**K-MEANS PYTHON APPLICATION**

**EXECUTIVE SUMMARY**

**PROJECT OVERVIEW**

In this project, we developed a K-means clustering application using a free dataset from the AWS Registry of open data. The data set is labeled us\_daily.csv from a harmonized collection of data of COVID-19 reported cases (COVID-19 Harmonized Dataset) managed by Talend/ Stitch. It contains statistics and diverse metrics across multiple regions from the United States. The purpose of this project is to cluster data using the K-Means algorithm and give insights into the results of the clustering.

**DATA OVERVIEW**

**Data Selection:**

The dataset us\_daily.csv was selected from a harmonized collection of reported COVID-19 cases (COVID-19 Harmonized Dataset) managed by Talend/ Stitch from the Registry of Open Data AWS. It includes metrics such as positive and negative test increases, death increases, and hospitalized increases which make it suitable for clustering analysis to identify patterns and trends.

**The Application Development:**

* The dataset was uploaded using pandas into PyCharm.
* The features relevant to the clustering were selected i.e. positiveIncrease, negativeIncrease, deathIncrease, and hospitalizedIncrease.
* Data cleaning was done by handling missing values by filling them with the mean of the respective columns.
* The K-Means algorithm was implemented to cluster the data into 3.
* In the original dataset cluster labels were added to indicate where each data point belongs.
* For the visualization, a 2D scatter plot was used using different colors to represent different clusters.

**The Environment:**

* PyCharm was used for the execution of the Python code.
* The libraries used are Pandas for data handling, Scikit-learn for machine learning, and Matplotlib for visualization.

**UNIT TESTING**

This is to ensure that the application performed as expected. The reasons were:

* To verify that the dataset was loaded correctly.
* To ensure that the missing values were handled properly.
* To confirm that the K-Means clustering was performed correctly with 3 clusters.
* To check that the cluster labels were added to the dataset.

**CHALLENGES**

* Due to missing values, the data cleaning process was time-consuming.
* It was managed by filling in the Mean of respective columns.
* Features relevant to clustering were chosen to represent the COVID-19 analysis.
* There was repetitive testing of the clusters and visual inspection of clustering results.

**KEY FINDINGS AND INSIGHTS**

* During clustering distinct groups of Covid-19 data were shown in the different patterns and trends revealing the spread and impact of the virus. i.e. daily counts or rates, population movement, etc.
* Actionable insights from the mean values of each group indicated by the cluster centers helped in decision-making.
* The visuals showed a clear understanding of how different states or periods were grouped on the selection feature.

**CONCLUSION**

The K-Means Clustering application successfully clustered the Covid-19 data into meaningful groups, thus giving insightful information on the patterns and trends of the pandemic. The project has shown that the K-Means algorithm is effective in clustering large datasets and pointed out the importance of data preprocessing and feature selection in achieving meaningful results.

**Reference:**

*COVID-19 Harmonized Data*. (n.d.). [Data set]. <https://registry.opendata.aws/talend-covid19/>

GeeksforGeeks. (2022, July 8). *K means Clustering - Introduction*. GeeksforGeeks. <https://www.geeksforgeeks.org/k-means-clustering-introduction/>