Advanced Mobility & Cloud Computing

Project Title: Visualization & Analysis of Corona Virus Data

Author: Sandeep Singh

Abstract

During the times of pandemic when Corona virus is spreading at such an enormous rate and tons of cases are being registered every day, I will be working on some Covid-19 data sets in order to visualize and analyze it.Data can be analyzed state wise, age wise and there are several other factors like age, recovered cases, deaths and most common symptoms amongst the patients. Also, we will be discussing about some of the reasons for the spread of Corona Virus along with its fatality rate, infectious rate and incubation period. The tools that are used for the analysis are the Jupyter Notebook and Amazon web Services like the Amazon SageMaker Studio.

Keywords: Covid-19, corona, SARS, Horseshoe, Genetic mutation, Ticking time bomb, MERS, CoV-2

Table of Contents

1. Introduction

Corona Virus also known as COVID-19(SARS-CoV-2) was first found in Wuhan city of china and this virus is caused by the Horseshoe Bats as it is a bat borne virus. The life cycle of this virus is 14 days which is also known as the incubation period and it comes under the SARS family (Severe Acute Respiratory Syndrome). Shockingly, the first case of SARS-CoV was found in Guangdong, China in November 2003 and it was less lethal at that time as the genetic structure of the virus was not so complex. Also, there were only 12 cases of Corona virus back in 2003. Hence, it was easily cured by the doctors and then Research paper was published by some scientists on Corona virus in 2007 stating that it’s spread was controlled without much effort and tried to warn everyone that if the structure of the virus undergo genetic mutation it will lead to far more dangerous form of virus and it would be very hard to find a cure or a vaccine for it.

Corona virus belongs to the first cluster of the SARS-CoV and its virus strain of SARS (SARS-CoV-2) which is contagious in human beings starting its onset from the Huanan seafood market. This virus was initially believed to be from the family of MERS (Middle East Respiratory Syndrome) which was discovered in 2012 in the desserts of Saudi Arabia and it is caused by the contact of Dromedary camels. But when there was some analysis done for the initial patients of corona virus then it was found that just 9 out of total patients had contact with the camels. Also, this zoonotic virus affects 75% of males whereas Corona virus affects both males and females equally. Hence, we concluded that MERS was not the case instead it was some new strain of virus new.

This virus enters the body of the host with the help of binding to the ACE2(Angiotensin Converting Enzyme 2) and it is done by coming in close contact or contact with the respiratory droplets of the infected patient .Corona was referred to as “Ticking Time Bomb” in that research paper back in 2007.

2. Implementation

This project will be performed with the help of **Anaconda** which uses python as a programming language. Also, it is referred as packaging manager of python which is provided by **conda.** It is comprised of various python libraries. Jupyter is the presentation layer which is used to create notebooks to run the kernel as it's an interpreter that can be implemented in Spark, Python and R. Jupyter is used for analytical work as a virtual notebook. Kernel is stored in the Amazon Cloud and it may be stored on one of the local machine known as localhost, so we will try to implement it on cloud using Amazon S3 web services so that it can it virtually accessed anywhere.

2.1 Understanding the Architecture

Before discussing about the architecture let us first discuss about some terminologies that is crucial and was used in the project.

Pandas: Data Analysis, manipulation and cleaning is done with the help of pandas. Mostly pandas is used to read and manipulate csv files by using Dataframe features which is basically a 2D table that is generated from the spreadsheet of csv files.

NumPy(Numerical Python): It is a library of python which is used to compute large amount of multidimensional data which has array and matrices like structures and it is mainly used in machine learning . Further this can be used to do operations using the indexes of the array or list and implementing different functions onto them. It consumes less memory as compared to Pandas.

Matplotlib: It is a library of python which is used of visualization of data that we get from the 2D dataframes. Pyplot is an API which works on the axes and these axes are defined on the base of indexes of the 2D dataframe. It allows the use of Pyplot to create some figures and bar charts just like the functionalities and working that we get in MATLAB.

Now, let's discuss some important concepts about the Jupyter Architecture

Jupyter is composed of Kernel which is used for computation for more than one frontend or we can say sockets. As we are using python language, so it is referred to as IPython Kernel which takes input in the form of **JSON** (JavaScript Object Notation) messages and executes it and the result is returned to the Client. Hence, we came to a conclusion that there exists two-process model that is implemented in the case of Jupyter. And as JSON is language independent which makes easier for the user to interact with the Jupyter.

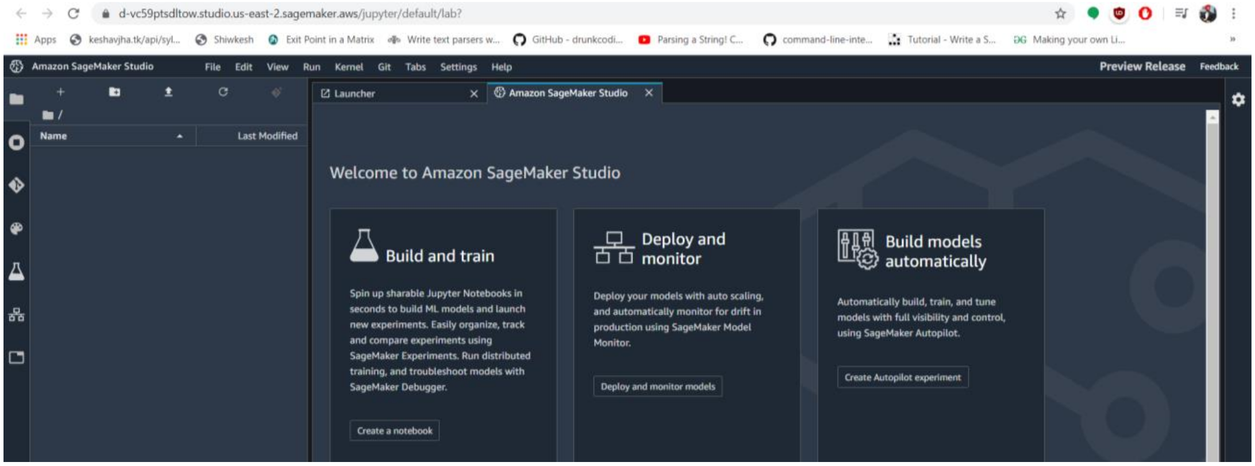
2.2 Establishing Connection with AWS

While setting up a virtual server in cloud, we need to create and initialize an EC2 instance in the AWS.EC2 stands for Amazon Elastic Compute Cloud which is a virtual server and rather than creating a local instance of Jupyter Notebook we will learn to set it up on cloud in order to virtualize it and our work can be stored on the cloud. Now we will be discussing about the steps that we performed in order to set up EC2 instance. It is used for machine learning and data analysis.

Amazon SageMaker Studio

Available in preview

Loading the JupyterServer application default...



d-vc59ptsdltow.studio.us-east-2.sagemaker.aws/jupyter/default/lab?

Welcome to Amazon SageMaker Studio

Build and train

Spin up sharable Jupyter Notebooks in seconds to build ML models and launch new experiments. Easily organize, track and compare experiments using SageMaker Experiments. Run distributed training, and troubleshoot models with SageMaker Debugger.

Deploy and monitor

Deploy your models with auto scaling, and automatically monitor for drift in production using SageMaker Model Monitor.

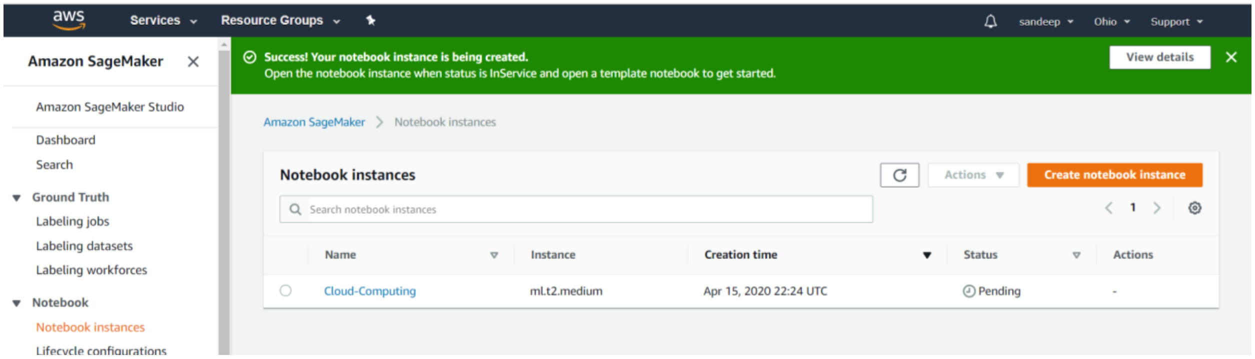
Build models automatically

Automatically build, train, and tune models with full visibility and control, using SageMaker Autopilot.

The steps are as follows:

1.Once we login into Amazon Aws services we will be opening Notebook Instances and then clicking on creating an instance.

2.After entering the name of the instance (here we have used the name of instance as Cloud Computing), then we will select the type to be m1.t2.medium as the data load is not so much and we will not be needing more Memory and CPU to work with.



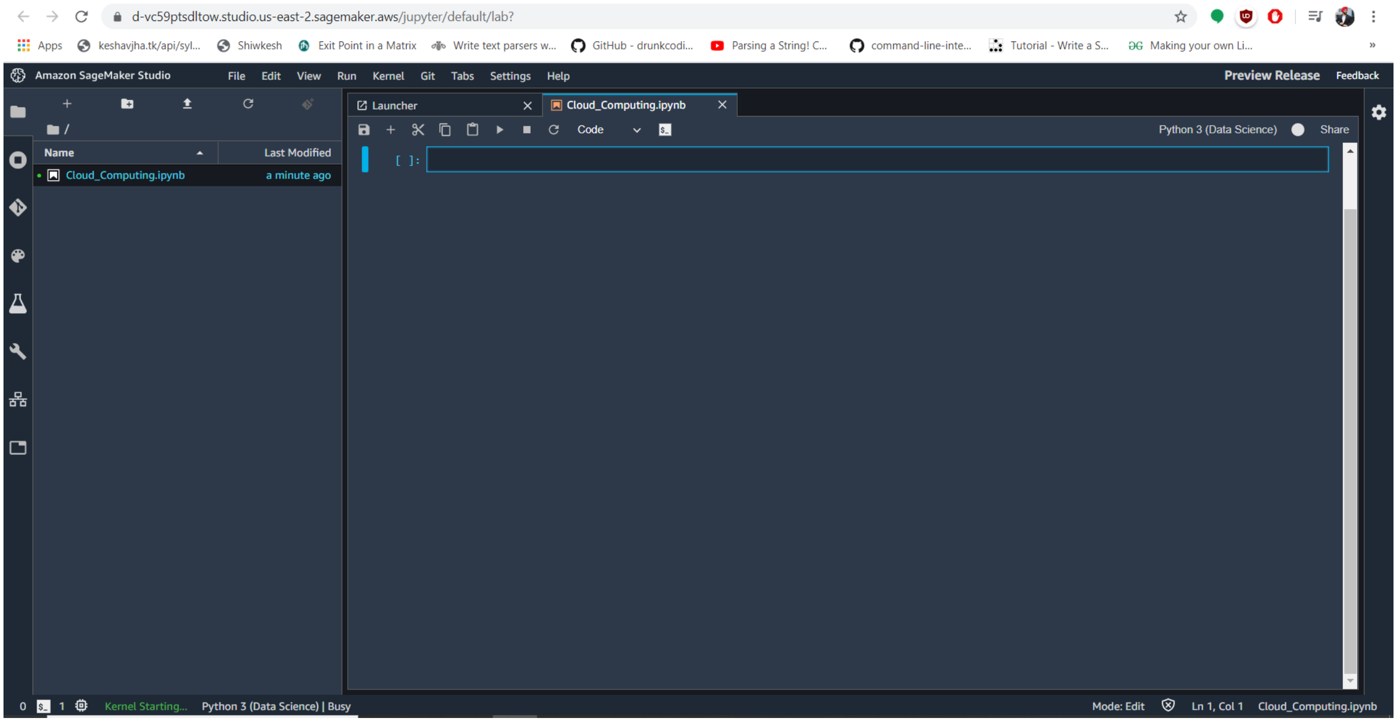
Success! Your notebook instance is being created.
Open the notebook instance when status is InService and open a template notebook to get started.

Notebook instances

Name	Instance	Creation time	Status	Actions
Cloud-Computing	ml.t2.medium	Apr 15, 2020 22:24 UTC	Pending	-

3.Medium instance allows us to access 2 virtual CPUs and memory up to 4-5 Gigabytes.

4.After creating the instance we can start the instance in the AWS SageMaker Studio.



Name	Last Modified
Cloud_Computing.ipynb	a minute ago

0 1 Kernel Starting... Python 3 (Data Science) | Busy Mode: Edit Ln 1, Col 1 Cloud_Computing.ipynb

5. This is how our Cloud Computing Kernel(with extention .ipynb) looked like and the address of the virtual server can be noticed in the address bar and all the computation with the CSV files were done in these notebook cells.

2.3 Data Collection

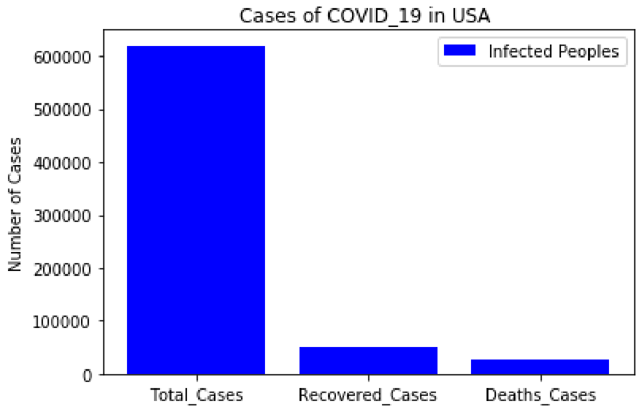
Data sets were collected from the official site of CDC(Centers for Disease Control and Prevention) and WHO(World Health Organization).The data was not filtered and the type of file in which the data was stored had an extension .CSV and they were up to date until 15th April 2020 so data reflected after that won’t be covered in the analysis . Kaggle is also turned out to be a good platform to get data sets and information that we were not able to conclude otherwise. These data files will be cleaned using any programming language like python.

2.4 Data Cleaning

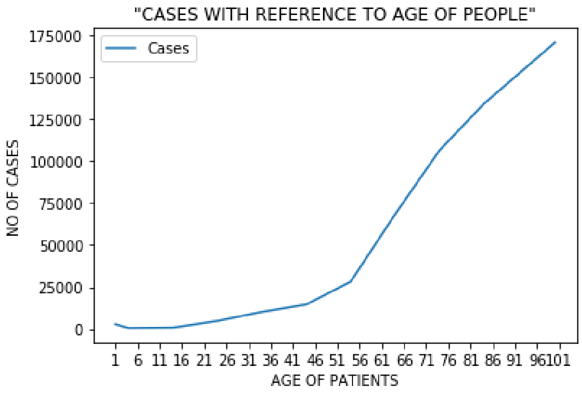
Data Cleaning comprises of removing any unwanted and useless attributes. Here in this project we have done this step with the help of creating dataframes in python and dropping unwanted columns and storing the required attributes into list that is later used of analysis. Also, we could have used online openCSV parser, but its performance is degraded exponentially in worst case scenario’s when the number of rows and columns increases. Hence, the column names were renamed for labeling so that the data can easily recognized by the naïve audience.

2.5 Data Analysis

Let us first start with the analysis of these two plots where Plot 1 refers to the case history of COVID-19 and Plot 2 refers to the cases associated with each age group for better understanding of how age factor affects the COVID-19 cases.



Plot 1

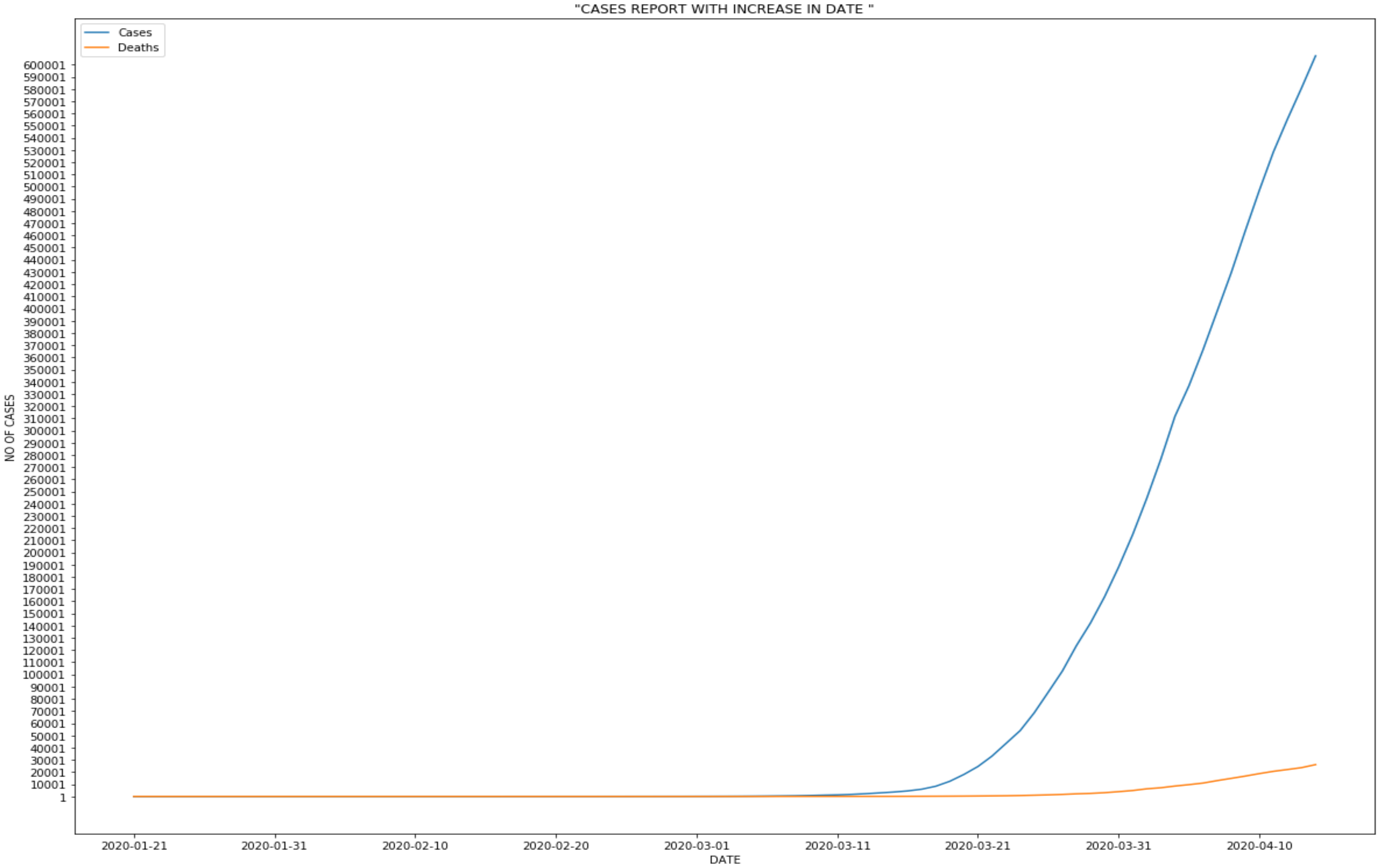


Plot 2

Plot 1: In this case we are trying to show the number of infected cases until April 15th and then the number of recovered patients and deaths. As in numbers the Total cases were 618923 and the recovered patients were close to 9% (49998) of the total cases and the deaths constitutes about 5% (27112) of the total number of cases.

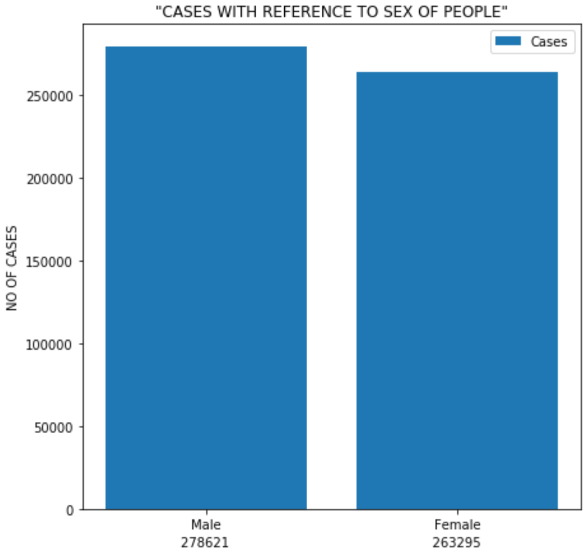
Plot 2: This figure portraits the increasing number of cases as the age number increases. Here we can notice that the rise in the number of cases is not so steep until the age increases after 40. It was found that about 70-92% of the infected people lies in this range Hence this was concluded that that the immune system of people above 40's is not so strong to defend the virus and they are more prone to virus. It is similar with the death cases too.

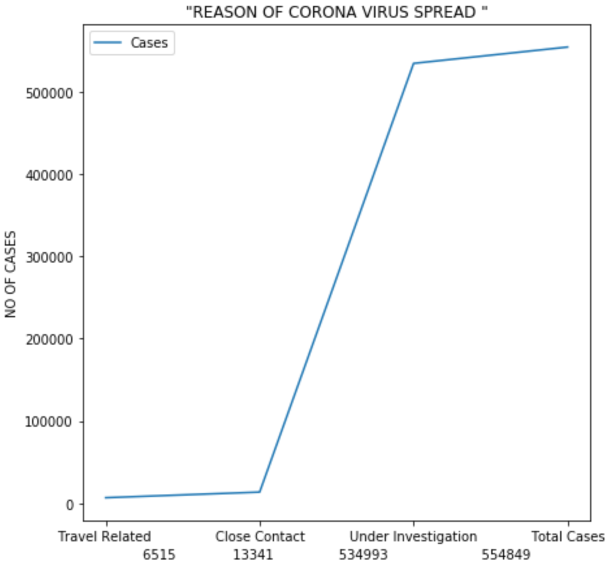
Now we will compare the positive COVID-19 cases vs the deaths with reference to the date in the next analysis.



Plot 3:

This representation shows us the number of infected cases that came in with reference to date and alongside it we are noticing the number of death cases. The peak in the registered cases was observed in between March 11th and March 21st. Probably this is the time when lockdown was imposed by the government of United States of America. 13th March was the day when National Emergency was declared due to the Covid-19 pandemic and first state to implement lockdown was Alabama on March 13 followed by Alaska on March 19. This is referred as the acceleration phase of the virus and stop or control it social distancing was imposed by the government.





Plot: 4

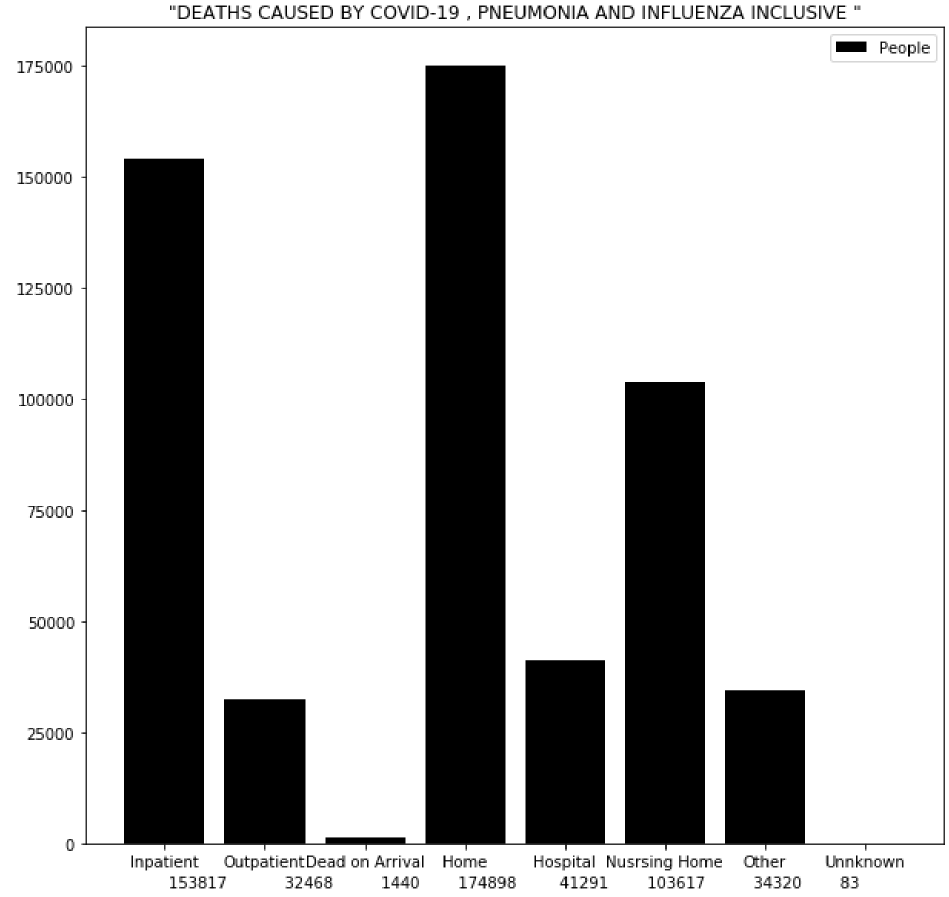
Plot: 5

Plot: 4

Unlike MERS zoonotic virus which is known to affect males up to 75%, COVID-19 affects both the genders equally. Hence, we concluded that 51.14% affected patients were males and rest of them were females. So, the ratio of infected cases in both genders were pretty much same but when it comes to the number of deaths in genders then it was found that females have more immunity genes due to the X chromosome. Hence, number of deaths in males were noticed to be higher than the females.

Plot: 5

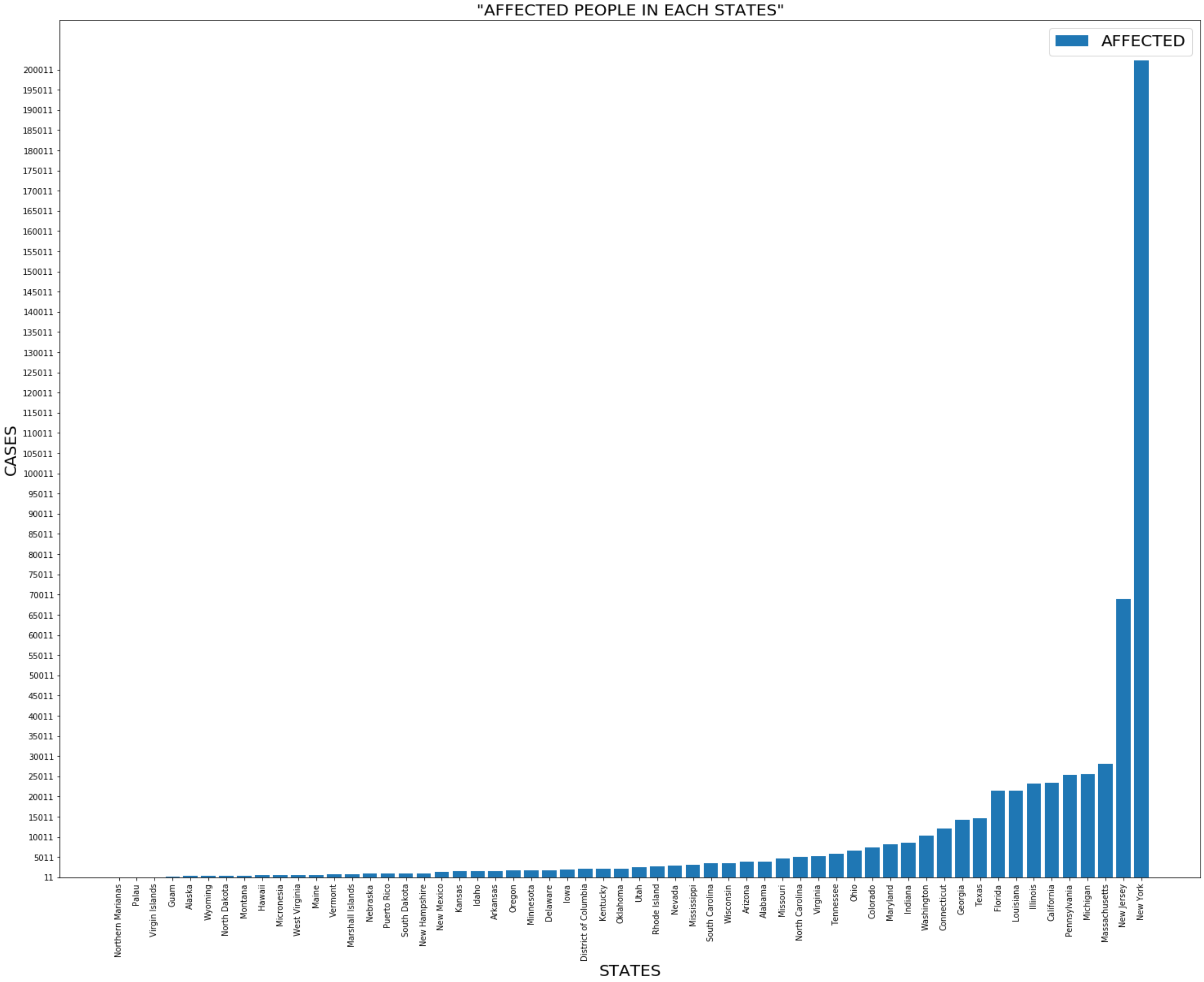
This figure describes us about the reasons that the Corona Virus was spread and now that whole medical field is focused to save the patients and find the cure of the virus, there wasn't much to conclude about the spread. We found that close to 2% of cases were due to travelling abroad and 3% of cases were due to close contact with the infected patient. Rest of them are under investigation to determine what could be the source of the virus. Also, there some extra information that I would like to add about the spread of virus. Major reason for spread of virus is encountering the cough droplets of the infected patient. For that one must be in the range of 1.8 meter to 6 feet of the infected person. Another possible way for contact with virus is indirect contact and this can be happening through steel (3-4 days), cardboard (1 day) and copper (4 hours - 1 day). Also, this was found that this virus cannot breathe in water and it only doubles in suitable heat environment once it enters the body of a person.



Plot: 6

Plot: 6

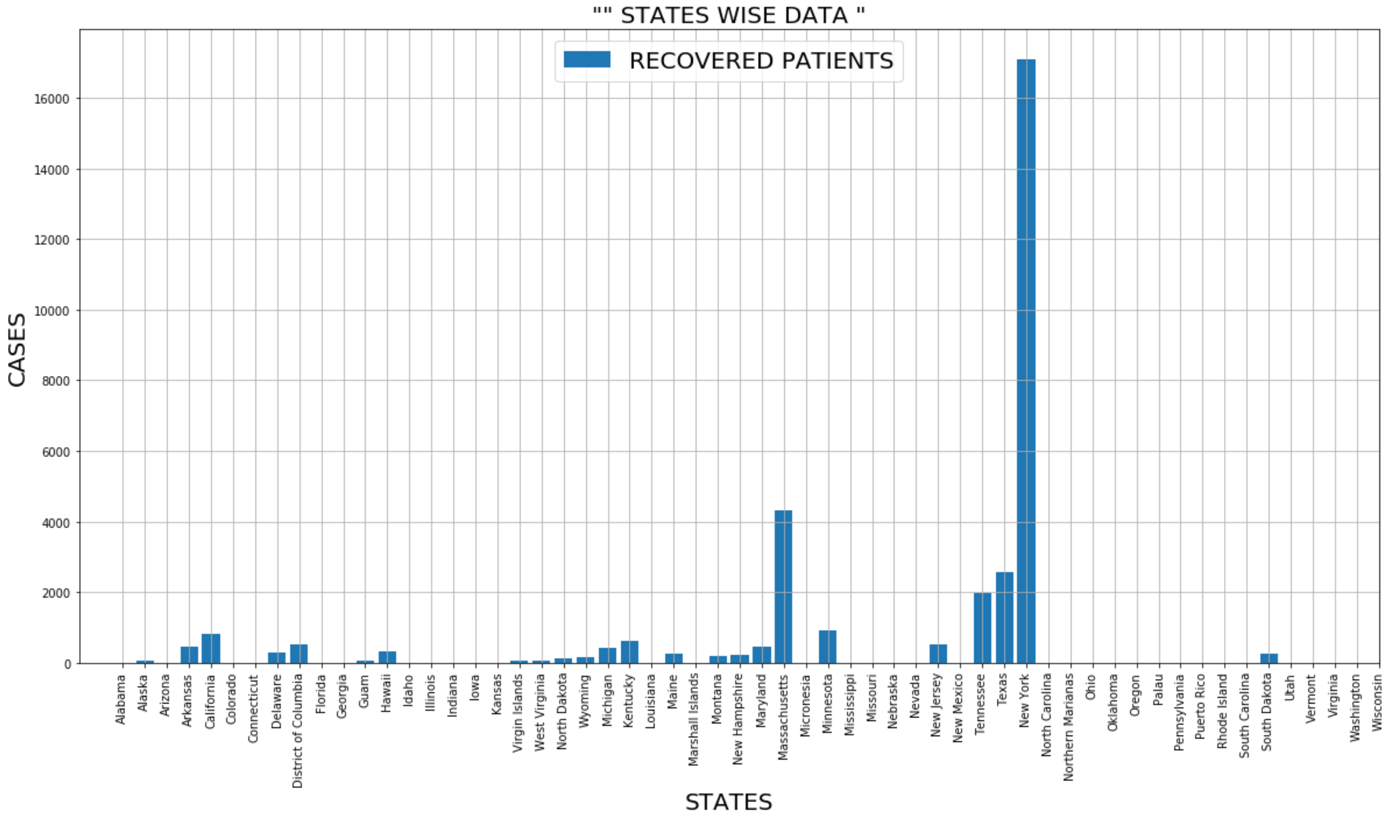
Here we are depicting the place of deaths due to COVID-19. Reason of deaths are COVID-19 , pneumonia and influenza as initially pneumonia was also included as a symptom of virus but later on when more number of cases came in then we were limited to only 3 symptoms and those were Fever, Shortness of breath and Cough. As we can notice that maximum deaths were noticed in home (174898) and inpatient cases (153817). This shows that there is severe shortage of beds in hospitals and it was reported that most of patients were sent home after giving them specific medications as there were no space to keep them and there was no assurance that it was a case of corona virus as the incubation period is of 14 days and it is not possible for any medical facility to notice some patient for this long. So, guidelines were given to head towards medical facilities only if any alerting signs were noticed like chest pain, difficulty in breathing and bluish lips. Later we will discuss a little about medications also.



Plot: 7(a)

Plot: 7(a)

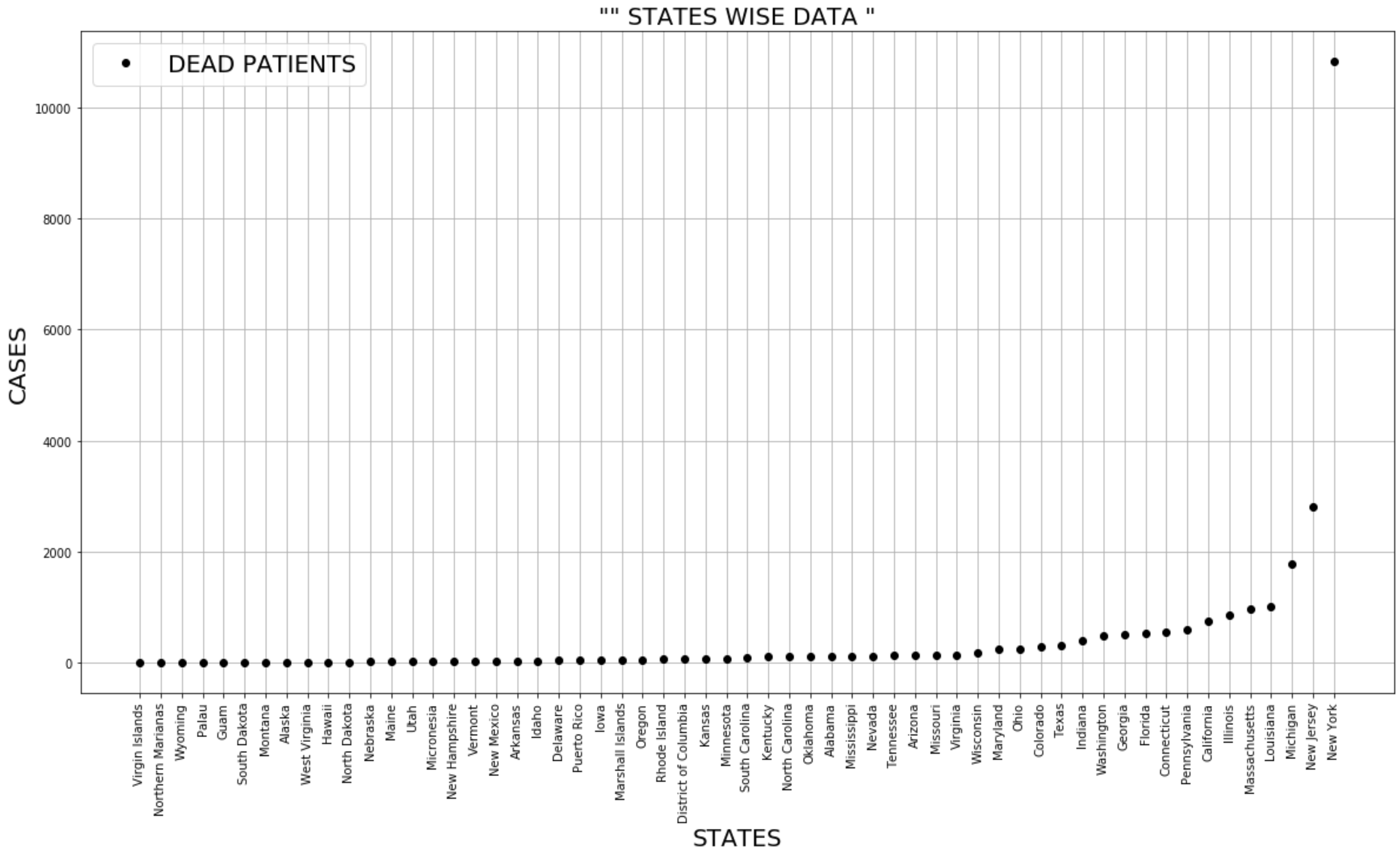
This figure represents the number of confirmed cases in each state of United States of America. It can be analyzed that the number of cases is directly proportional to the population of a state and inversely proportional to area per square mile in which people are residing. Top states that has maximum cases of COVID-19 is New York (202208), New Jersey (70000) and Massachusetts (30000). Whereas the minimum number of cases numbered to be 11 from Northern Marianas which is even less than 1% of total number of cases. The number of cases in New York is 33.7% of the total confirmed cases in USA. This was until April 15th but now the situation is worsened as the maximum number of deaths per day is noted to be 731 and maximum number of cases has gone to an extend of 3563 cases per day which is highest than any other state in the entire world. Another reason is that the lockdown and social distancing was implemented a little late in the such states which lead to the rapid increase in the cases as the Infectious Rate of COVID-19 lies between 1.9 to 3.9.



Plot: 7(b)

Plot: 7(b)

This figure shows us the number of Recovered patients in each state. Recovered patient data was not found for few states but major states have their recovered data reported. We can conclude that about 49% (17089) of total recovered patients are from New York and 13.1% (4316) of total recovered patients are from Massachusetts. Additionally, I would like to add that after the medication received from India that contains Hydroxychloroquine there has been some relief in recovery of patients. About 3.6 million tablets were delivered to USA. Minimum number of recovered patients were noticed from Virgin Island that is 43. Now we will discuss about the number of dead patients in the next figure.



Plot: 7©

Plot 7©

This figure denotes the number of dead patients from various states of USA. The black dots are representing the number of patients in Here we can see that New York has the greatest number of dead patients that is 42% (10834) followed by New Jersey 11% (2805), Michigan 6% (1768) and Louisiana 4% (1013). Least number of dead patients are from Virgin Islands that is 1. And it is expected that there will be rise in the number of death cases in future.

3. Failure of Prediction Models

When it comes to judging the correctness of models that are used to predict COVID-19 patients status like scale/curve of the virus or predicting the number of dead , recovered and confirmed cases then it will give incorrect results as we cannot keep count or include human behavior in our model.

- Human behavior is unpredictable, we cannot predict what might happen next so judging the spread of virus is impossible.
- We cannot predict whether social distancing is being followed or not or whether people is staying at their homes or not. These factors which are majorly responsible for the spread of virus cannot be tracked or predicted.

Hence, it tells us that the results of prediction model will be false and what we can do is analyze the data until now rather than making a perfect model for prediction.

3.1 Scalability

Also, this unpredictable nature of human beings makes the prediction models unscalable. Even if we increase or reduce the input given to the models they won't perform as per the expectations and correct results won't be delivered. Earlier it was predicted that the total number of deaths in USA will be close to 1,00,000 but now that we have noticed that as per the latest data only 59000 people have died. So, we can conclude that these models are failing tremendously.

4.Medications Knowledge

Nowadays we can notice clearly that there is shortage of ventilators in USA so, there is company named GM (General Motors) who's taking initiative to make ventilators and this process is costing them around $1 billion dollars for just 40000 ventilators. Also, there was an import from India of medicine to deal with virus and its name is hydroxychloroquine which acts like an antiviral. Initially it was used to cure malaria but now it helps to prevent the virus from multiplying inside the cells. But it comes with some side effects too like nervous system and can lead to visual impairment.

5.Summary

In this project of visualization and analysis of COVID datasets we learned the use of new technology i.e. how to use Anaconda and Jupyter and performing that analysis on virtual server using one of the Amazon Web Services like instantiating an EC2 instance in SageMaker Studio. We were able to analyze state wise confirmed cases of corona virus, along with deaths and recovered patients. We drew some conclusions based on the percentage of affected patients of each gender and discussed some common reasons of the spread of virus along with the analysis of place of death of those patients. We finally concluded that the reason for the failure of prediction models is human intervention and that is why they are not scalable too. Link to datasets and Jupyter Notebook kernel with extension (.ipynb) shows all the computations done in Amazon Cloud : https://github.com/ssandeep858/Cloud-Computing-IN

6.References

[1] CDC official site for the datasets.

https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/index.html

[2] Treatments steps taken by the Government.

https://www.goodrx.com/blog/coronavirus-medicine-chloroquine-hydroxychloroquine-as-covid19-treatment/

[3] Origin of virus and other basic Information.

https://en.wikipedia.org/wiki/Coronavirus

[4] Steps for Setting up environment for the analysis.

https://chrisalbon.com/aws/basics/run_project_jupyter_on_amazon_ec2/

[5] WHO official site

https://www.who.int/health-topics/coronavirus#tab=tab_1

7.Acronyms

- SARS (Severe Acute Respiratory Syndrome)
- MERS (Middle East Respiratory Syndrome)
- JSON (JavaScript Object Notation)
- EC2 (Elastic Compute Cloud)