ABSTRACT

The World Bank houses tons of data on various indicators such as forest area and greenhouse gas emission. We apply the various clustering analysis algorithms such as **KMeans clustering** and **Agglomerative clustering** to identify natural patterns and groupings of all the countries in the world with regards to forest area and greenhouse gas emission.

We pull the string harder by doing a curve fitting on the greenhouse gas emission. Curve fitting helps us determine necessary trends which helps governments make informed decisions. We make good use of python essential libraries in doing the analysis.

INTRODUCTION

Cluster analysis, or clustering, is an unsupervised machine learning task. It involves automatically discovering natural grouping in data. It is this machine learning algorithm which companies like Amazon and Alibaba use for basket analysis and goods recommendation systems.

KMeans Clustering involves assigning examples to clusters in an effort to minimize the variance within each cluster. It is the most widely used clustering algorithm.

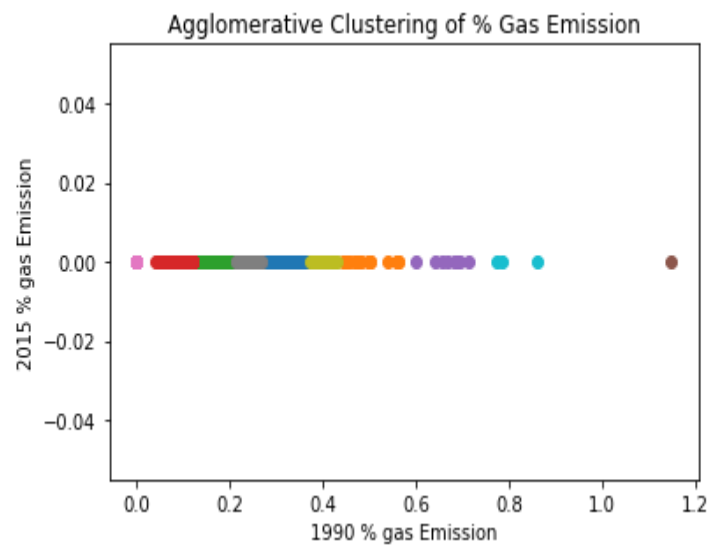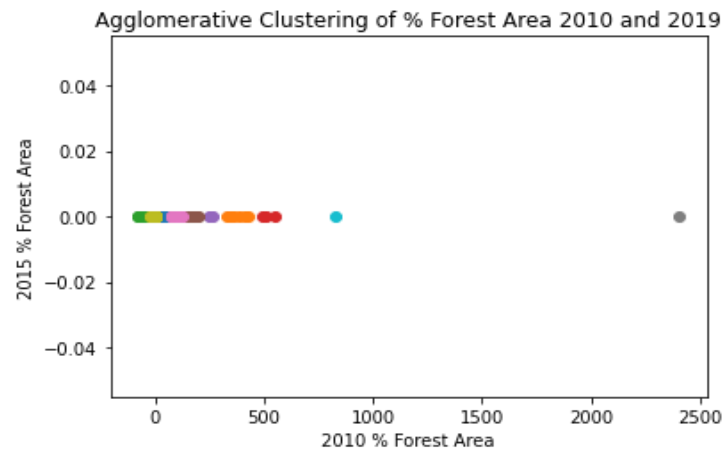Agglomerative clustering involves merging examples until the desired number of clusters is achieved.It is a part of a broader class of hierarchical clustering methods and you can learn more here.
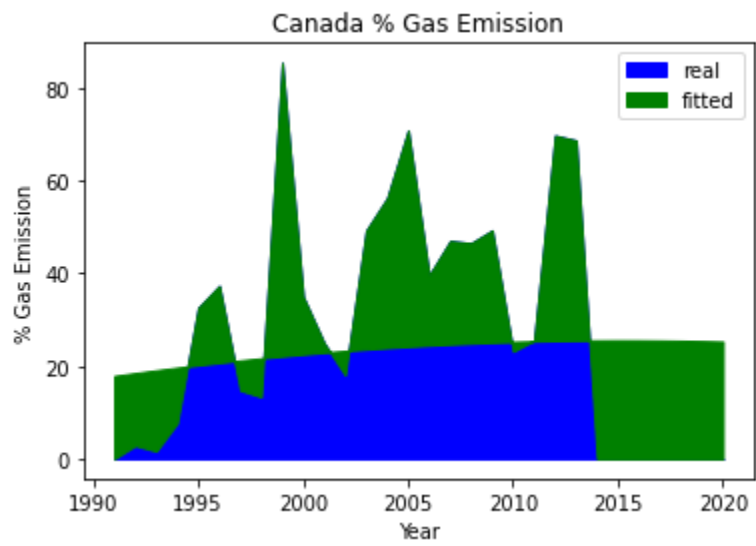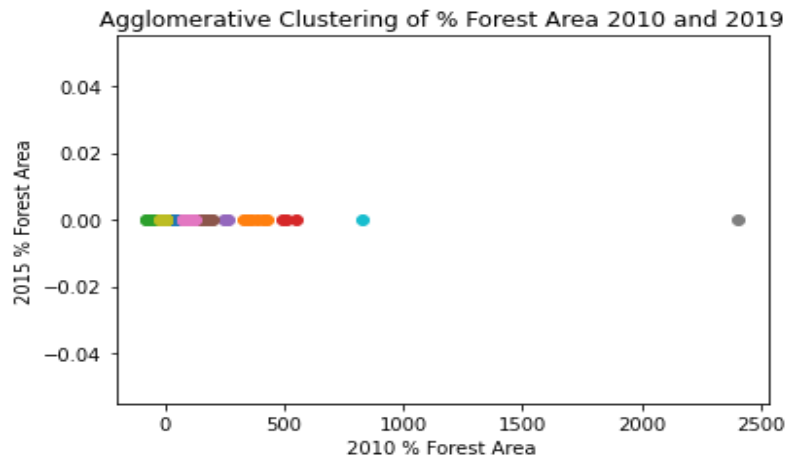
Curve fitting is the process of constructing a curve, or mathematical function, that has the best fit to a series of data points. It is very salient in doing future projection of values which help in making informed decisions.
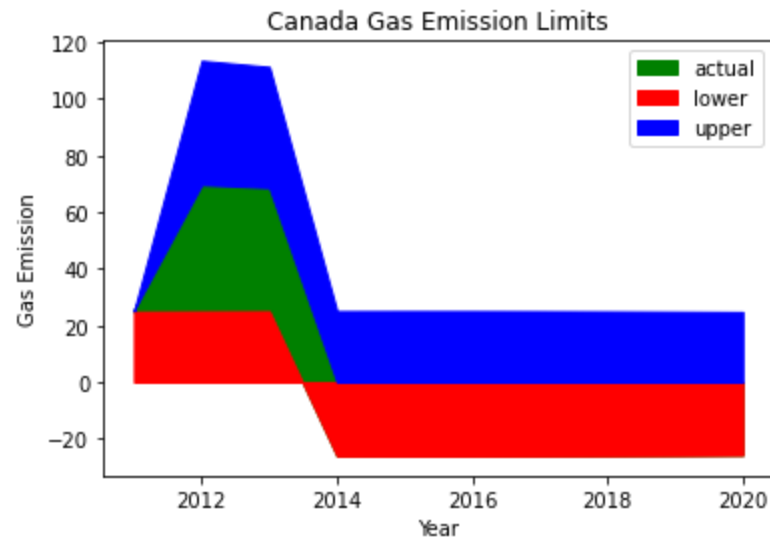
METHODOLOGY

1. Using python's fetch library, fetch the climate change dataset from the World Bank Data API and extract it using the zipfile module.
2. Load the csv file and perform data cleaning using pandas and numpy library.
3. Do clustering on the population growth rate in 2010 and 2019 for all the countries using the sklearn relevant functions.
4. Visualize the clusters using matplotlib .
5. Load the time series data of Canada on percentage gas emission and using sklearn do a curve fitting.
6. Visualize the real curve, fitted curve and error limits using matplotlib's pyplot function.

FINDINGS

Agglomerative Clustering of % Forest Area 2010 and 2019

Agglomerative Clustering of % Gas Emission

Agglomerative Clustering of % Forest Area 2010 and 2019

Canada % Gas Emission

Canada Gas Emission Limits

CONCLUSION

The percentage of greenhouse gas emissions of different countries varies from one country to another. The values for Canada vary greatly from that of China. This can be attributed to the exponential population growth rate in China and industrialization toll.