U.S. Covid-19 Vaccinations Analysis at County Level

1. INTRODUCTION

Coronavirus Disease (COVID-19) began in China and quickly spread over the world, resulting in a global pandemic. COVID-19 has now infected 264 million individuals globally, resulting in over 5.2 million deaths whereas the United States alone has 49 million infections and 0.7 million deaths. To control the situation Vaccination was developed rapidly during this time period, and people began to consume vaccinations throughout this period starting from Dec 11, 2020. There are several misconceptions circulating about the new vaccines, and various factors are driving the vaccination push to decline. In this project, we aim to find the factors that are influencing vaccination drive and identifying which components have an impact on vaccination consumption and predict the future vaccination trend.

The process of analysis involves looking for the factors that are playing an important role in vaccination drive For example, for checking the impact of age on vaccination consumption we will be running a different regression model to predict the importance of this factor in the vaccination decrease. Along with that, we will be using clustering analysis in order to find interesting facts in our end goal.

Before running into the analysis we need to pre-process the data. Preprocessing is necessary to ensure that the data is standardized and remove all the outliers that are not giving any information in findings and this step helps by giving a form that can be easily used to create a model. Certain factors from different categories can contain the same other values that
can prevent us from categorizing the factors correctly. For example, in some of the data sets the county name will be like Los Angeles County and in other data sets, it will be filled with different notation like LA county. In order to find analysis and interesting factors, we changed and mapped under one common notation.

Lastly, we have used different datasets that can help in extracting useful findings and meanings patterns from this raw data.

2. BACKGROUND

The national public health agency of the United States that is cdc[1] provides an integrated, county view of key data for monitoring the COVID-19 pandemic in the United States. It allows users to explore standardized data across the country, including vaccine data and some of the types of data sets as covid cases at counties, jurisdictions, and states, different covid vaccine brands, and its distribution among this entire period. We are using all these data in finding interesting analyses and predictions. CDC also provides a data visualization tool to find the basic feature present in the data sets and categorize them.

Categorization can help answer many interesting questions. For example, there is an interesting article [2] which tries to answer the is Driving the Decline in People's Willingness to Take the COVID-19

Vaccine in the United States? Here in this article author uses different factors in determining the result he claims that raising the covid cases and deaths is affecting the vaccination delivery and political comments are also affecting this drive.

Another interesting thing is that Vaccination trials were run among particular persons and here the race will play a huge role in vaccine accuracy so many people are declining it. So Many companies are taking this under consideration and trials were running on Hispanic and Black Americans as they have a higher complication rate, they are hoping to test the new vaccine on volunteers from those groups. So from this, we can say that race is also a major factor in vaccination delivery.

3 EXPERIMENTAL SETUP

In this section, we discuss our pipeline through which we identified valuable datasets, merged them with the original data, and used K-means Clustering for analysis and Machine Learning methods to gain insight into the vaccination.

3.1 Data collection

The most important part of the data science problem is identifying valuable data and getting that under one roof. To the end, we identified data sets from demographic data to economic datasets and even the education & employment data.

3.1.1 Datasets & Datasources.

During different steps of our study, we integrated multiple external datasets into our pipeline. In this section, we have an overview of them. Different datasets were merged under one dataset and then we divided the data into two main groups. The First Group contains each county covid vaccination details which would cover entire aspects. The second group contains county-specific details like demographic information, education, employment data, and economic data.

The first group of datasets is as follows:

Vaccination completed: Total number of people who are fully vaccinated (have the second dose of a two-dose vaccine or one dose of a single-dose vaccine) based on the jurisdiction and county where the recipient lives and also it contains all the information of each age group vaccination in three categories 12+,18+,65+. Vaccination delivery date starting from the initial date.

SVI category: Disparities in vaccination coverage by social vulnerability

Covid-19 virus: The number of historical cases and deaths presented on the CDC's website reflects the information provided by the counties. Thus, data may reflect either the date the case or death occurred or the date it was recorded in the state.

All the above data sets have been collected from cdc and covid act now
We look at second group datasets used to extract entities

- Education: Features such as "Population with high school degree or greater", "Population with bachelors degree", etc. are included in this dataset.
- Population: Including county-level yearly population change information such as births, death, and current population estimates along with domestic and international migration rates.
- Poverty: Similarly, this dataset includes household income per county for different years.
- Unemployment: Contains employment rate

per county for different years and different groups of people.
• County Category: Location of the county weather the county falls under metro or not.
• County FIPS: Having a list of US states, counties, and state abbreviations, we come up with a list of words to remove from the text similar to personal names.

All the second group data has been collected from different govt sites mentioned in the references section.

3.1.2 Preprocessing of the datasets.

Using all the data we obtained from the different sites, we generated an entire dataset. Now, while merging these different sets of data, we encountered several issues and resolved them.

1. One of the key issues is data set inconsistency and a lack of common factors across data sets, making combining them altogether more difficult than intended.
2. Another difficulty is that some counties do not receive enough information in a single direct hit. As a result, we obtained the information from their own website.
3. We have standardized everything under one roof because some of the data from the data sources is formatted differently than intended.
4. Each data set contains different way of representation of county name to link them under one category is bit challenging.
5. As some of the counties have far ahead of vaccination drive and some the counties are lag behind it so we normalized by applying different techniques like min_max scaling, logorithm so it would help in great visualization of data.
6. In this dataset Date is a very crucial field in order to predict model so date was formatted with different data types among the data sets. So they are bought among a single common data type.
7. Some of the data field contains categorical variables which is bad for the model so we applied a one-hot encoding to the integer representation so it benefited for the model.

4. RESLUTS

4.1 Analysis

In this section we will look into the different analysis done under this project which could give a greate insight of the covid vaccination trend and how it is effecting from different factor.

4.1.1 Covid Vaccination among counties

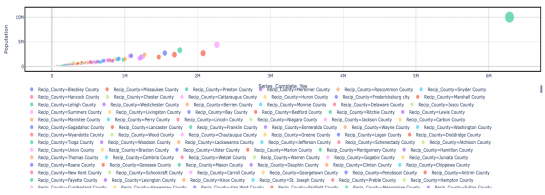


Fig 1 Covid Vaccination among counties vs population among counties.

From Figure1, we see that complete vaccination people was plot against the county population and each scatter point size was represented with complete vaccination among this the topmost county

that is leading in vaccination is 'Los Angeles County and the majority counties in united states were under 1M vaccination. And in the between 1M and top county i.e., LA most of the popular counties like Miami-Dade County fall under this category and doing best in delivering the vaccination.

4.1.2 Covid Vaccination trend among SVI category

SVI (Social Vulnerability Index) U.S. census variables to help local officials identify communities that may need support during serious situation. In covid times US administrative has been categorized SVI in 4 ways(1-4) numbers were 4 stands as highest risk and 1 is the Lowest risk Area.

Now the below Fig 2 represent the vaccination trend among these categories and the interesting thing is US Administration has given high priority to the SVI 4 counties and the vaccination is going at a high rate in these counties and the remaining counties were running slow.
Here, we used K-Means clustering to group the vaccination trend.

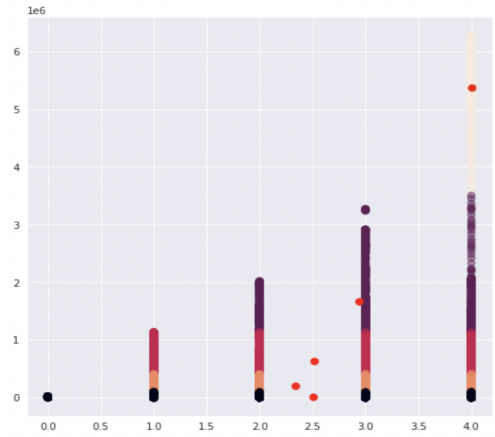


Fig 2. SVI and vaccination trend.

4.1.3 Counties Vaccination with highest and lowest

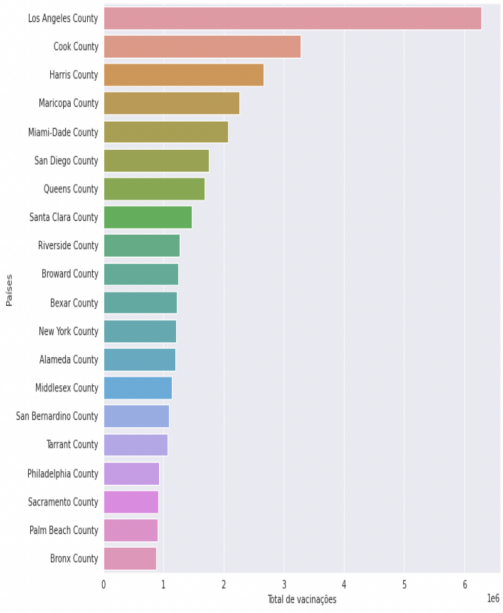


Fig 3 Highest 20 Vaccination counties

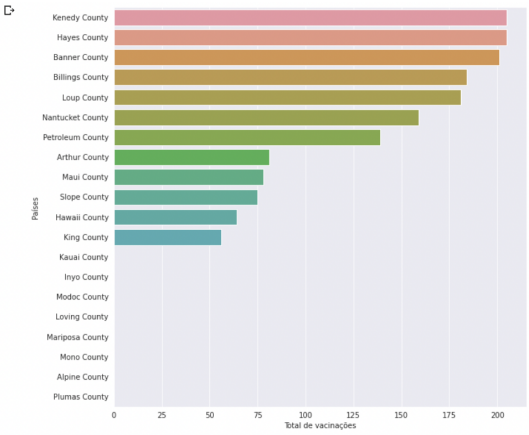


Fig 4: Lowest 20 Vaccination Counties

The above fig3,fig4 Represents the highest and lowest vaccination counties in the United States till Nov 28. Some of the counties with the Lowest vaccination didn't provide any details to CDC so we have eliminated them from our consideration. And these lowest counties has been started vaccination a very recently.

So For further analysis, we are considering these counties for deep analysis.

4.14 Vaccination trend from starting to till now

In the top counties, except for LA remaining counties run similarly in vaccination delivery.

In the Least vaccinated counties, 2 counties have been started delivering vaccination recently, and the remaining two counties have been started slowly when compared with the top counties.

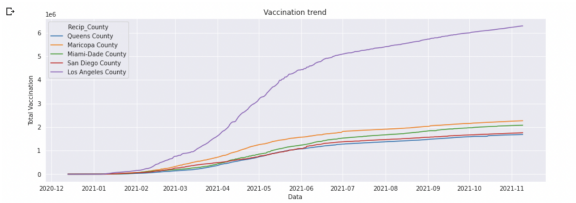


Fig 5 Covid Vaccination trend among the top 5 Counties

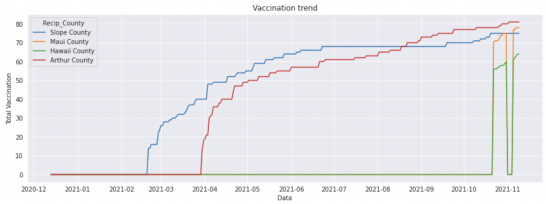


Fig 6 Covid Vaccination Trend among the least 5 Counties.

The below analysis will give a good insight into why they are running high and low in vaccination delivery.

They are different factors that are playing an important role in vaccination trends and they are listed as below:

- SVI-
    1. The reason behind the top counties running fast in vaccination trend is they are in categories 4 and 3 in the SVI index. so Administration is giving topmost priority to these counties as they are vulnerable so this is the main reason that trend is at running at peak this can be claimed from below scatter plots and the lowest counties are coming under the 1,2 SVI so this is one of the reasons behind vaccination trend.

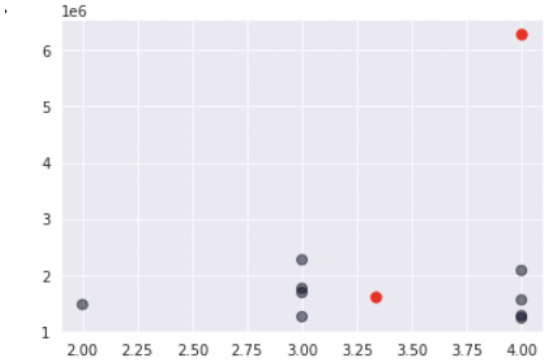


Fig7: SVI in top vaccination Counties

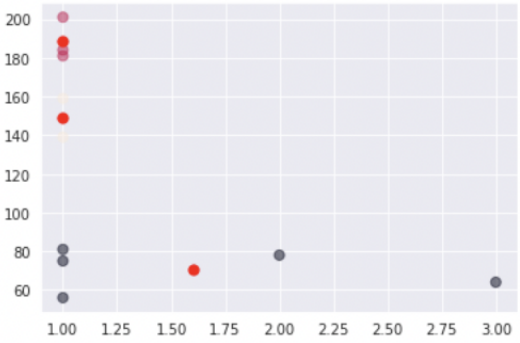


Fig 8: SVI in Lowest vaccination counties

- Covid Cases and deaths among these Counties

    1. As from the above SVI index, these top counties are having high covid cases and deaths when compared to the lowest counties. So people in the lowest counties might things that they are not having a covid impact in their area so they are showing less interest in having vaccination and people at the highest vaccination counties are fearing this covid cases and deaths and consuming the vaccination as quickly as possible. This Conclusion can be bought into consideration from the below fig.

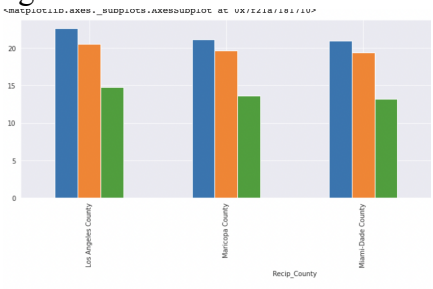


Fig9: Vaccination, Covid Cases and deaths in Top Counties

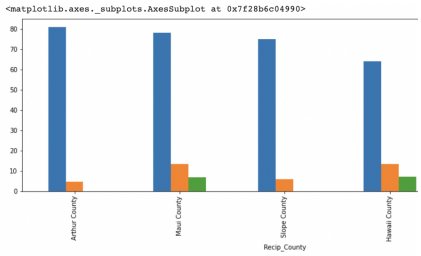


Fig10: Vaccination, Covid Cases and deaths in Lowest Counties

- Population and employed impact on covid vaccination

    1. One of the most interesting facts is that top counties are having the highest population and less employed and the lowest counties are having less population and a good number of the employed population so the lowest vaccination might be because the people are given top priority to work rather than the vaccination this might be the reason behind in vaccination decline and vice versa was accepted in the top counties and these conclusions were drawn from below bar graph.

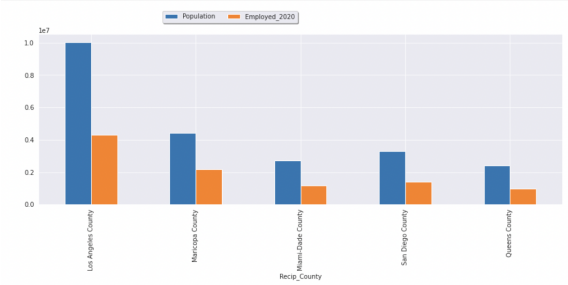


Fig 11: Population and Employed for the top county

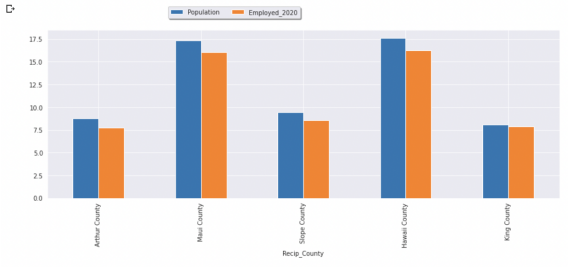


Fig 12: Population and Employed for the lowest county

- Correlation

From the above correlation, we can find those different parameters playing a great role in covid vaccination delivery. Population and full vaccination are strongly correlated so there is a positive side that people are showing interest in getting vaccinated and people with the highest level of education are standing towards vaccination consuming with a 0.74 correlation. Income is not playing a significant role in getting vaccinated and this means that there might be good Policies that should be taken by the state government such as giving free vaccination to lower-income households so that they can come up for vaccination.

4.2. Models

To start with, Regression is a statistical technique that is used to model the relationship of a dependent variable with respect to one or more independent variables.

We have used different Regressive models to find the factors affecting vaccination and below are the findings.

Baseline Model:

We have used Linear Regression as the baseline model. In this model, we have divided data into training and validation from the dataset itself as

```
validation_set=dataframe[(dataframe['YearMonth'] > '2021-09')]
train=dataframe[(dataframe['YearMonth'] > '2020-09') & (dataframe['YearMonth'] < '2021-09')]
```

Formula used for Linear Regression $Y=a+b*X + e$

The following features have been used for the Baseline model:

Date	County	State	Percentage Vaccinated people
Metro-status	Party	Estimated hesitant	Social Vulnerability Index (SVI)
Percent Hispanic	Percent non-Hispanic Asian	Percent non-Hispanic Black	Percent non-Hispanic White

The Recent vaccination period was used for the validation and here we have used some features that are strongly correlated and this baseline model was run without preprocessing of data so the RMSE was high:-

```
import sklearn
import math

mse = sklearn.metrics.mean_squared_error(list(validation_set.Series_Complete_Yes), y_pred1)

rmse = math.sqrt(mse)

print(rmse)
```

```
5022.356395183156
```

PyCaret Model:

Pycaret is an open-source, machine learning library in Python that helps you from data preparation to model deployment. It is easy to use and you can do almost every data science project task with just one line of code.

Train data and testing data were divided from the time series of vaccination the last 2months of available data was used for testing and the remaining data for the training.
The following setup function is used for the Pycaret Model:-

```
s = setup(data = train, test_data = test, target = 'Series_Complete_Yes', fold_strategy = 'timeseries')
```

In order to find the best regressive model we have used pycaret machine learning model with important and strongly correlated models and we have obtained the following details:-

```
best = compare_models(sort = 'MAE')
```

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
huber	Huber Regressor	136.4261	1.025008e+05	211.8339	1.0000	0.0005	0.0004	0.016
par	Passive Aggressive Regressor	138.1623	1.034666e+05	212.9290	1.0000	0.0005	0.0003	0.014
omp	Orthogonal Matching Pursuit	156.9712	9.204330e+04	205.9738	1.0000	0.0337	0.0004	0.011
en	Elastic Net	157.7142	9.221236e+04	206.3321	1.0000	0.0284	0.0004	0.013
lasso	Lasso Regression	157.9967	9.161947e+04	205.8499	1.0000	0.0312	0.0004	0.013

Arima Time Series Model:

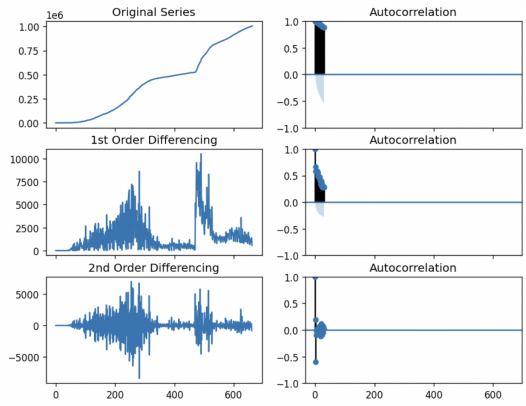
Arima is 'Auto-Regressive Integrated Moving Average' it is the time series analysis.

We Ran the Arima time series for Suffolk County and found the future vaccination trend.

The following step has been performed for ARIMA

To make the time series stationary to make it not depend on the time at which the series is analyzed. To make a non-stationary time series to stationary, we can do differencing operation for stabilizing the mean of the series where we calculate the differences in a back-to-back manner (that is, the difference between successive observations or subtract past one from the current observation value). This can aid in removing (or minimizing) trends and seasonality.

After analyzing the Autocorrelation plot, we have done it for second-order differencing to make the time series stationary.



To find the optimal value of differencing (we assumed d=2 previously by plotting the time series), we can use the Augmented Dickey-Fuller (ADF) test.

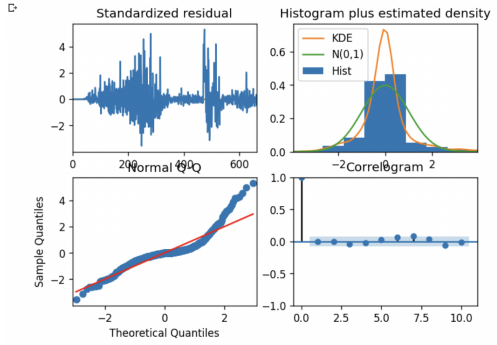
```
Performing stepwise search to minimize aic
 ARIMA(1,2,1)(0,0,0)[0] intercept   : AIC=11213.308, Time=0.50 sec
 ARIMA(0,2,0)(0,0,0)[0] intercept   : AIC=11598.109, Time=0.03 sec
 ARIMA(1,2,0)(0,0,0)[0] intercept   : AIC=11308.299, Time=0.07 sec
 ARIMA(0,2,1)(0,0,0)[0] intercept   : AIC=11224.980, Time=0.40 sec
 ARIMA(0,2,0)(0,0,0)[0]             : AIC=11596.112, Time=0.02 sec
 ARIMA(2,2,1)(0,0,0)[0] intercept   : AIC=11211.287, Time=0.28 sec
 ARIMA(2,2,0)(0,0,0)[0] intercept   : AIC=11265.666, Time=0.11 sec
 ARIMA(3,2,1)(0,0,0)[0] intercept   : AIC=11212.940, Time=0.32 sec
 ARIMA(2,2,2)(0,0,0)[0] intercept   : AIC=11211.859, Time=0.46 sec
 ARIMA(1,2,2)(0,0,0)[0] intercept   : AIC=11211.158, Time=0.23 sec
 ARIMA(0,2,2)(0,0,0)[0] intercept   : AIC=11216.391, Time=0.41 sec
 ARIMA(1,2,3)(0,0,0)[0] intercept   : AIC=11211.094, Time=0.45 sec
 ARIMA(0,2,3)(0,0,0)[0] intercept   : AIC=11210.173, Time=0.22 sec
 ARIMA(0,2,3)(0,0,0)[0]             : AIC=11208.027, Time=0.12 sec
 ARIMA(0,2,2)(0,0,0)[0]             : AIC=11214.114, Time=0.21 sec
 ARIMA(1,2,3)(0,0,0)[0]             : AIC=11209.028, Time=0.22 sec
 ARIMA(1,2,2)(0,0,0)[0]             : AIC=11209.157, Time=0.12 sec

Best model:  ARIMA(0,2,3)(0,0,0)[0]
Total fit time: 4.243 seconds
                               SARIMAX Results
==========================================================================================
Dep. Variable:                           y   No. Observations:                  664
Model:               SARIMAX(0, 2, 3)   Log Likelihood               -5600.014
Date:                Sat, 27 Nov 2021   AIC                          11208.027
Time:                        02:45:42   BIC                          11226.008
Sample:                             0   HQIC                         11214.996
                                  - 664
Covariance Type:                  opg
==========================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------
ma.L1         -0.8569      0.022    -38.858      0.000      -0.900      -0.814
ma.L2          0.2106      0.033      6.340      0.000       0.145       0.276
ma.L3         -0.1203      0.022     -5.409      0.000      -0.164      -0.077
sigma2      1.313e+06   3.77e+04     34.837      0.000    1.24e+06    1.39e+06
==========================================================================================
Ljung-Box (L1) (Q):                   0.03   Jarque-Bera (JB):               986.05
Prob(Q):                              0.86   Prob(JB):                         0.00
Heteroskedasticity (H):               2.37   Skew:                             1.17
Prob(H) (two-sided):                  0.00   Kurtosis:                         8.50
==========================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

The best model chosen is : ARIMA (p,d,q) = ARIMA(0,2,3).

As we assumed, second-order differencing

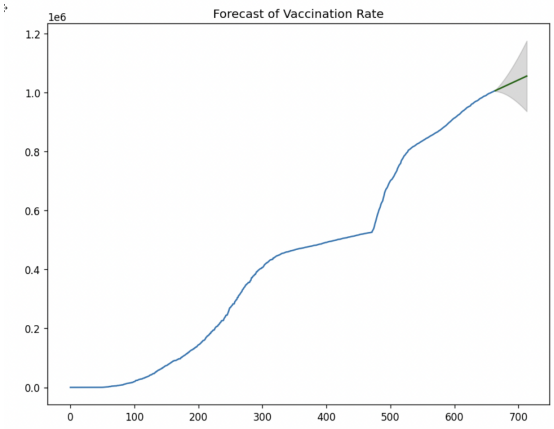
has to be performed for making the time series stationary in this scenario.

As the next step, perform the diagnostic check for the residual as given below:



we can forecast the vaccination rate now

The below plot shows the forecast of vaccination rate (green line)in Suffolk county based on the data we used



From the above graph, it will take another 100 days to get vaccinated completely at Suffolk county.

5 CONCLUSION

In this report, we have drawn the important features that are influencing the vaccination trend and we have seen how different factors are playing an important role and administration is giving importance to most SVI counties to protect the people's health. we have predicted the future trend for the particular county.

6. Reference:-

County view from CDC:

https://covid.cdc.gov/covid-data-tracker/#county-view

https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh

https://aspe.hhs.gov/reports/vaccine-hesitancy-covid-19-state-county-local-estimates

Birth ,Deaths, international mig, domestic….
https://www2.census.gov/programs-surveys/popest/datasets/2010-2020/counties/totals/

Factors:-

State Level Policies on covid vaccination: (Multiple Data sets for each policies)
https://www.kff.org/report-section/state-covid-19-data-and-policy-actions-policy-actions/

County level of Poverty, Population, Unemployment , Educational :
https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/

Vaccine Brand :
https://www.usatoday.com/in-depth/graphics/2021/01/14/covid-vaccine-distribution-by-state-how-many-covid-vaccines-have-been-given-in-us-how-many-people/6599531002/

https://www.cdc.gov/coronavirus/2019-ncov/vaccines/distributing.html

https://www.nytimes.com/interactive/2020/us/covid-19-vaccine-doses.html

Different data sets for county level metric:-

https://data.cdc.gov/browse?category=Vaccinations&page=2

https://www.nytimes.com/interactive/2020/us/covid-19-vaccine-doses.html

https://www.bloomberg.com/graphics/covid-vaccine-tracker-global-distribution/#us
https://www.washingtonpost.com/graphics/2020/health/covid-vaccine-states-distribution-doses/

Papers:
https://jamanetwork.com/journals/jama-health-forum/fullarticle/2773320