Statistics and Machine Learning
(20201:645:652:01)

Analysis and prediction of Covid-19

Professor
Prof. Patrick Shafto / Prof. Pushpi Paranamana

Group Members
Omkar Vilas Dhuri - Omkardhuri2508@gmail.com
Venkata Alekhya Varapula - varapula.alekhya@gmail.com
Deepal Rathod - deepal11rathod@gmail.com
Ashish Mohan - ashmct1995@gmail.com
Ashique Nawaz chowdhaury - anc.nawaz93@gmail.com

Introduction:

Coronaviruses are a large family of viruses which may cause illness in animals or humans. In humans, several coronaviruses are known to cause respiratory infections ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). COVID-19 is the infectious disease caused by the most recently discovered coronavirus. This new virus and disease were unknown before the outbreak began in Wuhan, China, in December 2019.

Goal:

To analyze and track the coronavirus outbreak and be aware of the actual stats that are affecting the areas that we live in and the whole world. Through this project we aim to perform analysis on available data on COVID-19. We plan to use classifiers, unsupervised learning and forecasting to help arrive at our mentioned objectives.

Objectives:

1) Through our EDA, we are predicting the survival rate of individuals in USA and worldwide.
2) The model with all the coefficients leading to death or even infection
3) Prediction of the curve (when it is going to end and when it will flatten)
4) To implement an unsupervised learning model (K-means clustering) to identify for underlying groups or categories existing in our data and gain possible insights.
5) Best Forecasting model that would predict the optimal future trends.

Target Audience:

Researchers, Statisticians, hospitals, common people

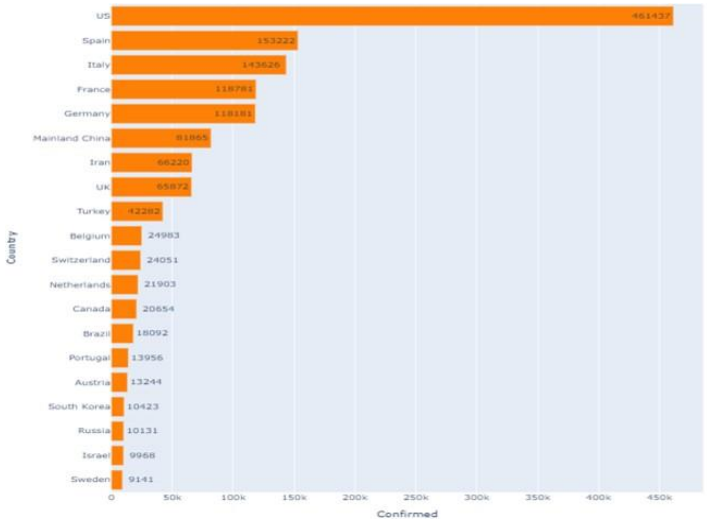
Dataset:

The dataset has been collated from Johns Hopkins GitHub repository and we have utilized 2 level of csv files to identify Exploratory Data Analysis and perform statistical modelling on the coronavirus outbreak data. It has over 18000 records with 23 columns over a period of January to April 2020 used for further analysis. The columns provide the total number of cases as well as detailed information on patients.

Data Cleaning:

- Derived new columns:
  - Active = Confirmed cases - Death cases - Recovered cases
  - Difference = Hospital visit date - exposure date
- Updated the blank values in a few columns with zeroes, as required for numerical calculation.
- Date formatting on all date fields to datetime.
- Created data frames with filters on specific countries and states for detailed analysis
- Replaced wrong values in ‘death’ column with appropriate flags.
- Dropped null values from 'Difference', 'age' and 'gender' columns.
- Replaced null values to unknown in ‘symptoms’ column.
- Replaced gender values to flags as required for modelling.
- Sorted the data by date field.

Exploratory Data Analysis:



This plot states "Confirmed Cases Worldwide" through Bar Graph divided by countries.

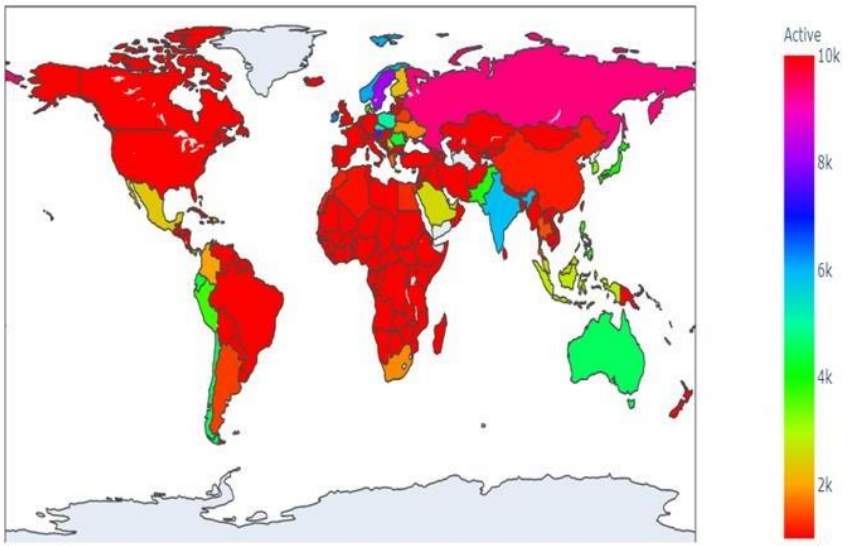
United States is the most affected country with a total count of 461.437k cases followed by Spain and Italy.

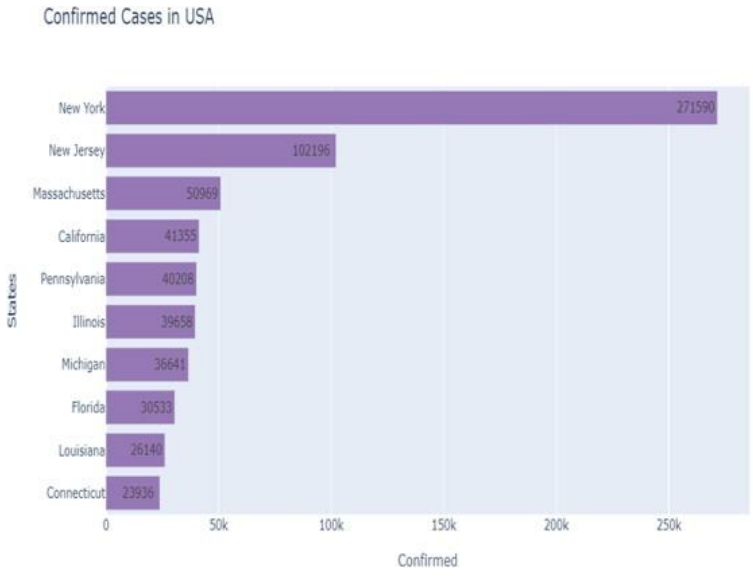
Sweden is the least affected with 9.141K cases in total.

This is a world map created through plotly graph which states "Active cases Worldwide".

United States has highest number of active cases with total cases of 419.549K, followed by other European Countries.

Here we can see that, China has very low Active Cases with only 1154 Cases which was the epicentre of Covid-19.





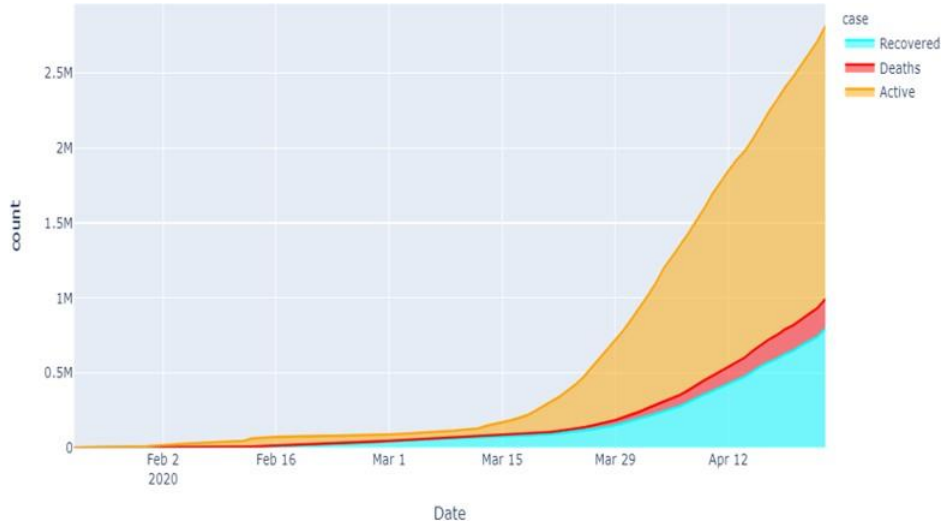
This plot states "Confirmed Cases in USA" which is divided into States.

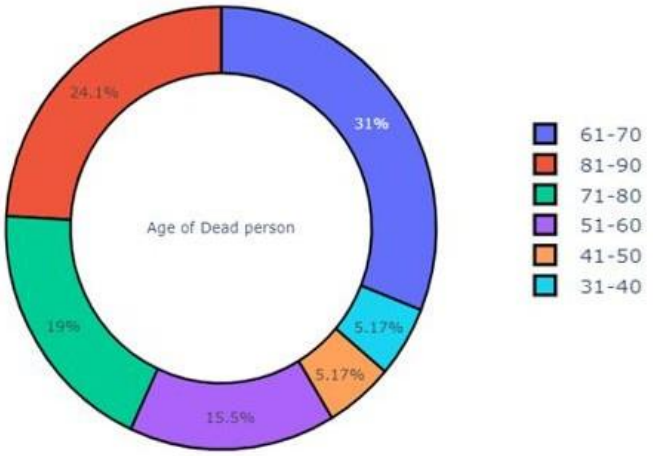
New York is the most affected state with total of 271590 followed by New Jersey.

Montana is the least affected state with only 453 cases.

Active cases raised drastically after March 24th 2020 and might see a big rise in the trend.

we can see there is slow increase in death and recovered cases as well after March 24th, 2020.





This plot states "Age of Dead Person" through iplot.

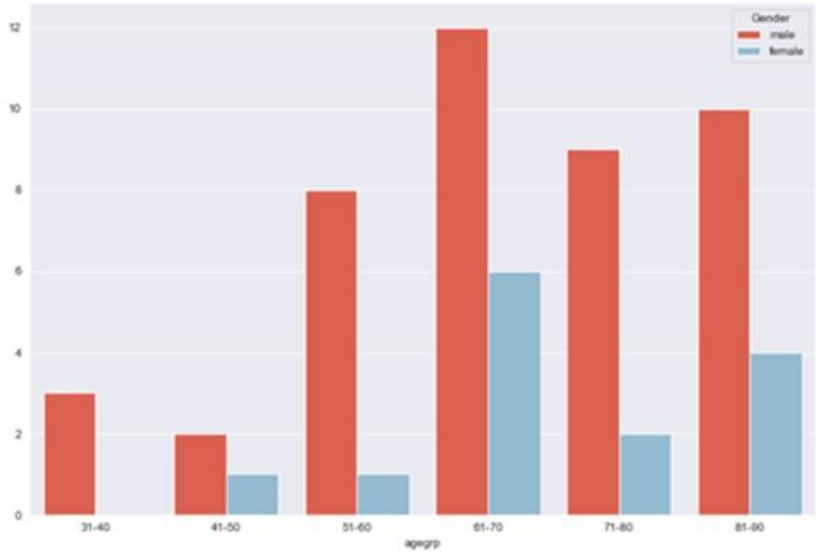
The age group of 61-70 are at higher end of the risk with a total death cases of 31% which is followed by 81-90.

31-50 age range is least with just 5% cases.

The bar graph shows the categorization of female and male and the total deaths for various age groups.

Deaths in male is higher compared to female at every age group.

The age-group of 61-70 have maximum number of deaths in both male and female.



Machine Learning:

Random Forest model

For our Random Forest model, we decided to use 'age', 'Difference', 'Gender' columns to predict the chances of survival or death for an infected person. We decided to choose this model because it is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. Also, it is a Multiclass Classifier, as using predict_proba method which is very efficient and checks for probability of each class.

```
#Defining X and Y
X = df[['age', 'Difference', 'gender']]
Y = df['death']
```

```
from sklearn.model_selection import train_test_split
X_Train, X_Test, Y_Train, Y_Test = train_test_split(X, Y, test_size = 0.25, random_state
```

```
from sklearn.ensemble import RandomForestClassifier
rfc = RandomForestClassifier(random_state=100)
```

Cross -Validation

Evaluating a classifier is more complicated than evaluating a regressor (t-test, F-test, etc.). One way to evaluate a model is to use cross-validation which breaks up the training data into different subsets, called K-folds. We found out the cross-validation score for our data, which showed a good score for the three folds of validation.

```
#For checking accuracy in rfc
from sklearn.model_selection import cross_val_score
cross_val_score(rfc,X_Train,Y_Train,cv=3,scoring='accuracy')
```

```
array([0.92792793, 0.9009009 , 0.9009009 ])
```

Error Metric:

The precision and recall for our data were as follows:

```
#Finding precision score
from sklearn.metrics import precision_score
precision_score(Y_Train ,Y_Pred1)
```

```
0.44
```

```
#Finding recall score
from sklearn.metrics import recall_score
recall_score(Y_Train ,Y_Pred1)
```

```
0.4074074074074074
```

We use a measure that combines precision and recall into a single variable called the F1 score, which is a harmonic mean.

A harmonic mean gives more weight to lower values, as a result a classifier can only get a high score if both values are high.

The F1 score obtained was:

We see our F1 score is 42 % which is distributed on both precision and recall.

```
from sklearn.metrics import f1_score
f1_score(Y_Train ,Y_Pred1)
```

```
0.4230769230769231
```

```
from sklearn.metrics import roc_auc_score

roc_auc_score(Y_Train,forest_scores[:,1])
```

```
0.8569958847736625
```

The roc_auc score was the following for our model: ROC Score is 85.7%

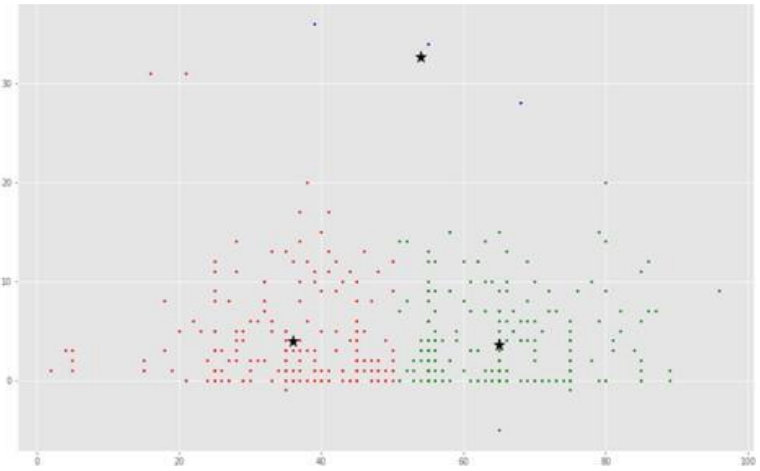
K-Means clustering:

On our dataset we aimed at finding existing clusters present in the data between two columns viz. Age & Difference.

We then implemented the K-Means algorithm on our data, by setting the number of clusters as 3, the following shows the random positions of the 3 cluster points in our data.

From our K means clustering 3 significant cluster groups were observed, out of which 2 were age groups of 20-50 and 50-100 takes around maximum of 20 days, to seriously fall ill which is described by 'Difference' attribute viz the difference between infected date and hospital visit date.

Additionally, there is one more group (outlier points), which takes more than 30 days and is scattered across all age groups.

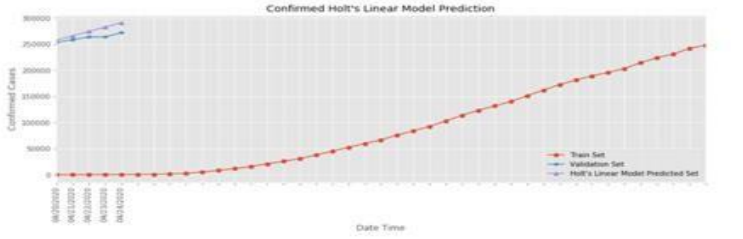


Forecasting Models:

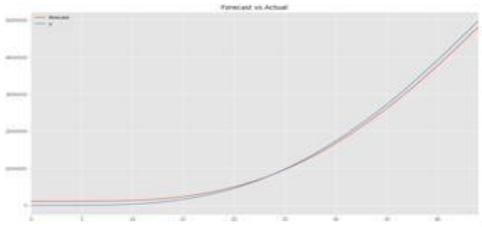
Holt's Linear Trend Model:

This method considers the trend of the dataset. The forecast function in this method is a function of level and trend.

Since New York is next target of whole world and it is highly diverse, hence we will forecast the infection rate in New York, and we see an increasing trend in New York as of now.



We could see our model is predicting pretty well. However let us try ARIMA model considering it might not have trend as well. But we have to convert time series data into stationary data for the same



ARIMA Model Results

Dep. Variable:	D.y	No. Observations:	45
Model:	ARIMA(0, 1, 0)	Log Likelihood	-579.460
Method:	css	S.D. of innovations	94651.746
Date:	Sun, 26 Apr 2020	AIC	1162.921
Time:	22:22:23	BIC	1166.534
Sample:	1	HQIC	1164.268

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.105e+05	1.41e+04	7.833	0.000	8.29e+04	1.38e+05

ARIMA model:

ARIMA describes the correlation between data points and considers the difference of the values. It aims to describe the autocorrelations in the data, and we use ARIMA considering in COVID 19 no seasonality and trend as well.

Conclusion:

- From our analysis and EDA, we identified USA would be the most affected country trending in larger number of confirmed and death cases in current situation followed by European countries.
- Through the RFC model we see that though death count is low now, but it would potentially increase in future without necessary steps. Age, gender, getting medical attention are not the only attributes to be considered for survival and we could improve our model by maximizing recall.
- From both Holts Linear Trend and ARIMA model we could forecast increasing number of cases. But Holts Linear Trend model is better for forecasting in COVID 19 case since we could see an increasing trend and fluctuations are irrelevant and the same could be endorsed by RMSE values of both the models.
- From our K means clustering we found that age group of 20-50 and 50-100 takes around maximum of 20 days (with maximum people in the group of 50-100) to seriously fall ill and is described by difference between infected date and hospital visit date.

References:

- John Hopkins Git repository: https://github.com/CSSEGISandData/COVID-19
- https://www.cdc.gov/coronavirus/2019-ncov/index.html
- https://www.analyticsvidhya.com/blog/2018/08/auto-arima-time-series-modeling-python-r/