BTECH IN COMPUTER SCIENCE AND ENGINEERING(ARTIFICIAL INTELLIGENCE)

MATHEMATICS FOR INTELLIGENT SYSTEMS PROJECT REPORT

PROJECT TOPIC : FORECASTING COVID-19 CASES

FACULTY : DR GEORG GUTJAHR

GROUP MEMBERS

AKSHAY KRISHNAN T(AM.EN.U4AIE21109)

MUHAMMED SHAJAHAN(AM.EN.U4AIE21144)

NAVNEETH SURESH(AM.EN.U4AIE21147)

ASWIN US(AM.EN.U4AIE21119)

NIRANJAN PRASANTH(AM.EN.U4AIE21148)

ACKNOWLEDGEMENT

We would like to express our sincere thanks to Dr Georg Gutjahr sir and Dr Gopakumar.G sir for their guidance in completing our project and report. We would also like to express our sincere gratitude to our beloved college Amrita Vishwa Vidyapeetham for providing and presenting us with the opportunity to work on this report.

ABSTRACT

The COVID-19 epidemic has spread to more than 200 countries and considered as an unprecedented public health crisis, which seriously affect people's daily life. In order to find out a better way to predict and forecast the epidemic situation, this project utilized machine learning and a series of regressions, including linear regression and polynomial regression. We tried to find out the model that fits the training data by calculating the mean square error (MSE), Mean absolute error(MAE) and Root mean square error. We hope our findings can serve as a valuable reference to forecast the trend of the COVID-19 epidemic situation.

I. INTRODUCTION

Covid-19 has caused a worldwide pandemic and has troubled all of us in our lives. We have heard many terms related to numbers like confirmed cases, deaths and recovery. To enhance the pertinence of government policies, researchers should examine the relevant conditions related to the spread of the epidemic. Furthermore, if researchers could forecast the spread of COVID-19 with an analysis of the number of deaths and recoveries, people in charge of decision making to fight the pandemic can use the result of forecasting to decide immediate actions.

In a nutshell analysing the trend of the Covid 19 cases is an important aspect in fighting Covid 19. Hence machine learning models could be applied to find future trends of Covid 19. This can help to make precautionary measures to prevent Covid 19. Hence for predicting the Covid 19 cases we will be using a Linear Regression model. We also visualized the dataset using bar graphs and pie charts. This can help better understanding the dataset.

Linear regression:

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable we wish to predict is called the dependent variable. The variable(s) we use to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimises the discrepancies between predicted and actual output values. There are simple

linear regression calculators that use a least squares' method to discover the best-fit line for a set of paired data. We can then estimate the value of the dependent variable from the independent variable.

REGRESSION

Regression analysis is a collection of statistical procedures for evaluating the relationships between a dependent variable ('outcome' or 'response' variable) and one or more independent variables ('predictors', 'covariates', 'explanatory variables' or 'features').

● **Dependent Variable**: The dependent variable is the main factor in regression analysis that we wish to predict or understand.

● **Independent Variable**: Independent variables are the elements that influence the dependent variables or are used to predict the values of the dependent variables.

Regression analysis is primarily used for two conceptually different purposes.

● Regression analysis is commonly used for prediction and forecasting, and it shares a lot of ground with machine learning.

● To infer causal links between the independent and dependent variables, regression analysis can be utilised.

There are over 15 types of regression:

● Linear Regression

● Polynomial Regression

● Logistic Regression

- Quantile Regression
- Ridge Regression
- Lasso Regression
- Elastic Net Regression
- Principal Components Regression
- Partial Least Squares Regression
- Support Vector Regression
- Ordinal Regression
- Poisson Regression
- Negative Binomial Regression
- Quasi Poisson Regression
- Cox Regression
- Tobit Regression

The most ordinary form of regression analysis is linear regression. It's a method in which the dependent variable is always the same. The dependent variable's relationship with the independent factors is linear.

Polynomial regression

Polynomial regression, E(y |x), describes the fitting of a nonlinear relationship between the value of x and the conditional mean of y.

In this project we have used

A quadratic regression is the process of finding the equation of the parabola that best fits a set of data.

The best way to find this equation manually is by using the least squares method. A quadratic regression is the process of finding the equation of the parabola that best fits a set of data.

As a result, we get an equation of the form:

$y=ax^2+bx+c$ where $a\neq0$.

That is, we need to find the values of a,b and c such that the squared vertical distance between each point

(x_i,y_i) and the quadratic curve $y = ax^2+bx+c$ is

minimal.

In the cubic regression model, we deal with cubic functions, that is, polynomials of degree 3.

The cubic regression function takes the form:

$y = a_0x^0+ a_1x^1 + a_2x^2 + a_3x^3$,

where a, b, c, d are real numbers, called coefficients of the cubic

regression model

As you can see, we model how the change in x affects the value of y. In other words, we assume here that x is the independent (explanatory) variable and y is the dependent (response) variable.

II. DATASET

We collected cumulative daily data from various countries from from 22nd January 2020 to 15th March 2020 involving three variables such as the number of confirmed diagnoses, the number of people who recover and who died and etc.

ie, 3 datasets used for this model which are of confirmed cases, deaths and recoveries. The dataset consists of confirmed cases, deaths and recoveries from 22nd January 2020 to 15th March 2020. The other columns of the dataset includes Countries, Provinces, Latitudes and Longitudes.

a) Confirmed cases: https://drive.google.com/file/d/1i1ZdEmn2pEI4y8G2QMOkdxGPtt71mrRf/view?usp=sharing
b) Deaths: https://drive.google.com/file/d/15kqh3wEeilsR0nlLCl2V2kQtgWHMd2Yk/view?usp=sharing
c) Recovered: https://drive.google.com/file/d/1P9ivCZ8Jf3OcC3fH-M2XO2CFWaH3NV5M/view?usp=sharing

III. METHODOLOGY

1) Importing the necessary libraries which include pandas, numpy, matplotlib, sklearn
2) Loading the datasets and creating dataframes using python library.
3) Extracting required keys from the dataset.
4) Creating a future forecasting arrays using numpy arrays.
5) Differentiating countries and provinces from the dataset.
6) Removing NaN (Not a Number) values from the dataset.
7) Visualizing the data of confirmed cases using bar graphs, pie charts etc.
8) Building the Linear Regression model. Splitting the dataset into training and test dataset. Fitting the dataset using *linear.fit()* and predicting using .predict() method and finally display the Mean absolute error and Mean squared error.
9) Displaying the comparison between confirmed cases and predicting output from the Linear regression model.
10) Displaying the comparison between deaths and predicting output from Linear regression model.
11) Displaying the comparison between recoveries and predicting output from Linear regression model.

Types of Errors:

1) **The Mean Squared Error** (MSE) or Mean Squared Deviation (MSD) tells you how close a regression line is to a set of points.

- It does this by taking the distances from the points to the regression line (these distances are the "errors") and squaring them.
- The squaring is necessary to remove any negative signs.
- It also gives more weight to larger differences. It's called the mean squared error as you're finding the average of a set of errors. The lower the MSE, the better the forecast.
- MSE formula = $(1/n) * \Sigma(\text{actual} - \text{forecast})^2$

n = number of items,

Σ = summation notation,

Actual = original or observed y-value,

Forecast = y-value from regression.

2) **The mean absolute error**

The mean absolute error is a way to measure the accuracy of a given model.

- Mean absolute error is a loss function used for regression.
- The loss is the mean over the absolute differences between true and predicted values.
- It is calculated as:
- $MAE = (1/n) * \Sigma|yi - xi|$

where:

Σ: Sumamtion Notation

y_i: The observed value for the i^{th} observation

x_i: The predicted value for the i^{th} observation

n: The total number of observations

- In general, the lower the value for the MAE the better a model is able to fit a dataset. When comparing two different models, we can compare the MAE of each model to know which one offers a better fit to a dataset.

3) Root mean square error

- Root mean square error tells us the average distance between the predicted values from the model and the actual values in the dataset.
- It is a way to assess how well a regression model fits a dataset.
- The lower the RMSE, the better a given model is able to fit a dataset.
- RMSE = $\sqrt{\Sigma(P_i - O_i)^2 / n}$
- Here:

  Σ = summation notation

  P_i = is the predicted value for the i^{th} observation in

  the dataset

  O_i = is the observed value for the i^{th} observation in

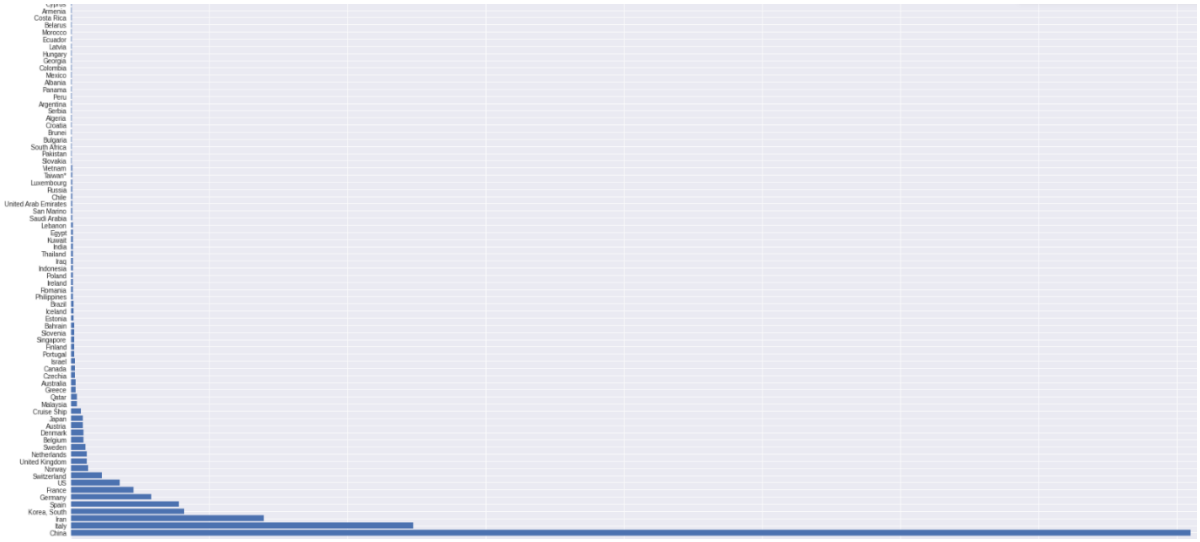
  the dataset

  n = Sample size

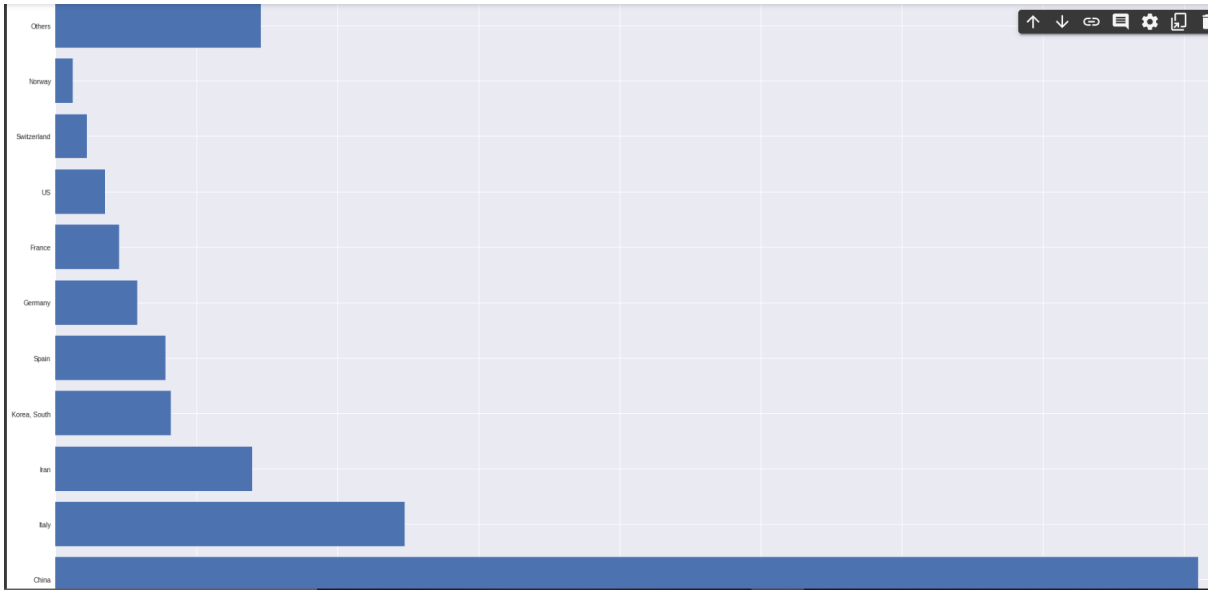
IV. RESULTS

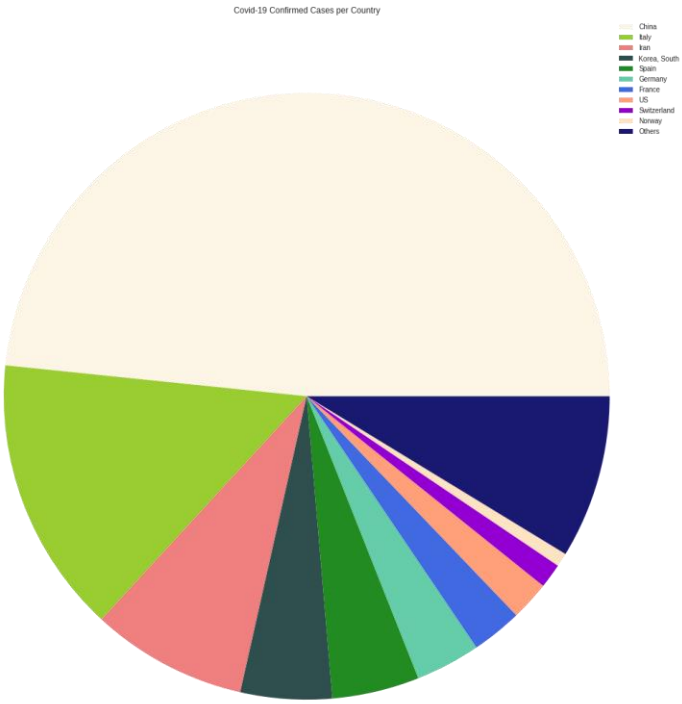
1) Visualization of the dataset

All Countries:



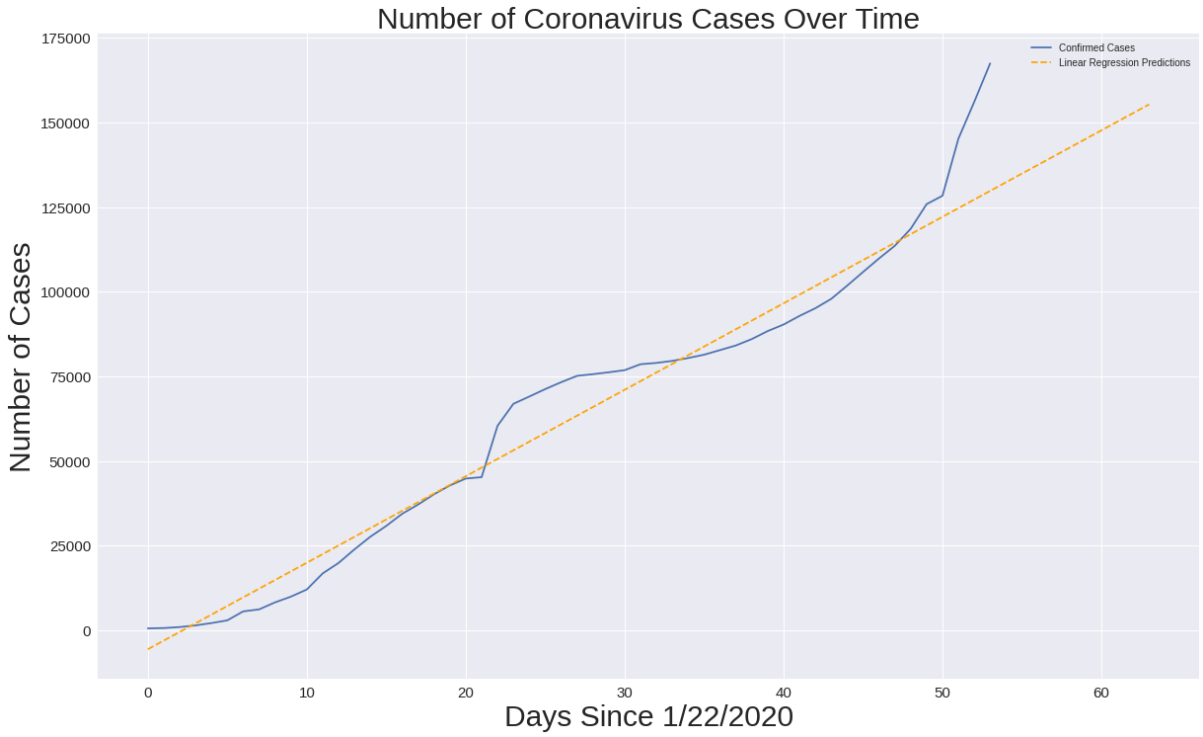
Countries with major cases:



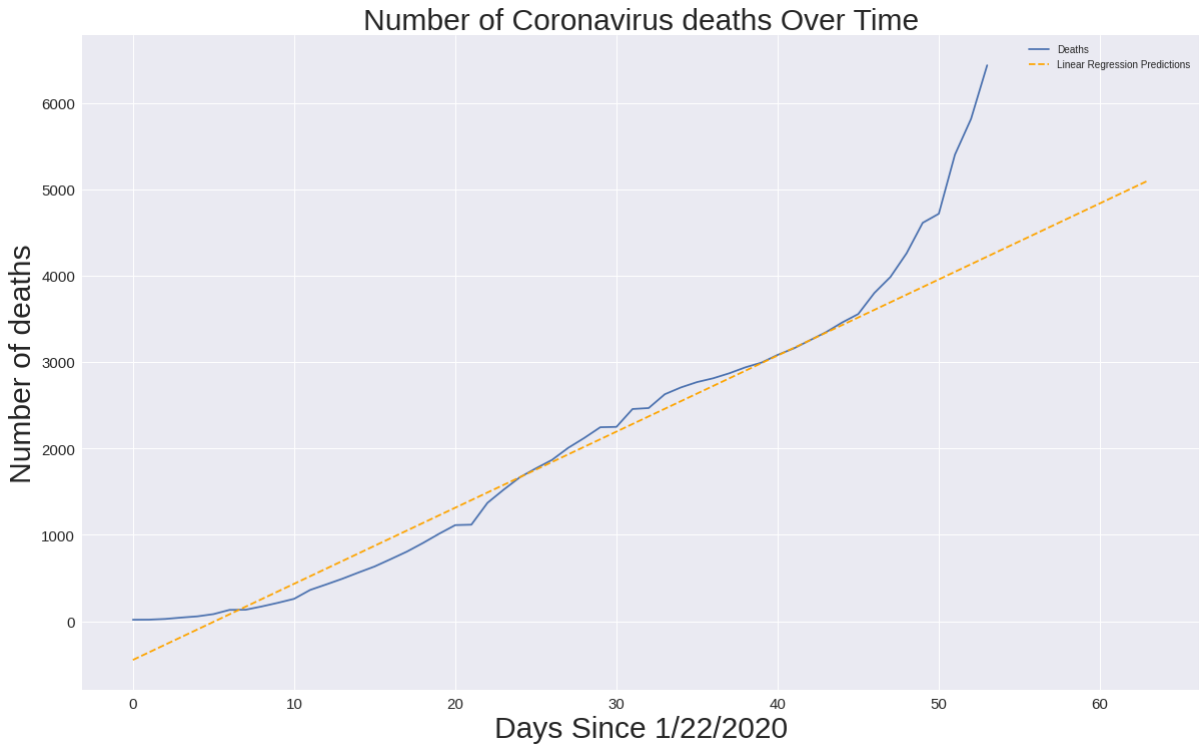


2) Linear Regression Model Graph

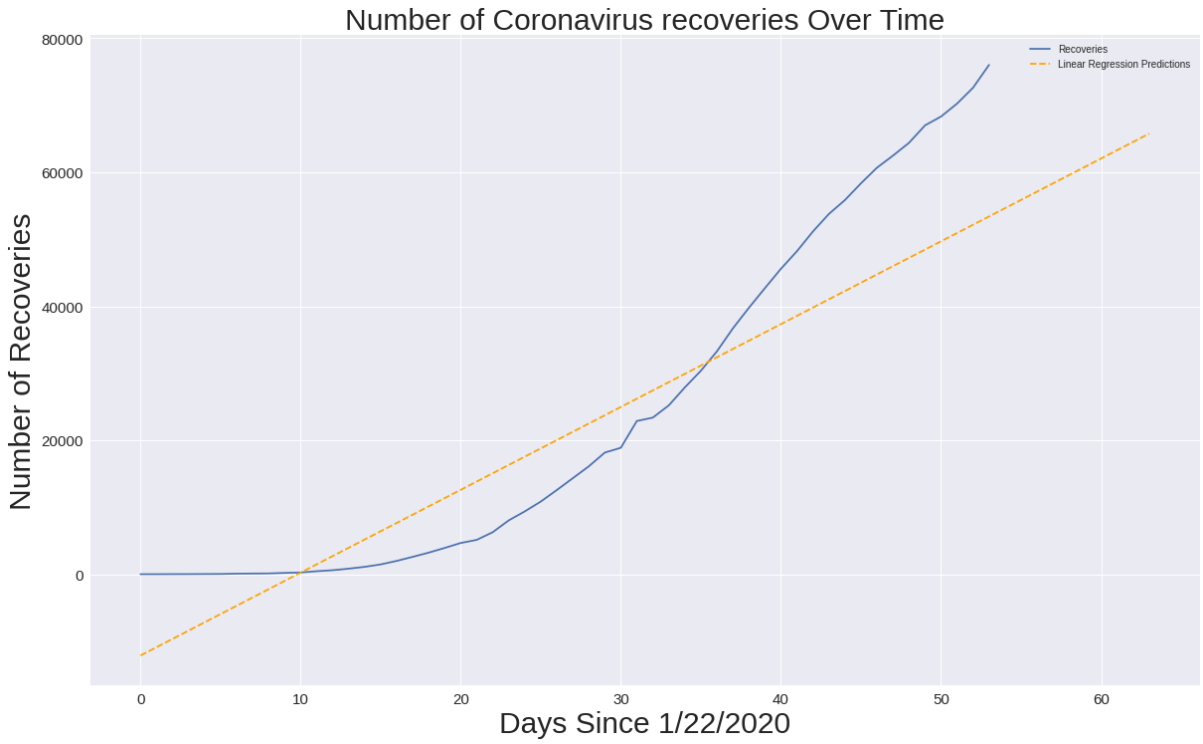
Covid cases prediction:



Covid deaths prediction:



Covid recoveries prediction:



Error Calculated:

Linear Regression:

```
MAE: 11965.537037037033
MSE: 307996364.0108404
RMSE:  17549.82518462336
```

Polynomial Regression (2nd Degree)

```
 MAE: 10419.971004746621
 MSE: 169131996.17456436
 RMSE:  13005.075785037332
```

Polynomial Regression (3rd Degree)

```
MAE: 8574.477890075488
MSE: 94808210.52774853
RMSE:  9736.950781828391
```

Polynomial Regression (4th Degree)

```
MAE: 1831.0198382906772
MSE: 6115553.428541872
RMSE:  2472.964502078805
```

V. DISCUSSION

Statistical models are important techniques for evaluating infectious disease data analyses in real time. We used the Linear regression model in this paper to evaluate the epidemic data. This prediction model will speculate the advance situation that is coming in days and effective measures are to be more enhanced to flatten the curve. Using Python will help to make the model better understandable and having a good dataset will improve the model and will give better predictions. We can also use more advanced algorithms to better predict the cases.

Limitations of the Model: Limitations of the model can be thought of in terms of gathering more independent variables or information, ways to find the number of contact tracing cases. If the number of contact tracing cases are been reduce, it will indirectly reduce the number of daily active cases.

VI. CONCLUSION

- From this project we have understood that our system doesn't predict the case, deaths or recoveries entirely accurately using linear regression. Although our salary prediction project ran with small inconsistencies and predicted linear regression models based on several parameters and their relationship with each other, increasing the number of our primary parameters would make the system more efficient. As we increase the order of polynomial and apply it to the model we can increase the accuracy and reduce the errors. If we introduce a new variable "contact tracing" can also increase the efficiency of model.

VII. APPENDIX

https://colab.research.google.com/drive/1LvPrNkT0swc6vQ-tMBxcV-9TNV2j5rYF#scrollTo=D34Qxxl7mx8S

VIII. BIBLIOGRAPHY

- https://www.kaggle.com/learn
- https://pandas.pydata.org/docs/
- https://numpy.org/doc/
- https://www.kaggle.com/datasets
- https://docs.python.org/