19 virus is evolving with numerous variant names, with Omicron being the most recent. It is important to get the characteristics of how the variant is spreading, the rate of vaccination and other attributes like recovery and deaths. In the analysis, we have designed a process flow where data is obtained from EU/EAA Covid website.We have used AWS cloud resources,MongoDB atlas cluster and python programming language for data cleaning ,preprocessing and visualization**

Index Terms—**MongoDB Atlas, AWS S3 bucket, MySQL RDBMS,Python,Pandas,GIT EU/EAA Covid datasets**

I. Introduction

Corona virus is an RNA virus with a large closed structure that is commonly seen in animals and can cause minor damage to human respiratory systems. Later in 2003, it mutated to SARS-CoV, a dangerous virus, and subsequently to SARS-CoV-2, a life-threatening virus, which was detected internationally in January 2020, causing a global pandemic and first appearing in China in December 2019. SARS-CoV-2 has a sub-mutation called Omicron, which has been discovered to be extremely dangerous even after receiving the full vaccination dose, and people are anticipated to receive a booster injection to combat the current version. The individual infected with virus can be a asymptomatic and symptoms which include common cold, fever, highly infectious pneumonia, organ failures, loss of taste and smell and death. The mortality rate for SARS-CoV-2 is more than SARS-CoV mean chances of death is more in SARS-CoV-2 infection. This infection is easily transmitted through air, water, and surfaces, and once infected, the virus replicates quickly. The pandemic phase may last until human bodies create potent antibodies to fight viruses, due to the emergence of new variants with sophisticated complex structures.

We have taken the datasets which are related to different departments of Covid-19 to check vaccination ratios of first dose and second dose by region wise, deaths, ICU in hospitals availability, change in variants. All of these questions prompted us to conduct a thorough investigation of the current impact and potential future revelations. We have collected the date of different data of Covid-19 like deaths occurrences, vaccination, variants and ICU hospital beds availability by region wise and other characteristics. To understand and get the insights of all these we have chosen the mentioned datasets above. Also, to identify some of the research questions like "Which company providing more vaccines shots, which region to supply more of first and second doses and other etc. Weekly testing rate and tests conducted, along with weekly new cases. Tests done in different countries and regions. Also to check corona deaths and calculation of current active cases.

To get the insights of these research questions, we have gone through different parameter of the data. The parameters affect these can be first dose, second dose , regions, vaccine company for finding out the highest number of vaccines by regions and vaccine company popularity.

Also more related research have been made on similar topic with the data which have some insights of what have already happened, what can be noted for further insights were captured from this research. In the below modules , we will discuss in detail methodology used for this research and abstracting all insights which will answer the research questions in the form of visualizations in the end.

II. RELATED WORK

Before proceeding to the analysis part a literature survey was made.In one paper **[1]** the focus was on the exploratory data analysis of different countries based on the factors like confirmed cases, the number of deaths, recovered cases as well as comparing the mortality rate against the recovered rate for more than 221 countries. This study also incorporates machine learning algorithms have been incorporated to evaluate and visualize the rise and count of cases in a given area.

In another paper **[2]** the study was conducted using knowledge discovery in database(KDD) methodology. Here big data visualization and visualization analytics tools have been incorporated for analysis and visualization of Covid-19 data.

In another paper **[3]** advanced visualization techniques have been incorporated to study the impact of Covid-19 in different countries. This study focused on different factors like population, population density, median age, human development index, number of cases, and number of deaths. Also, linear regression and R values were used to showcase observed trends.

In this paper **[4]** machine learning models like k-nearest neighbors(kNN) and linear regression have been incorporated to study and predict the future situation regarding Covid-19.

In another paper **[5]** time series forecasting models have been incorporated for the analysis of Covid-19 data. This paper focused on forecasting the impact of coronavirus on the world and individuals. The intention behind this study is to help society to handle the disease across the country.

This paper **[6]** focused on the classification of Covid-19 patient data using various machine learning algorithms like decision tree, random forest, support vector machines, and k-nearest neighbors.

The objective of this paper **[7]** is to implement exploratory data analysis on the Covid-19 dataset for inferring various information from it. The idea of this study is to understand the patterns and insights of the effect of the pandemic.

In this study **[8]** a spatial data science system has been presented for the analysis of the Covid-19 data where more focus is given on the spatial data analytics across various locations. This study intends to help the users to have a better understanding of related confirmed cases of Covid-19.

III. Methodology

In this project, we have used python programming language exhibited database and analytics procedures we have also used benefits of AWS resources. To start the project we have collected the JSON data from European Centre for Disease Prevention and Control website which is open source**[9]**. These are the datasets we chose to work on –

1) Daily number of new reported Covid-19 cases and deaths by EU/EEA country
2) SARS-CoV-2 variants in the EU/EEA
3) Hospital and ICU admission rates and current occupancy for Covid-19
4) Testing for Covid-19 by week and country
5) Covid-19 vaccination in the EU/EAA

A. Dataset 1:Daily number of new reported Covid-19 cases and deaths by EU/EEA country

This data set contains the data regarding the number of new cases and deaths reported per day and per country in the EU/EEA. This data set encompasses 7991 rows and 11 columns. The variables of this data set are 'dateofreport',' day',' month',' year',' cases',' deaths',' country',' geoid','countrycode','population2020' and 'continent'.

B. Dataset 2: SARS-CoV-2 variants in the EU/EEA

This data set gives information on the total number and percentage of variants by week and country. The data set has 5487 rows and 13 columns.The variables of this dataset are 'country','countrycode','source','newcases','numbersequenced', 'percentcasessequenced','validdenominator', 'variant' and 'numberdetectionsvariant'.

C. Dataset 3: Hospital and ICU admission rates and current occupancy for covid-19

This data set contains the data about the hospitalization and intensive care admission rates and current occupancy for covid-19 by country. The dataset encompasses 3675 rows and 7 columns. The variables of the data set are 'country',' indicator',' date','yearweek',' source', and 'value'.

D. Dataset 4:Testing for covid-19 by week and country

This data set contains the data regarding the test conducted for covid-19 by week and country. The data set has 6533 rows and 13 columns. The variables of the data set are 'country', 'countrycode',' level',' region','regionname','newcases','testsdone', 'population','testingrate','positivityrate', 'testingdatasource' and 'yearweek'.

E. Dataset 5: Covid-19 vaccination in the EU/EAA

This dataset contains the data regarding Covid-19 vaccination in the EU/EEA. The dataset encompasses 4899 rows and 11 columns.The variables of the data set are 'YearWeekISO','FirstDose','FirstDoseRefused', 'SecondDose',DoseAdditional1','UnknownDose', 'NumberDosesReceived','NumberDosesExported', 'Region' ,'TargetGroup'and 'Population'.

F. Process Flow

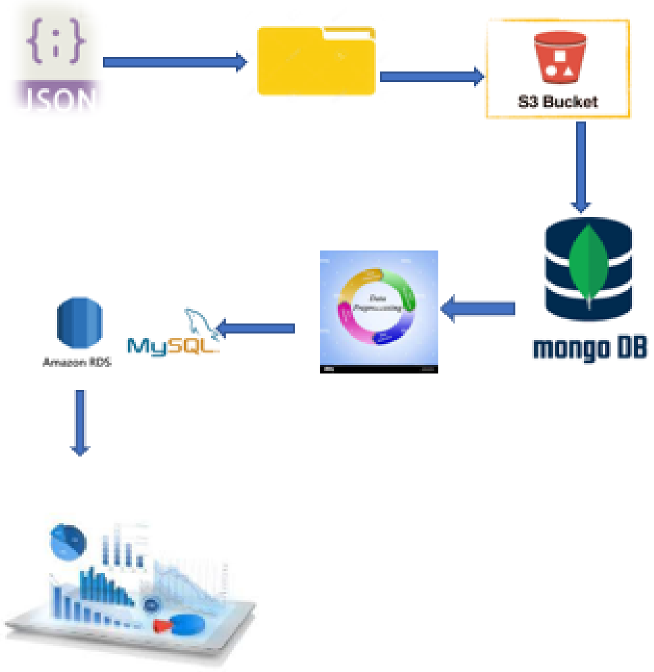

Fig. 1. Process Flow

To store these file we have used Amazon S3 bucket which is a public cloud storage resource. We have created 'ACCESS' key and the' SECRET' key and gave permission to access the Amazon S3 bucket resources. By using 'ACCESS' key and

‘SECRET’ key and python library ‘boto3’ to upload files from local to cloud S3 bucket named ‘covid-19-eu-bucket’. Later we have created a NOSQL Shared MongoDB cluster named ‘Cluster-Covid19’ in MongoDB Atlas to upload files from the S3 bucket. Here each JSON file is stored as a collection in the cluster. The file in the collection are fetched from these clusters and converted to python data frames using python. The data cleaning process was taken of each of the dataframes. Below are the data cleaning steps performed-

1) Removing unuseful columns
2) Renaming columns to useful name
3) Removing empty rows
4) Splitting column of year-week to year and week

These cleaned dataframes are stored in structured AWS MYSQL RDBMS. We have created security group for MYSQL instance creation and in this instance, multiple tables are created for each dataframe table named ‘E_Covid_Deaths’,’ EU_Variants’,’ EU_Hospital_ICU’,’ EU_Tests’,’ EU_Vaccines’.

These tables are accessed by python library ‘mysql.connector’ by giving hostname, username and password.Further, these accessed tables are created to data frames and analysis were done.For Analysis purposes we have plotted various interesting graphs using python libraries names‘plotly’,’matplotlib’,’seaborn’ etc.The entire project updated in GitHub repository. ’https://github.com/niveditavh/DAP_Project/’.

IV. RESULT

The data from AWS MySQL RDBMs was fetched and put in several dataframes using Python. After that, they were used to make visualizations.

1) Daily number of new reported Covid-19 cases and deaths by EU/EEA country: In a time series line graph, the week number was plotted against the number of cases. At the beginning and end of the year, the number of cases skyrocketed. This could be due to a variety of factors. It’s possible that a new strain has spread over the continent at this time, or that it’s the holiday season, when people would rather be outside than inside. The bar plot gives an overview the number of cases in each country. We can see from the graph that France and Germany had the most Covid-19 cases, followed by Italy and Spain. The Schengen Area encompasses all of these countries. The fact that these countries have a larger population could possibly be a factor in the increased number of cases.

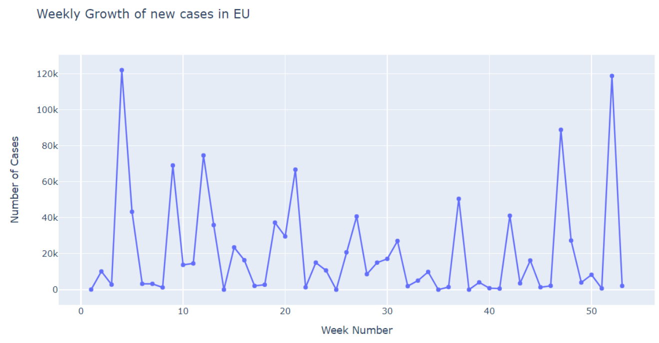

Fig. 2. Week wise Active cases

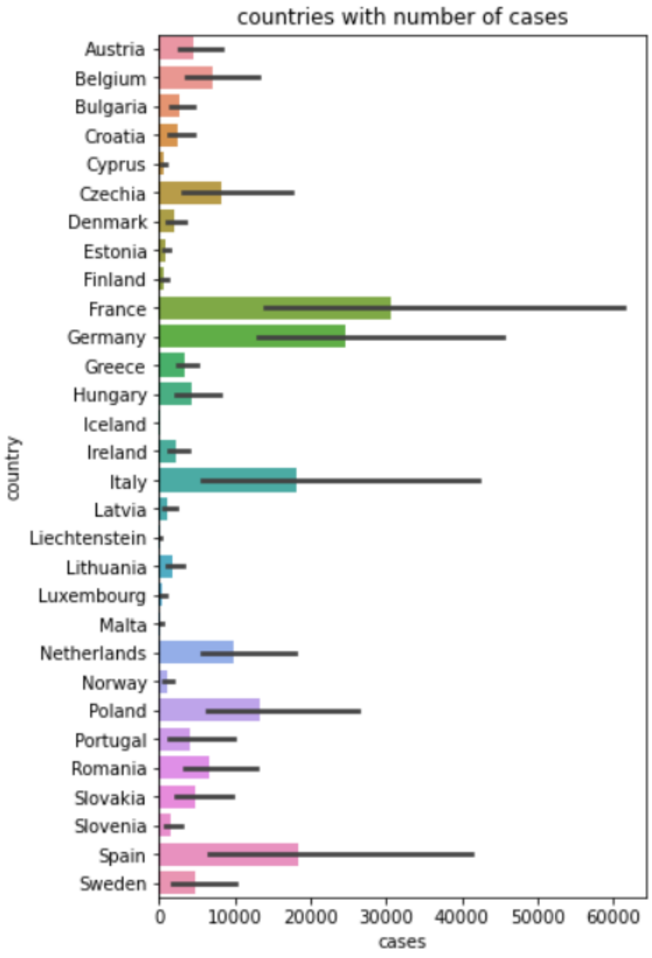

Fig. 3. Countries with Cases

A graph was created using the number of deaths data, and we can see from the graph that the countries with the most cases had the most fatalities, which is logical. A significant finding from this graph is that Poland has a greater percentage of deaths to cases than other countries. According to a 2017 study, high-income citizens in Poland have better access to healthcare than low-income citizens, despite the fact that the majority of the population had insurance**[10]**.

A. *SARS-CoV-2 variants in the EU/EEA*

There were many variants that spread across the continent, according to the Covid-19 variant data and graph. All of the mutations had above 4 million cases reported. The Delta variant, B.1.617, was the most widely spread variant. The spreading rate of this variant was extremely high, as evidenced by the graph. Delta is at least 50% more transmissible than the alpha (Kent) variant, which was first discovered in the United

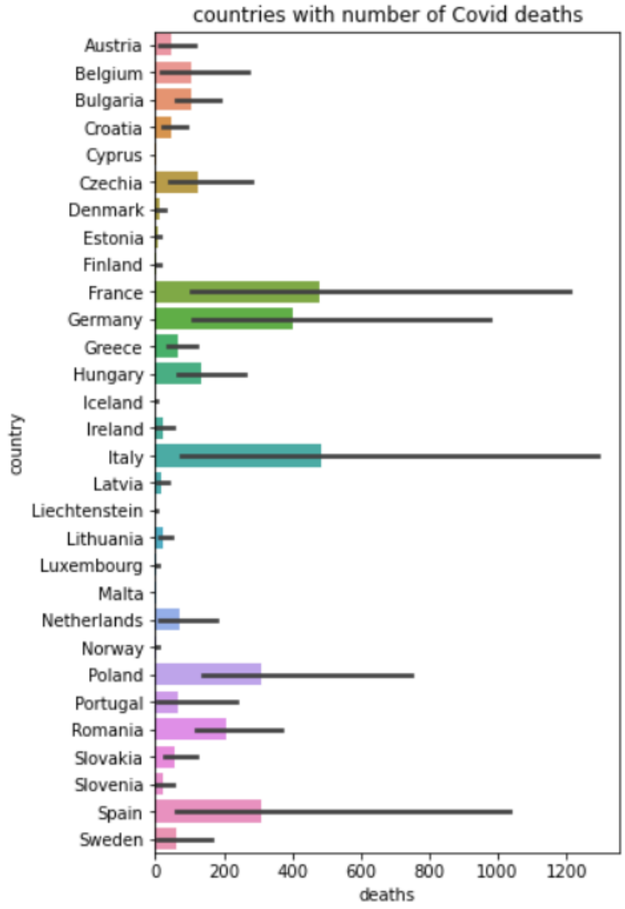

Fig. 4. Countries with Deaths

Kingdom. B.1.617.2 was the original name for Delta. When the World Health Organization adopted a new naming scheme on May 31, 2021, it was called delta. In October of 2020, the delta variant was discovered for the first time in India. We may correctly assert that the timing of detection was prior to the start of vaccination in India**[11]**.

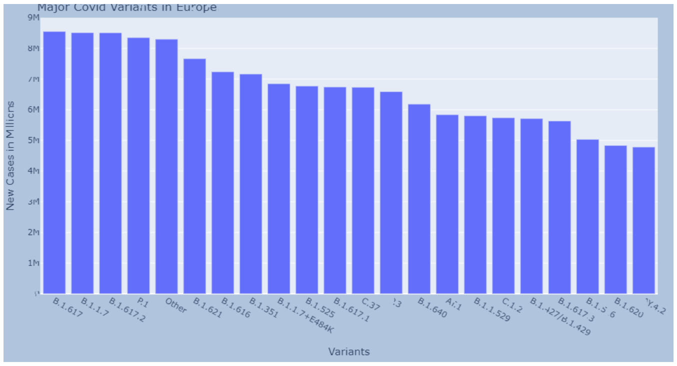

Fig. 5. Countries with Covid variant

B. Hospital and ICU admission rates and current occupancy for Covid-19

This dataset contained information on the number of hospital and ICU cases in various European countries. The number of ICU cases per 100k per week was examined using a bubble chart. Malta was found to be unique in terms of the number of ICU cases per 100,000 people. Malta has a population of only half a million people. In 2021, Malta became the world leader in vaccine rollout, and the country's regulations were extremely rigorous, even after other European governments lifted theirs**[12]**. According to the violin plot of hospital cases reported per 100k of the population, Slovenia and Spain had a very high number of hospital cases reported per 100k of the population in both years. In 2021, the number of hospitalizations in the Netherlands and Portugal was lower than in 2020. Following the first wave, the Netherlands established extremely strict restrictions. In Slovenia, ICU admissions increased dramatically in 2021 compared to 2020. Slovenia has the worst Covid-19 statistic in Europe in 2021, as evidenced by the increase of ICU admissions in that year. Slovenia's vaccination program is moving at a snail's pace. According to statistics from November 2021, hardly half of the population has been vaccinated**[13]**. Slovenia also had the most Covid-19 ICU cases per 100,000 people in the years 2021 and 2020, followed by the Netherlands and Sweden. In comparison to Slovenia, the Netherlands and Sweden have larger populations. Because the Netherlands is one of the most popular tourist destinations, the spread and cases are well understood. Norway

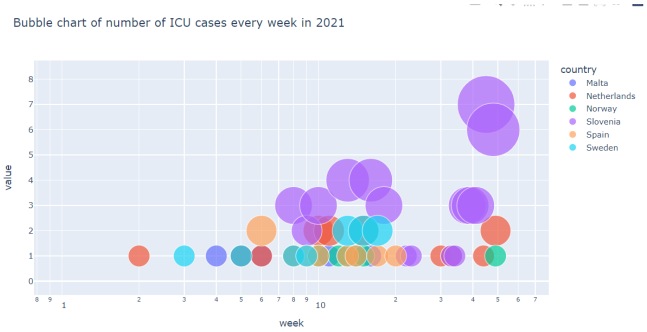

Fig. 6. Bubble chart of ICU cases per 100k every week

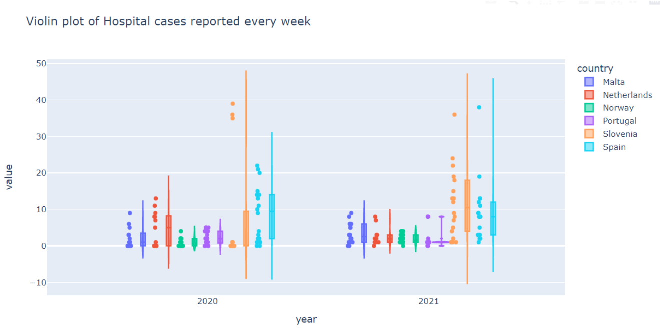

Fig. 7. Violin plot of Hospital cases reported every week in 2020 and 2021

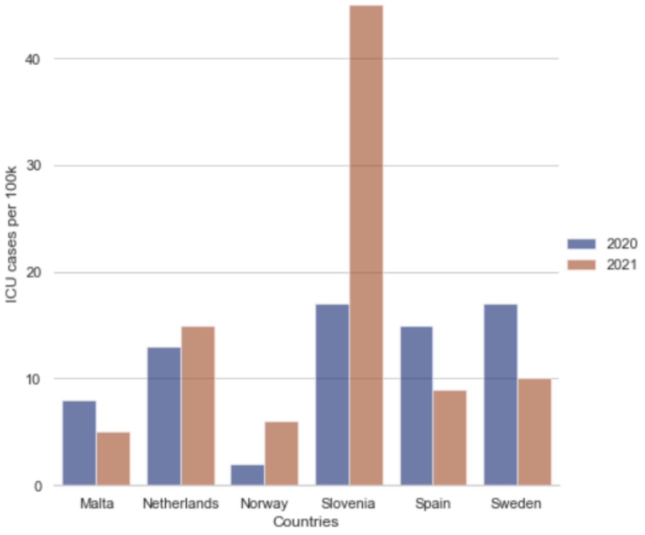

Fig. 8. ICU cases in 2021 vs 2020

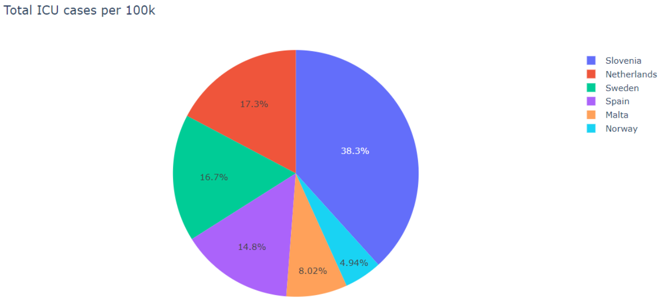

Fig. 9. Total ICU cases reported per 100k

has very few ICU cases, as indicated by the histogram of hospital cases vs. ICU admissions per 100k. Malta has a low number of ICU cases relative to hospitalized cases, which could be attributable to the fact that Malta has a high life expectancy of 80 years**[14]**. Citizens in Malta may have better health and living conditions than people in other countries.

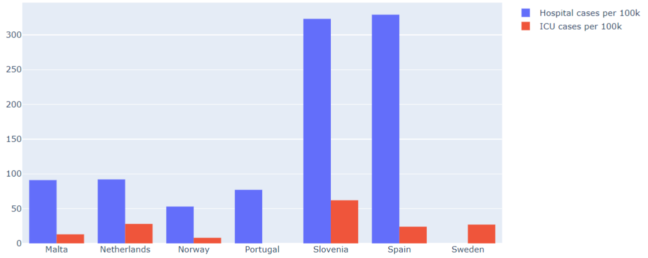

Fig. 10. Hospital cases vs ICU cases

C. Testing for Covid-19 by week and country

The word cloud image gives an insight of different countries which were taken to account for Covid-19 testing. The data of Covid-19 testing provided useful information on the test rate, positivity rate, and new cases, among other things.

Fig. 11. Different countries of Europe

A stacked bar plot showing the number of tests performed in various European countries was created, and it was discovered that Italy performed the most tests, followed by the Netherlands. As other countries with higher number of cases like Germany, Spain and France were not present in the data it was not possible to estimate the tests done in these countries. For example, the number of cases and deaths in Italy was extremely high. This meant that the spread was wide, and thus the testing was extensive.

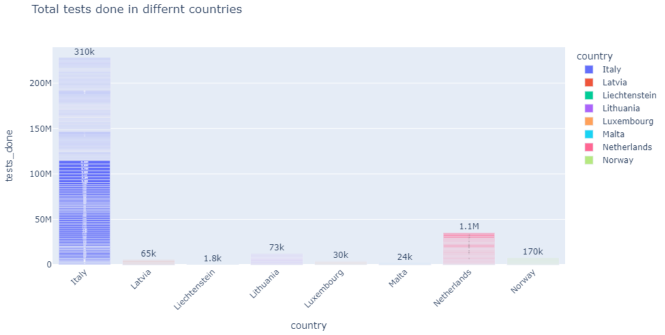

Fig. 12. Tests done in different countries of Europe

From the bar graph of testing rate in different countries it is observed that even though the Netherlands had higher number of cases but the testing rate is very less. The demand for booking a test is very high in the Netherlands. The Netherlands was maxing out its coronavirus testing capacity due to the limitations of the countries health services**[15]**. Italy had the highest testing rate as cases were very high in the country.

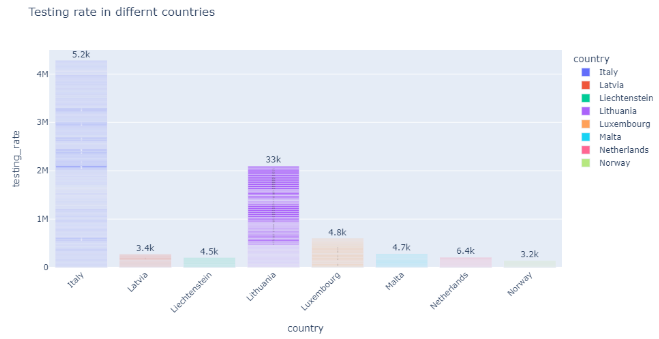

Fig. 13. Testing rate in different countries of Europe

According to the bar graph depicting positivity rates in each country, Italy had the highest positivity rate and the highest testing rate. The efforts of the healthcare professionals were applauded by the general public. In terms of population, all of the other countries in the graph are far smaller. Surprisingly, the rate of test positivity was lowest in Norway.

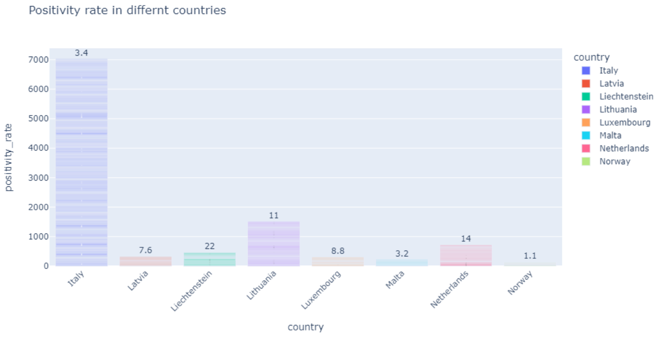

Fig. 14. Positivity rate in different countries of Europe

D. *Covid-19 vaccination in the EU/EAA*

Vaccination provides a brief overview of regions vaccinated by the first and second doses where regions include Austria, Belgium, Bulgaria, Czechia, Finland, Poland, Slovakia. Also there are many vaccine companies that are providing vaccines in different regions like Moderna, Astra zeneca,Novavax, Johnson and Johnson. It is observed that Czechia region is in the top for getting first and second doses. Its observed that Comirnaty of Pfizer/BioNTech is top company that provided first and second doses in EU region. At the end of 2021, world faces most deadly variant name omicron. From the visualization is can be said Czechia could be the region with maximum of getting first and second doses. The vaccine Comirnaty would be the most supplied vaccine booster for omicron

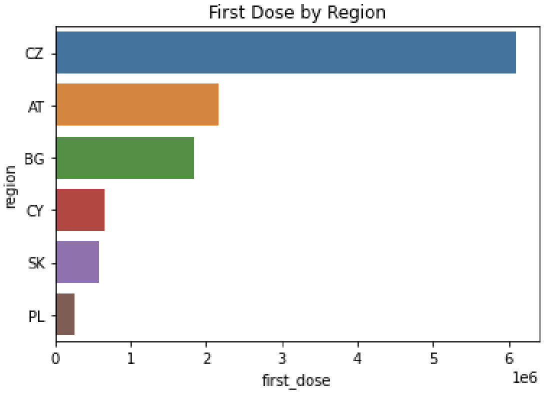

Fig. 15. First dose in Top regions

Bar graph is plotted between First Dose and Region, where it is clear that Czechia region has the highest first doses followed by Austria, Bulgaria, Cyprus, Slovakia and Finland. Its is also observed Austria, Bulgaria has almost similar doses received and Cyprus, Slovakia received almost the same first doses.

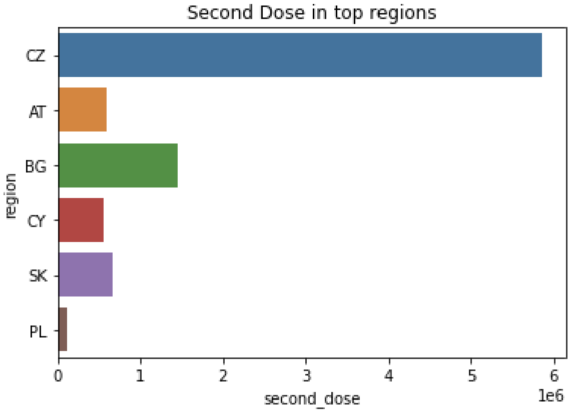

Fig. 16. Second dose in Top regions

Bar graph is plotted between Second Dose and Region, where it is clear that Czechia region has the highest first doses followed by Austria, Bulgaria, Cyprus, Slovakia and Finland. Its is also observed Austria, Bulgaria has differences from the first doses received. Where Bulgaria gets the highest in receiving second dose.

Bar graph is plotted between Vaccine and First Dose, where it is clear that Comirnaty vaccine has the highest first doses followed by AstraZeneca, , Novavax, Johnson and Johnson, Moderna. It is observed that AstraZeneca, Johnson and Johnson, Moderna almost had similarity in providing the first dose.

Bar graph is plotted between Vaccine and Second Dose, where it is clear that Comirnaty vaccine has the highest first doses followed by AstraZeneca, , Novavax, Johnson and Johnson, Moderna. It is observed that AstraZeneca, Moderna almost had similarity in providing the second dose and Johnson and Johnson was not used. Sputnik developed

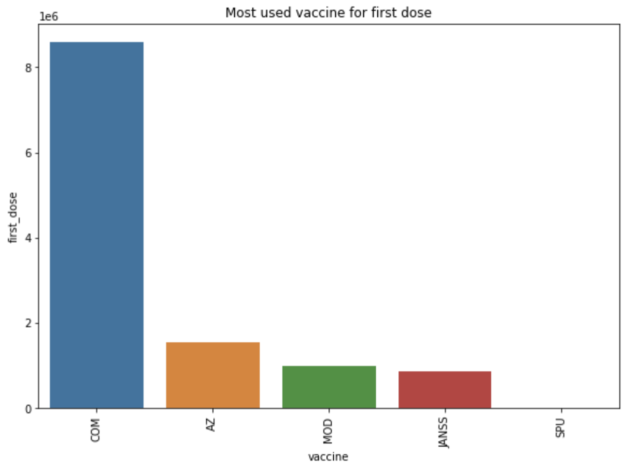

Fig. 17. Most used vaccine in Top regions for First Dose

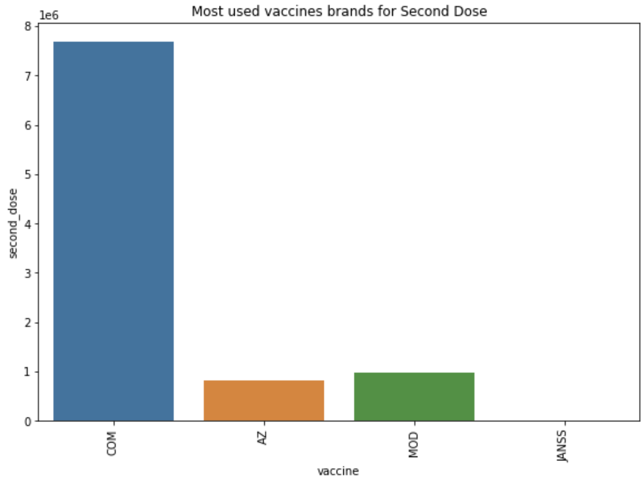

Fig. 18. Most used vaccine in Top regions for Second Dose

by Russia is the vaccine that is used in minimal for first and second doses.

V. Conclusion and Future Work

The Covid-19 data was analyzed, and it was discovered that the number of cases in the Schengen region of Europe was extremely high. The number of Covid-19 cases increased dramatically at the beginning and end of the year. Despite the fact that hospitalization rates were relatively high in several of the countries, ICU admission rates were very low. Testing was quite low in several countries compared to cases, and positivity rate followed a similar trend. The Pfizer vaccination was found to be the most popular in Europe, whereas Sputnik was the least popular.In future the parameters identified in these datasets can subsequently be used to create machine learning and deep learning models that predict coronavirus cases and mutations, as well as which regions are most likely to be affected.

References

[1] S. K. Saini, V. Dhull, S. Singh and A. Sharma, ”Visual Exploratory Data Analysis of COVID-19 Pandemic,” 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), 2020, pp. 1-6, doi: 10.1109/ICRAIE51050.2020.9358331.

[2] C. K. Leung, Y. Chen, C. S. H. Hoi, S. Shang, Y. Wen and A. Cuzzocrea, ”Big Data Visualization and Visual Analytics of COVID-19 Data,” 2020 24th International Conference Information Visualisation (IV), 2020, pp. 415-420, doi: 10.1109/IV51561.2020.00073.

[3] H. Raj and R. K. Mishra, ”Data Analysis of Novel Coronavirus Based on Multiple Factors,” 2020 Seventh International Conference on Information Technology Trends (ITT), 2020, pp. 135-139, doi: 10.1109/ITT51279.2020.9320887.

[4] A. Abdullha and S. Abujar, ”COVID-19: Data Analysis and the situation Prediction Using Machine Learning Based on Bangladesh perspective,” 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), 2020, pp. 1-8, doi: 10.1109/iSAI-NLP51646.2020.9376812.

[5] S. Maurya and S. Singh, ”Time Series Analysis of the Covid-19 Datasets,” 2020 IEEE International Conference for Innovation in Technology (INOCON), 2020, pp. 1-6, doi: 10.1109/INOCON50539.2020.9298390

[6] I. M. Putra, I. Tahyudin, H. A. Awal Rozaq, A. Yahya Syafa’at, R. Wahyudi and E. Winarto, ”Classification Analysis of COVID19 Patient Data at Government Hospital of Banyumas using Machine Learning,” 2021 2nd International Conference on Smart Computing and Electronic Enterprise (ICSCEE), 2021, pp. 271-274, doi: 10.1109/ICSCEE50312.2021.9498020.

[7] J. DSouza and S. Velan S., ”Using Exploratory Data Analysis for Generating Inferences on the Correlation of COVID-19 cases,” 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-6, doi: 10.1109/ICCCNT49239.2020.9225621.

[8] S. Shang, C. K. Leung, Y. Chen and A. G. M. Pazdor, ”Spatial Data Science of COVID-19 Data,” 2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2020, pp. 1370-1375, doi: 10.1109/HPCC-SmartCity-DSS50907.2020.00177.

[9] European Centre for Disease Prevention and Control. 2021. Download COVID-19 datasets. [online] Available at: ¡https://www.ecdc.europa.eu/en/covid-19/data¿ [Accessed 15 December 2021].

[10] “An Overview of Healthcare in Poland.” The Borgen Project, 19 Aug. 2020, borgenproject.org/healthcare-in-poland/.

[11] Page, Michael Le. “Indian Covid-19 Variant (B.1.617).” New Scientist, www.newscientist.com/definition/indian-covid-19-variant-b-1-617/.

[12] “Malta Cannot Learn from Other Countries’ COVID Mistakes - Fearne.” Times of Malta, timesofmalta.com/articles/view/malta-cannot-learn-from-other-countries-covid-mistakes-chris-fearne.879284. Accessed 26 Dec. 2021.

[13] Maček, Sebastijan R. “Slovenian ICU Wards Overflowing as Cases Continue to Surge.” Www.euractiv.com, 15 Nov. 2021, www.euractiv.com/section/politics/short_news/slovenian-icu-wards-overflowing-as-cases-continue-to-surge/. Accessed 26 Dec. 2021.

[14] “Average Life Expectancy by Country.” Worlddata.info, 2017, www.worlddata.info/life-expectancy.php.

[15] Moses, Claire. “The Netherlands Is Maxing out Its Coronavirus Testing Capacity.” The New York Times, 17 Nov. 2021, www.nytimes.com/2021/11/17/world/europe/covid-testing-netherlands.html. Accessed 26 Dec. 2021.