Naan Mudhalvan

IBM PROJECT

Applied Data Science(phase 3-Development)

COVID-19 VACCINE ANALYSIS

Edit with WPS Office

PROBLEM STATEMENT

- Forecasting of time taken for completing 100% total vaccinations of particular region over the time period.
- By this, vaccine manufacturing companies get to know the prior requirements of vaccine which helps to produce the vaccines in large scale and complete the vaccination drive with in calculated time.

Edit with WPS Office

DATASETS AND ITS ATTRIBUTES

1. Vaccinations(data)
2. Vaccine_Names(vaccine)

Edit with WPS Office

Attribues of vaccinations dataset :

- 'iso_code'
- 'continent'
- 'location'
- 'date'
- 'total_cases'
- 'new_cases'
- 'total_deaths'
- 'new_deaths'
- 'total_deaths_per_million'
- 'total_tests'
- 'new_tests'
- 'positive_rate'
- 'total_vaccinations'
- 'people_vaccinated'
- 'people_fully_vaccinated'
- 'total_boosters'
- 'new_vaccinations'
- 'total_vaccinations_per_hundred'
- 'people_vaccinated_per_hundred'
- 'people_fully_vaccinated_per_hundred'
- 'total_boosters_per_hundred'
- "population'
- 'population_density'
- 'median_age'
- 'aged_65_older'
- 'aged_70_older'
- 'human_per_capita'

Edit with WPS Office

Attributes of Vaccine_Names dataset :

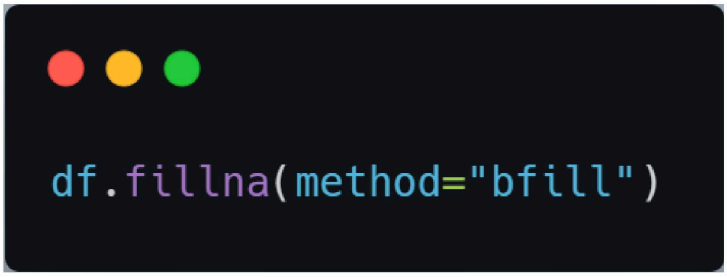
- 'location'
- 'date'
- 'vaccine'
- 'total_vaccinations'

- The Attributes present and one of the appropriate and relevant large dataset found was the reason for choosing these datasets.
- The Attributes having null percentage of more than 85 are removed. Attributes which are most frequently used and require for analysing and visualizing are selected .

Edit with WPS Office

DATA PRE-PROCESSING

Edit with WPS Office



- The 'BFILL' fills the missing values backward so they are replaced with the next value.
- All the missing values of all tuples are now replaced by appropriate values

Edit with WPS Office

	iso_code	continent	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	...	female_sm
0	AFG	Asia	Afghanistan	2020-02-24	5.0	5.0	NaN	NaN	NaN	NaN	...	
1	AFG	Asia	Afghanistan	2020-02-25	5.0	0.0	NaN	NaN	NaN	NaN	...	
2	AFG	Asia	Afghanistan	2020-02-26	5.0	0.0	NaN	NaN	NaN	NaN	...	
3	AFG	Asia	Afghanistan	2020-02-27	5.0	0.0	NaN	NaN	NaN	NaN	...	
4	AFG	Asia	Afghanistan	2020-02-28	5.0	0.0	NaN	NaN	NaN	NaN	...	

5 rows × 67 columns

Out[64]:

	location	date	vaccine
0	Argentina	2020-02-24	Moderna
1	Argentina	2020-02-25	Oxford/AstraZeneca
2	Argentina	2020-02-26	Sinopharm/Beijing
3	Argentina	2020-02-27	Sputnik V
4	Argentina	2020-02-28	Moderna
...	...	...	...
38667	European Union	2020-12-16	Oxford/AstraZeneca
38668	European Union	2020-12-17	Pfizer/BioNTech
38669	European Union	2020-12-18	Sinopharm/Beijing
38670	European Union	2020-12-19	Sinovac
38671	European Union	2020-12-20	Sputnik V

38672 rows × 3 columns

Merging of two datasets(data and vaccine)

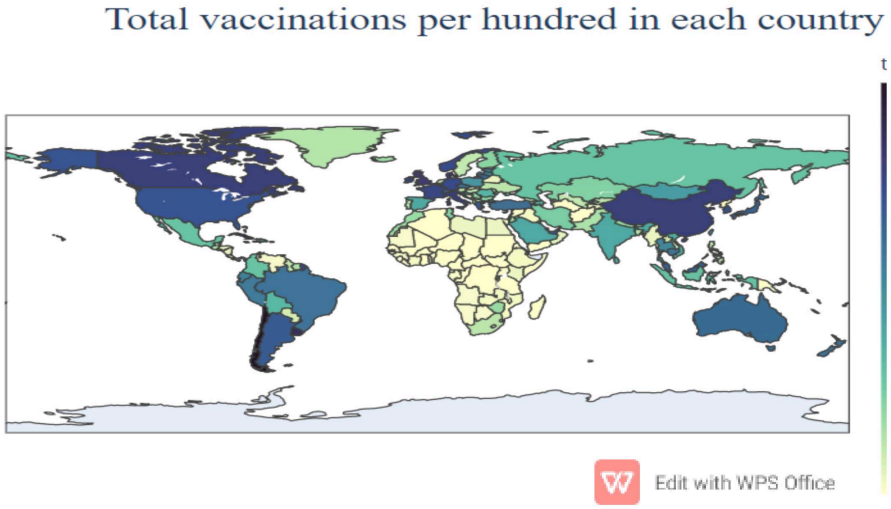
Edit with WPS Office

One hot encoding

vacci_moderna	vacci_oxford_ast	vacci_pfizer_bio	vacci_sinopharm_	vacci_sputnik_v	vacci_cansino	vacci_johnson&jo	vacci_novavax	v
0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	
0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	

Edit with WPS Office

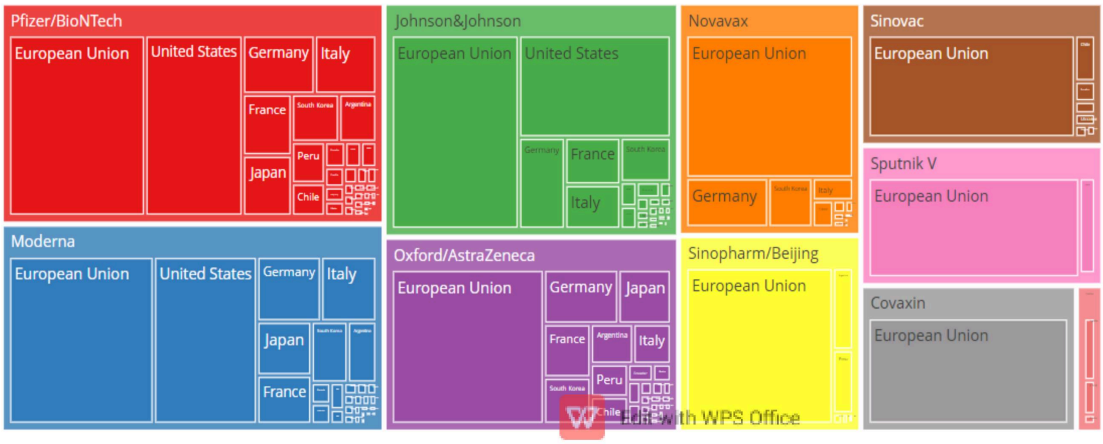
Compare the total vaccinations per hundred over all the regions of the world ?



North America,South America and countries with higher gdp have higher total vaccination per hundred

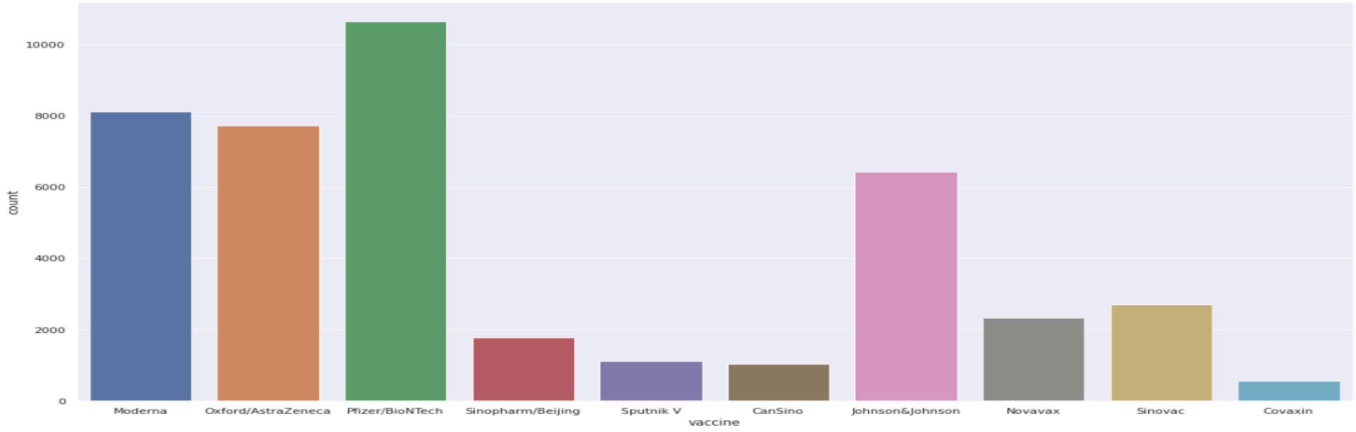
Which Vaccines are used in a region?

Total Vaccinations per country grouped by Vaccines



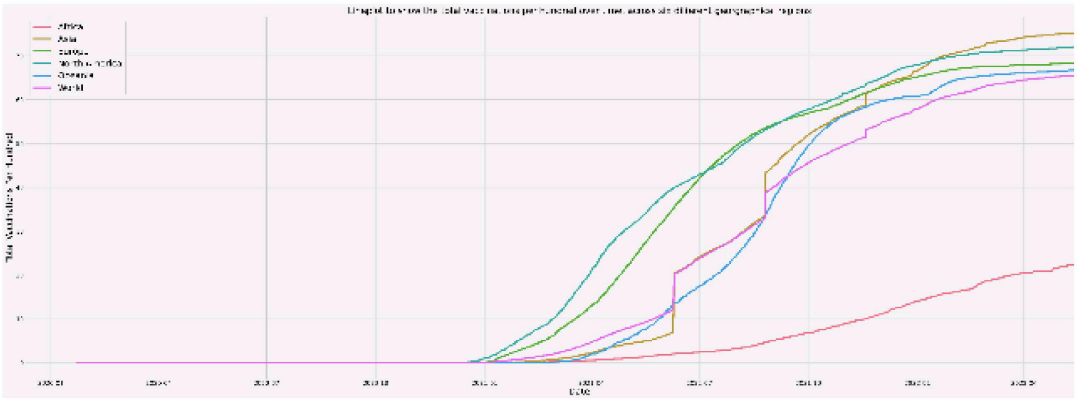
Vaccines like pfizer, Moderna are used by many countries where as vaccines like sinovac and sputnik are not approved in many countries

What is the most used vaccines in our data ?



In this analysis, x-axis is labelled using vaccine names and y-axis is labelled using count vaccines.Pfizer/BioNtech is the most widely used vaccine over the world and can be easily suggested in the future years also for the additional manufacturing of this vaccine.

Analyse the trend of vaccinations across the regions?

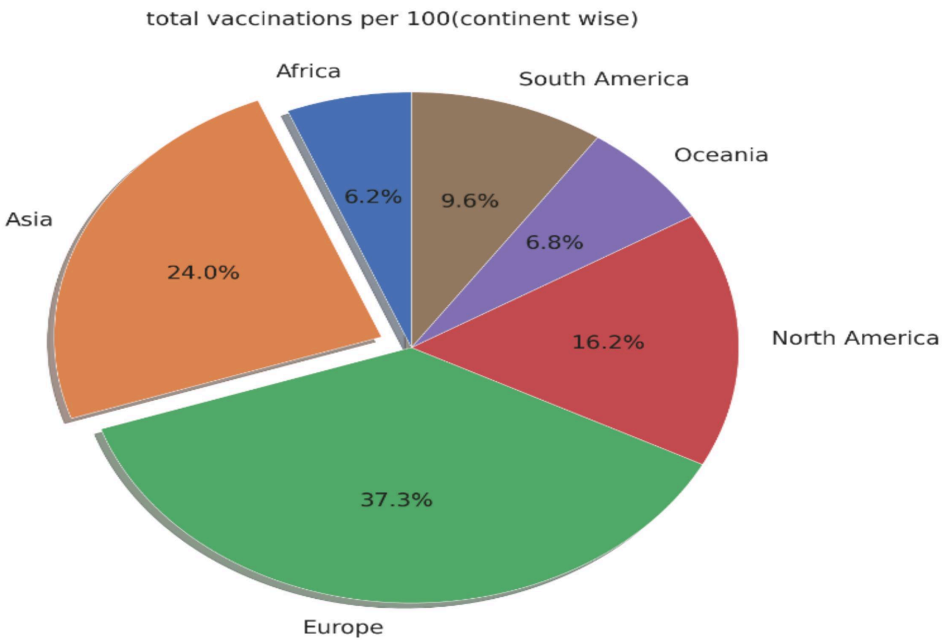


This shows the trend of vaccinations over the time.Similarly we are going to predict the time taken for completion of fully vaccinated region and to calculate the average time taken for completion of vaccination across fully vaccinated region.

Edit with WPS Office

Data analysis and visualisation

Edit with WPS Office

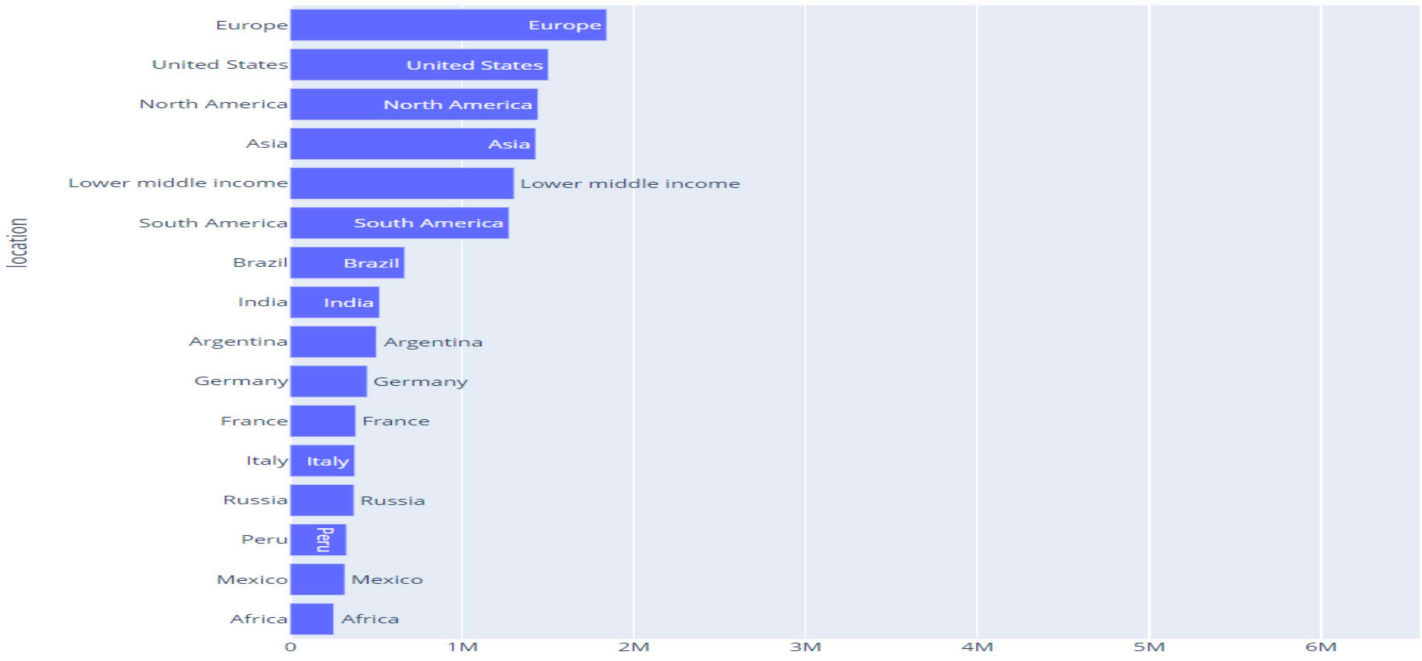


Asia and Europe have around 50 percent of all vaccinations.

Edit with WPS Office

gdp_per_capita

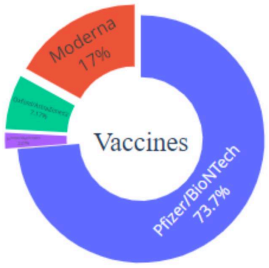
Germany, Switzerland, Qatar, Luxembourg has highest gdp per capita

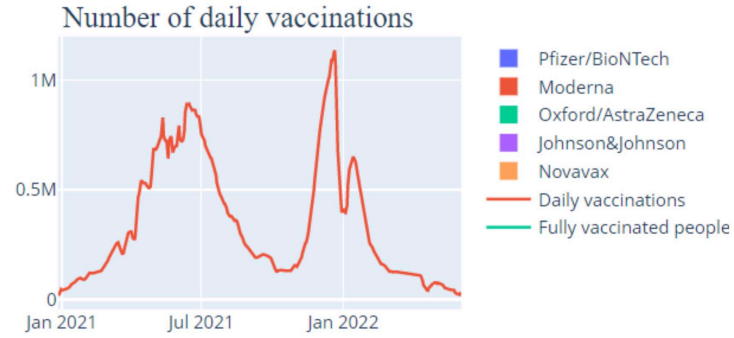


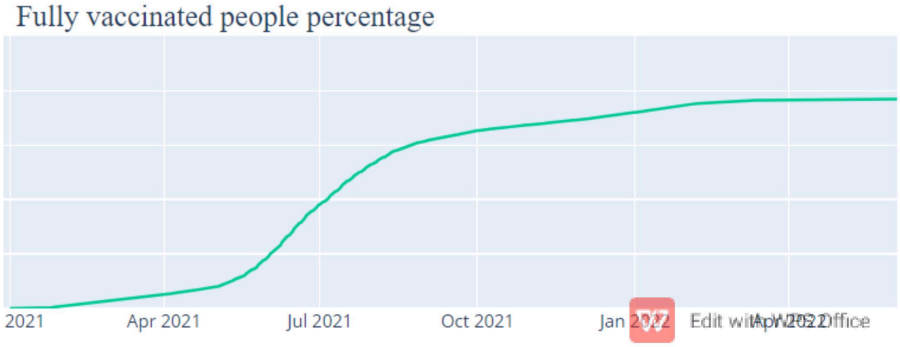
European Union has the highest new deaths

Edit with WPS Office

Germany abstract informations







MODEL BUILDING

Edit with WPS Office

Time series forecasting using Huber regressor

- Time Series Analysis is the way of studying the characteristics of the response variable with respect to time, as the independent variable.
- our dataset contains data points which typically consists of successive measurement made from the same source over a time interval.

Edit with WPS Office

PyCaret is an open-source machine learning library that automates machine workflows.

This library helps in deciding which model is more efficient for a given method.

```
setup = caret.setup(data = train , test_data = test , target = 'Target' , fold_strategy = 'timeseries'
                , remove_perfect_collinearity = False , numeric_features = ['Series' , 'Window_mean' , 'Shift1']
                    , fold = 5 , session_id = 51)
```

Edit with WPS Office

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
huber	Huber Regressor	0.1336	0.0398	0.1678	0.9974	0.0075	0.0066	0.0220
ard	Automatic Relevance Determination	0.1532	0.0564	0.1879	0.9974	0.0077	0.0071	0.0100
tr	TheilSen Regressor	0.1635	0.0597	0.2079	0.9957	0.0100	0.0089	0.3400
lar	Least Angle Regression	0.1897	0.0892	0.2351	0.9959	0.0094	0.0085	0.0100
lr	Linear Regression	0.1898	0.0892	0.2351	0.9959	0.0094	0.0085	0.6200
ransac	Random Sample Consensus	0.1898	0.0892	0.2351	0.9959	0.0094	0.0085	0.0140
br	Bayesian Ridge	0.1945	0.0938	0.2397	0.9957	0.0095	0.0086	0.0100
omp	Orthogonal Matching Pursuit	0.2811	0.1646	0.3257	0.9907	0.0119	0.0113	0.0100
kr	Kernel Ridge	0.3965	0.3938	0.4589	0.9839	0.0205	0.0197	0.0220
ridge	Ridge Regression	0.5305	0.6235	0.6077	0.9555	0.0307	0.0302	0.0080
mlp	MLP Regressor	1.1472	4.1281	1.3262	0.7803	0.0724	0.0667	0.0820
par	Passive Aggressive Regressor	1.5214	4.2526	1.6684	0.4432	0.1103	0.1131	0.0120
en	Elastic Net	1.9403	14.9515	2.2686	0.2728	0.1553	0.1288	0.0140

- We can observe that by testing all possible models we are getting Huber regression fits best.

Edit with WPS Office

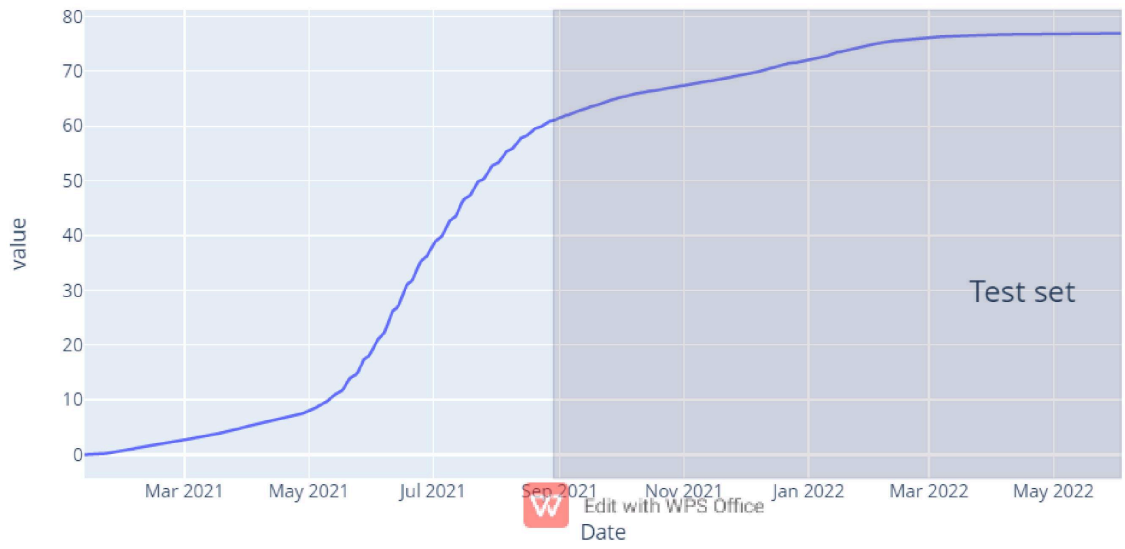
Huber Regression

In statistics, Huber loss is a particular loss function (first introduced in 1964 by Peter Jost Huber, a Swiss mathematician) that is used widely for robust regression problems — situations where outliers are present that can degrade the performance and accuracy of least-squared-loss error based regression.

- Huber regression is a type of robust regression that is aware of possibility of outliers in a dataset and assigns them less weight than other examples in the dataset.

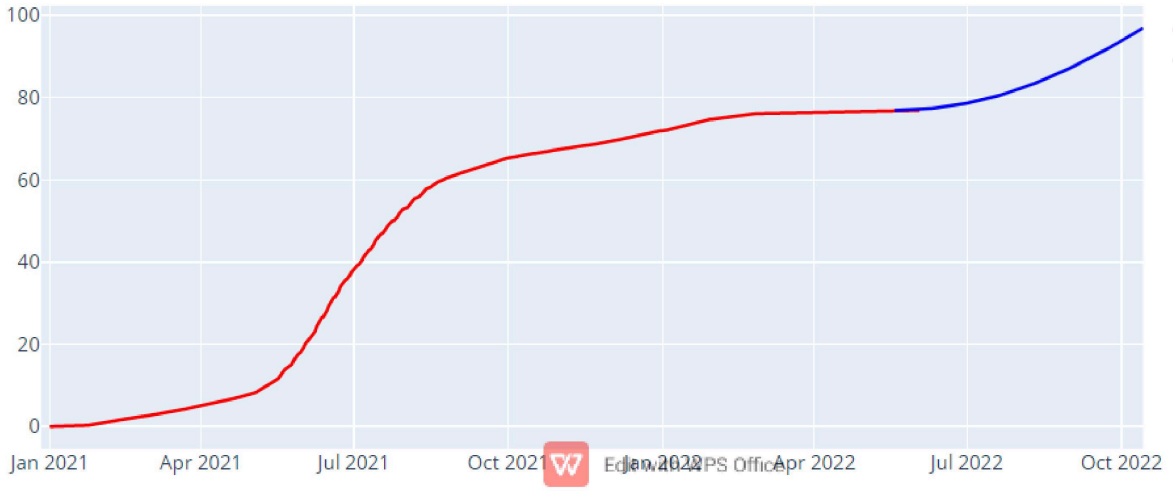
Edit with WPS Office

Given Data of vaccination per hundred of Germany



We can observe that about 76% of Germany's population has been vaccinated

After forecasting using TS



After forecasting we can observe the at the end oct 2nd week vaccination completes in germany

CONCLUSIONS:

- European union is one of the best example for the region which has many deaths and also which produced high number of total vaccinations. Also the gdp of countries of this region has not much different after pandemic.
- Comparing the Root-mean-square error (rmse) and coefficient of discrimination(R2) values of models, Huber regression is selected with least RMSE of 0.1678 and R2 value of 0.9974 as the best model to predict the total vaccinations of particular region over the time period.
- According to our model,Total Vaccinations of Germany might be completed by 2nd week of October.

Edit with WPS Office

Future Scope

- Governments can improve the vaccination facilities and availability of vaccines to different locations by referring demand of vaccines and by referring most used vaccines across different countries.
- The synthesis of current research will be helpful to researchers analysing historical trends in COVID-19 pandemic and individuals interested in better understanding and advocating for understanding and advocating for underserved communities across the globe.

Edit with WPS Office