CLUSTERING & FITTING
ADS 2

By

Ehime Aisagbonhi

MSc-Data Science

21031782

19/01/2023

University of Hertfordshire

1. ABSTRACT

Two key data analysis methods that can be utilised to learn more about the employment rate from the World Bank data are clustering and fitting. Fitting is a way of supervised learning that can be used to determine the parameters of a model that best reflect the data, whereas clustering is a form of unsupervised learning that can be used to group similar observations together.

Clustering can be used to group comparable observations in the case of employment rate data based on variables like age, education, or industry. K-means clustering, for instance, might be used to put people with comparable employment rates and demographic traits together. By doing so, it may be possible to spot patterns and trends in the data that are not immediately obvious.

Fitting, on the other hand, can be used to estimate the parameters of a model that best describe the relationship between employment rate and other factors such as GDP or education level.

2. INTRODUCTION

Overall, clustering and fitting are effective methods for extracting information from unemployment data and using that information to guide policy decisions. These methods can be combined with other data science methods like machine learning and data visualisation to provide a thorough knowledge of the data and spot patterns and trends that might not be immediately obvious..

3. METHODOLOGY

The collection and cleaning of the unemployment data is the first phase. This might entail addressing outliers, eliminating duplicate or missing values, and changing variables. examining data The data should then be investigated to understand its distributions and properties. To find patterns and trends in the data, this could entail making visualisations like scatter plots. The following stage is to combine comparable observations using clustering algorithms.

4 . RESULTS

The first task was to write the Function for reading using pandas and loading the data in the file. And viewing the data information to be analyzed and used.Respresenting a group of data from the list of countries Specified shows the GDP rate for the respective countries and also the unemployment rate of total females as a percentage.

The dataset grouped to be used for this findings

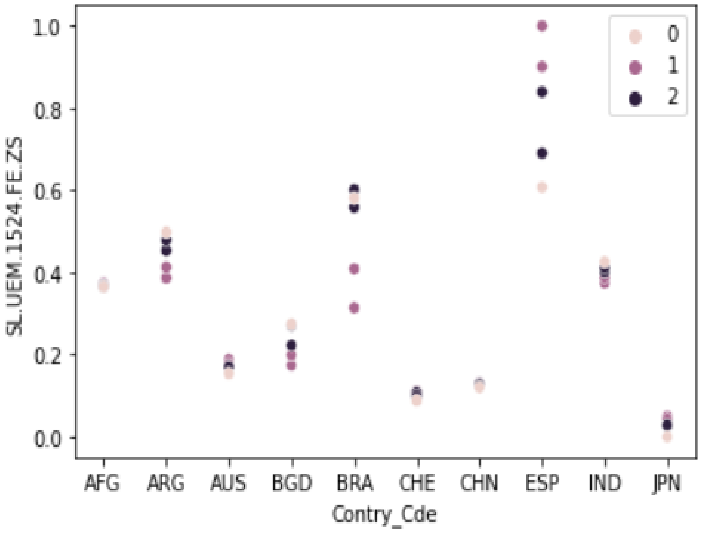
1-Unemployment of total female as a percentage

2-Current GDP in USD

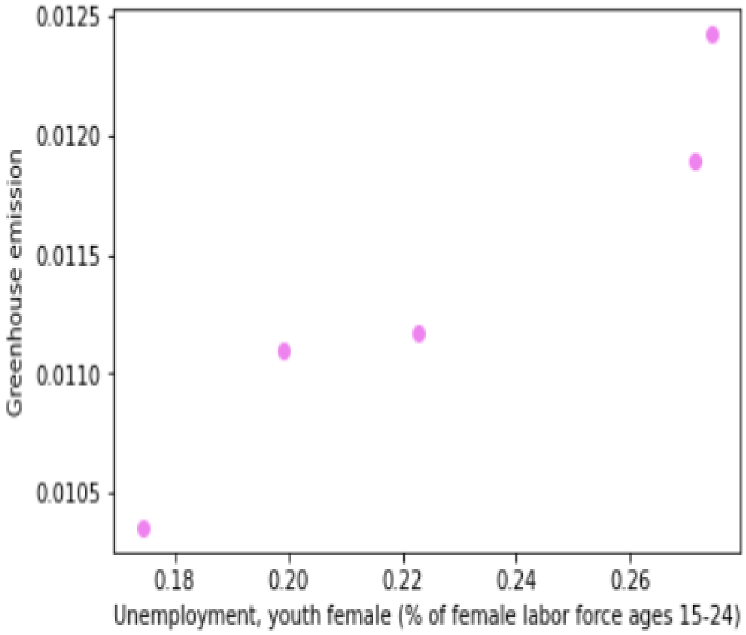
3-Greenhouse gas rate of emission

4-Electricity production from renewable sources, excluding hydroelectric (kWh)

4.1 Clustering the dataset on Unemployment, youth female (% of female labor force ages 15-24) The plot below shows the rate of unemployment from the list of countries extracted from our data set. This data shows the age range of unemployed youths on the y-axis of the plot while the list of countries on the X-axis. With this plot we can deduce that ESP(Spain has the highest level of older unemployed females. Looking at the plot we can also notice that Japan has the lowest female youth unemployment.

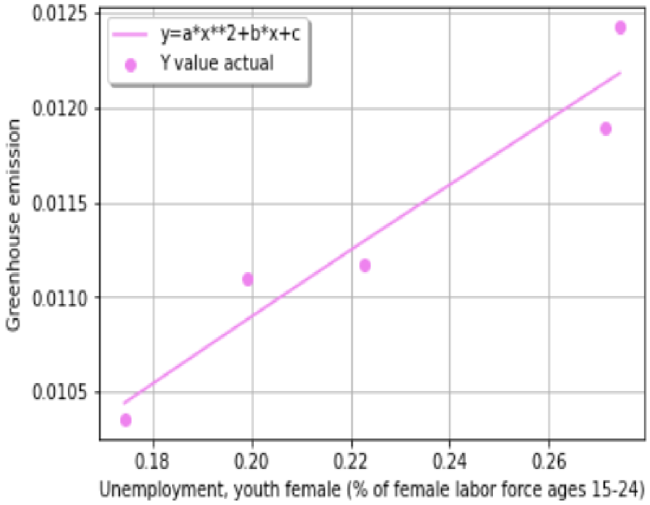


Visualisation between Unemployment, youth female (% of female labor force ages 15-24) and greenhouse emission.It is important to note that the relationship between these three factors is complex and may be influenced by a variety of economic, social, and political factors, so any visualization should be accompanied by a thorough analysis and interpretation of the data.The plot represents the percentage of the female labor force ages 15-24 years showing a high rate of older female youth on the high side.

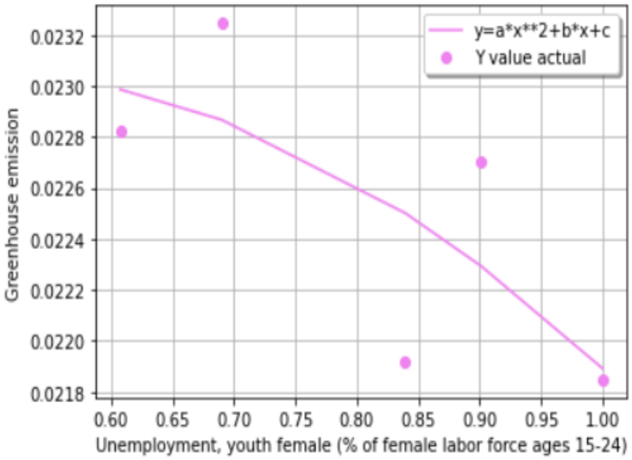


The plots below describe the fitting for each of the countries identified as having the highest, medium, and the lowest rate of unemployment on youth females.

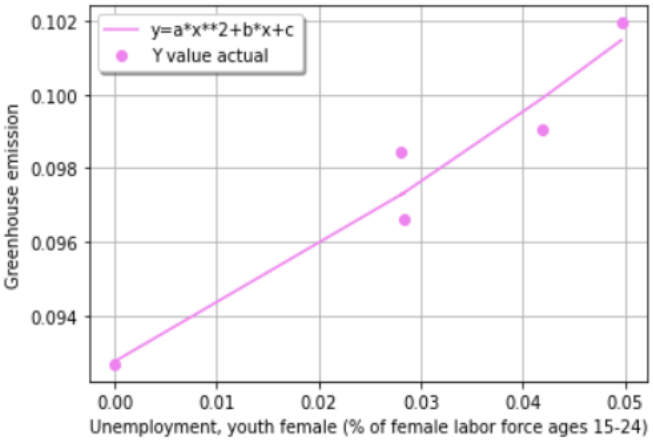
curve_fit implementation for the country Bangladesh which has a medium Unemployment, youth female (%)



#curve_fit implementation for the country Spain which has a high Unemployment, youth female (%)



curve_fit implementation for the country Japan which has a low Unemployment, youth female (%)



5. **DISCUSSIONS**

COMPARISON ANALYSIS

The clustering has been done on the basis of Unemployment, youth female (%) which has given us three segments which are High, Low, and Medium. For the country with High Unemployment of youth female (%), the relationship between greenhouse gas emission and Unemployment, youth female (%) is indirect. For the country with Medium and Low Unemployment of youth female (%), the relationship between greenhouse gas emission and Unemployment, youth female (%) is direct.

6**. CONCLUSION**

In order to analyse the data, K-means clustering was employed. It is a potent and popular method for perusing and comprehending sizable datasets. It is a straightforward method that may be used to find links and patterns in data as well as to cluster together similar data points. The fact that K-means clustering is simple to comprehend, use, and analyse is its main benefit. The algorithm is constrained by the assumption of spherical clusters and the requirement to predetermine the number of clusters, for example. Furthermore, K-means can be sensitive to the first beginning points, therefore it might take several runs with various initial configurations to get reliable results. Despite these drawbacks, K-means clustering is still a crucial instrument in determining and comparing large data sets.

7. REFERENCE

***** *https://www.kdnuggets.com/2019/09/hierarchical-clustering.html*

•*https://www.analyticsvidhya.com/blog/2019/05/beginners-guide-hierarchical-clustering/*

•*https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_clustering_algorithms_hierarchical.htm*

•*https://www.geeksforgeeks.org/hierarchical-clustering-in-data-mining/*