Data Science Certification Course - Python

Certification Project

Certification Project

New DG Food Agro are a multinational exporter of various grains from India since nearly 130 years. But their main product of exporting since early 1980s has been Wheat. They export wheat to countries like America, Afghanistan, Australia etc. They started seeing varying exports of sales year on year for various countries. The reason that was theorized by them had a lot of natural causes like floods, country growth, population explosion etc. Now they need to decide which countries fall in the same range of export and which don't. They also need to know which countries export is low and can be improved and which countries are performing very well across the years.

The data provided right now is across 18 years. What they need is a repeatable solution which won't get affected no matter how much data is added across time and that they should be able to explain the data across years in less number of variables.

Objective: Our objective is to cluster the countries based on various sales data provided to us across years. We have to apply an unsupervised learning technique like K means or Hierarchical clustering so as to get the final solution. But before that we have to bring the exports (in tons) of all countries down to same scale across years. Plus, as this solution needs to be repeatable we will have to do PCA so as to get the principal components which explain max variance.

Implementation:

1) Read the data file and check for any missing values
2) Change the headers to country and year accordingly.
3) Cleanse the data if required and remove null or blank values
4) After the EDA part is done, try to think which algorithm should be applied here.
5) As we need to make this across years we need to apply PCA first.
6) Apply PCA on the dataset and find the number of principal components which explain nearly all the variance.
7) Plot elbow chart or scree plot to find out optimal number of clusters.
8) Then try to apply K means, Hierarchical clustering and showcase the results.
9) You can either choose to group the countries based on years of data or using the principal components.
10) Then see which countries are consistent and which are largest importers of the good based on scale and position of cluster.