Submitted By:

Ishan Bhatt

(180006050)

**Project coordinator: Dr Chris Mayhew**

**Project Supervisor: Michael Thompson**

COVID19 Visualisation and Prediction using Tableau and Python

Table of Contents

[**1.Executive summary** 2](#_Toc38392022)

[**2.Introduction** 2](#_Toc38392023)

[**3. Aim of project** 2](#_Toc38392024)

[**4.Methodology:** 3](#_Toc38392025)

[**5.Data description:** 3](#_Toc38392026)

[**6. Data Preprocessing:** 5](#_Toc38392027)

[**7. Data Modeling for prediction** 5](#_Toc38392028)

[**8. Technology and tools** 8](#_Toc38392029)

[**9. Experimental Analysis** 8](#_Toc38392030)

[**10. Result analysis:** 9](#_Toc38392031)

[**11. List of problems while working on project** 18](#_Toc38392032)

[**12. Different phases of project** 19](#_Toc38392033)

[**13. Conclusion** 19](#_Toc38392034)

[References 20](#_Toc38392035)

# **1.Executive summary**

Corona virus proved one of the most dangerous viruses in 2019 and it affected many people worldwide that cause many deaths in all countries. However, because of the COVID-19 many countries suffered and hence, it impacts globally including economically. And there is no medicine of COVID-19. This virus mostly affected in all countries, not only this, but also impacts many businesses all around the world. Due to this virus also rapidly spread from affected people, all countries closed the borders including some of import and export materials. Therefore, Covid-19 is major epidemic globally. However, this analysis shows the statistics of overall countries including confirmed cases, deaths, and recoveries. Furthermore, it shows the various visualization in tableau and use python for different analysis. Python is also used to build prediction model. This analysis also includes some prediction techniques and features for example linear regression techniques, polynomial features, and svm prediction. But, before this, python requires various libraries including matplot, pandas, and scikit-learn which is used for machine learning and prediction. This all techniques used for forecasting future confirm case using linear regression and svm prediction.

# **2.Introduction**

This analysis includes overall information and statistics of various symptoms affected by covid-19 and causes overall deaths, case of people who affected by covid-19, and recoveries from the virus. Nevertheless, this requires multiple data and lots of filtration and preprocessing. In some dataset I used python for preprocessing of data and in some of dataset I used tableau for basic filtration. Tableau has many inbuilt functions for example removing null values and process data based on requirements. I also create world map affected by COVID-19 including all the mandatory information. I also made different two dashboards which contains different visualization graphs of covid-19. In addition, I used python to build prediction model and some analysis to count overall confirmed cases, recoveries, and death group by different unique countries. And tableau for visualize statistics of overall countries, basic overview of covid-19 and what extent different countries affected by covid-19.

# **3. Aim of project**

The major aim of this project is to know the basic statistic of covid-19, visualize different graphs, and predict future confirmed cases. Because from this visualization any people can understand the statistics easily even medical staff and non-technical person easily understand the graphs. However, the major challenge is to process various data based on requirements and made interactive dashboard.

# **4.Methodology:**

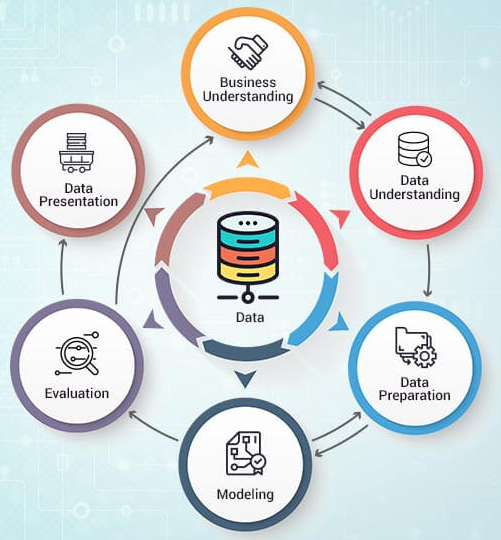


Figure 1:CRISP-DM Methodology (Analytics and business intelligence – The knowledge factory. (n.d.). The Knowledge Factory – Learn – Share – Care.)

This methodology is one of the most popular in data analytics; however, this is the basic step to start any data analytics approach. Nevertheless, this methodology contains Six different stages which is shown in the Figure 1. In which, the first and second step are to design data based on business understanding and data understanding. After that the data has been process based on requirement which is known by Data Preparation. In addition, we create model to further analysis for future prediction. And then data evaluation comes to the picture; in this, different data sources gathered for analysis and some filtration needed as per business and data requirements. At the last, the main stage comes into the frame – Data Presentation which is helpful to understand the data easily to the technical person as well as non-technical person.

# **5.Data description:**

I found these datasets from different sources – GitHub and Kaggle. There are two datasets used in this project as first is mostly used for visualizing first dashboard. This dataset contains the all basic information for each country such as Age, Sex, city, province, countries, Geo location, location history, etc. This data has 33 attributes; however, only few attributes selected for the analysis based on feature selection and total 13174 instances. The other dataset consists of confirmed case, deaths case, and recoveries which mostly based on timeseries data as this includes all the number of cases date wise. In python, apart from prediction model, I gathered three attributes which is confirmed case, deaths, and recoveries. And count total number of cases group by countries.

1. **Dataset 1:**

|  |  |
| --- | --- |
| Feature | Description |
| ID | Identification number |
| Age | Age of patient |
| Sex | Gender |
| City | City |
| Province | Province |
| Country | Country |
| Wuhan or not | Travel from Wuhan or not 0 means Wuhan and 1 means not Wuhan |
| Latitude | Geo location |
| Longitude | Geo location |
| Symptoms | Symptoms detected in COVID-19 test |
| Lives in Wuhan | Stay in Wuhan or not |
| Travel history | Travel location |
| Additional Information | Such as contacted with positive case, arrived date and location |
| Admin1 | Which contains district based on geo location |
| Admin2 | Which contains district based on geo location |
| Admin3 | Which contains district based on geo location |
| Geo resolution | Admin1, Admin2, Admin3 |

Figure 2: data description

The data has to many unwanted attributes that’s why I did not include here.

1. **Dataset2:**

|  |  |
| --- | --- |
| Feature | Description |
| Province/State | Province and state |
| Country/Region | Country and region |
| Lat | Geo location |
| Long | Geo location |
| Dates: 1/22/20 – 4/17/20 | This data has been updated daily. So, latest date varies. |

Figure 3: data description

There are total three data – confirmed case, deaths, recoveries which has same attributes. And this data is gathered in python for advance analysis for better understanding of the data in one frame.

# **6. Data Preprocessing:**

Data preprocessing is the process in which viewer can easily understand data and remove unwanted information for ease analysis. However, this is the basic understanding. From the perspective of this project, I removed some unwanted data and null values in excel and tableau for reducing time in analysis and particular tools such as tableau and python. However, other in other dataset the data has been removed which contains null values and remove some countries which has no information about the confirmed case, deaths, and recoveries.

# **7. Data Modeling for prediction**

In this project, I build prediction model in python. However, python has inbuilt functions and libraries such as Matplotlib for visualization, Scikit learning for machine learning, NumPy, and pandas. And made prediction model – svm prediction and Polynomial regression which contains linear regression model. All model is used for predicting future confirmed case by split dataset into train and test for accurate result. However, both models have different algorithm.

1. **SVM Prediction**

Support vector Machine (SVM) is one of the popular algorithms which is known as supervised machine learning techniques. It gives high accuracy and reliable model compared to another model. This model is not only used in prediction but also apply in face recognize application and fetch information from text related data and classification. This model is relatively simple and perform well compared to another model.

Model 1:

After the step of the data processing, it is easy to implement and create model based on requirement as it gives faster result because of the data preprocessing.

A screenshot of a social media post

Description automatically generated

Figure 4: SVM prediction model

Here is the model of SVM prediction. It has various codding parts and functions; in which gamma is link with accuracy of the model. Here I used low value of the gamma because sometimes we face issue of overfitting data if we use high value of gamma. I split data into train and test data for prediction. And made plot for understanding of MAE and MSE.

A screenshot of a social media post

Description automatically generated

Figure 5: Result of the SVM prediction

1. **Polynomial Regression Model**

It is form of linear regression model, but we can say that multiple regression model. The function of linear regression and polynomial regression is almost same. It is mostly used in organizations and very difficult to implement because of it requires in depth knowledge for proper implementation. However, the output of the two model is totally different. Amongst these models, polynomial regression is very accurate and that is the reason why it is popular.

Model 2:

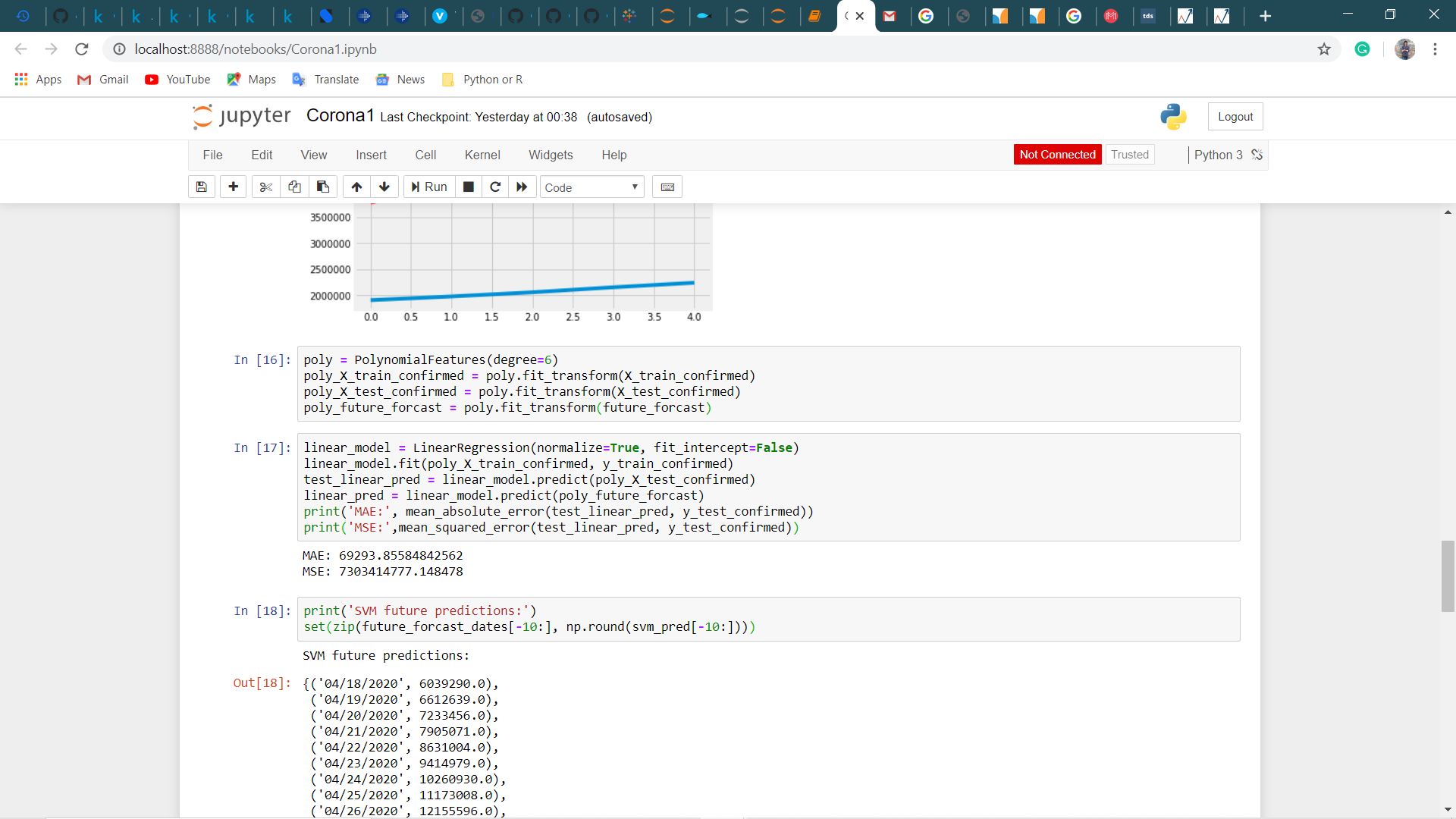


Figure 6: Polynomial model

Here is the polynomial regression which also includes degree of polynomial features; in addition, this is the reason why polynomial regression is more accurate than linear regression. It also shows the MAE and MSE, if it is too low than its consider as an overfitted data, however, it is totally depends on the data.

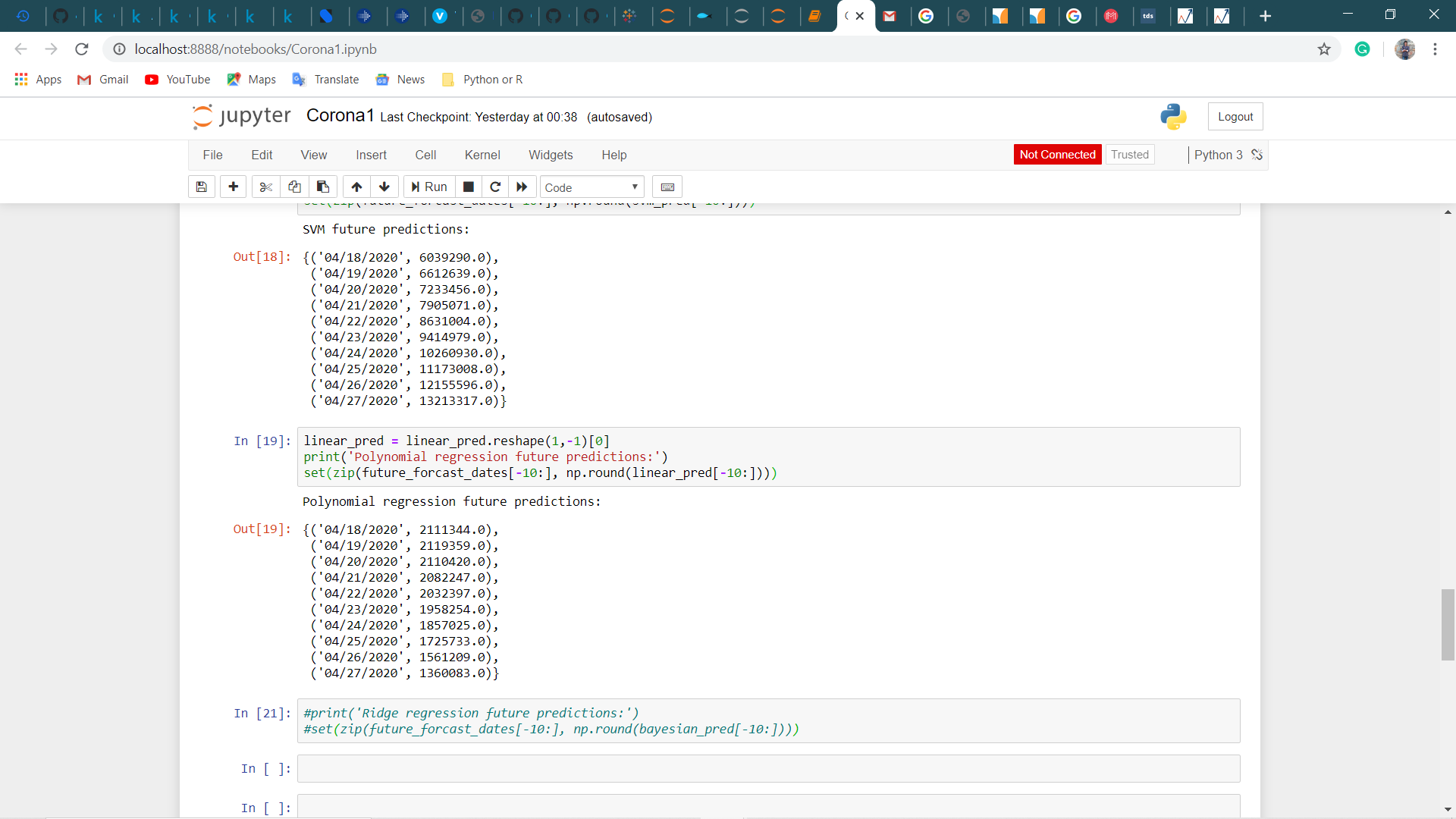


Figure 7: Result of the polynomial model

Timeseries data requires the reshape as we do not know the dimension and axis of the data. This picture shows the results of the next predicted confirm cases.

# **8. Technology and tools**

There are various technologies and tools used in this project.

1. Tableau: Tableau used for visualisation and from the diverse graphs I made dashboards to demonstrate the data including world map.
2. Python: I used python to predict future confirmed case and gathered data from various sources to make better visualisation graphs and for filtration.
3. Excel: excel used for filtering some data based on requirements.

# **9. Experimental Analysis**

There are different tools used in this analysis for better insights, predict model and make better visualisation and all the analysis done using python, excel, and tableau on laptop with 2.6GHz processor and 16GB of ram. I used two different machine learning algorithms to predict future case and collect data from different datasets to make dashboard in tableau. The database contains 13,174 instances with 33 attributes. Nevertheless, all attributes are not necessary to predict model and visualisation that is why I delete some attributes based on requirements for analysis as it is easy to process in specific tools and make accurate model. In this first model I used SVM prediction model which shows future number of confirmed cases till updated date available in dataset. However, this dataset updated daily depends on the new cases globally. The other model is polynomial regression analysis which also shows the total number of confirmed cases, the both models predict the totally different result as this is just for experimental analysis. The other model is based on linear regression. However, polynomial regression is the accurate model compared to linear model. In the other side, I made two dashboards in which dashboard 1 includes the overall statistics of covid-19 such as age, sex, countries, and travel history. And another dashboard illustrates the number of recoveries, number of deaths, and total confirm worldwide cases.

# **10. Result analysis:**

This section consists the output of the analysis of tableau and python including codding part.

**a.)**

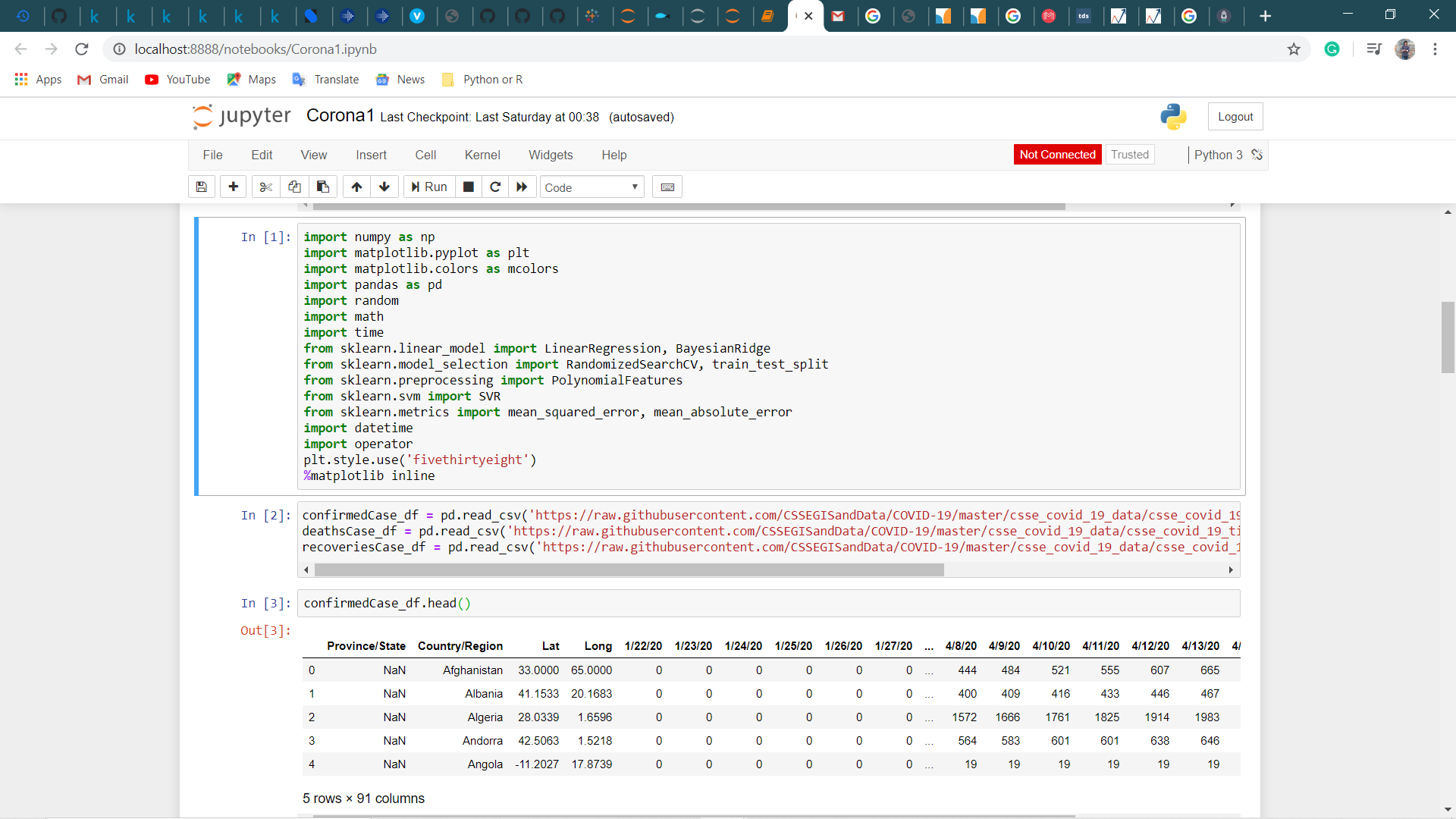


Figure 8: libraries and functions used in python

This figure includes the different libraries and function for analysis and predicting model as it mandatory to analysis and prediction model. It also shows the first 5 results of the data.

**b.)**

**A screenshot of a social media post

Description automatically generated**

Figure 9: Output of polynomial regression

This figure shows the result of predicted future confirm cases. As it also depicts in future the cases will be decrease day by day.

**c.)**

A screenshot of a social media post

Description automatically generated

Figure 10: Output of SVM future prediction

This figure shows the result of SVM future prediction. However, it is not accurate model compared to polynomial regression as it shows huge amount of future cases.

**d.)**

**A screenshot of a computer

Description automatically generated**

Figure 11: Codding part of total number of cases for tableau visualization

This part shows the total number of confirmed cases, total number of deaths, and total number of recoveries till the date of 14th of April group by unique countries.

A screenshot of a computer

Description automatically generated

Figure 12: Output of total number of cases

This shows the result of the previous coding part as it shows the US has highest number of confirmed cases and number of deaths. Whereas Canada is second highest in terms of confirmed cases. In addition, China recovers maximum number of patients. This result also shows various gradients from this we can easily understand which country has the highest number of cases and lowest number of cases.

**e.) Tableau Analysis 1**

All the visulization graphs are done in tableau which is one of the most popular visulization tool. The description of graphs are already write in the tableau for better understanding.

**A screenshot of a cell phone

Description automatically generated**

Figure 13: Sum of number of records for each age

**A screenshot of a cell phone

Description automatically generated**

Figure 14: Number of records group by symptoms which is provided in data

**A picture containing screenshot

Description automatically generated**

Figure 15: Number of records group by history location

**A screenshot of a cell phone

Description automatically generatedDashboard 1:**

Figure 17: Overall statistics of COVID-19

This dashboard shows overall information about the covid-19 including sex, age, and travel history location. All the graphs contain the legends; so, the any user can understand easily.

**f.) Tableau Analysis 2**

**A screenshot of a cell phone

Description automatically generated**

Figure 18: Total Number of confirmed cases group by countries

**A screenshot of a cell phone

Description automatically generated**

Figure 19: Total Number of deaths group by countries

**A picture containing screenshot

Description automatically generated**

Figure 20: Total Number of Recoveries group by countries

**A picture containing text, food

Description automatically generated**

Figure 21: World map filter by number of cases

**Dashboard 2**

A screenshot of a map

Description automatically generated

Figure 22: Overall scenario of covid-19

This dashboard represents the different graphs of the total number of cases and world map affected by covid-19.

# **11. List of problems while working on project**

* Problems while installing software such as tableau
* Looking for perfect IDE to work in python
* Find difficulty to deal multiple source of data
* Difficulty in python coding
* Dealing with removing null values and filtration in data
* Load multiple data in tableau for data insights
* Dealing with tableau for perfect visualization and create dashboards

# **12. Different phases of project**

|  |  |  |
| --- | --- | --- |
| **Phase** | **Summary of Phase** | **Key Deliverables** |
| Requirements of tools for project | * Installing resources to create environment. * Install Visualization tool - Tableau * Install Anaconda 3 for python | * Python coding and Tableau visualization |
| Execution | * Preprocessing of data * Prediction Regression model * Gathered all data for better understanding | * Analysis in python * Collect data from multiple sources |
| Outcome | * create best prediction algorithms * create dashboards for ease understanding | * Made two dashboards * Predict future confirm cases in python |

# **13. Conclusion**

By putting all the efforts, I successfully able to analyze, predict and visualize covvid-19 data. This summary of the data and prediction absolutely meets standard business needs. I create interactive tableau dashboards; in which every user can understand easily including non-technical person. In python environment I use polynomial regression and SVM prediction machine learning techniques to predict future confirm cases. However, this can be improved by changing some coding to make better model as it is not accurate model. I can also predict future deaths and recoveries by simple modification in code; but I don’t want any complexity in project. At the end I gathered data from multiple sources to count total number of cases – confirmed, deaths, and recoveries group by countries to make dashboards in tableau for advance analytics. In next version, I can also predict future deaths and recoveries in python and make interactive web app for visualization; so that anyone can able to see data insights and impact on covid-19 to the different countries.

# References

1. *(Tutorial) support vector machines (SVM) in scikit-learn. (n.d.). DataCamp Community. https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python*

*2. Analytics and business intelligence – The knowledge factory. (n.d.). The Knowledge Factory – Learn – Share – Care.* (n.d.). Retrieved from The knowledge factory: https://theknowledgefactory.org/index.php/analytics/

3. 1.4. Support vector machines — scikit-learn 0.22.2 documentation. (n.d.). scikit-learn: machine learning in Python — scikit-learn 0.16.1 documentation. Retrieved April 21, 2020, from <https://scikit-learn.org/stable/modules/svm.html>

4. CSSEGISandData/COVID-19. (n.d.). GitHub. <https://github.com/CSSEGISandData/COVID-19>

5. Novel corona virus 2019 dataset. (n.d.). Kaggle: Your Machine Learning and Data Science Community. <https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset>

6. Polynomial regression. (2017, October 5). Mindmajix. <https://mindmajix.com/polynomial-regression>

7. Tran, N. (2019, March 20). Machine learning: Polynomial regression with Python. Medium. <https://towardsdatascience.com/machine-learning-polynomial-regression-with-python-5328e4e8a386>

8. Tutorial: Understanding linear regression and regression error metrics. (2019, June 10). Dataquest. <https://www.dataquest.io/blog/understanding-regression-error-metrics/>

**14. Appendix**

A screenshot of a social media post

Description automatically generated

A screenshot of a social media post

Description automatically generated