ASSIGNMENT 1

ANDREW ID: parmenin

18-788: Big Data Science

3/27/23

Niyomwungeri Parmenide ISHIMWE

I, the undersigned, have read the entire contents of the syllabus for course 18-788 (Big Data Science) and agree with the terms and conditions of participating in this course, including adherence to CMU's AIV policy.

Signature: **Niyomwungeri Parmenide ISHIMWE**

Andrew ID: **parmenin**

Full Name: **Niyomwungeri Parmenide ISHIMWE**

LIBRARIES USED

- import numpy as np
- import pandas as pd
- import matplotlib.pyplot as plt
- import warnings
- # %pip install haversine
- from haversine import haversine
- from scipy.optimize import curve_fit
- from sklearn.metrics import r2_score
- from sklearn.metrics import mean_squared_error
- from sklearn.model_selection import train_test_split
- from sklearn.neighbors import KNeighborsRegressor
- from sklearn.tree import DecisionTreeRegressor
- warnings.filterwarnings("ignore")

QUESTION 1:

The two historical monthly datasets for each of Rwanda's thirty districts—RwandaDistrictRainfall.csv and RwandaDistrictVegetation.csv—were required to be loaded into the environment (Jupyter notebook). They provide data that is generated from satellite imaging data, such as measures of rainfall and the improved vegetation index. This was done by downloading them and using the pandas' read_cv function[1] to read from them.

QUESTION 2:

It was now required to graph two, time series for both rainfall and vegetation index with 6x5 subplots each for all the districts. The data frame was preprocessed by first transposing it and then adding the date range to it to make it easier to depict the time. The subplots were plotted for both rainfall and vegetation using the matplotlib.pyplot's subplots function[2] for plotting and subplots_adjust[3] for adding margins and the following two graphs were produced.

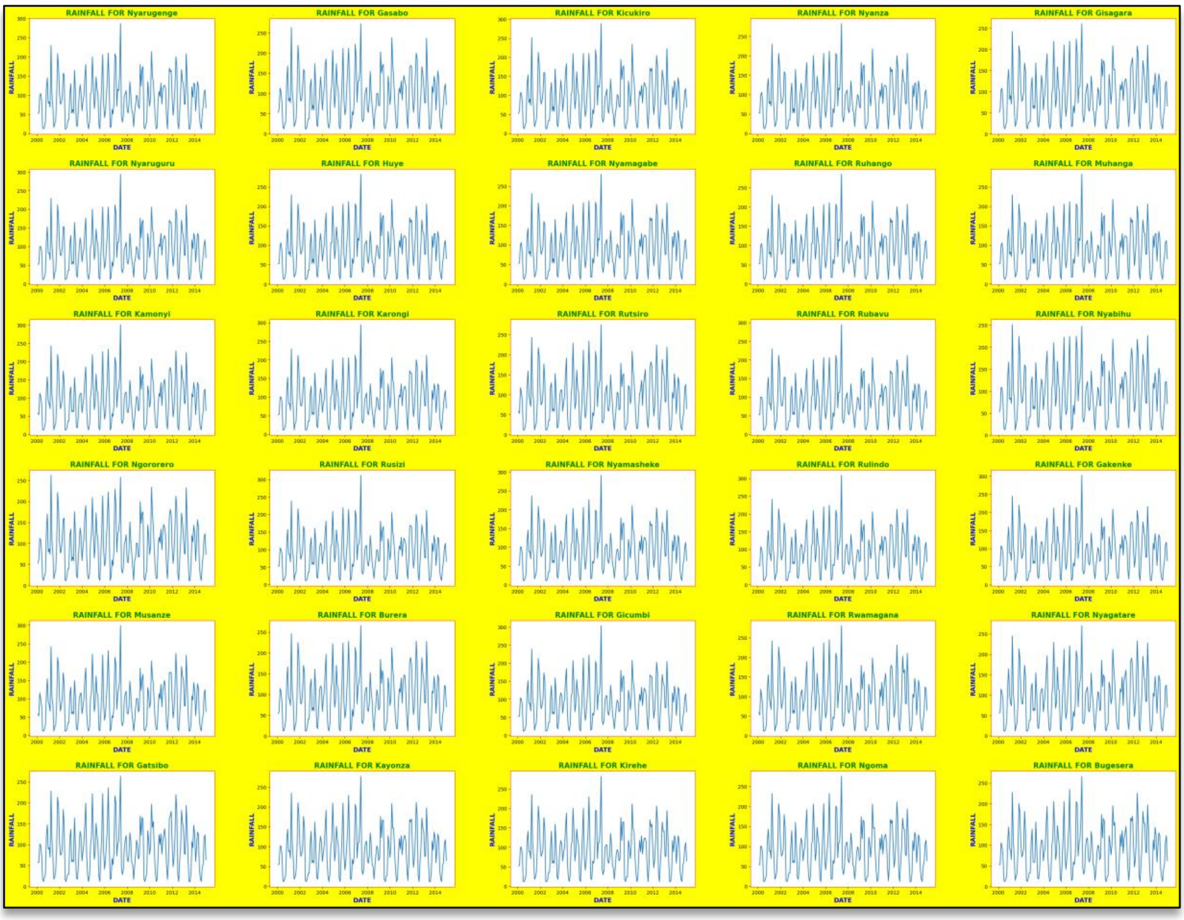


Figure 1: **The subplots for rainfall in 30 districts**

We can infer from the graph that the highest quantity of rain had fallen first, in 2007 and second, in 2001 in most of the districts of the country.

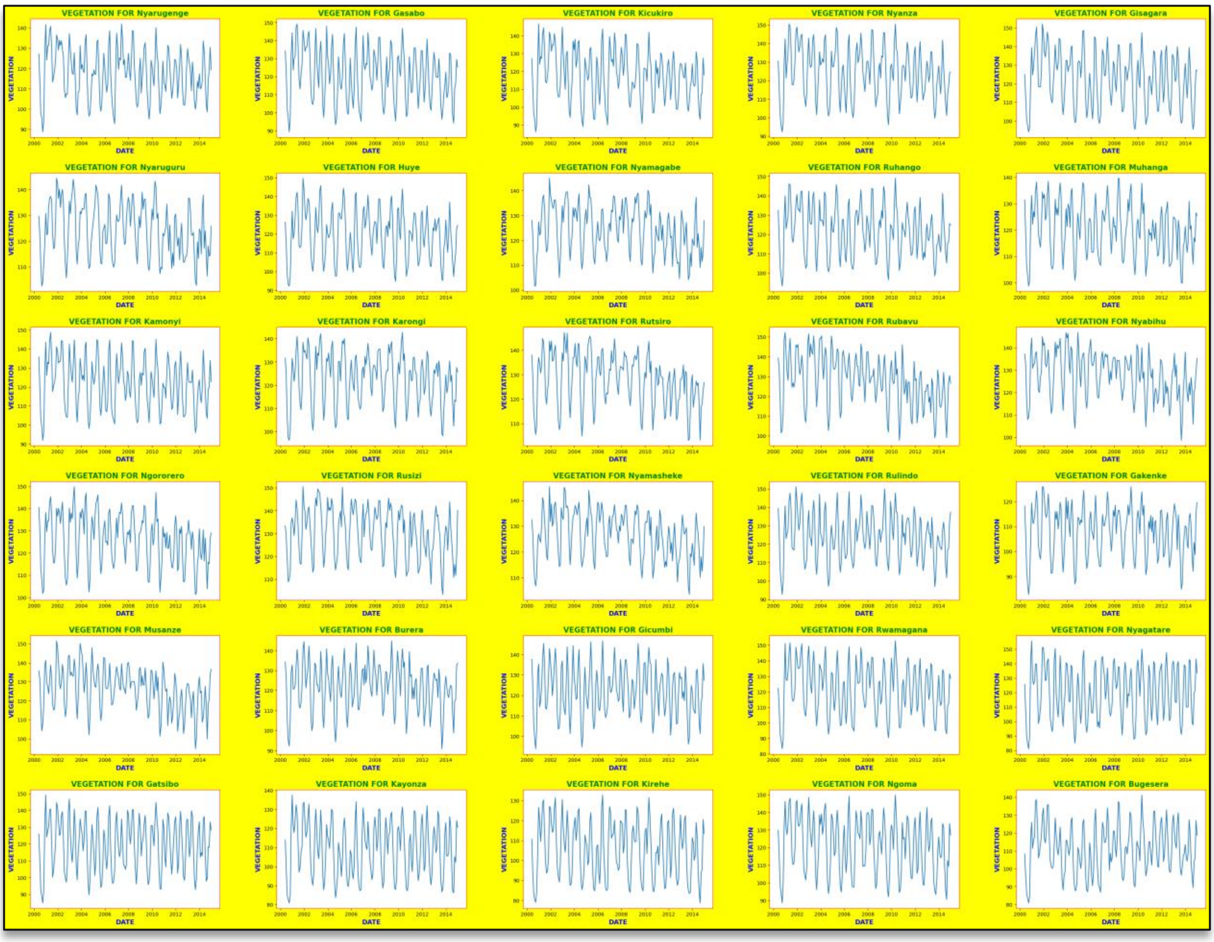


Figure 2: The subplots for vegetation index in 30 districts

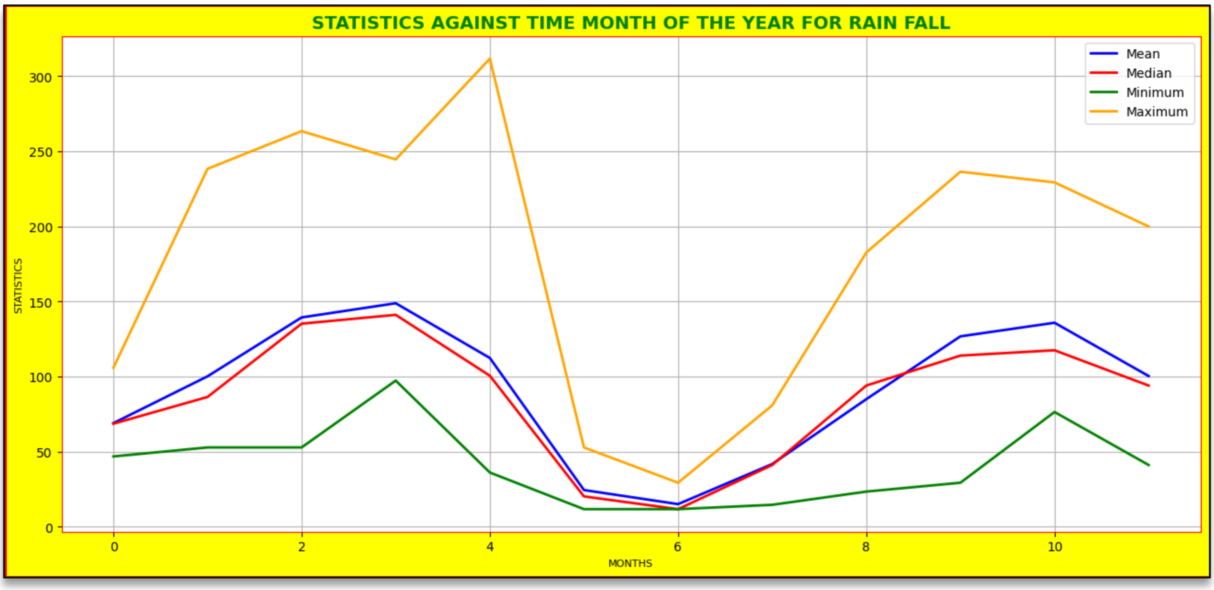
We can infer from the graph that, as the highest quantity of rain had fallen first, in 2007 and second, in 2001 in most of the districts of the country causing the vegetation index to go high in 2001 and 2002 as well, even though the lowest quantity of vegetation index was found in the year 2001 for the most of the districts.

QUESTION 3:

It was required to calculate the mean, median, minimum, and maximum of both the rainfall and vegetation index for each month of the year, i.e., 12 monthly values, and to plot them against the month of the year. These were calculated and stored inside the following data frame and plotted below.

	Mean	Median	Minimum	Maximum
0	69.083333	68.70	46.9	105.9
1	100.239778	86.50	52.9	238.4
2	139.459333	135.30	52.9	263.5
3	148.915556	141.20	97.4	244.7
4	112.469111	100.65	36.2	311.8
5	24.529333	20.30	11.8	52.9
6	15.152667	11.80	11.8	29.4
7	41.829778	41.20	14.7	80.9
8	85.049111	94.10	23.5	182.7
9	126.871111	114.05	29.4	236.5
10	135.926667	117.60	76.5	229.4
11	100.438444	94.10	41.2	200.0

<u>Figure 3:</u> The mean, median, minimum, and maximum data frame for rainfall

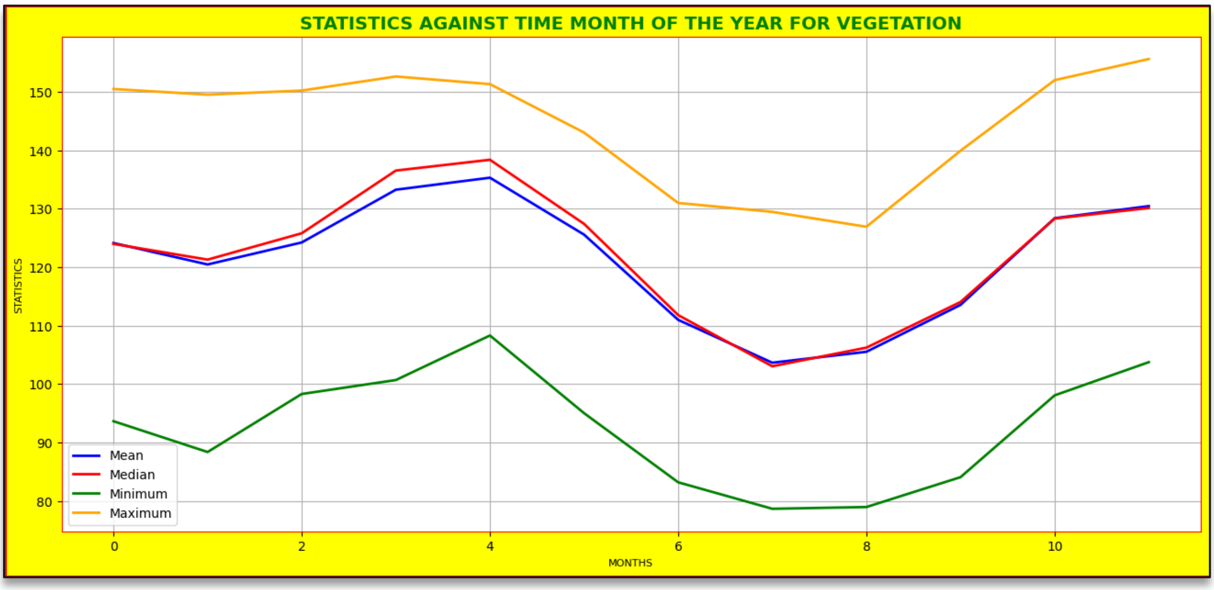


<u>Figure 4:</u> The mean, median, minimum, and maximum graph for rainfall

We can infer from the above plot and data frame that the low precipitation in Rwanda is seen in the month of July and the highest rainfall is seen in the month of April.

	Mean	Median	Minimum	Maximum
0	124.140874	123.987221	93.649422	150.495856
1	120.469059	121.296982	88.375662	149.525034
2	124.236352	125.807144	98.306588	150.224410
3	133.260526	136.536959	100.691685	152.635095
4	135.315459	138.384247	108.315401	151.328737
5	125.576479	127.417296	95.024470	143.043877
6	110.991302	111.805282	83.192432	130.992456
7	103.655364	103.049909	78.661577	129.470863
8	105.528631	106.236392	78.962225	126.923562
9	113.553146	114.049997	84.074041	139.933869
10	128.385639	128.324193	98.071576	152.008141
11	130.453648	130.134722	103.750315	155.625355

<u>Figure 3:</u> The mean, median, minimum, and maximum data frame for the vegetation index



<u>Figure 4:</u> The mean, median, minimum, and maximum graph for the vegetation index

From the above data frame and graph, we can infer that the low vegetation index in Rwanda is seen in the month of August and the highest vegetation index is seen in the month of May.

QUESTION 4:

It is needed to compute the correlation coefficient (C) for rainfall between all pairs of districts and create a graph that shows the correlation against distance in kilometers. The graph should use the C(d) = C0exp(-ad) model to fit the data and plot a curve that illustrates the decline in correlation with distance. Finally, we needed to estimate the values of C0 and the decay constant (a) for the model.

This was done by first, loading the "RwandaDistrictCentroidsLongitude_Latitude.csv" data set, calculating the correlation coefficient **C** for rainfall between each pair of districts, and computing the distance **d** between the pair. This was performed with the help of the haversine library which calculates the great-circle distance between two points on the Earth's surface[4] and was found to be **34.693078998302425 km**. After that, the following graph is made to show the correlation values versus the distance.

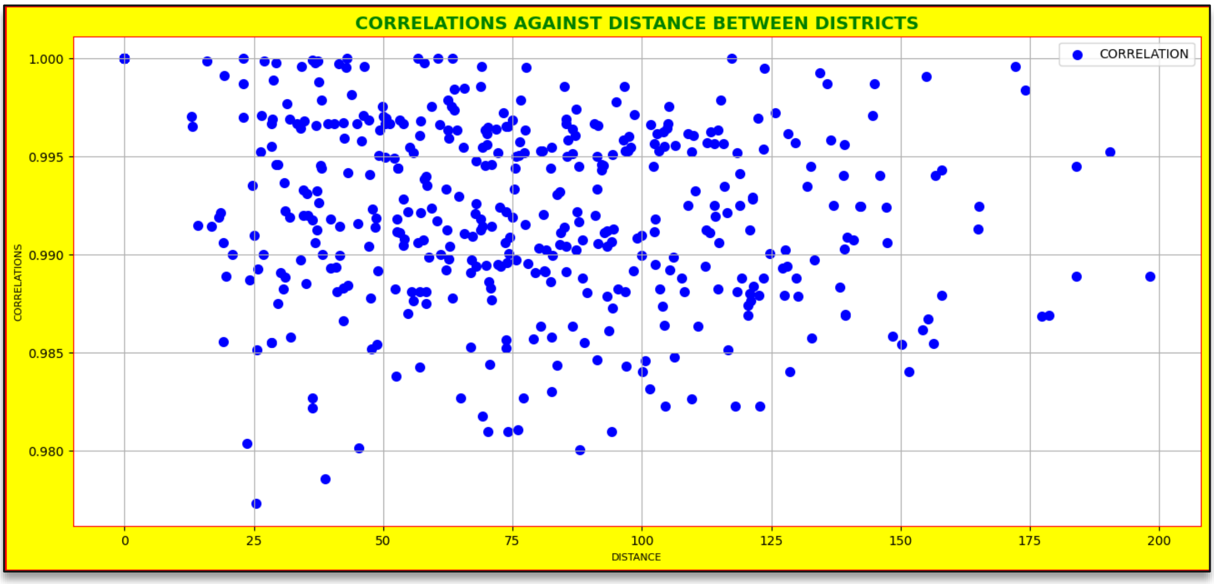
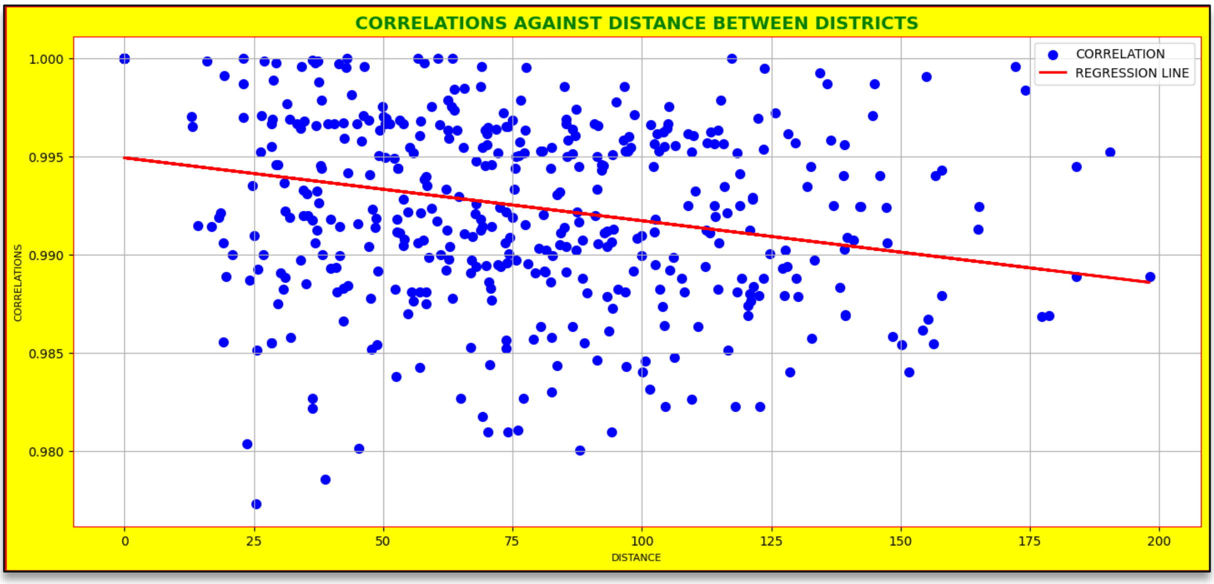


Figure 5: **The graph to show the correlation values versus the distance.**

Next, the model **C(d) = C0exp(-ad)** is fit, and the **curve_fit**[5] function is fit to the model, distances and correlations to provide the params **(C0)** and the decay constant (**a)** are estimated and they are **0.9949362375041936** and **3.234477915376383e-05** respectively.

Finally, this curve is plotted below to show how quickly the correlation declines with distance.

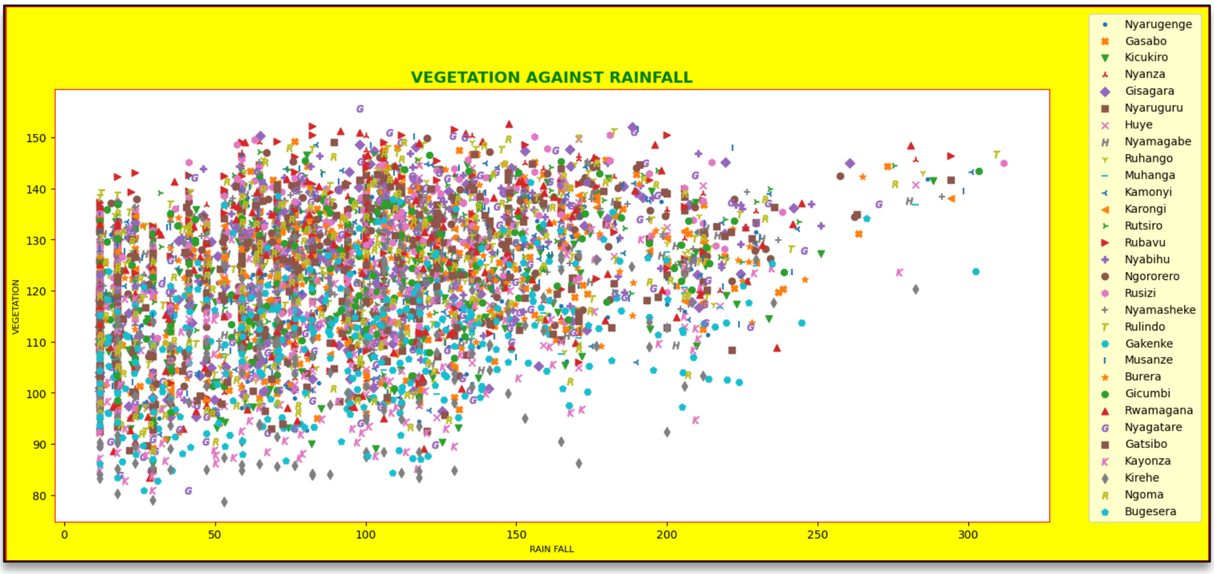


<u>Figure 6:</u> The graph shows how quickly the correlation declines with distance.

From the graph, we may deduce that the more remote the districts are from one another, the less their data are correlated, and the closer the districts are to one another, the more correlated their data are.

<u>QUESTION 5:</u>

To answer this question, it was required to synchronize the dates for both the rainfall and the vegetation index, produce a scatter plot for the same months, and add a caption to the graph using colors and symbols. This was done by first dropping the first four rows in the rainfall data set to match the shape with the vegetation index. Then synchronizing the dates by making the date column the index. Then, the following labeled plot is sketched to show the vegetation against rainfall.



<u>Figure 7:</u> The graph for vegetation index against rainfall

We can infer from the graph that there is a strong correlation between rainfall and vegetation.

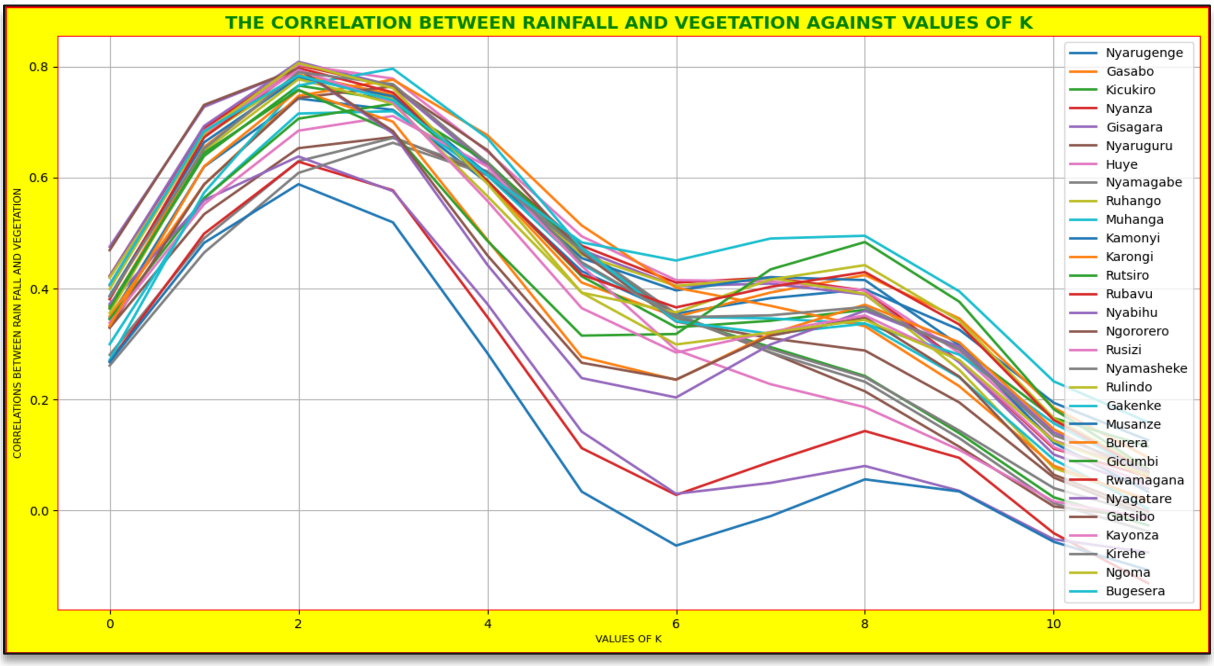
<u>QUESTION 6:</u>

It was necessary to change the rainfall time series to provide a new feature that makes better predictions of the vegetation index by testing rainfall in the month t-k against the vegetation index in the month t. For each district, the correlation between the rainfall time series and the vegetation index was calculated after the rainfall time series had been altered by delaying it by k [0:12].

Then, it was found that it takes one month to see the considerable effect of rain on the vegetation in each district since the optimal K is 1 i.e., giving the highest correlation for all districts as shown in the code.

<u>QUESTION 7:</u>

This time, it is needed to find the optimal value of n for each district by using the correlation, considering moving averages of rainfall over the last n months ranging from 1 to 12. To see a long-term trend, the rainfall time series was transformed using simple moving averages to smooth out short-term fluctuations.



<u>Figure 8:</u> The correlation vs values of k

By averaging over the last n months with windows of k ranging from 1 to 12, it is found that the K that gives the highest correlation between the smoothened rainfall index and vegetation index for each district is 2 as shown in the above graph as it has been repeated many times.

<u>QUESTION 8:</u>

It is asked for evidence for using a quadratic model to describe how the vegetation index varies with rainfall. This was accomplished by looking at the correlations between the rainfall index and the vegetation index, the rainfall index and the delayed rainfall index, the rainfall index, and the simple moving average rainfall index, and by testing the same correlation with linear, quadratic, and cubic regression. Again, use modified R-squared, RMSE, and R-squared to measure the effectiveness of the performance indicators. The tables below present the outcomes.

	LINEAR MODEL	QUADRATIC MODEL	CUBIC MODEL
R^2 FOR RAINFALL	0.109453	0.116198	0.118972
R^2 FOR DELAYED RAINFALL	0.388732	0.446825	0.449767
R^2 FOR SMA FOR RAINFALL	0.453645	0.471145	0.471967

<u>Figure 9:</u> The r-squared table.

	LINEAR MODEL	QUADRATIC MODEL	CUBIC MODEL
ADJUSTED R^2 FOR RAINFALL	0.109284	0.116030	0.118805
ADJUSTED R^2 FOR DELAYED RAINFALL	0.388615	0.446719	0.449662
ADJUSTED R^2 FOR SMA FOR RAINFALL	0.453540	0.471043	0.471866

<u>Figure 10:</u> The adjusted r-squared table.

	LINEAR MODEL	QUADRATIC MODEL	CUBIC MODEL
RMSE FOR RAINFALL	13.197640	13.147566	13.126912
RMSE FOR DELAYED RAINFALL	10.944063	10.411039	10.383314
RMSE FOR SMA FOR RAINFALL	10.358588	10.191343	10.183412

<u>Figure 11:</u> The root mean squared error table.

From the results above in tables, it can be inferred that the best model can be cubic as it has a low rate of errors and hence good performance compared to quadratic and linear as shown by RMSE and larger r square and adjusted r squared scores.

<u>QUESTION 9:</u>

To choose the optimum transformation, it was asked to utilize cross-validation (train test split), combining moving averages and delays of the monthly measurements, and presenting tables of results for adjusted r-squared, RMSE, and r-squared. The train_test_split[6] function was used to test models on the out-of-sample data while computing performance metrics for rainfall, delayed rainfall, and SMA rainfall, and the results are presented below.

	LINEAR MODEL	QUADRATIC MODEL	CUBIC MODEL
R^2 FOR RAINFALL	0.096467	0.098395	0.098415
R^2 FOR DELAYED RAINFALL	0.386211	0.438682	0.440549
R^2 FOR SMA FOR RAINFALL	0.435076	0.446277	0.447589
R^2 FOR SMA FOR DELAYED RAINFALL	0.249434	0.268109	0.270667

<u>Figure 12:</u> The r-squared table.

	LINEAR MODEL	QUADRATIC MODEL	CUBIC MODEL
ADJUSTED R^2 FOR RAINFALL	0.095610	0.097540	0.097560
ADJUSTED R^2 FOR DELAYED RAINFALL	0.385626	0.438146	0.440015
ADJUSTED R^2 FOR SMA FOR RAINFALL	0.434534	0.445745	0.447059
ADJUSTED R^2 FOR DELAYED SMA RAINFALL	0.248710	0.267403	0.269963

Figure 13: The adjusted r-squared table.

	LINEAR MODEL	QUADRATIC MODEL	CUBIC MODEL
RMSE FOR RAINFALL	13.081383	13.067421	13.067272
RMSE FOR DELAYED RAINFALL	10.848912	10.374837	10.357573
RMSE FOR SMA FOR RAINFALL	10.247044	10.144953	10.132926
RMSE FOR SMA FOR DELAYED RAINFALL	11.827096	11.679034	11.658604

Figure 14: The root mean squared error table.

Again, from the results above tables, we can infer from the above tables that the best model is the cubic model. This is because it has a low rate of errors and hence good performance compared to quadratic and linear as shown by RMSE and larger r square and adjusted r squared scores.

QUESTION 10:

It was asked to describe the optimal model to recommend for predicting the vegetation index by considering linear, nonlinear, and nonparametric models. For this, the decision tree, KNN regressor, cubic, quadratic, and linear models are selected, and the results are presented below.

	LINEAR MODEL	QUADRATIC MODEL	CUBIC MODEL	DECISION TREE MODEL	KNN REGRESSOR MODEL
R^2 SCORE	0.435076	0.446277	0.447589	0.225599	0.401993
RMSE SCORE	10.247044	10.144953	10.132926	11.997385	10.542822

Figure 15: The r^2 and RMSE scores for all models.

Based on Q9, the SMA for rainfall is the best feature as it has a low RMSE. By looking at the above table, we can infer that the cubic model is the best-performing model as it has the lowest root mean squared error and the highest r-squared score.

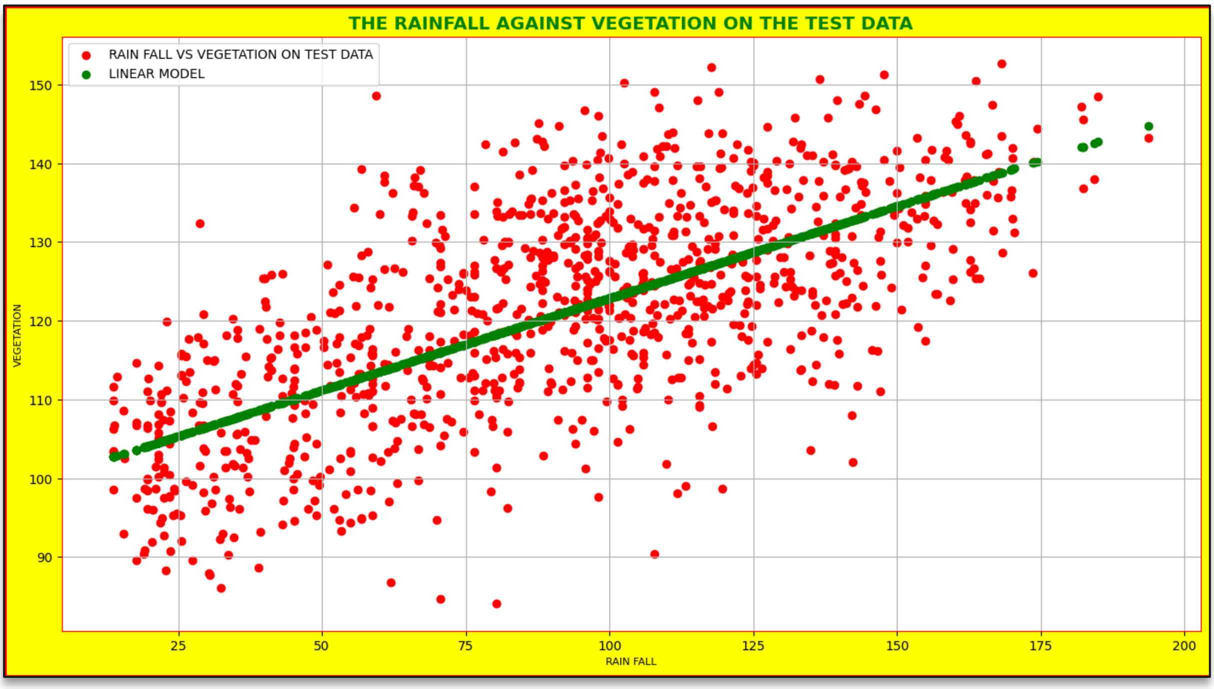


Figure 16: A linear model for vegetation vs rainfall

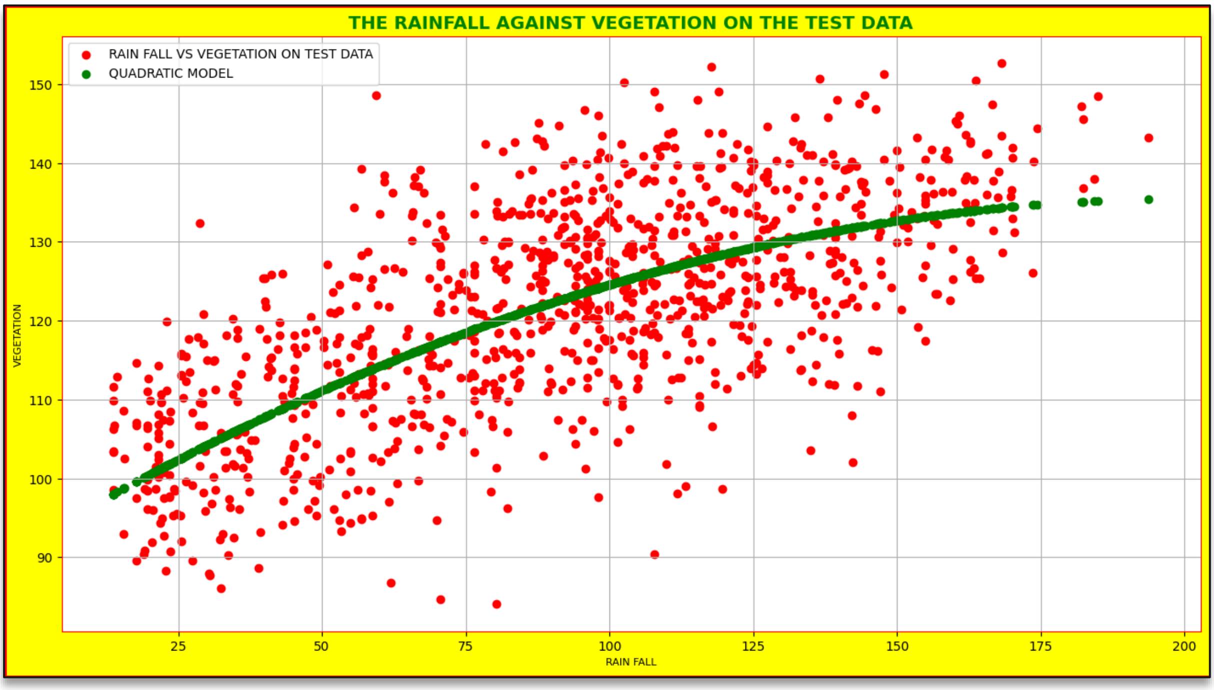


Figure 17: Quadratic model for vegetation vs rainfall

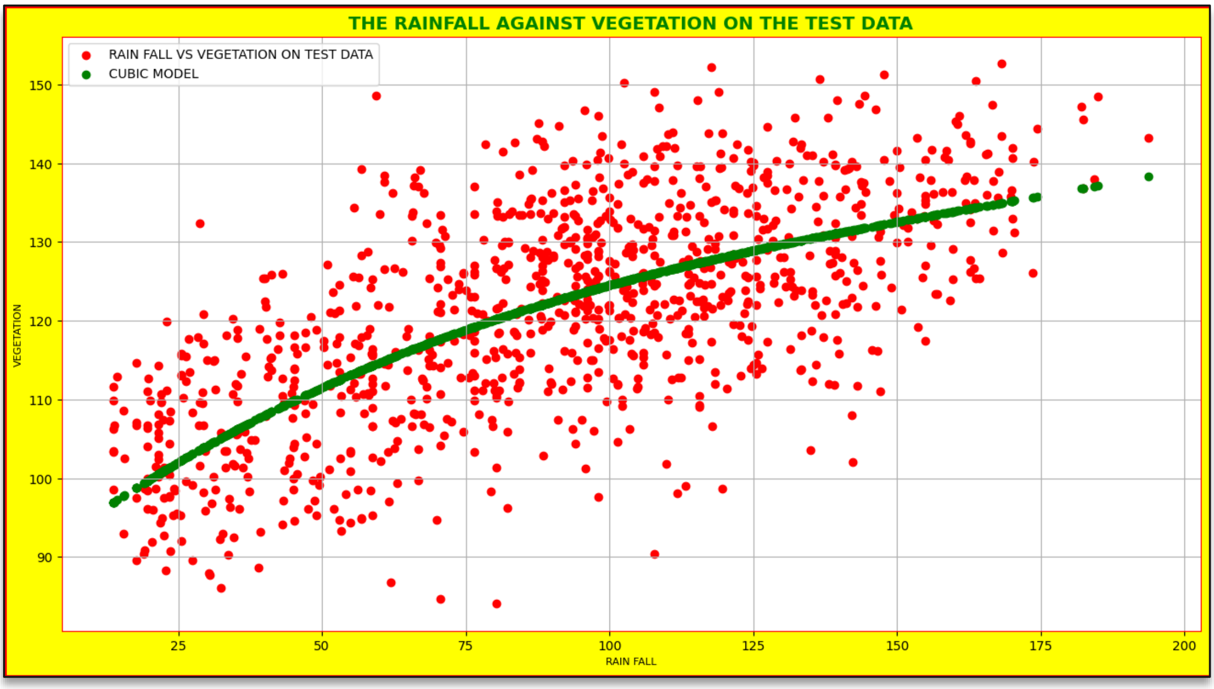


Figure 18: Cubic model for vegetation vs rainfall

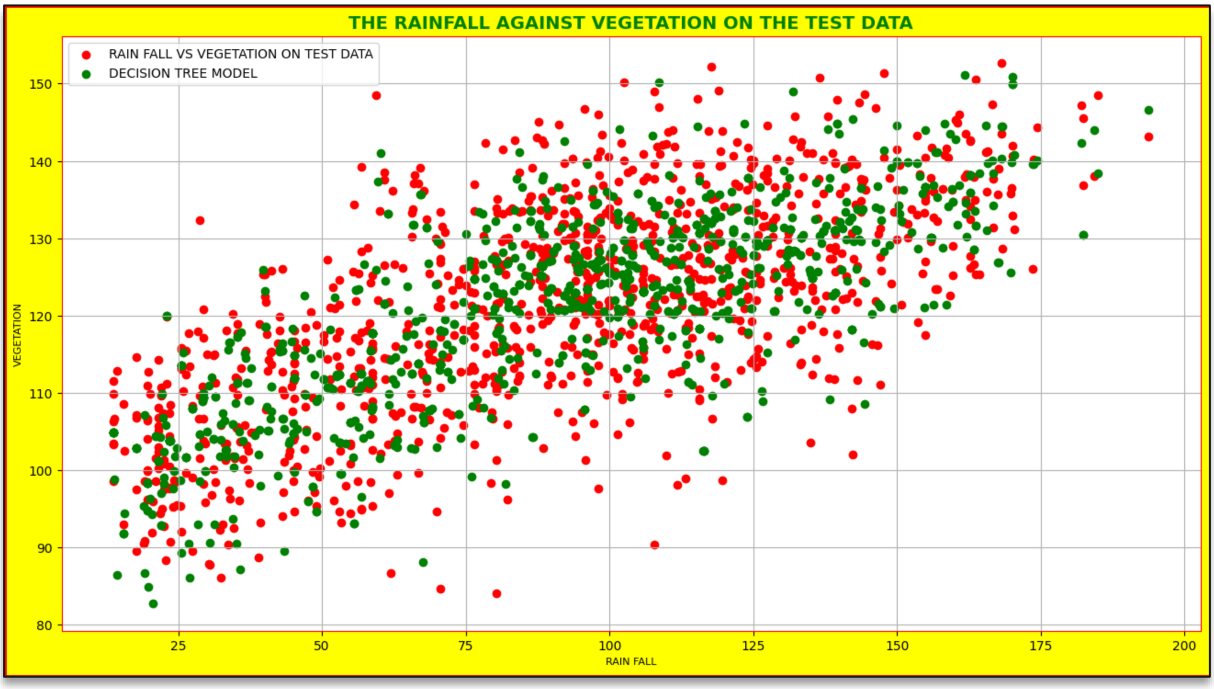


Figure 19: Decision tree model for vegetation vs rainfall

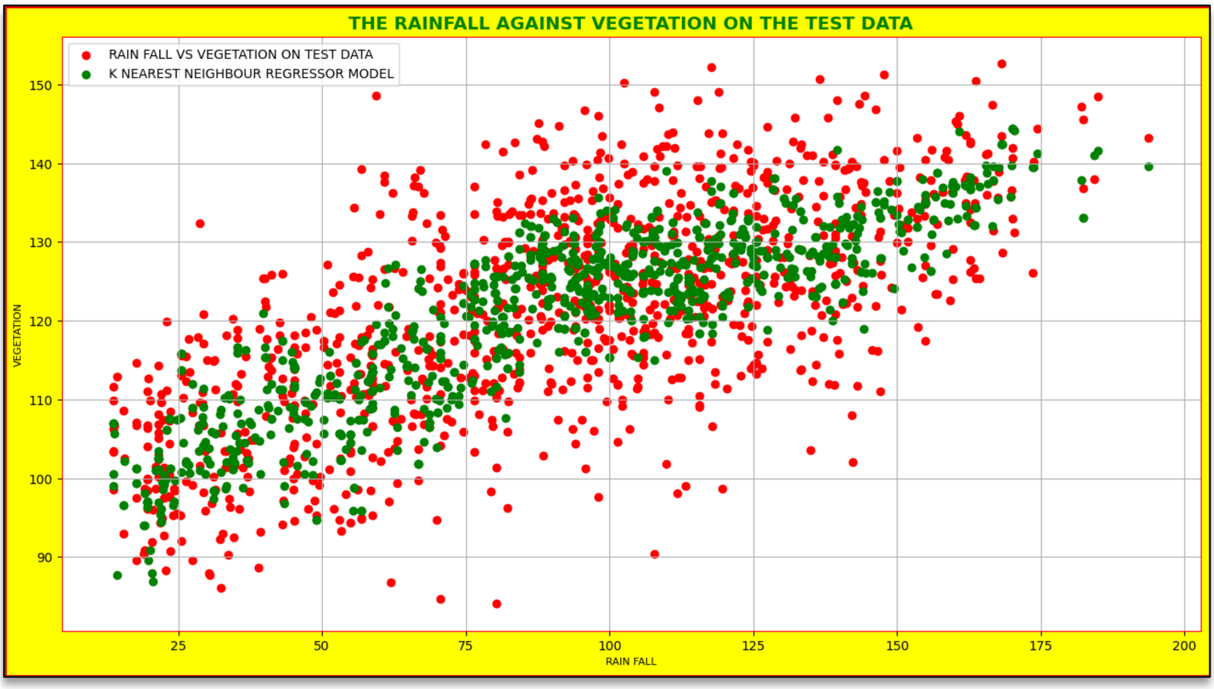


Figure 20: KNN model for vegetation vs rainfall

REFERENCES

[1] ‘pandas.read_csv — pandas 1.5.3 documentation’. https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html (accessed Mar. 25, 2023).

[2] ‘17. Creating Subplots in Matplotlib | Numerical Programming’. https://python-course.eu/numerical-programming/creating-subplots-in-matplotlib.php (accessed Mar. 25, 2023).

[3] ‘Matplotlib Subplots_adjust - Python Guides’, Sep. 16, 2021. https://pythonguides.com/matplotlib-subplots_adjust/ (accessed Mar. 25, 2023).

[4] ‘haversine: Calculate the distance between 2 points on Earth.’ Accessed: Mar. 25, 2023. [Online]. Available: https://github.com/mapado/haversine

[5] ‘scipy.optimize.curve_fit — SciPy v1.10.1 Manual’. https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.curve_fit.html (accessed Mar. 25, 2023).

[6] ‘sklearn.model_selection.train_test_split’, *scikit-learn*. https://scikit-learn/stable/modules/generated/sklearn.model_selection.train_test_split.html (accessed Mar. 25, 2023).