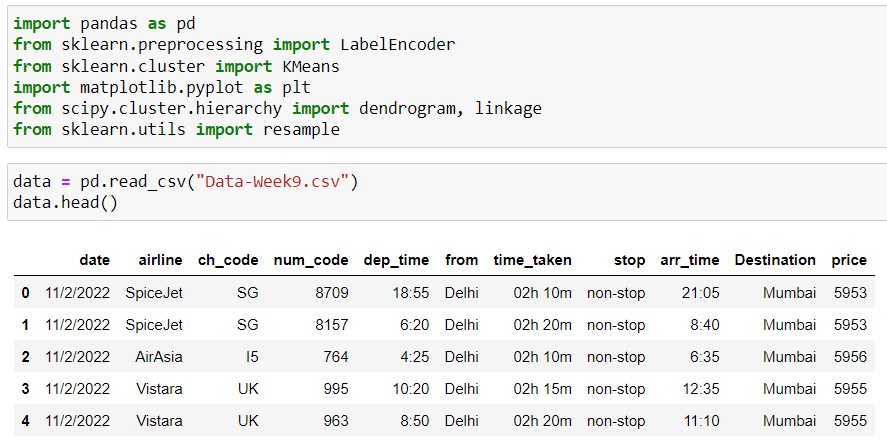
**Name:** Aravind Kumar Kaspe **Banner ID:** 001291145

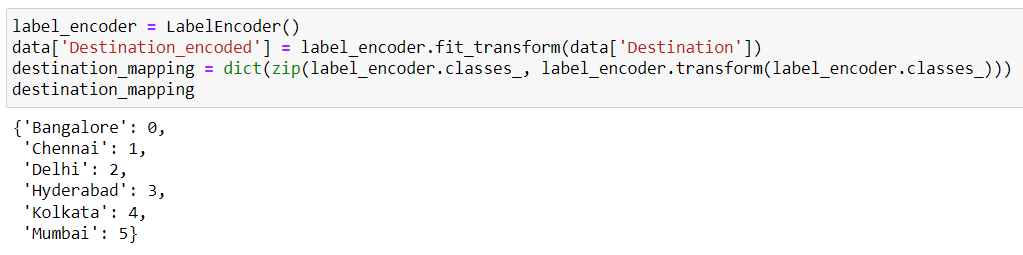
**ASSIGNMENT 6**

**CSCI 5930 – Homework 6: K-Mean**

1. **For data of week 9 flight ticket prices prediction, copy the code and results.**

****

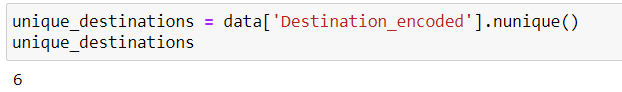
**a) Use the below link to convert “Destination” string attribute to numeric features.**[**https://www.geeksforgeeks.org/how-to-convert-categorical-string-data-into-numeric-in-python/**](https://www.geeksforgeeks.org/how-to-convert-categorical-string-data-into-numeric-in-python/)

****

**b) X1: Destination (after convert), X2: price**

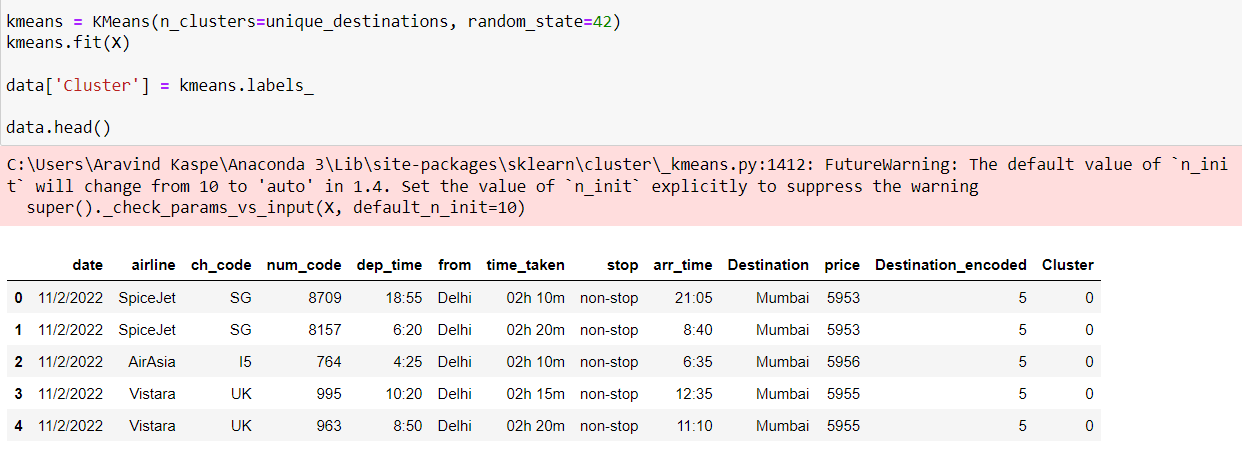
****

**c) Find the number of destinations from the Excel file for k. What is k?**

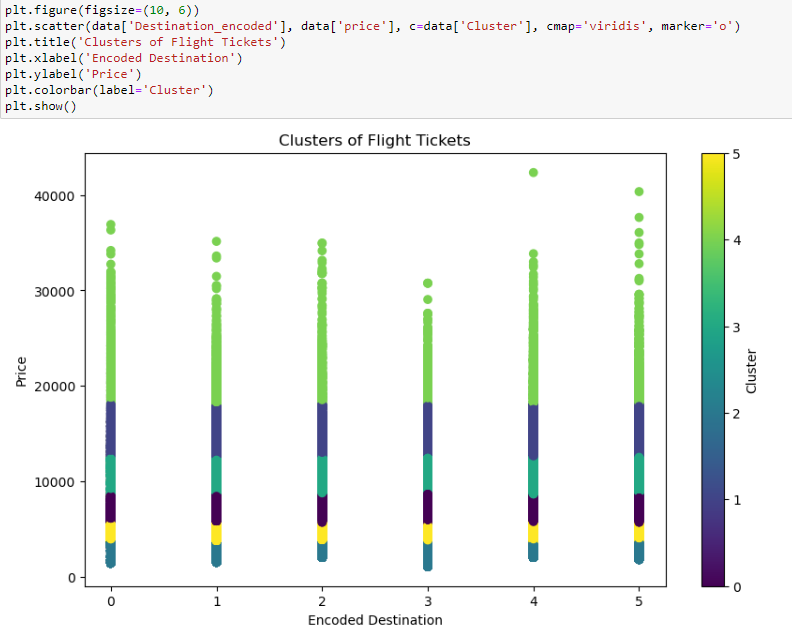
****

The dataset contains 6 unique destinations. Therefore, we'll use k =6 for the k-means clustering.

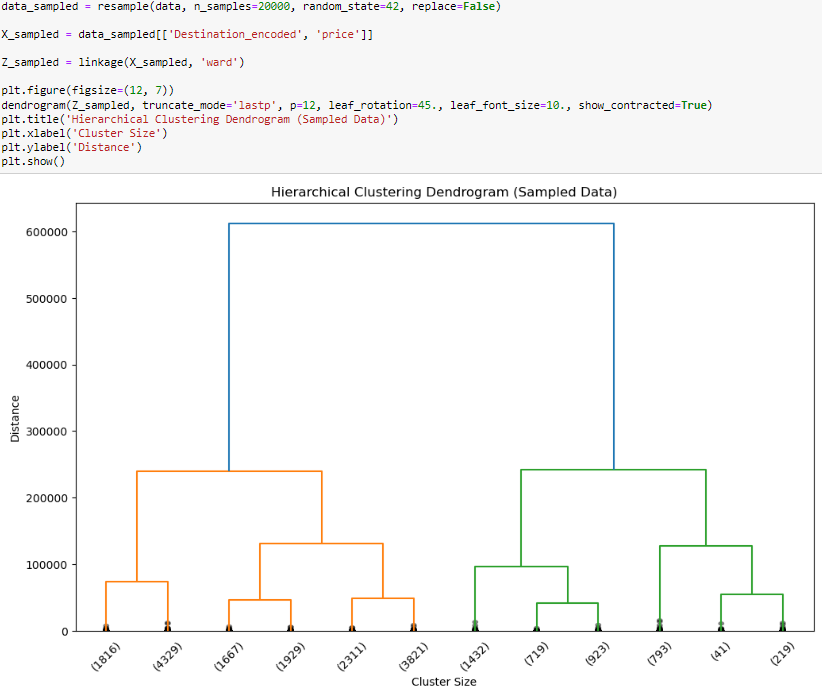
**d) Apply k-mean clustering to predict X variables.**

****

**e) Make the graph.**

****

**f) Build hierarchical clusters.**

****

**g) Interpret the results.**

The dendrogram from hierarchical clustering reveals a potential structure within the flight ticket pricing data where a substantial dissimilarity is observed, suggesting two main clusters. The vertical distances in the dendrogram reflect the dissimilarities between data points, with a particularly prominent division indicating that the dataset may naturally divide into two significantly distinct groups. However, this is a high-level view that could be refined by choosing a different cut-off value for the dendrogram, which would result in a greater number of smaller clusters.

On the other hand, the scatter plot from k-means clustering presents a more nuanced breakdown, delineating six distinct clusters based on ticket prices and encoded destinations. Each cluster color represents a grouping as determined by the k-means algorithm, with vertical spread within clusters indicating the price range variability for tickets associated with particular destinations. Notably, some destinations, marked as 0 and 4, exhibit a substantial spread in ticket prices, reaching towards the higher end of the price spectrum, signifying a broader variance in pricing. Similarly, destinations marked as 2 and 5 show a diverse range of prices, though not stretching to the upper extremes seen in the former group.

Together, the dendrogram and the k-means scatter plot offer complementary perspectives. The dendrogram suggests a broad classification into two main clusters, while the scatter plot provides a detailed subdivision into six clusters, giving more granular insight into the pricing distribution relative to the destinations. The methodology chosen for clustering thus depends on the analytical goals, whether seeking a general overview or detailed segmentation of the data.