APPLIED DATA SCIENCE

COVID-19 VACCINES ANALYSIS USING DATA SCIENCE

PHASE-1

DATA SCIENCE

Data science is an interdisciplinary field that encompasses a set of techniques, processes, and methods for extracting valuable insights, knowledge, and patterns from data. It combines elements of statistics, computer science, Domain expertise, and data engineering to collect, analyze, and interpret large and complex datasets. The ultimate goal of data science is to use data to inform decision-making, solve problems, and drive improvements in various domains, including business, healthcare, finance, and more. Data scientists use a combination of data analysis, machine learning, data visualization, and domain-specific expertise to uncover meaningful information from data and provide valuable insights for organizations and individuals.

TOOLS USED

1. ***Python***

   Python is often used as a support language for software developers, for build control and management, testing, and in many other ways.

2. ***Python idle***

   IDLE can be used to execute a single statement and create, modify, and execute Python scripts. IDLE provides a fully-featured text editor to create Python scripts that include features like syntax highlighting, auto completion, and smart indent.

3. ***PyCharm***

   PyCharm is a dedicated Python Integrated Development Environment (IDE) providing a wide range of essential tools for Python developers, tightly integrated to create a convenient environment for productive Python, web, and data science development.

4. ***A Kaggle dataset***

   It also known as a Kaggle Kernel is a set of data provided by companies, students, and alum.These data sets allow competitors to work through problems, or use them as a practice simulation.

5. ***MATLAB***

In Data Science, MATLAB is used for simulating neural networks and fuzzy logic. Using the MATLAB graphics library, you can create powerful visualizations. MATLAB is also used in image and signal processing.

6. ***Excel***

Excel is a powerful analytical tool for Data Science. While it has been the traditional tool for data analysis, Excel still packs a punch. Excel comes with various formulae, tables, filters, slicers, etc. You can also create your own custom functions and 2 formulae using Excel. While Excel is not for calculating the huge amount of Data, it is still an ideal choice for creating powerful data visualizations and spread sheets.

7. ***Jupyter***

Project Jupyter is an open-source tool based on IPython for helping developers in making open-source software and experiences interactive computing. Jupyter supports multiple languages like Julia, Python, and R.

8. ***Matplotlib***

Matplotlib is a plotting and visualization library developed for Python. It is the most popular tool for generating graphs with the analyzed data. It is mainly used for plotting complex graphs using simple lines of code. Using this, one can generate bar plots, histograms, scatterplots etc.

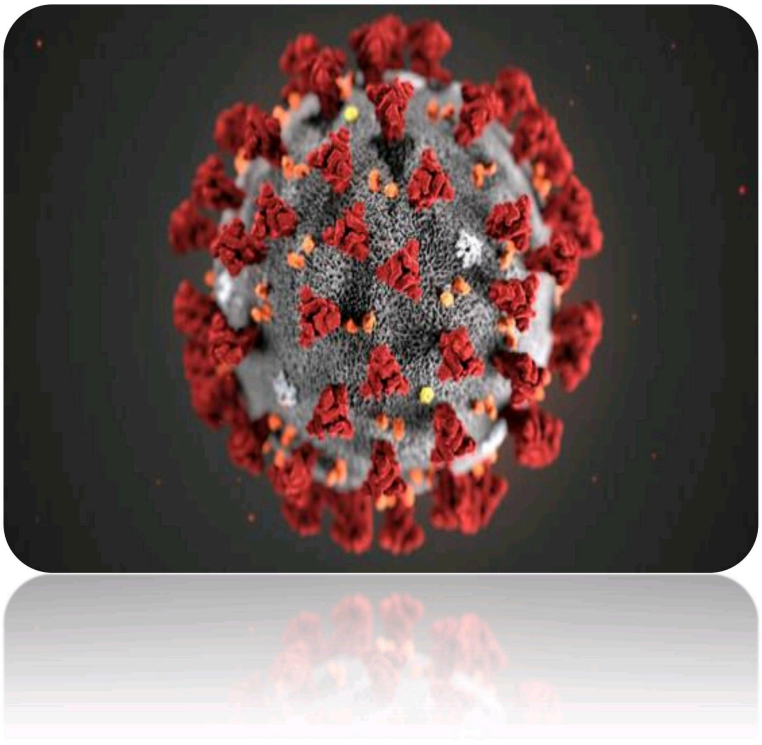
9. ***Sciket-learn***

Scikit-learn is a library-based in Python that is used for implementing Machine Learning Algorithms. It is simple and easy to implement a tool that is widely used for analysis and data science. Scikit-learn makes it easy to use complex machine learning algorithms. It is therefore in situations that require rapid prototyping and is also an ideal platform to perform research requiring basic Machine Learning. It makes use of several underlying libraries of Python such as SciPy, Numpy, Matplotlib, etc.

INTRODUCTION

- The COVID 19 pandemic is a disaster in the health sector that has hit the whole globe by reaching over 2.7 Million of death toll. With the advent of the Covid19 vaccines, estimating the vaccination rate is an important issue. Most of the vaccines developed by scientists often cause curiosity regarding the manufacturing process and the process of distributing these vaccines. This is due to the large number of researchers, practitioners, statisticians, and medical personnel who have tried to keep up with the spread of the virus in various countries in the world using various methods.

- The data is processed to show some important information, such as countries that started vaccination for the first time and countries with the highest vaccinations. In addition, the types of vaccines offered and used by countries in the world. The prediction of the number of vaccines to be given is also presented using data science approach, i.e., linear regression. This problem is presented in the following arrangement.



PROBLEM DEFINITION

- The problem is to conduct an in-depth analysis of COVID-19 vaccine data, focusing on vaccine efficiency, distribution, and adverse effects. The goal is to provide insights that aid policymakers and health organizations In optimizing vaccine deployment strategies. This project involves data Collection, data preprocessing, exploratory data analysis, statistical analysis, And visualization.

DESIGN THINKING

1. *DATA COLLECTION:*

- COVID-19 vaccine data sets are collected from **kaggle** which does collect the data from health organisations, government databases, and research publications.

LINK: **https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress**

2. *DATA PREPROCESSING:*

- The downloaded raw data is read and collected based on several new fields, namely **country, Iso_code, date, Total number of vaccines, total number of people vaccinated, total number of people fully vaccinated, daily vaccinations, total vaccinations per hundred, vaccines used in the country and the vaccine, etc..** which is the vaccine scheme used in a particular country. At this stage, an important procedure is carried out, namely **DATA CLEANING**, which is an important step for data analysis.

- The case that often occurs is the appearance of the Not-A-Number (NaN) value which can be resolved by changing the value to 0. In addition, empty rows marked with a value of 0 and also repeating columns can be resolved by deleting the column manually, directly, or by deleting unwanted rows.

DATA SET:

country_vaccinat...y_manufacturer

country_vaccinations_by_...

	A	B	C	D	E	F	G	H	I	J	K	L
1	location	date	vaccine	total_vaccinations								
2	Argentina	2/29/2020	Moderna	2								
3		2/29/2020	Oxford/A	3								
4	Argentina	2/29/2020	Sinopharm	1								
5	Argentina	2/29/2020	Sputnik V	20481								
6	Argentina	2/30/2020	Moderna	2								
7	Argentina	2/30/2020	Oxford/A	3								
8	Argentina	2/30/2020	Sinopharm	1								
9	Argentina	2/30/2020	Sputnik V	40583								
10	Argentina	2/31/2020	Moderna	2								
11	Argentina	2/31/2020	Oxford/A	3								
12	Argentina	2/31/2020	Sinopharm	1								
13	Argentina	2/31/2020	Sputnik V	43388								
14	Argentina	1/1/2021	Moderna	2								
15	Argentina	1/1/2021	Oxford/A	5								
16	Argentina	1/1/2021	Sinopharm	1								
17	Argentina	1/1/2021	Sputnik V	43513								
18	Argentina	1/2/2021	Moderna	2								
19	Argentina	1/2/2021	Oxford/A	6								
20	Argentina	1/2/2021	Sinopharm	1								
21	Argentina	1/2/2021	Sputnik V	46824								
22	Argentina	1/3/2021	Moderna	2								
23	Argentina	1/3/2021	Oxford/A	6								
24	Argentina	1/3/2021	Sinopharm	1								
25	Argentina	1/3/2021	Sputnik V	47266								
26		1/4/2021	Moderna	2								
27	Argentina	1/4/2021	Oxford/A	6								
28	Argentina	1/4/2021	Sinopharm	1								
29	Argentina	1/4/2021	Sputnik V	57726								
30	Argentina	1/5/2021	Moderna	2								
31	Argentina	1/5/2021	Oxford/A	6								
32	Argentina	1/5/2021	Sinopharm	5								
33	Argentina	1/5/2021	Sputnik V	68445								
34	Argentina	1/6/2021	Moderna	2								
35	Argentina	1/6/2021	Oxford/A	6								
36	Argentina	1/6/2021	Sinopharm	8								
37	Argentina	1/6/2021	Sputnik V	78551								
38	Argentina	1/7/2021	Moderna	2								
39	Argentina	1/7/2021	Oxford/A	7								
40	Argentina	1/7/2021	Sinopharm	8								
41	Argentina	1/7/2021	Sputnik V	91771								
42	Argentina	1/8/2021	Moderna	2								
43	Argentina	1/8/2021	Oxford/A	7								
44	Argentina	1/8/2021	Sinopharm	8								
45	Argentina	1/8/2021	Sputnik V	124452								
46	Argentina	1/9/2021	Moderna	2								
47	Argentina	1/9/2021	Oxford/A	7								
48	Argentina	1/9/2021	Sinopharm	9								

HANDLING MISSING VALUES:

- Identify missing data in your dataset.
- Decide how to handle missing values :
  Either by removing rows with missing data,
  filling them with a default value(e.g. ,mean, median ,or mode),
  or using more advanced techniques like interpolation.

ENCODING CATEGORICAL FEATURES:

- Convert categorical variables into numerical representations that machine learning models can understand.
- For nominal variables(categories without a specific order), you can use one-hot encoding or label encoding.
- For ordinal variables(categories with a meaningful order),use ordinal encoding.

FEATURE SCALING:

- Normalize or standardize numerical features to bring them to a similar scale.
- Common techniques include Min-Max scaling(scaling to a specific range) or Z-score normalization (scaling to have mean=0 and standard deviation=1).

HANDLING OUTLIERS:

- Identify and decide how to deal with outliers.
- You can choose to remove them or transform them using techniques like log transformation.

FEATURE ENGINEERING :

- Create new features or modify existing ones to better represent the underlying patterns in the data.
- This can involve creating interaction terms, polynomial features, or other domain- specific transformations.

DATA SPLITTING:

- Split your data set into training, validation, and test sets to evaluate your model's performance properly.

3. ***EXPLORATORY DATA ANALYSIS:***

DATA OVERVIEW:

- Begin the EDA by providing a summary of the dataset, including its size, Number of features, and data types.
- Highlight any key variables of interest that will be central to the analysis.

DESCRIPTIVE STATISTICS:

- Calculate and present summary statistics for numerical features.
- This should include measures of central tendency(mean, median) And dispersion(standard deviation, range).For categorical variables, provide Frequency counts and percentages.

DATA DISTRIBUTION:

- Visualize the distribution of key variables using appropriate plots and Charts.
- Histograms, box plots, and density plots can reveal the underlying distribution patterns and potential outliers.

EXPLORATORY VISUALIZATION:

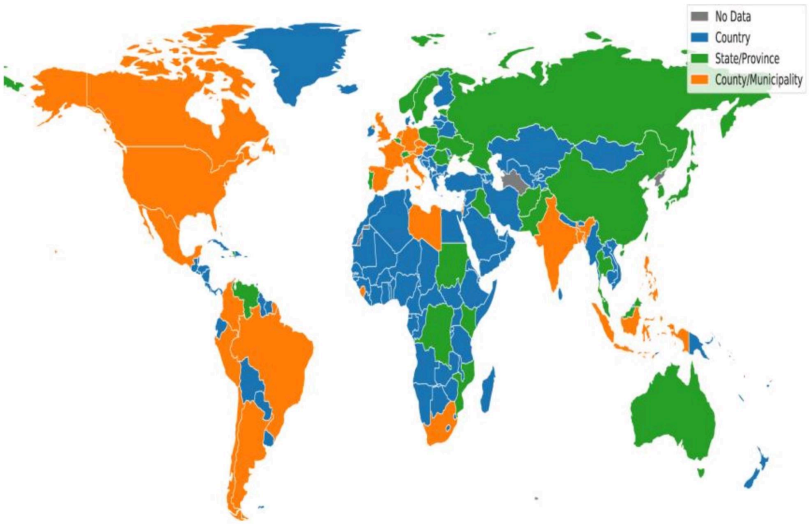
- Create visualizations that provide insights into the data.
- Examples:

- Time series plots:
  If applicable, plot trends over time related to vaccine distribution, case numbers, or public sentiment.
- Word clouds:
  Visualize frequently occurring words in text data to identify common themes.
- Scatter plots:
  Explore relationships between variables, such as vaccine coverage and disease incidence.

4. ***DATA VISUALIZATION:***

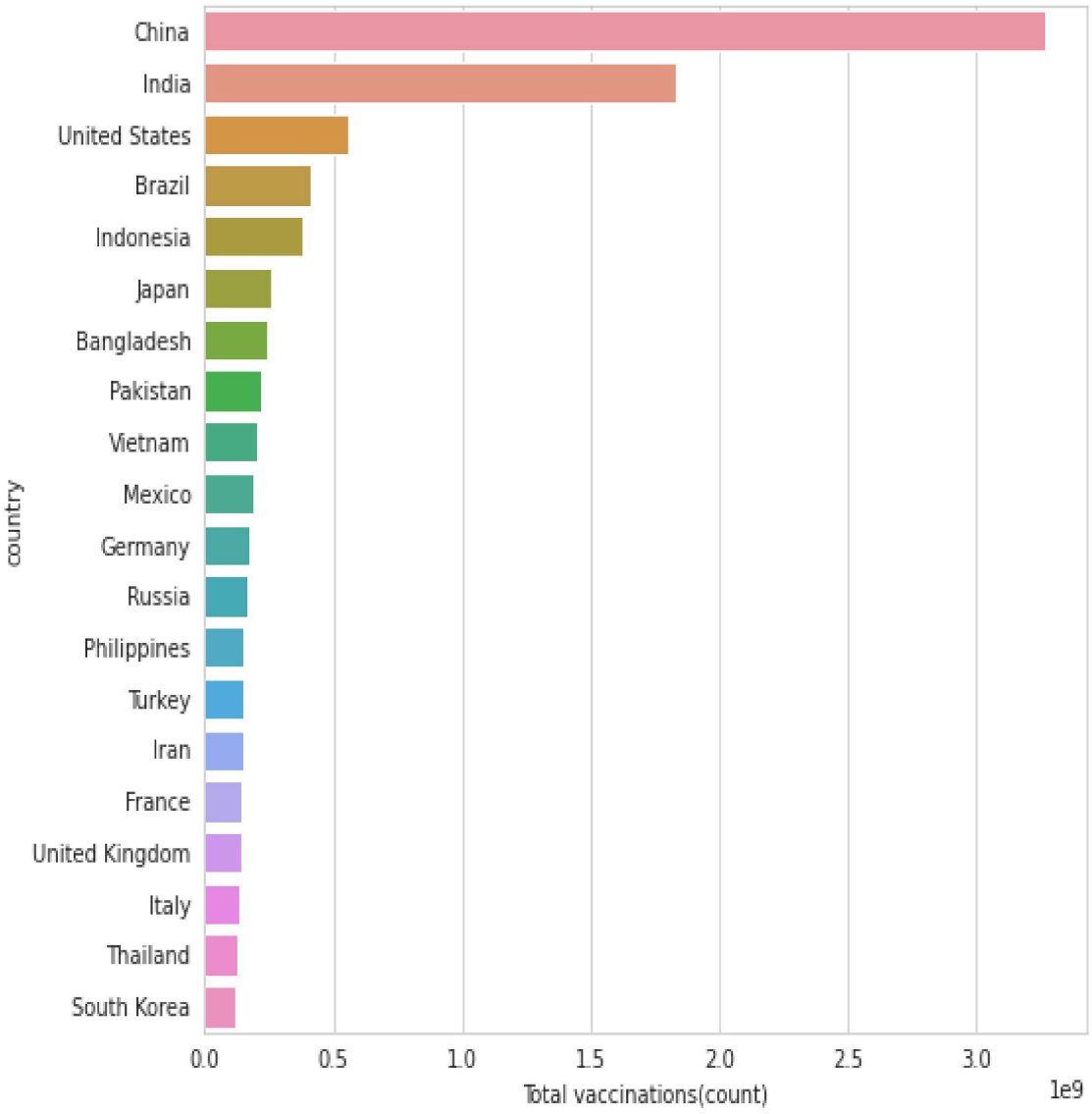
- Data visualization is the representation of data through use of common Graphics ,such as charts, plots, info graphics, and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

- This paper provides a data visualization and analysis of the COVID-19 vaccination program. Important information such as which countries have the highest vaccination rates and numbers. In addition to the types of vaccines used and used by countries in the world, an info graphic on the geographic distribution of vaccine use is also shown. To model the obtained data, daily vaccination rates were modeled by linear regression in which five sample countries with different vaccination ranges were processed using data science approach, namely, linear regression. The modeling results show a gradient coefficient that represents an increase in vaccine rates. The prediction results showed that the highest rate of increase in daily vaccination was 1,826,126 additional vaccines per day.



CODE:

```
x= df.groupby("country")["Total_vaccinations(count)"].mean().sort_values(ascending=
False).head(20)
sns.set_style("whitegrid")
plt.figure(figsize= (8,8))
ax= sns.barplot(x.values,x.index)
ax.set_xlabel("Total vaccinations(count)")
plt.show()
```



5. ***INSIGHTS*** :

- **Insight Generation**
  Summarize the most significant findings and trends identified during the analysis. Highlight insights related to vaccine efficacy, safety, distribution, public sentiment, and disease impact.

- **Identifying Patterns**
  Discuss any recurring patterns or correlations uncovered in the data. Explain their implications in the context of COVID-19 vaccines.

- **Public Perception**
  Analyze public sentiment trends and the factors influencing vaccine hesitancy or acceptance. Consider insights from social media data and surveys.

- **Vaccine Variants**
  Assess the adaptability of existing vaccines to emerging variants of the virus and implications for future vaccination efforts.

- **Hypothesis Validation**
  Review and validate hypotheses formulated during the analysis phase. Discuss the extent to which the data and analysis support or refuse these Hypotheses.

6. ***RECOMMENDATIONS:***

- **Public Health Recommendations**
  Develop recommendations for public health authorities, health care providers, and policymakers.
  These recommendations may include vaccination strategies, communication plans, and targeted interventions.

- **Vaccine Distribution Strategies**
  Offer insights into optimizing vaccine distribution to ensure equitable access, especially in underserved populations or regions.

- **Safety Monitoring**
  Recommend strategies for ongoing safety monitoring and reporting of adverse events associated with COVID-19 vaccines.

- **Future Research Directions**
  Suggest areas for future research and analysis, such as the evaluation of booster shots, long-term vaccine impact, and vaccine adaptability to new variants.

- **Actionable Insights**
  Emphasize actionable insights that can lead to tangible improvements in vaccination campaigns, public perception, and overall COVID-19 response efforts.

- **Communication**
  Prepare clear and concise reports, presentations, or documents that convey the insights and recommendations effectively to various stakeholders

7. ***CONCLUSION***

Data visualization and analysis of the COVID-19 vaccination program are provided. Important information such as which countries achieve the highest vaccination rates and numbers. In addition to the types of vaccines used and used by countries in the world, an info graphic on the geographic distribution of vaccine use is also shown. To model the data held, daily vaccination rates were modeled by linear regression in which five sample countries with different vaccination ranges were processed using linear regression. The modeling results show a gradient coefficient that represents an increase in vaccine rates. The prediction results showed that the highest rate of increase in daily vaccination was 1826126 additional vaccines per day achieved by the United States.