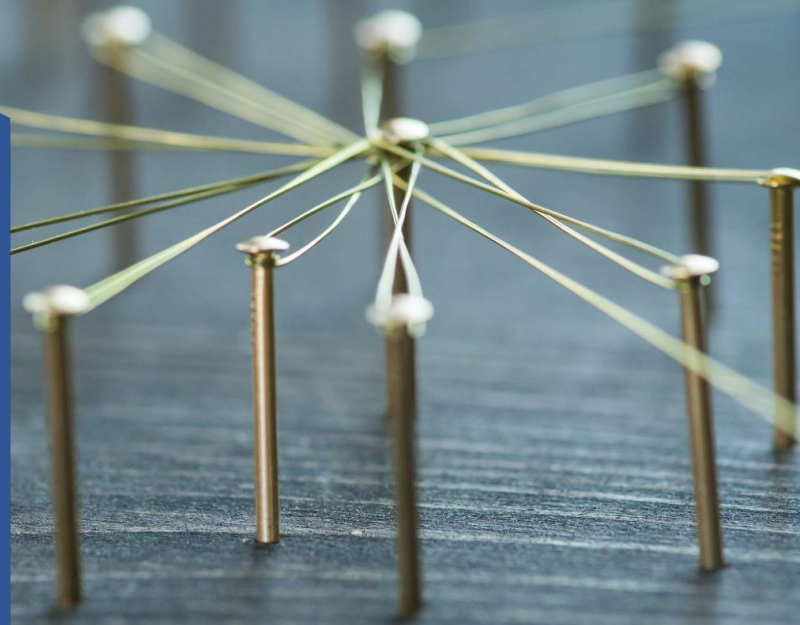


Mixture of Functional Linear Models

Upmanyu Singh
Shikha Sharma
Sanchit Vijay



Data Collection

- Used the library 'wbgapi' for collection of the data.
- Dependent variable or response variable is CO2 per capita.
- GDP per capita is the covariate.
- Data consist of years from 1994 to 2021 for 56 countries.

Data Preprocessing

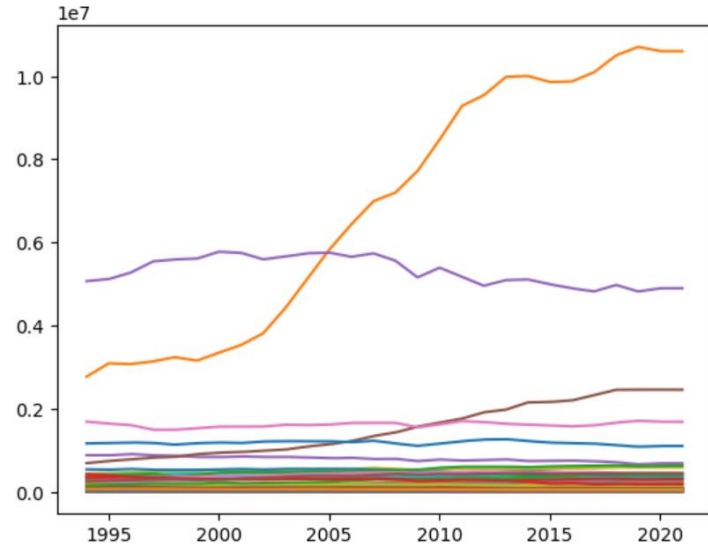
- Removing the columns with more than 25 null values
- Filling nan with knn imputation technique
 - A new sample is imputed by finding the samples in the training set “closest” to it and averages these nearby points to fill in the value.

Functional Data Analysis



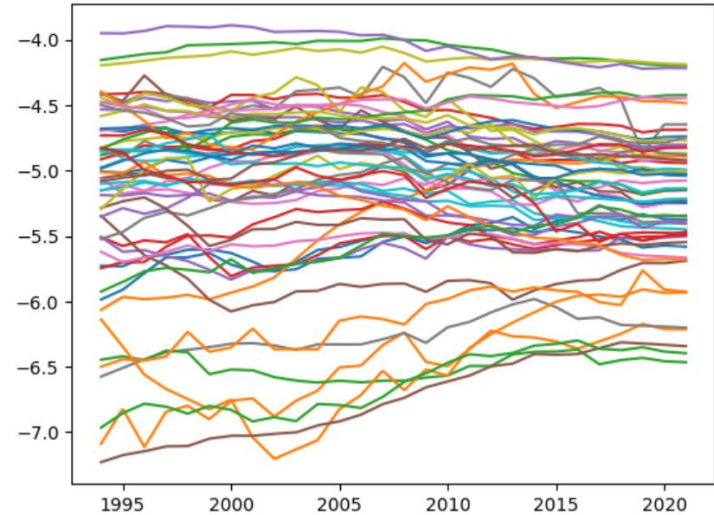
Graph of CO2 emission vs year

- The graph doesn't explain the variability of the CO2 for every year. It's because there is a high difference between the values.



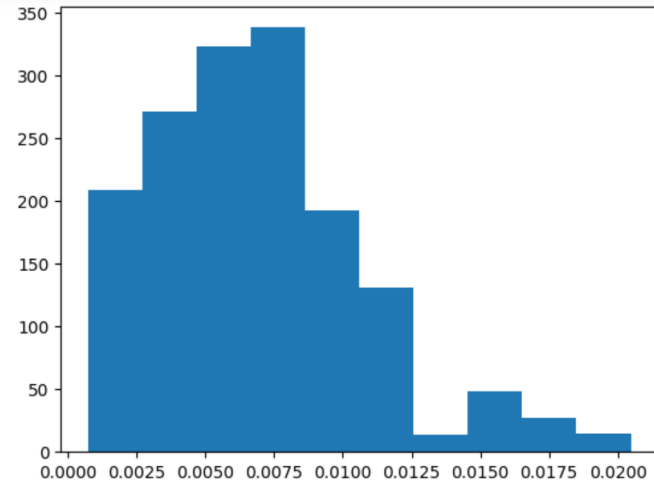
Taking Log of CO2 per capita explains the variability across years

- So by taking the log of co2 per capita, we can see variability can be seen.



Normality of Response variable (CO2 per capita)

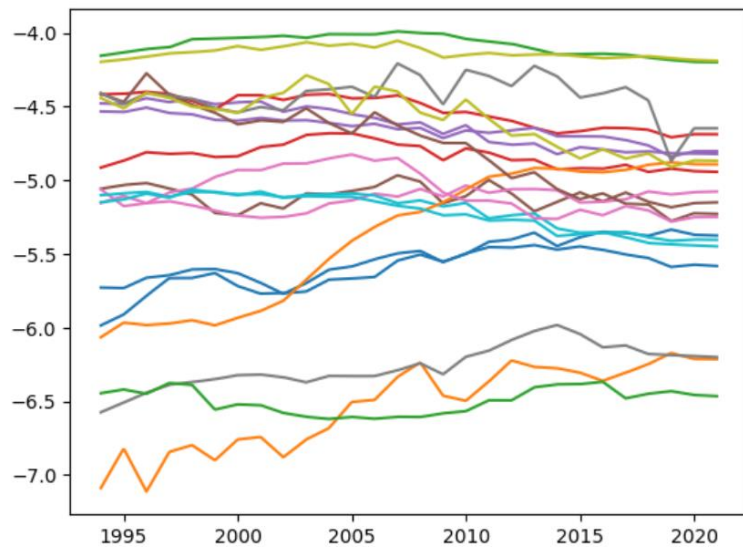
- Response variable is co2 per capita and it is rightly skewed



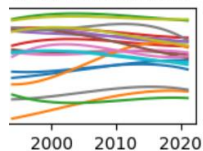
Basis Function

- The linear basis function models are used when the relationship between the inputs and the target is non-linear.
- This is a generalization of linear regression that essentially replaces each input with a function of the input.
- The equation will be:
 - $y(x,w) = w_0 + w_1\phi(x_1) + w_2\phi(x_2) + \dots + w_D\phi(x_D)$

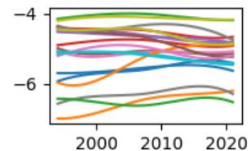
Basis Function



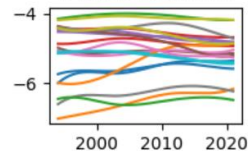
1 basis functions



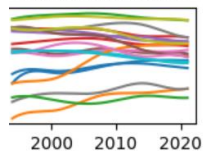
2 basis functions



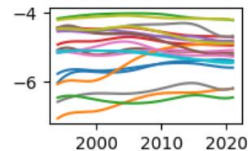
3 basis functions



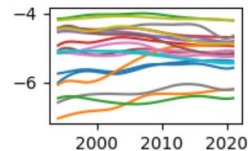
4 basis functions



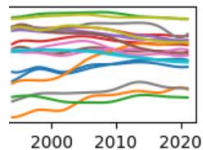
5 basis functions



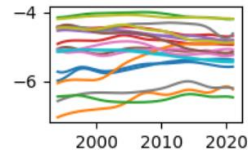
6 basis functions



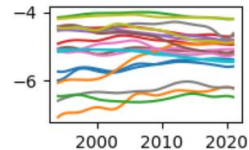
7 basis functions



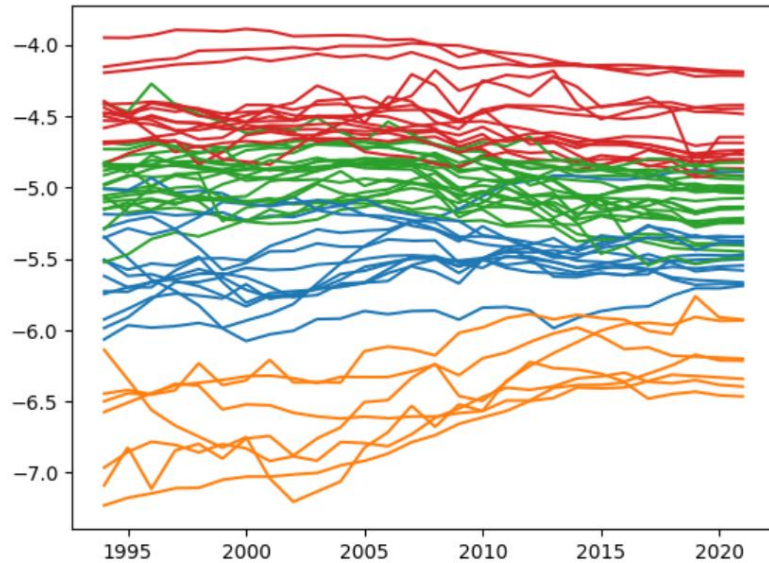
8 basis functions



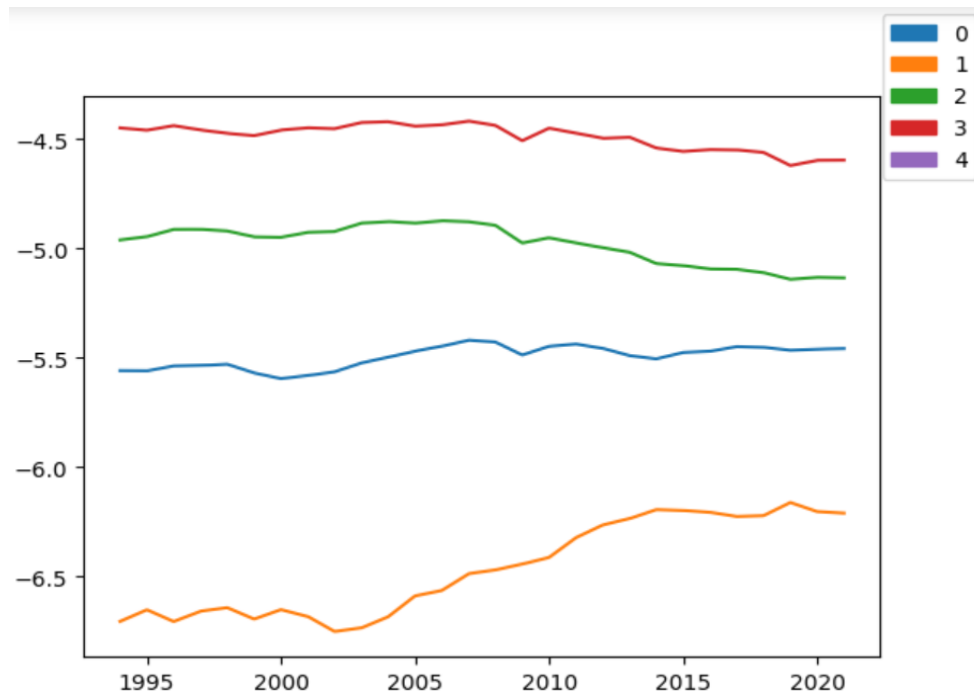
9 basis functions



Clustering



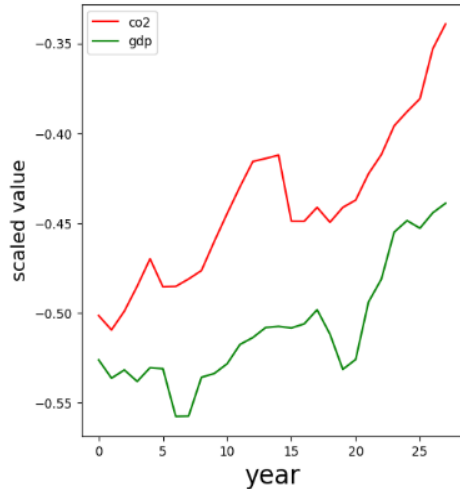
Clustering



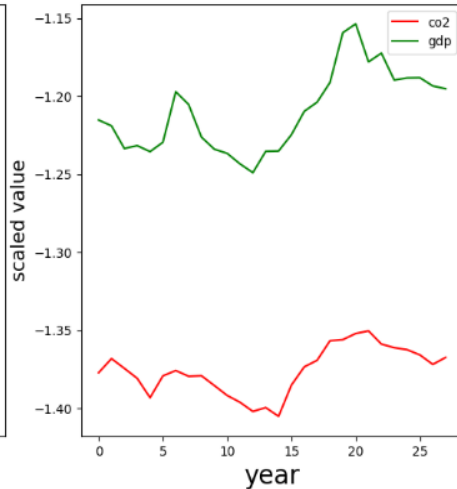
Each group showing different relation of co2-gdp

Clusters are formed by grouping the cluster and taking the mean of GDP and co2 per capita.
Cluster 4 contains the developed countries and **Kazakhstan**.

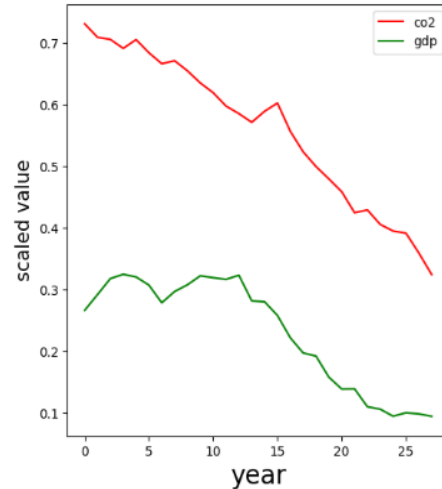
Cluster1



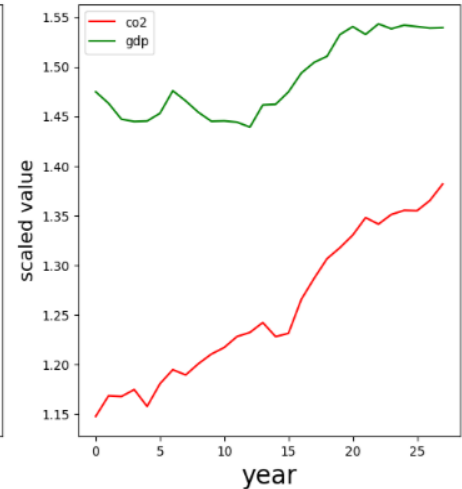
Cluster2



Cluster3

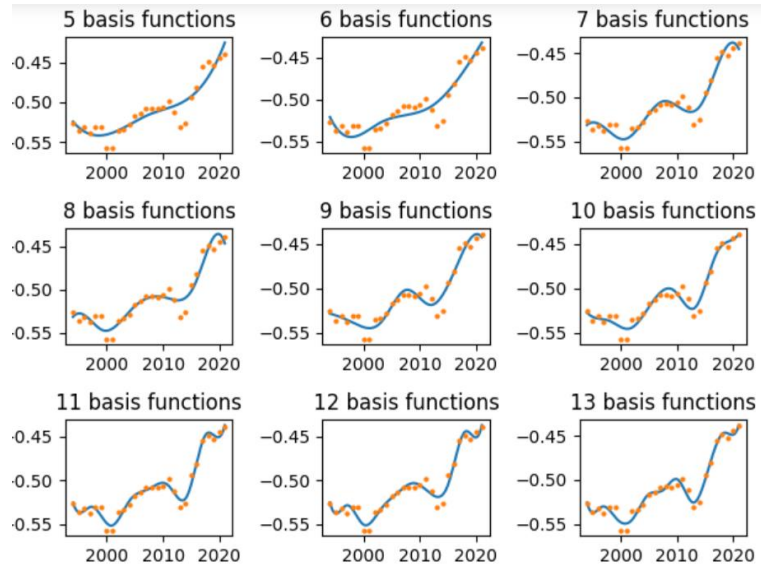


Cluster4



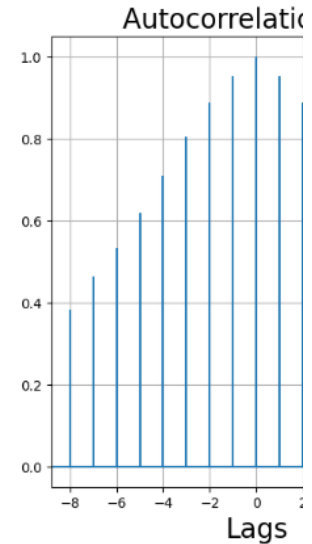
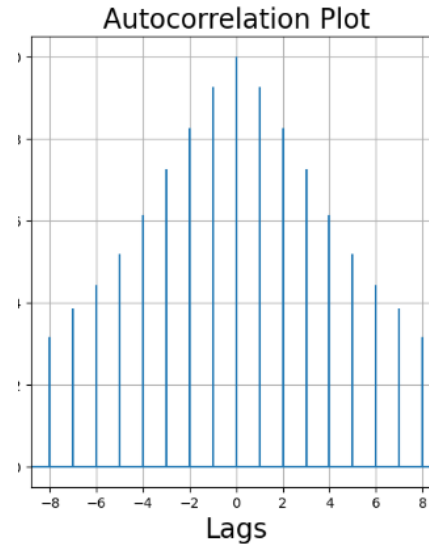
Basis representation of the first cluster

- This is the basic representation of cluster1 and when you see that if you are increasing the basis function number the function tends to overfit with the scatter plot.

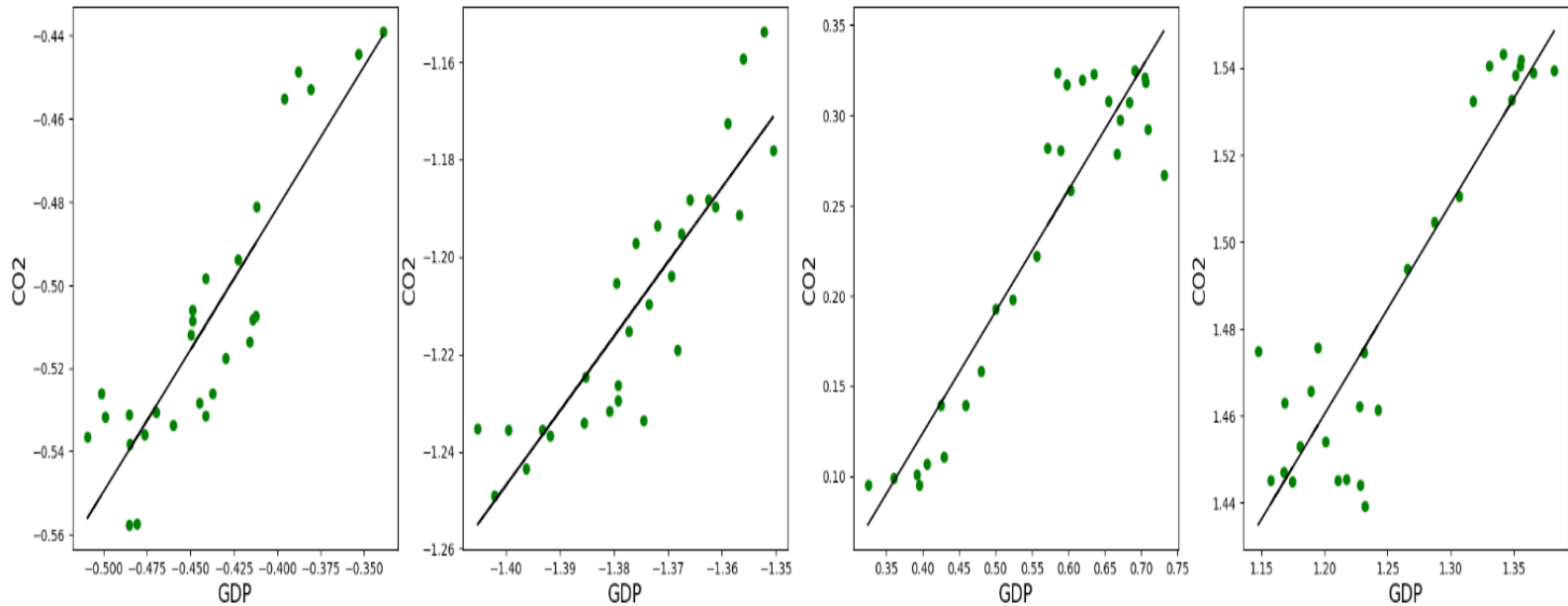


Autocorrelation plot

By selecting the basis function equal to 9 we plot the autocorrelation plot of each cluster and we can see that each cluster is having a correlation with the lagging term.

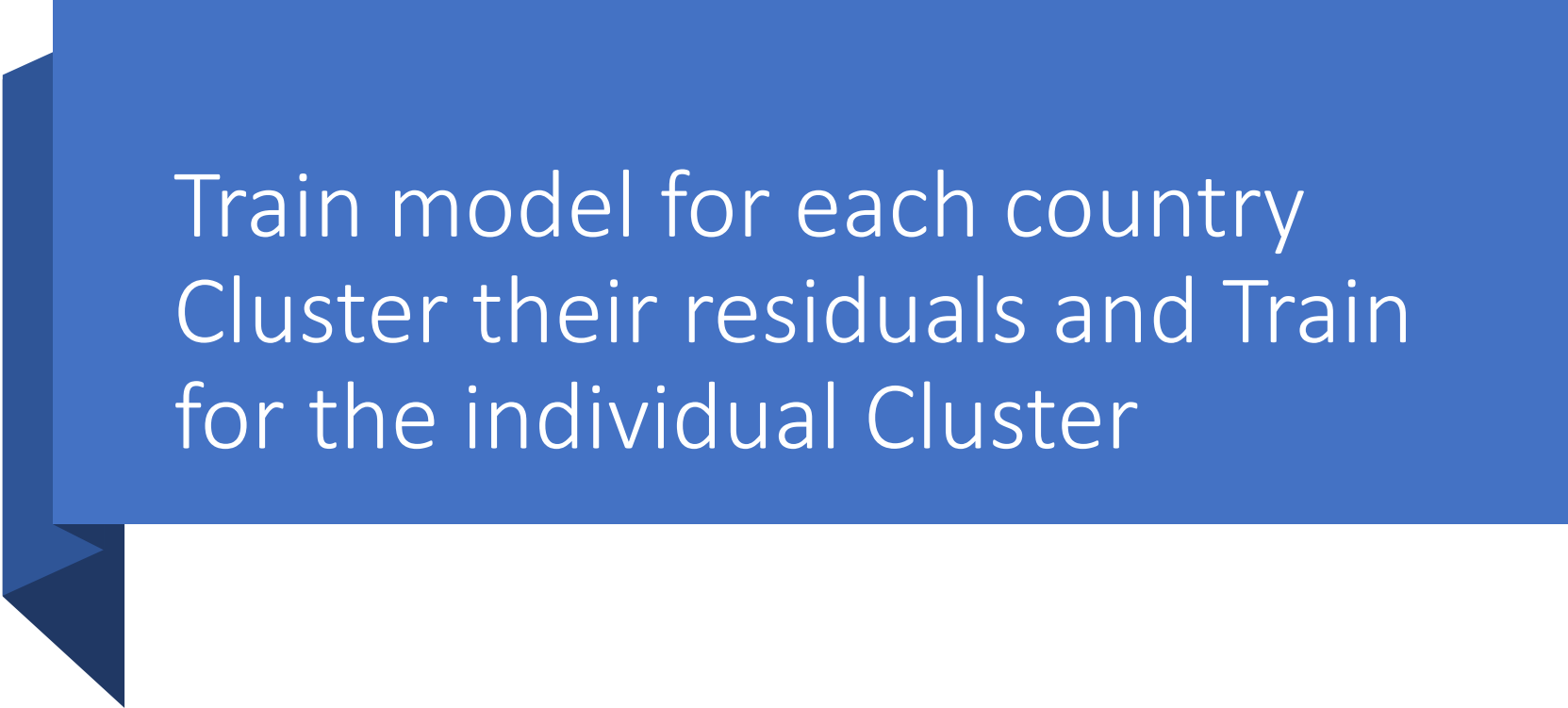


Regression plot for each cluster



Root Mean
squared error
for each
cluster

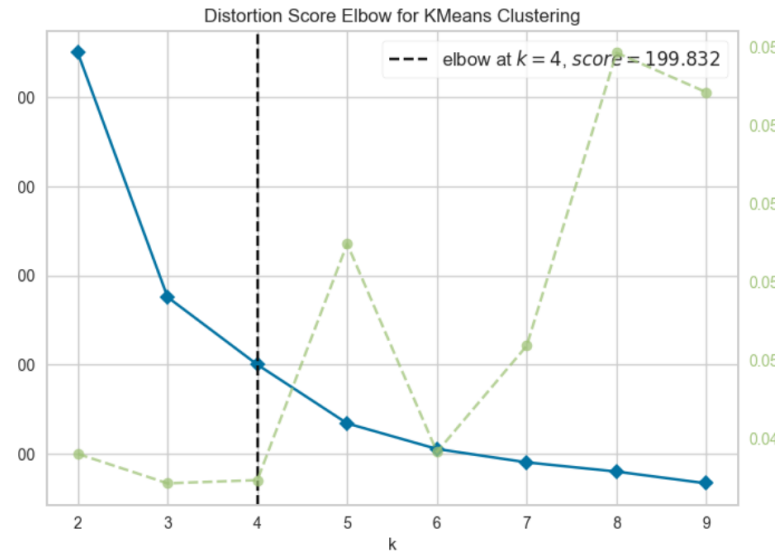
Cluster	RMSE
1	0.015
2	0.011
3	0.032
4	0.015



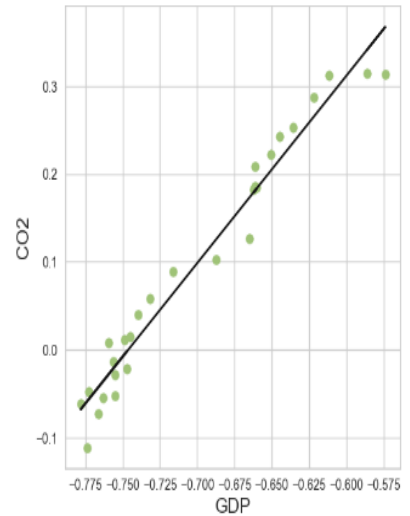
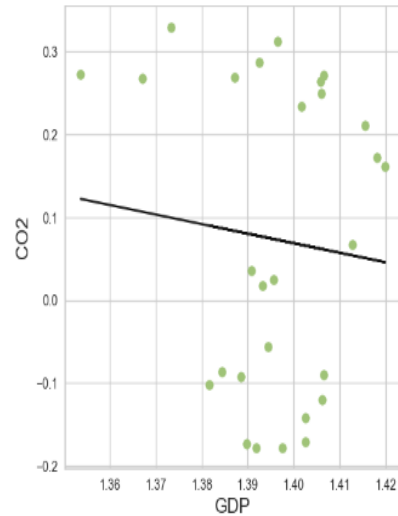
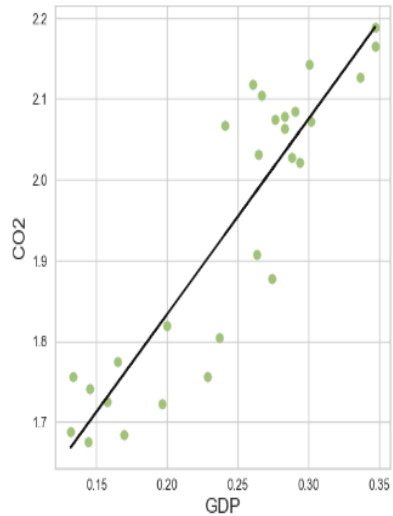
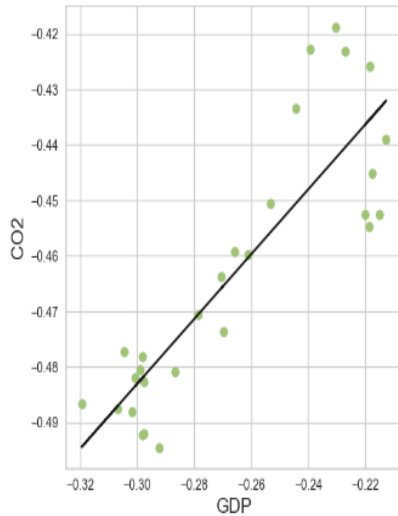
Train model for each country
Cluster their residuals and Train
for the individual Cluster

Elbow curve for finding the optimal value of N (K-means)

- Optimal cluster = 4



Train model for each cluster again (Linear Regression)

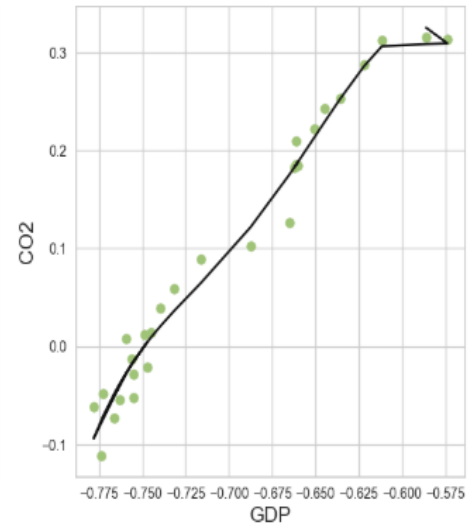
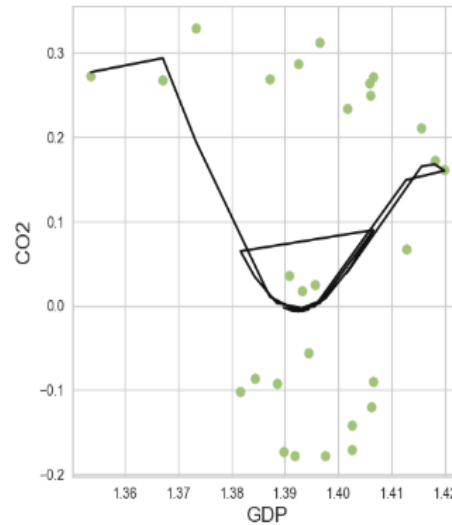
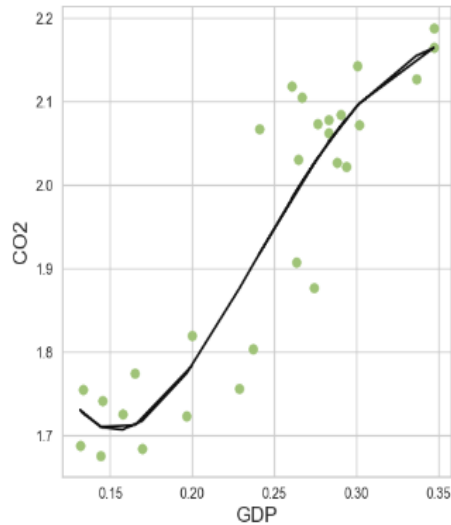
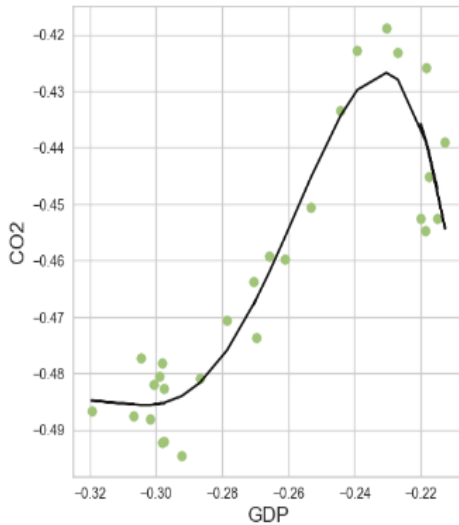


Root Mean
squared error
for each
cluster

Cluster	RMSE
1	0.011
2	0.07
3	0.18
4	0.026

Train model for each cluster again (Polynomial Regression)

Degree=4



Root Mean
squared error
for each
cluster

Cluster	RMSE
1	0.007
2	0.068
3	0.16
4	0.021

Challenges faced

- Understanding of the functional linear model
- Took time to understand how to implement the mixture of functional linear model
- Basis function and why it is important for functional model
- Fitting the best fit model for the data took a little more time

Improvements

- We could have back-test the model on historical data and different countries for the validation of the model
- Should have tried different regression other than linear regression like SVM regression.



Thank you!