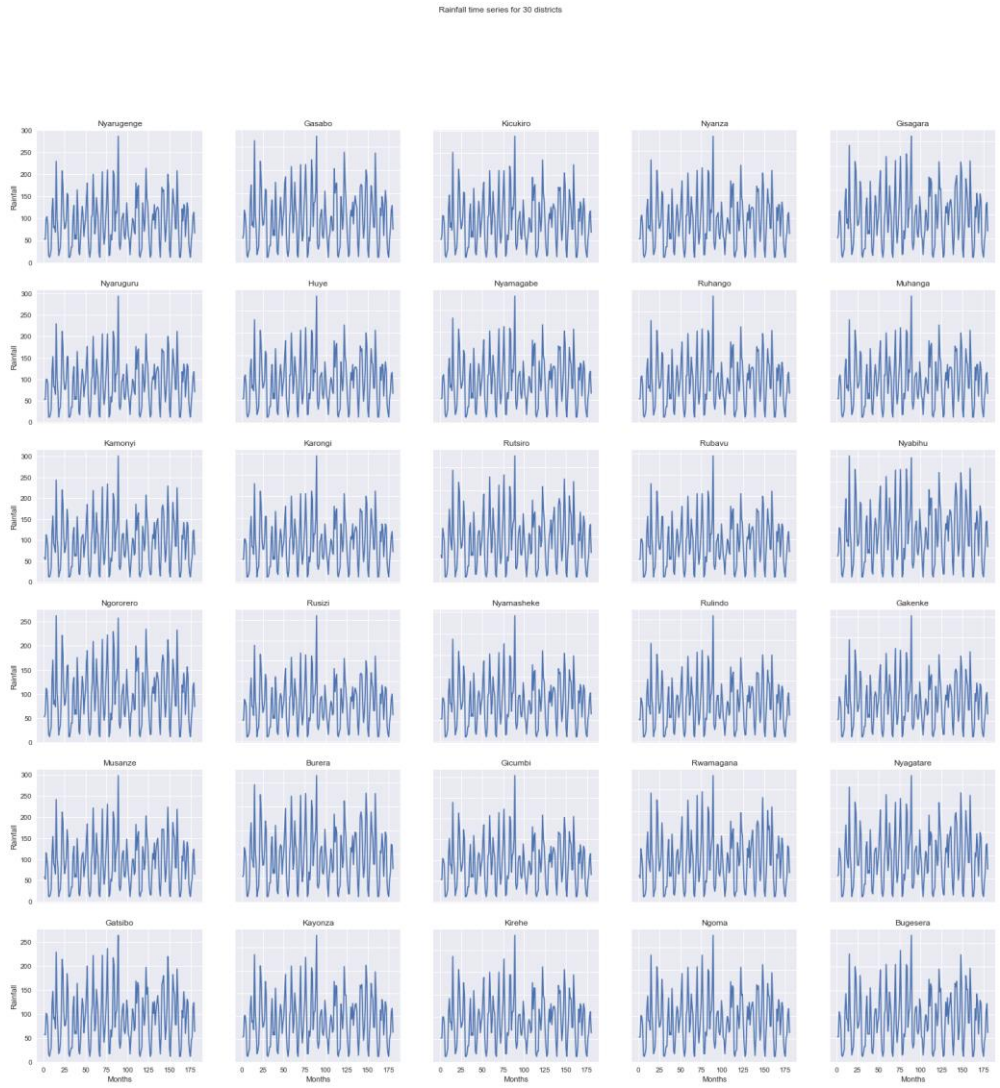
# BDS-Assignment1

Author: Jingyi Wu(jingyiw2)

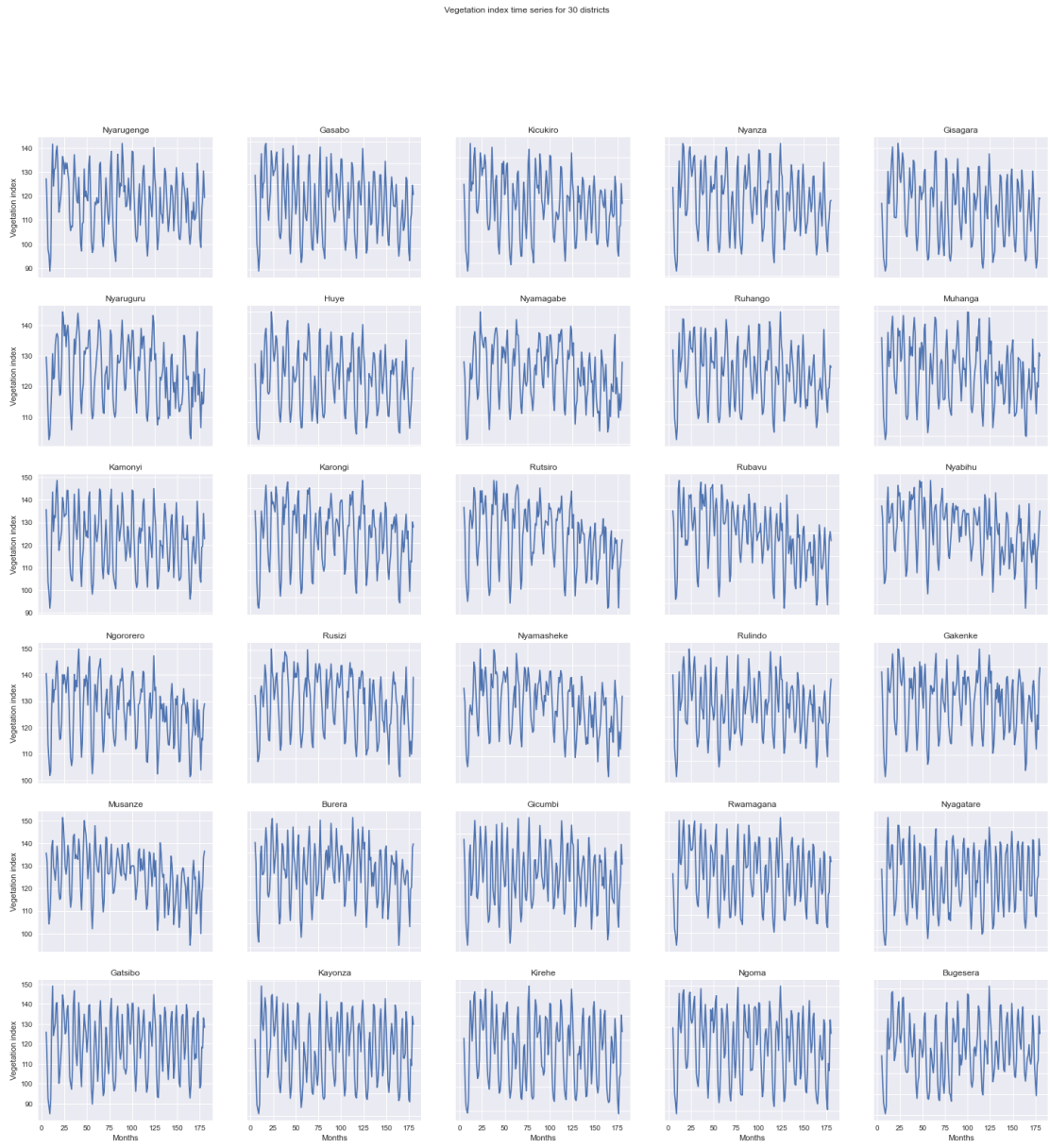
**Used library:**

Numpy
Pandas
Matplotlib
Seaborn
Haversine
Sklearn
Xgboost
Scipy

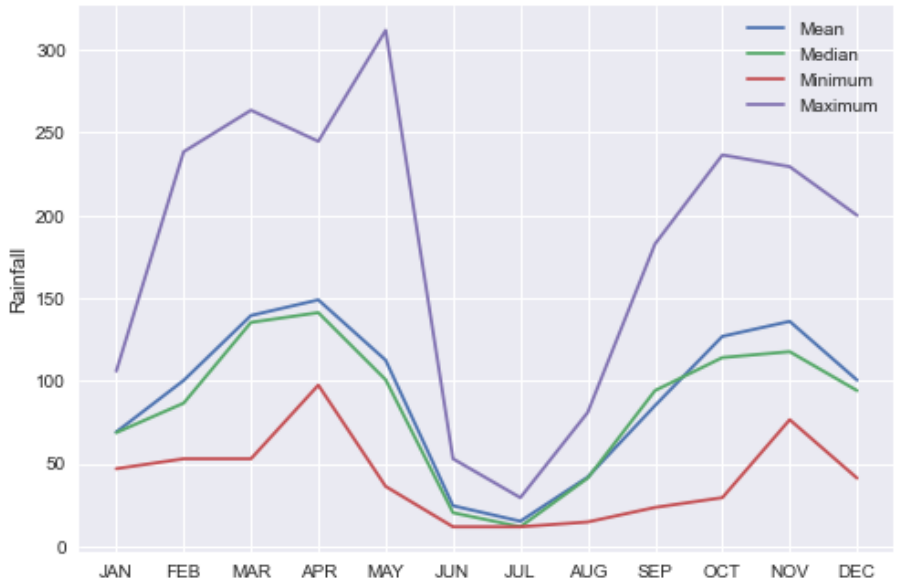
1. In this question, we need to load the two datasets provided to data frame. We get `rainfall` and `VI` as the result.
2. In Q2, we plot the two time series data of 30 districts. We get rainfall data as below. The rainfall data are quite similar for 30 districts and majority of values fall between 50~200.



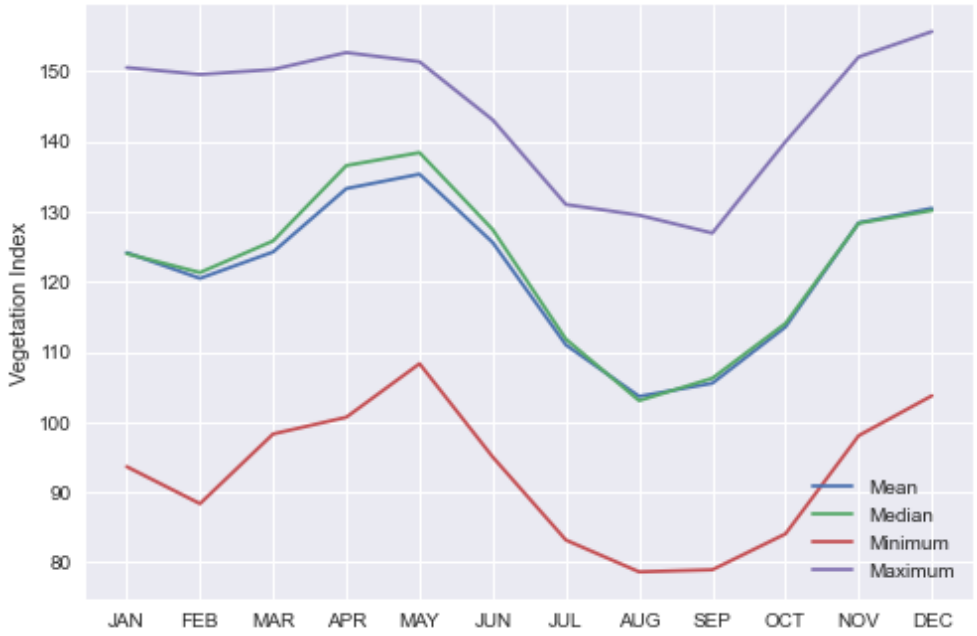

The vegetation index as plotted as below. Also, quite similar among districts. Majority of values vary between 110~140.




3. In Q3, we calculate statistics of these two variables for each month. As for rainfall, average of rainfall reaches highest in April and November, two rainy months for Rwanda, and reaches the lowest in July, dry season (June, July, August). There is abnormally high value in May, 2007 in Rusizi.



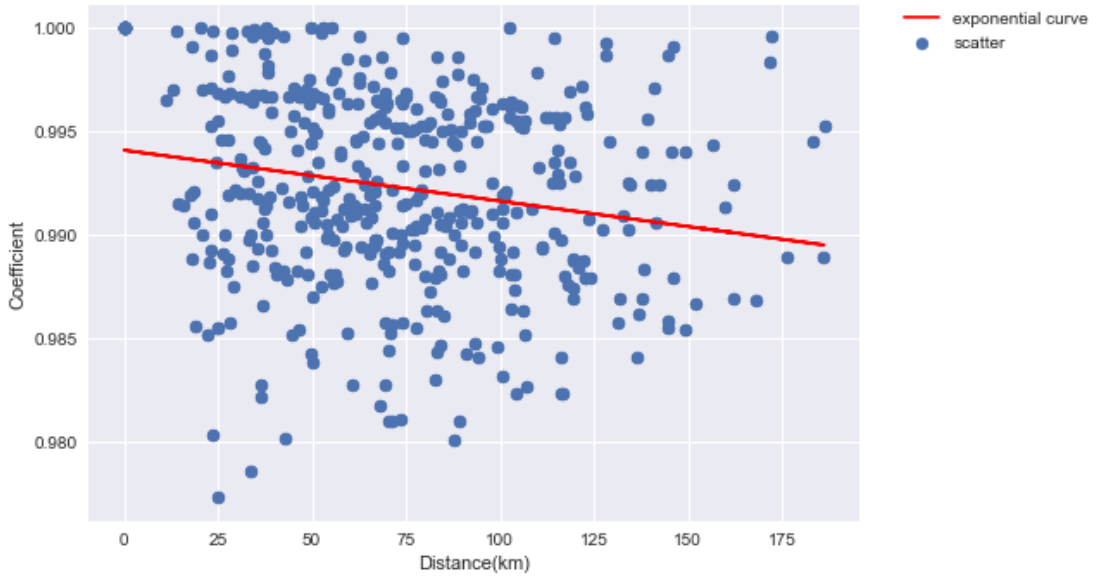

For vegetation index, there are two peaks in May and December, which implies a delayed impact of rainfall on vegetation index. The index is generally smallest in August.



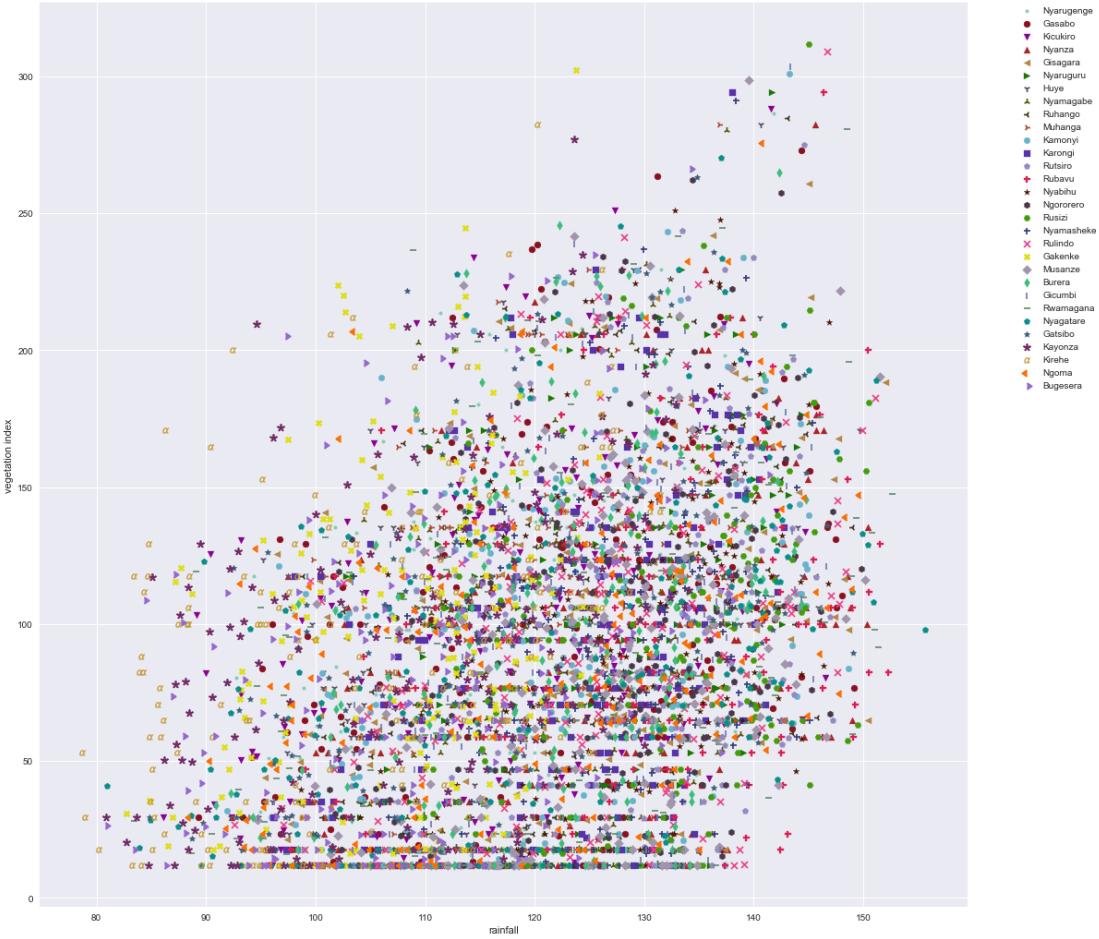

4. In Q4, we are required to research the relationship between correlation of district pairs and their distance, and figure out the parameters.

   After loading the coordinate data, we use the haversine module to calculate the distance. Then we fit the C(d) = C0exp(-ad) model to the data points and plot it on the scatter plot. In general, districts in Rwanda are strongly correlated to each other. With distance being larger, the correlation declines a little, but not too much. The estimated C0 is about 0.994 and decay constant a is

-2.4826529745717737e-05. Note that the R2 value for this model is only 0.038, showing the fitting result is not satisfying.

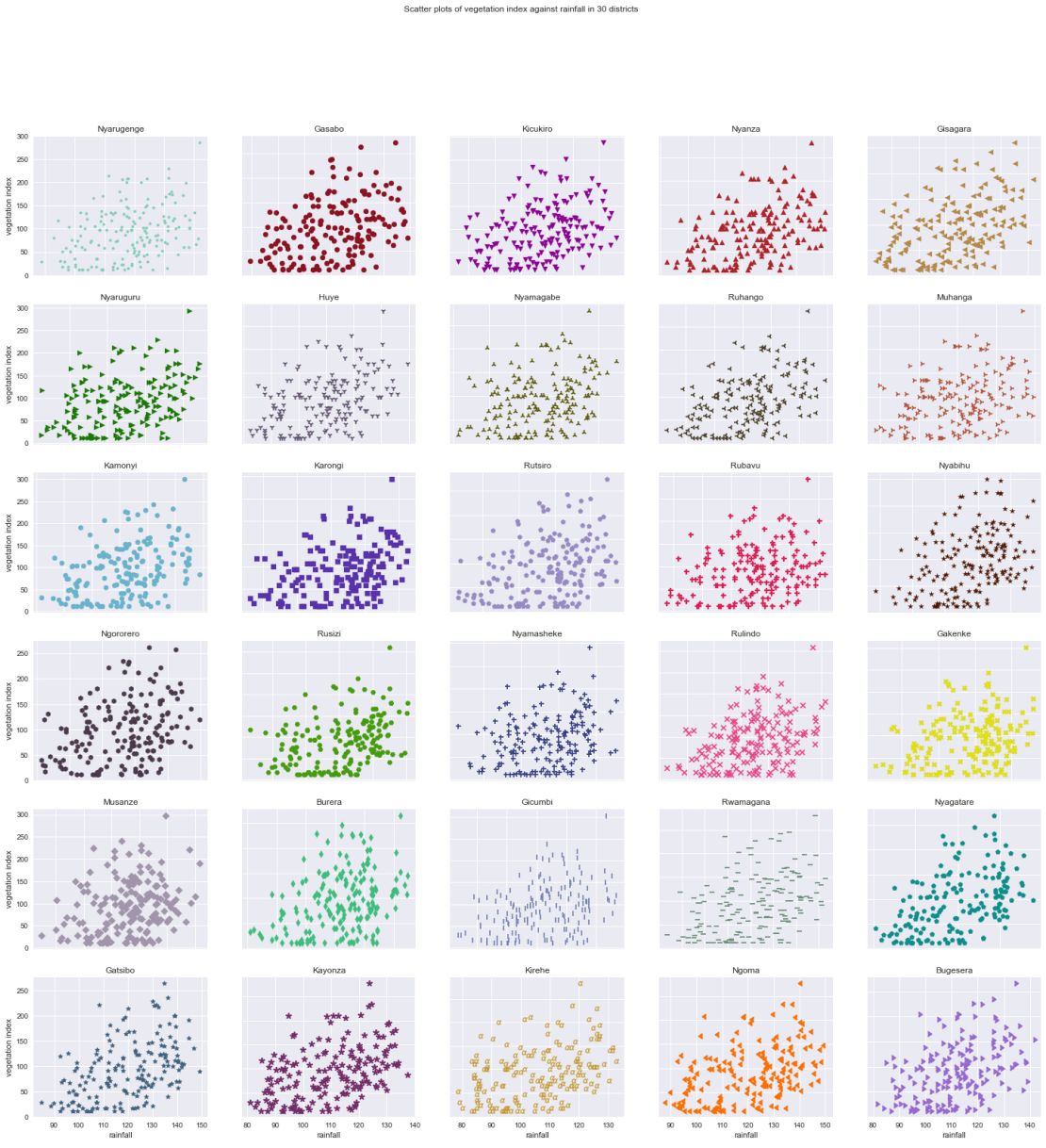



5. In Q5, we make scatter plot of vegetation and rainfall in 30 districts.




Although we use different color and symbol to distinguish between

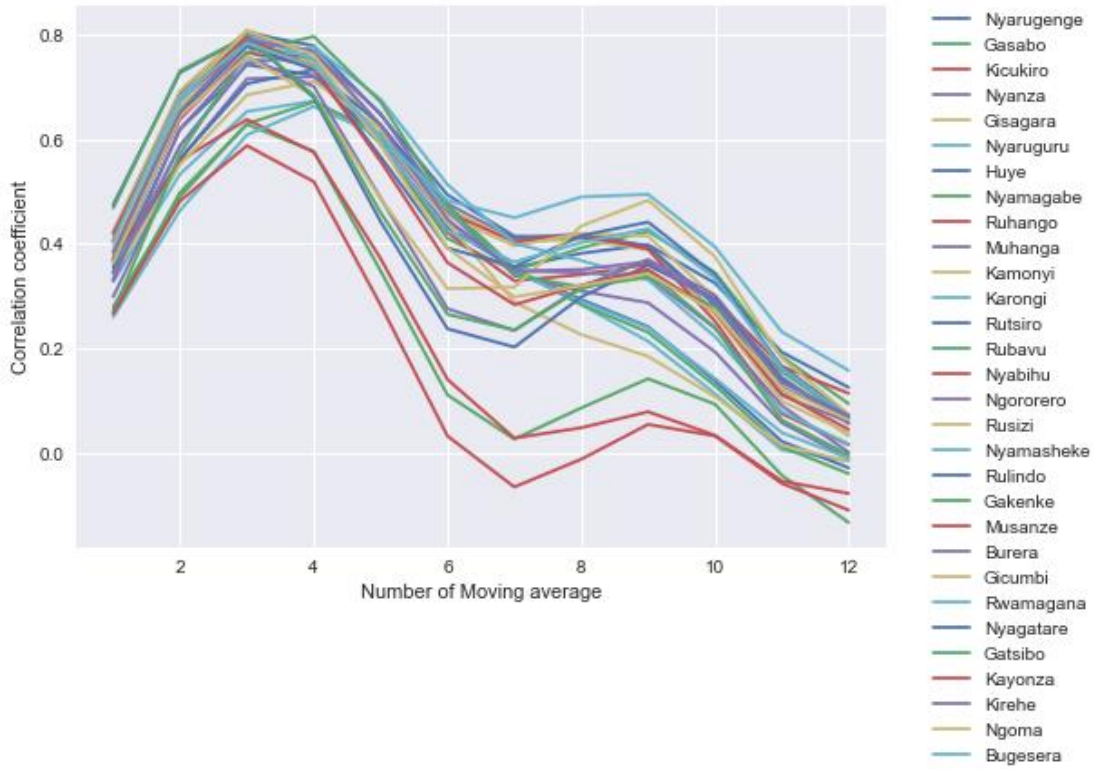
30 districts, it seems quite messy and hard to find out insights. Therefore, I made a 6*5 subplots to see each district respectively, which is clearer. We can see from the plot that the patterns are quite similar in 30 districts. There is slightly positive correlation pattern between vegetation index and rainfall.




6. In Q6, we transformed rainfall variable to delayed rainfall and find out the optimal shift number for each district. The result is: Except for Musanze, 1 is the best shift period. Therefore, the optimal k here is 1.

7. In Q7, we use SMA on rainfall variable to capture the relationship, and plot the correlation coefficient against moving

window size for each district.




The consensus in SMA turns out to be 3, so optimal n here is 3.

8. Q8 requires us to apply linear/quadratic/cubic regression model to capture the correlation. Also try different transformed rainfall at the same time. The result is as below.

R-squared value

| Variable | Linear | Quadratic | Cubic |
|---|---|---|---|
| Rainfall | 0.10945257623706 | 0.1161975165734 | 0.11897204790563 |
| Delayed Rainfall | 0.38797887040735 | 0.446699179065 | 0.449685709769 |
| SMA Rainfall | 0.450929714412 | 0.4690852180417 | 0.469938931742 |

Adjusted R-squared value

| Variable | Linear | Quadratic | Cubic |
|---|---|---|---|
| Rainfall | 0.10928384804 | 0.1160300663113 | 0.11880512332205 |
| Delayed Rainfall | 0.3878629133915 | 0.1160300663113 | 0.449581444083 |
| SMA Rainfall | 0.4508256844223 | 0.46898462789738 | 0.46983850334686 |

RMSE value

| Variable | Linear | Quadratic | Cubic |
|---|---|---|---|
| Rainfall | 13.19763984287 | 13.147565874387 | 13.1269124996877 |
| Delayed Rainfall | 10.9408472512 | 10.40275580575 | 10.3746425391401 |
| SMA Rainfall | 10.3629097842 | 10.1901401087 | 10.1819439136425 |

We can conclude that the best model here is using SMA and cubic model.

9. Use cross validation (train test split) to test if SMA and shift combined is a better transformation. Use window_size=3 and shift=1 from conclusions above and apply regression model. The results are as below.

R-squared value

| Variable | Linear | Quadratic | Cubic |
|---|---|---|---|
| Rainfall | 0.10741096 | 0.12009401927 | 0.1245250833 |
| Delayed Rainfall | 0.27284784248 | 0.369964131017 | 0.376388805465 |
| SMA Rainfall | 0.38291647316 | 0.398467784698 | 0.400395033994 |
| Delayed SMA | 0.37780995531 | 0.4285018953749 | 0.42697508724736 |

Adjusted R-squared value

| Variable | Linear | Quadratic | Cubic |
|---|---|---|---|
| Rainfall | 0.10658295416 | 0.11927777996 | 0.1237129544 |
| Delayed Rainfall | 0.2721733043 | 0.119277779956 | 0.375810316416 |
| SMA Rainfall | 0.38234403946 | 0.397909777077 | 0.399838814174 |
| Delayed SMA | 0.3772162625188 | 0.4279565727559 | 0.42642830775046 |

RMSE value

| Variable | Linear | Quadratic | Cubic |
|---|---|---|---|
| Rainfall | 13.23973308 | 13.145332911 | 13.1121922522 |
| Delayed Rainfall | 11.949950409 | 11.1233625359 | 11.0665029930 |
| SMA Rainfall | 11.00843116 | 10.8688326041 | 10.851407307149 |
| Delayed SMA | 11.61278181918 | 11.12966608779 | 11.1445231189418 |

We can see that combining two transformations together generates a **more satisfying** outcome (**Adjusted R-squared wise**). But SMA cubic performs better **RMSE-wise**. The best model here is delayed SMA Quadratic. **But** we will also consider SMA in Q10 as comparison.

10. In Q10, apart from the previous models, using the best outcome from Q9 research – SMA and delayed SMA, I tried decision tree regressor and a fancier xgb regressor. The results are as below.

R-squared value

| Variable | Linear | Quadratic | Cubic | Decision Tree | Xgb |
|---|---|---|---|---|---|
| SMA Rain | 0.38291647 | 0.398468 | 0.400395 | 0.41279840021 | 0.406857492 |
| Delayed SMA | 0.37780995531 | 0.42850120 | 0.42697508724736 | 0.4225682246 | 0.390857768 |

Adjusted R-squared value

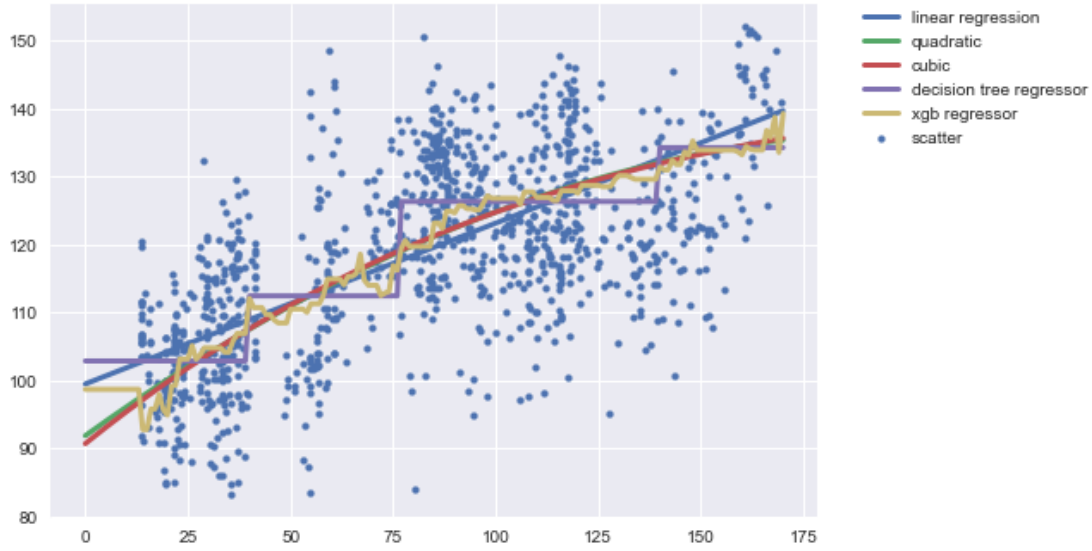
| Variable | Linear | Quadratic | Cubic | Decision Tree | Xgb |
|---|---|---|---|---|---|
| SMA Rain | 0.382344039 | 0.397909777 | 0.39983881 | 0.412253686 | 0.406307267 |
| Delayed SMA | 0.377216263 | 0.427956573 | 0.42642831 | 0.422017240 | 0.3902765253 |

RMSE value

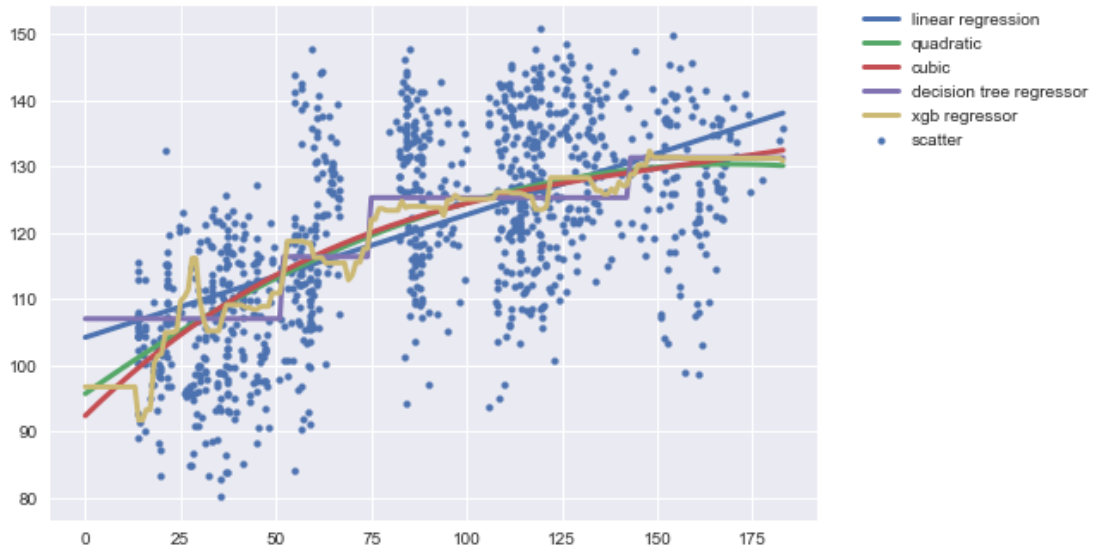
| Variable | Linear | Quadratic | Cubic | Decision Tree | Xgb |
|---|---|---|---|---|---|
| SMA Rain | 11.008431 | 10.8688326 | 10.8514073 | 10.738585258059 | 10.7927714179 |
| Delayed SMA | 11.612782 | 11.1296662 | 11.1445231 | 11.18729466 | 11.4903720960 |

From the above charts, we can see that delayed SMA in Quadratic regression performs best, since we always consider R2 and adjusted R2. But Decision tree model using SMA performs best when we consider RMSE, so it can also be used in predictions. The plot of test data and fitted curves are as below.

Using SMA variable:




Using delayed SMA:


150
140
130
120
110
100
90
80
0
25
50
75
100
125
150
175
linear regression
quadratic
cubic
decision tree regressor
xgb regressor
scatter