BIG DATA SCIENCE ON COVID-19 DATA

IBRAHIM T , 20

S_7 B.Tech CSE

Government Engineering College, Wayanad

January 7, 2022

Overview

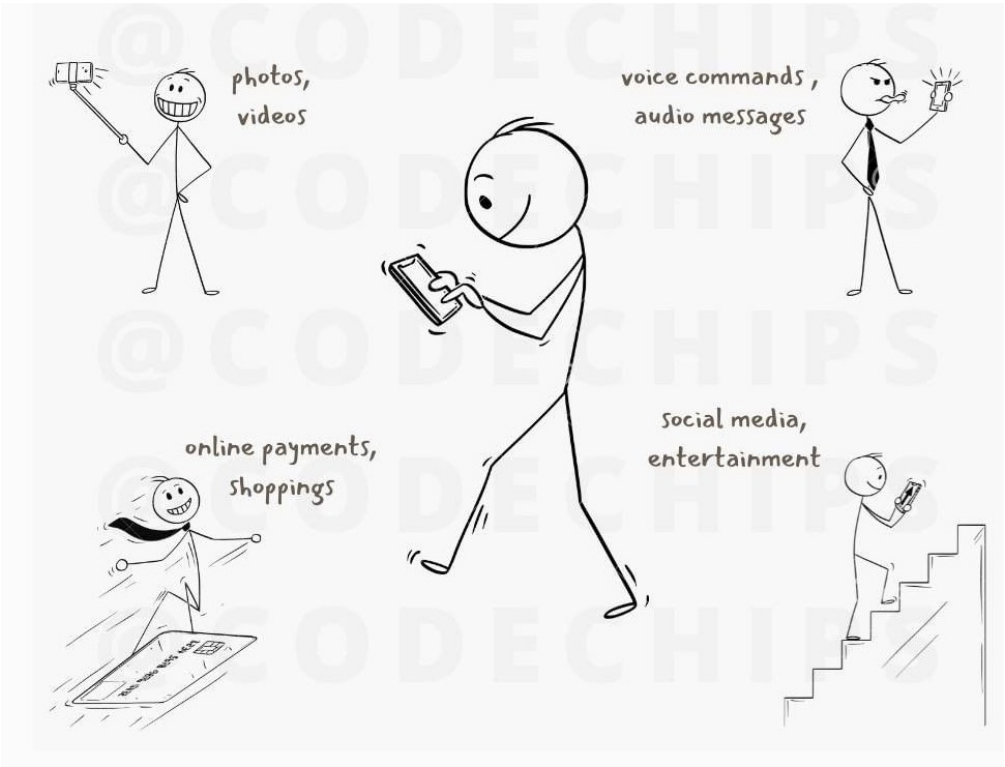
Introduction

- In earlier days processable data can be easily collected

- Data easily collected and stored in excel sheets
- But, it's not same now.Data incresing day by day

Introduction

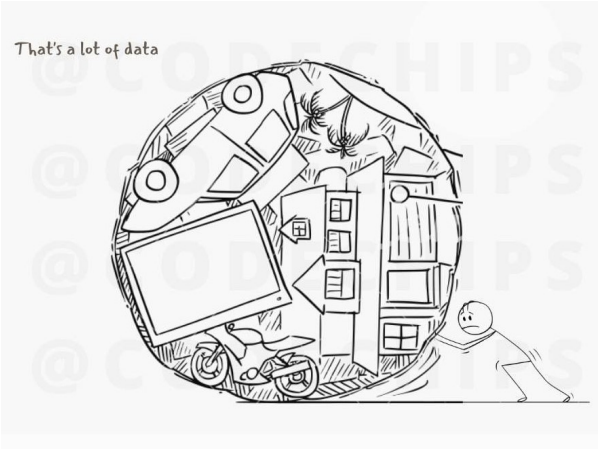
- Data of different format collected from different sources.



- Data can be any format video,audio,text etc..
- Data sources can be from different platforms.

Introduction

- It is estimated that,more than 2.5 Quintilian data created everyday.



- Big data is data that contains greater variety,arriving in increasing volumes and with more velocity.
- Big data is larger,more complex data sets, especially from new data sources.

RELATED WORKS

DESCRIPTION	MERITS	DEMERITS
1.Impact analysis using Crawler technology and data visualization		
■ Analysis and visualization of the data extracted from specific economic website.The variation of the changes easy to understand through visualization.	■ Easily understandable in visualization. ■ Extracting details of any website easily.	■ Limitations on digging inside. ■ Limitations over bot can't access. ■ Spamming website undetectable. ■ creating correlation require more efforts.

Related Works

DESCRIPTION	MERITS	DEMERITS
2.Spatial data science on Covid-19 data		
■ Spatial data science system for analyzing big covid-19 epidemiological data, with focus on the spatial data analysis among different geographic locations.	■ Intensity of cases can be easily detected geographically. ■ Sub diving locations based on the cases and taking actions on each sections.	■ Spatial data exploration is not easy. ■ Mistakes are common in geographical mapping.

Related Works

DESCRIPTION	MERITS	DEMERITS
3.A Data Science Solution for Mining Patterns from Uncertain Big Data		
■ Mining different variety of patterns,arriving correlations and making some predictions are prescribed in this paper.	■ We can predict the coming outcomes and issues. ■ Patterns helps us to understand information easily.	■ Prediction require prefect training set,which is difficult to make. ■ Accuracy of prediction depends upon our patterns.

5 V's of big data

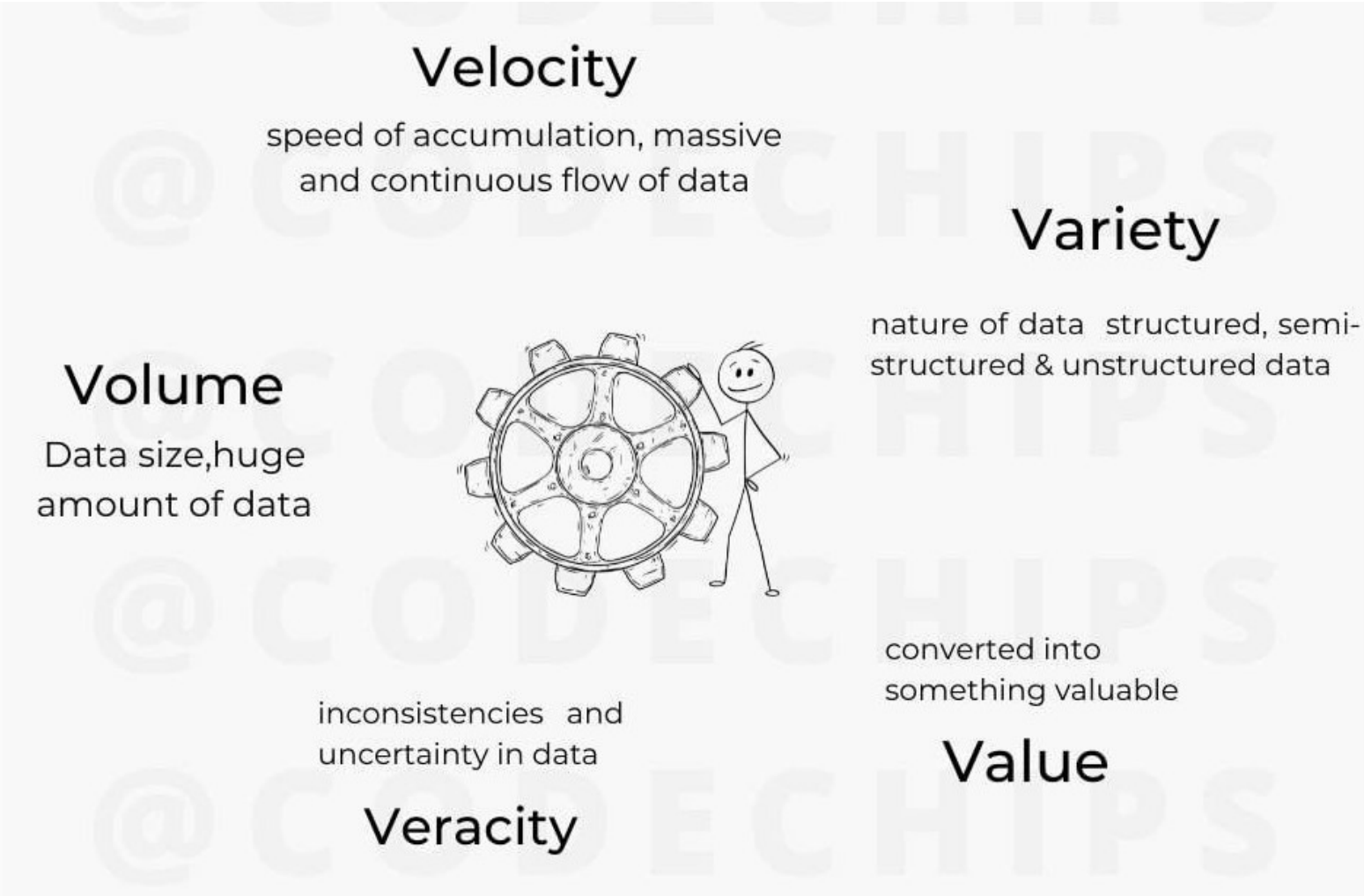


Figure: 5 V's of big data

5 V'S OF BIG DATA

- Volume : Size and amount of data.
- Veracity : Inconsistencies and uncertainty in data.
- Value : Converted into something valuable.
- Variety : nature of data.Structured and unstructured.
- Velocity : Speed of accumulation,massive and continues flow of data.

DATA ANALYSIS

- Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

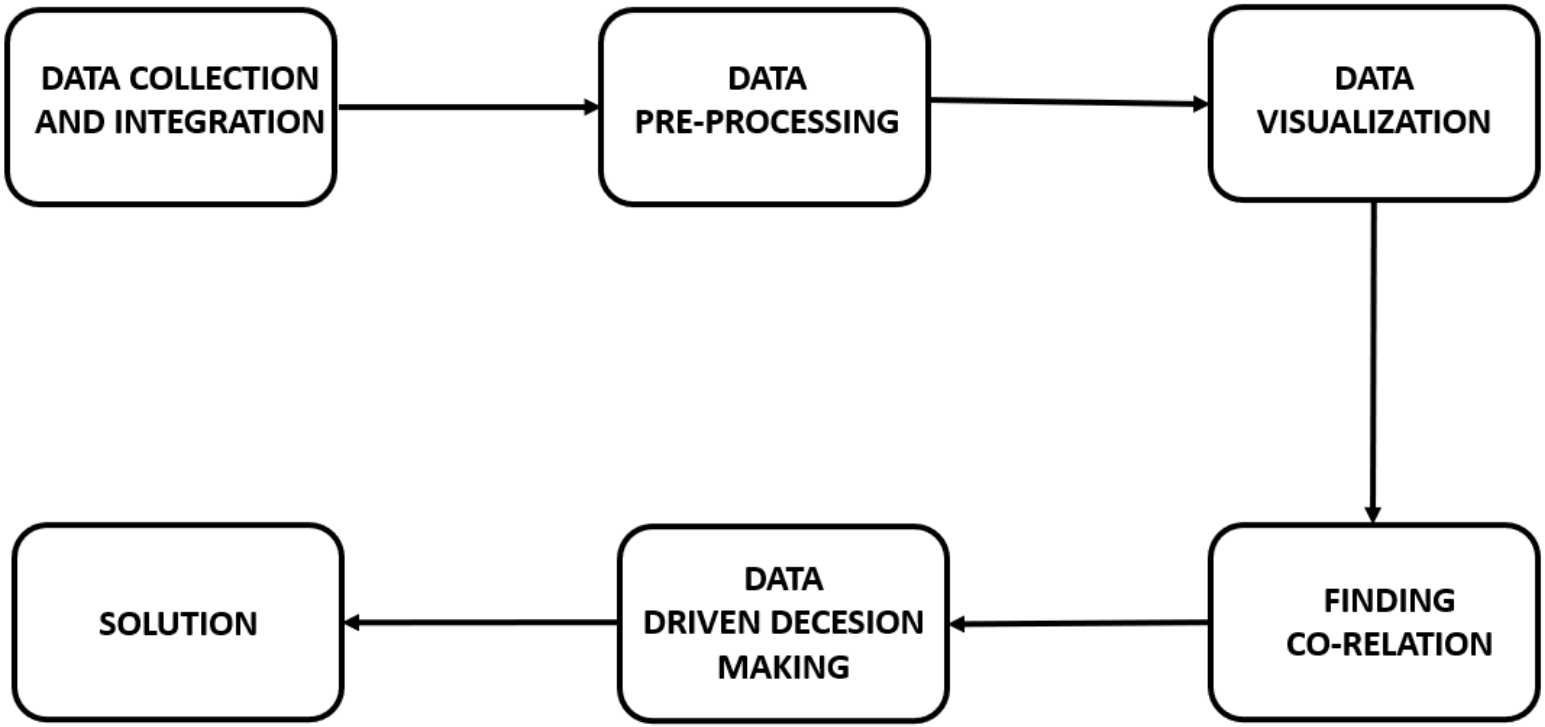


Figure: Work flow

DATA ANALYSIS ON COVID-19 DATA

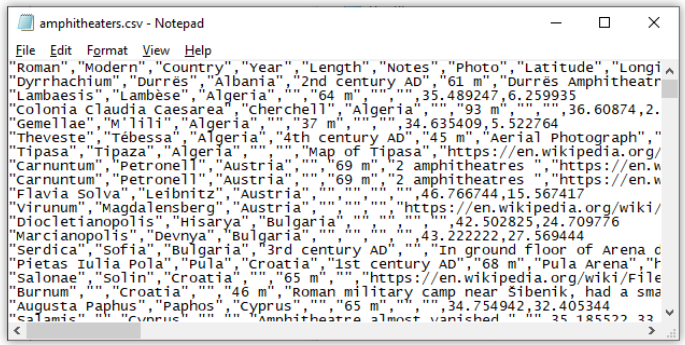
DATA COLLECTION AND INTEGRATION

- Data collection is process where we collect data from different resources.
- It is one of the difficult step for any data analyst to get a consistent and reliable data.
- All data sets are sourced from COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University which is updated daily by them.
- Integration part is done when we extract the data from different sites to our notebook.

DATA PRE-PROCESSING

- Data pre-processing is the process of transforming raw data into an understandable format
- We will get CSV (comma separated values) files, where we need to make to a readable form.
- By the help of python library PANDAS we can easily do that.

CSV file :



Pre-processed file :

```
In [3]: confirmed_df.head()
```

Out[3]:

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/
0	NaN	Afghanistan	33.93911	67.709953	0	0	0	0
1	NaN	Albania	41.15330	20.168300	0	0	0	0
2	NaN	Algeria	28.03390	1.659600	0	0	0	0
3	NaN	Andorra	42.50630	1.521800	0	0	0	0
4	NaN	Angola	-11.20270	17.873900	0	0	0	0

DATA PRE-PROCESSING

[4]:

```
# For world Vaccination Dataset
usa_vac = vac_data[vac_data['country'] == 'United States']
uk_vac = vac_data[vac_data['country'] == 'United Kingdom']
ger_vac = vac_data[vac_data['country'] == 'Germany']
ita_vac = vac_data[vac_data['country'] == 'Italy']
fra_vac = vac_data[vac_data['country'] == 'France']
chi_vac = vac_data[vac_data['country'] == 'China']
rus_vac = vac_data[vac_data['country'] == 'Russia']
isr_vac = vac_data[vac_data['country'] == 'Israel']
uae_vac = vac_data[vac_data['country'] == 'United Arab Emirates']
can_vac = vac_data[vac_data['country'] == 'Canada']
jpn_vac = vac_data[vac_data['country'] == 'Japan']
ind_vac = vac_data[vac_data['country'] == 'India']
ino_vac = vac_data[vac_data['country'] == 'Indonesia']
mal_vac = vac_data[vac_data['country'] == 'Malaysia']
ban_vac = vac_data[vac_data['country'] == 'Bangladesh']
nig_vac = vac_data[vac_data['country'] == 'Nigeria']
phi_vac = vac_data[vac_data['country'] == 'Phillipines']
vie_vac = vac_data[vac_data['country'] == 'Vietnam']
egy_vac = vac_data[vac_data['country'] == 'Egypt']
```

Figure: Data-pre-processing

DATA PRE-PROCESSING

[5]:

```
#For Indian Vaccination Dataset
df2=state_vac
df2 = df2.rename(columns= {'Updated On':'Date','Total Doses Administered':'TotalDoses','Male(Individuals
df2.Date = pd.to_datetime(df2.Date, format="%d/%m/%Y")
df3=india
df1=state
df2 = df2[df2['State'] !='India']
df2 = df2.rename(columns= {'Updated On':'Date','Total Doses Administered':'TotalDoses','Male(Individuals
df2.Date = pd.to_datetime(df2.Date, format="%d/%m/%Y")
df2_2=df2[df2['Date']=="2021-08-9"]
df2_2.dropna()
df2_1 = df3[df3['Date']=='2021-08-11']
```

Figure: Data pre-processing

DATA VISUALIZATION

- Data visualization is the process of translating large data sets and metrics into charts, graphs and other visuals.
- The resulting visual representation of data makes it easier to identify and share real-time trends, outliers, and new insights about the information represented in the data.

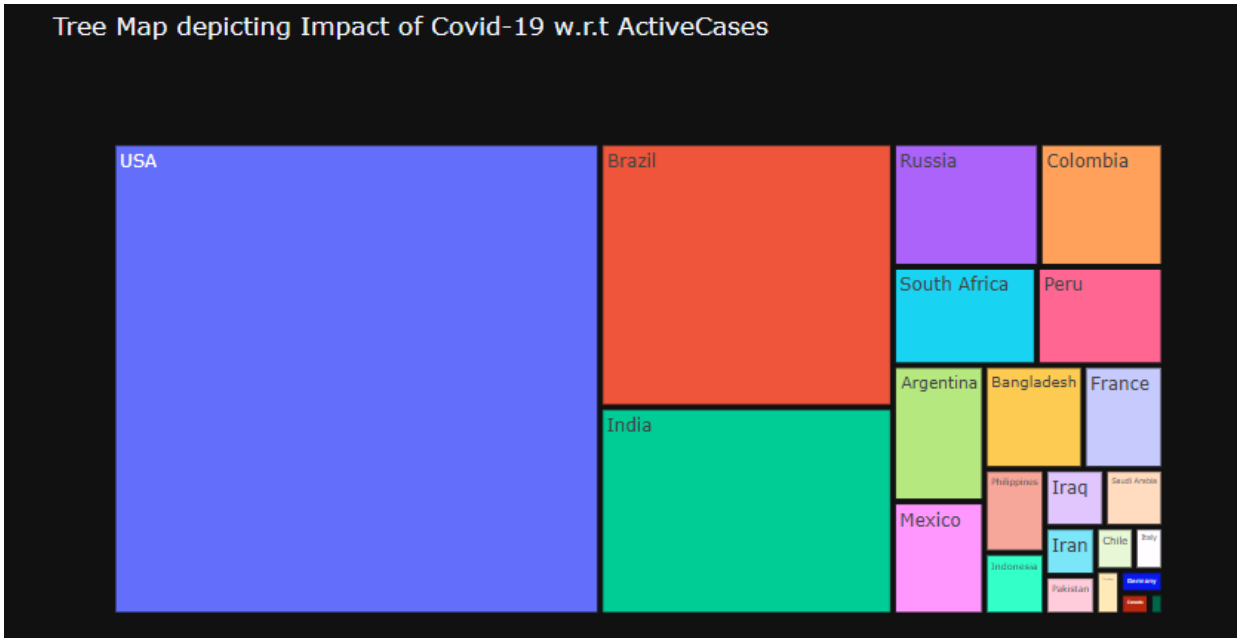


Figure: Tree map to depict active cases

DATA VISUALIZATION

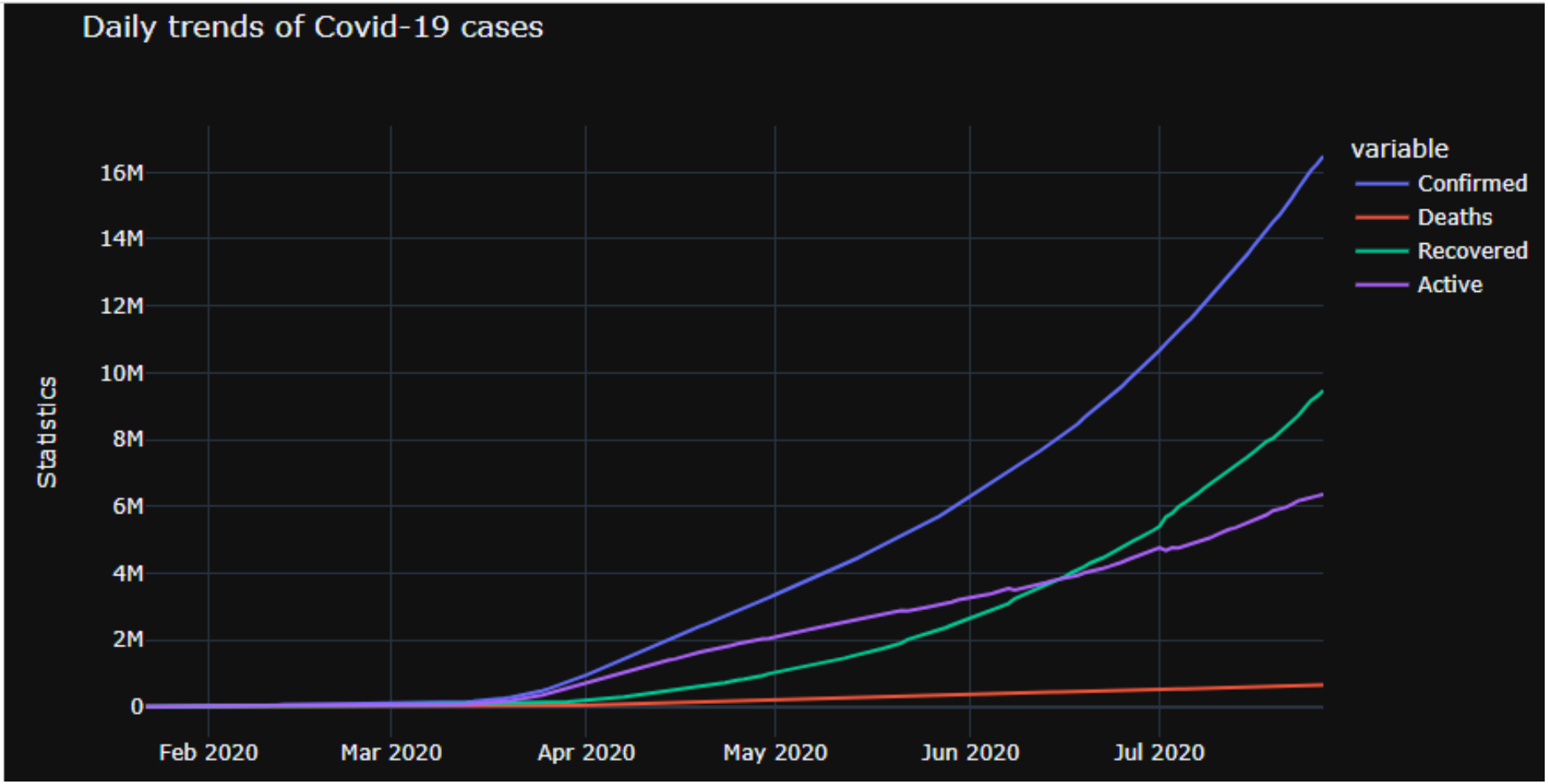


Figure: Daily trends of covid-19 globally

DATA VISUALIZATION

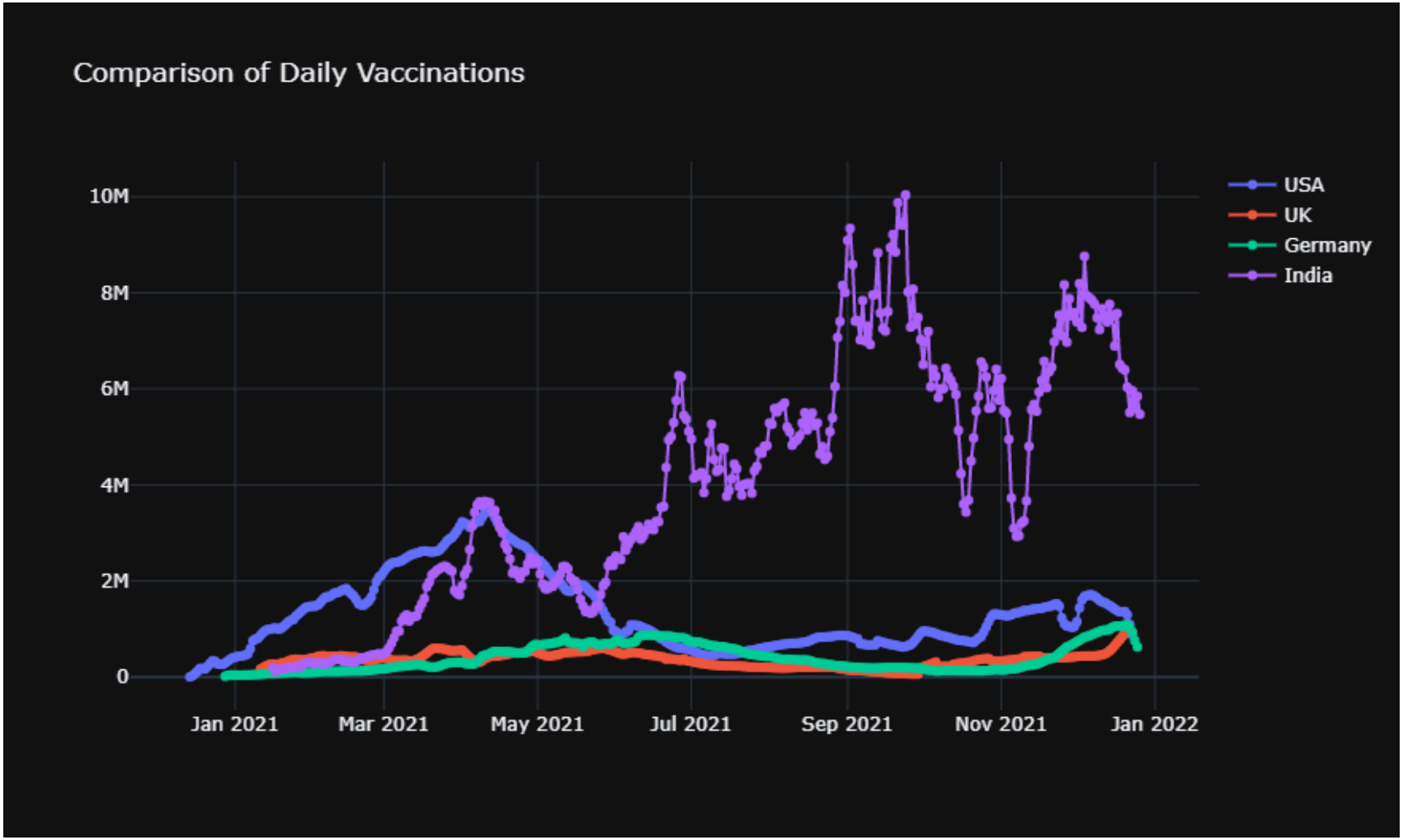


Figure: Vaccination progress

CO-RELATION

- Relation between two data can be easily depicted using visualization where the outcome is known as co-relation.
- Correlation help us to reach data driven decision making.
- From above visualization some of the co-relations are
  - From tree map the highly covid-19 affected country is USA and then Brazil.
  - From the line graph the increase of case is clearly visible. The active cases overtaken by the recovered cases.
  - The line graph shows the daily increment in the vaccination numbers in different countries. From a period of September to November the daily vaccination rate India has been decreased. Since people thought, the pandemic has been vanished.

DATA DRIVEN DECISION MAKING

- The co-relations we obtained are benefited to make decisions.These decisions are called Data driven decisions.
- From above co-relations, some of the decisions are
  - The peoples from USA, Brazil and other highly affected countries are restricted to many less affected countries.
  - The increase of recovery denotes the medical procedures are effective, so the countries can enhance more on the same procedure.
  - Give more awareness about vaccine to people, since people slowly stopped vaccination.

CONCLUSION

Big data and it's analysis is a long process where we should care every minute detail about data.Every stage demands the analyst critical thinking.It's been good journey to explore about data and do the analysis.

References I

[1] Olga Ormandjieva Alaa Alsaig, Vangular Alagar.

A critical analysis of the v-model of big data.

IEEE TrustCom/BigDataSE, 2018.

[2] Carson K. Leung Adrienne V. Pind Karl E. Dierckens, Adrian B. Harrison.

A data science and engineering solution for fast k-means clustering of big data.

IEEE Trustcom/BigDataSE/ICESS, September 2017.

[3] Fan Jiang Carson Kai-Sang Leung.

A data science solution for mining interesting patterns from uncertain big data, 2017.