## Corona virus in US

On January 31, the first 2 novel coronavirus cases in the UK, the first 2 cases in Russia, and the first case in Sweden and in Spain were reported. Canada reported its 4th case. On Jan. 31, the United State [declared Coronavirus a Public Health Emergency](https://www.hhs.gov/about/news/2020/01/31/secretary-azar-declares-public-health-emergency-us-2019-novel-coronavirus.html) issued 14 days quarantine rules for US citizens entering the US from China (mandatory if entering from the Hubei province). issued an order to deny entry to foreigners who have travelled to China within the past two weeks. On January 30, the World Health Organization declared the coronavirus outbreak a Global Public Health Emergency. On January 30 CDC confirmed the first US case of human to human transmission.Germany, Japan, Vietnam and the United States have reported cases in patients who didn't personally visit China, but contracted the virus from someone else who had visited Wuhan, China. These cases of human to human transmission are the most worrisome, according to the WHO Wuhan (the city where the virus originated) is the largest city in Central China, with a population of over 11 million people. The city, on January 23, shut down transport links. Following Wuhan lock down, the city of Huanggang was also placed in quarantine, and the city of Ezhou closed its train stations. This means than 18 million people have been placed in isolation. The World Health Organization (WHO) said cutting off a city as large as Wuhan is "unprecedented in public health history. and praised China for its incredible commitment to isolate the virus and minimize the spread to other countries.

**Problem Definition**

This is the data repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). Also, Supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL).

**Coronavirus** is a family of viruses that can cause illness, which can vary from common cold and cough to sometimes more severe disease. **Middle East Respiratory Syndrome (MERS-CoV)** and **Severe Acute Respiratory Syndrome (SARS-CoV)** were such severe cases with the world already has faced. **SARS-CoV-2 (n-coronavirus)** is the new virus of the coronavirus family, which first discovered in 2019, which has not been identified in humans before. It is a contiguous virus which started from **Wuhan** in **December 2019**. Which later declared as **Pandemic** by **WHO** due to high rate spreads throughout the world? Currently (on the date 20 May 2020), this leads to a total of 300K+ Deaths across the globe, including 90K+ deaths alone in USA.The dataset  is provided to identify the deaths and recovered cases.

This data set contains an aggregation of each USA State level data. Let’s see what the different features this data set has are

* **Province\_State** - The name of the State within the USA.
* **Country\_Region** - The name of the Country (US).
* **Last\_Update** - The most recent date the file was pushed.
* **Lat** - Latitude.
* **Long\_** - Longitude.
* **Confirmed** - Aggregated confirmed case count for the state.
* **Deaths** - Aggregated Death case count for the state.
* **Recovered** - Aggregated Recovered case count for the state.
* **Active** - Aggregated confirmed cases that have not been resolved (Active = Confirmed - Recovered - Deaths).
* **FIPS** - Federal Information Processing Standards code that uniquely identifies counties within the USA.
* **Incident\_Rate** - confirmed cases per 100,000 persons.
* **People\_Tested** - Total number of people who have been tested.
* **People\_Hospitalized** - Total number of people hospitalized.
* **Mortality\_Rate** - Number recorded deaths \* 100/ Number confirmed cases.
* **UID** - Unique Identifier for each row entry.
* **ISO3** - Officialy assigned country code identifiers.
* **Testing\_Rate** - Total number of people tested per 100,000 persons.
* **Hospitalization\_Rate** - Total number of people hospitalized \* 100/ Number of confirmed cases.

Please go through the project: https://github.com/AshleshBR/Dynamic\_projects-M20/blob/main/Dynamic%20project-9(covid\_US).ipynb

**Data Analysis**

After processing of data and exploring the various variables in the data set we can come to the conclusion that there are null values and some of the columns have very low correlation with target so I have decided to drop those columns. Coluimns with categorical data have been labelled Summary statistics shows that

* Columns like lat, long, recovered ,incident rate, people tested, people hospitalized, mortality rate, testing rate, Hospitalization Rate have null values we need to fill those values.
* There are total 58 province states. Country region is US only so we can drop this column
* Last date is same for all the rows so we can drop that also
* Iso 3 has 6 categories.
* By looking at the values of mean, median , and 75% and 100% of the each column,we can say that skewness and outliers are present in some of the columns

**Univariant analysis:**

Let’s start with univariant analysis of the variables which has high correlation

In the first univariant analysis I have tried to study the distribution of active, recovered ,confirmed and deaths cases recorded in various states .From the graph we can see that density of confirmed ,active, recovered and deaths are high at the begining of graph .only 22% cases are recovered 75% is still active.so only few states are highly infected(more than 1lakh),so we can say that covid spread is in its starting stage. Next analysis of People Tested and People Hospitalized and it is found that Out of total people tested only 1.53 per cent of people hospitalized

# Bivariant analysis

In bi variant analysis I have performed analysis of confirmed, active, recovered and deaths at various provinces and it is found that covid spread started first at province “36” which has maximum cases and deaths covid spread started just now in province 2 which has least cases for now. During testing rate at various provinces it is found that Testing rate is maximum at province 45 and minimum at 44 Average testing rate in country is 3894.

After performing bi variant analysis I have decided to check the correlation .skewness, and outliers using boxplot .

**Pre-processing Pipeline**

Data pre-processing is a predominant step in machine learning to yield highly accurate and insightful results. Greater the quality of data, the greater is the reliability of the produced results. Incomplete, noisy, and inconsistent data are the inherent nature of real-world datasets. Data pre-processing helps in increasing the quality of data by filling in missing incomplete data, smoothing noise, and resolving inconsistencies.

* Incomplete data can occur due to many reasons. Appropriate data may not be persisted due to a misunderstanding, or because of instrument defects and malfunctions.
* **Noisy data** can occur for a number of reasons (having incorrect feature values). The instruments used for the data collection might be faulty. Data entry may contain human or instrument errors. Data transmission errors might occur as well.

There are many stages involved in data pre-processing,

**Data cleaning** attempts to impute missing values, smooth out noise, resolve inconsistencies, removing outliers in the data.

**Data integration** integrates data from a multitude of sources into a single data warehouse.

**Data transformations**, such as normalization, may be applied. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements.

**Data reduction** can reduce the data size by dropping out redundant features. Feature selection and feature extraction techniques can be used.

Lets see how we have done our data pre-processing in this data set.

* Using the the box plot we can see that there are lots of outliers in various columns and hence we removed those outliers using the z score method
* Using the hist plot we can check the skewness. usually skewness + or – 0.5 is acceptable anything above that needs to be corrected. While removing the skewness it is important to consider only independent variable and not to consider categorical variable. By applying these rules we have applied power transformation function to remove the skewness.

Data standardization is this process of making sure that your data set can be compared to other data sets. It’s a key part of research, and it’s something that everyone who uses data should consider before they even collect, clean, or analyse their first data point.in our data set we have standardised the value using StandardScaler function,science there are not categorical independent feature we can apply for whole x variable and also it I have kept in mind that I have to predict that both recoveries and deaths for different instances so I have separated x1,y1 for recoveries prediction and x2 and y2 for deaths prediction.

Now we have already separated the feature variables and target variables data is now ready for model fitting

**Building Machine Learning Models**

Building machine learning models that have the ability to generalize well on future data requires thoughtful consideration of the data at hand and of assumptions about various available training algorithms. Ultimate evaluation of a machine learning model’s quality requires an appropriate selection and interpretation of assessment criteria.

At this stage, you develop an understanding of your problem which you are trying to solve. Now your data is also in its usable shape. Now it’s time to select and train your machine model. There are many models that you can select according to your business objectives. The step of selection of models includes algorithms of prediction, classification, clustering, deep learning, linear regression, and so forth. Now you will be required to train datasets to operate smoothly. The step of training your machine model involves several algorithms and techniques. The outcome machine model can be used for evaluation to check whether it meets the operational and business requirements.

In our case the problem type in both the case is regression type and we have imported and applied all the regression algorithm and picked the best result.we got the best result in RandomForestRegressorn algorithm for prediction of recoveries and decision tree for prediction of deaths. we have checked for any over fitting using k fold cross validation technique. It is a technique for assessing how the results of a [statistical](https://en.wikipedia.org/wiki/Statistics) analysis will [generalize](https://en.wikipedia.org/wiki/Generalization_error) to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how [accurately](https://en.wikipedia.org/wiki/Accuracy) a [predictive model](https://en.wikipedia.org/wiki/Predictive_modelling) will perform in practice.

In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (called the [validation dataset](https://en.wikipedia.org/wiki/Validation_set) or testing set). The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like [overfitting](https://en.wikipedia.org/wiki/Overfitting" \o "Overfitting) or [selection bias](https://en.wikipedia.org/wiki/Selection_bias) and to give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem).after performing the cross validation I have got nearly same accuracy as of trained data set hence I have proceeded to tune the parameters of algorithm and search for best parameter combination using grid search cv function. It is also called hyper parameter tuning Different model training algorithms require different hyper parameters . A hyper parameter is a [parameter](https://en.wikipedia.org/wiki/Parameter) whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned. Grid search cv and randomized search are the two famous hyper parameters used.in our case I have used grid search and extracted the best combination of parameters for model.

After the parameter tuning not it is time to predict the data and compare the results with labelled data .the model is then evaluated using different evaluation metrics. This step involves the evaluation of the machine models using a model metric approach, quality measurements, datasets, and matrix calculations. This phase is the quality assurance of a machine learning approach. This is done with the help of metrics like r2 score.

**Concluding Remarks**

After getting the satisfied results we can save the model for production. I use pickle function to save the model

So this is how I have built the model for the prediction of recoveries and deaths cases across US due to covid-19.

Thanks you

Ashlesh B R