**Optimizing Smart City Public Transportation Through Data and Technology**

**Assignment 3**

**DS-670-HYB2-23WNTR**

**Capstone: Big Data & Bus Analysis**

**Reda Mastouri**

**19/12/2023**

**Team Members:**

Kethana devi kadiyala

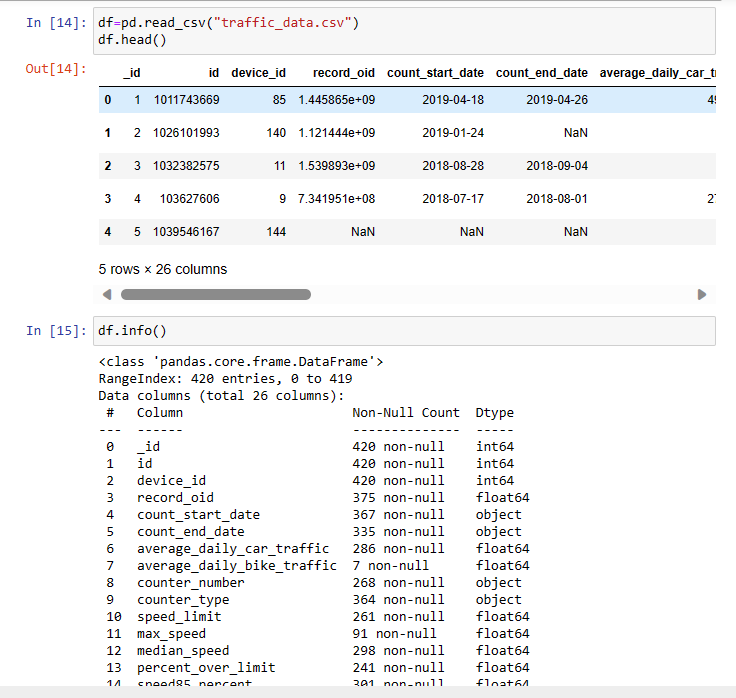
Venkata Sai ram Rohith Mechineni

Deva Sai Vikas Vakkalagadda

**Report on Clustering and Analysis**

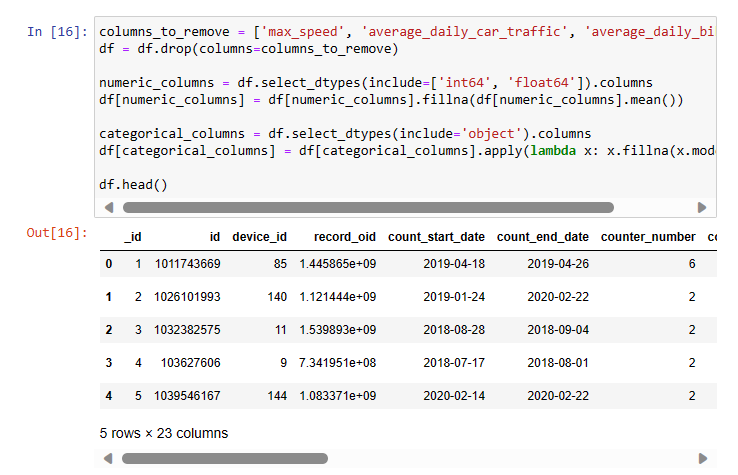
Dataset Overview:

* **Dataset Information:**
  + RangeIndex: 420 entries, 0 to 419
  + Data columns (total 24 columns)
  + Data types: int64, float64, object
  + Memory usage: 77.2+ KB

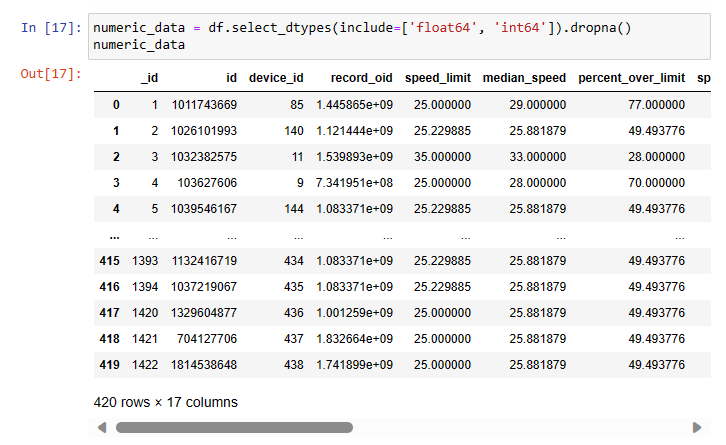


Data Preprocessing:

1. **Columns Removed:**
   * The columns 'max\_speed', 'average\_daily\_car\_traffic', 'average\_daily\_bike\_traffic' were removed to focus on relevant features for clustering.

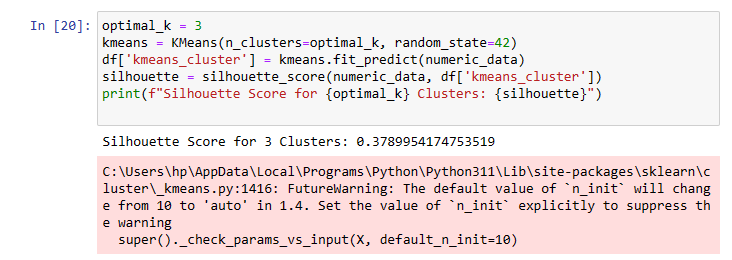


1. **Handling Missing Values:**
   * Null values in numeric columns were filled with the mean to maintain data integrity.
   * Null values in categorical columns were filled with the mode for a representative replacement.



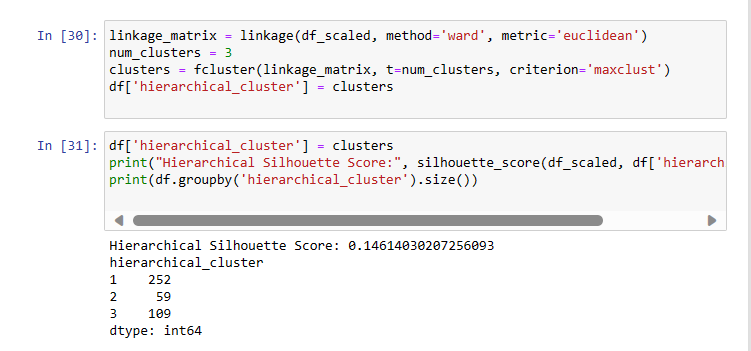
K-Means Clustering:

1. **Determination of Optimal K:**
   * Utilized the Elbow Method to find the optimal number of clusters (K=3).
   * Identified a point where additional clusters do not significantly reduce inertia.
2. **K-Means Clustering:**
   * Applied K-Means clustering with K=3 to group data points.
   * Silhouette Score for K-Means: 0.379, indicating a reasonable separation between clusters.



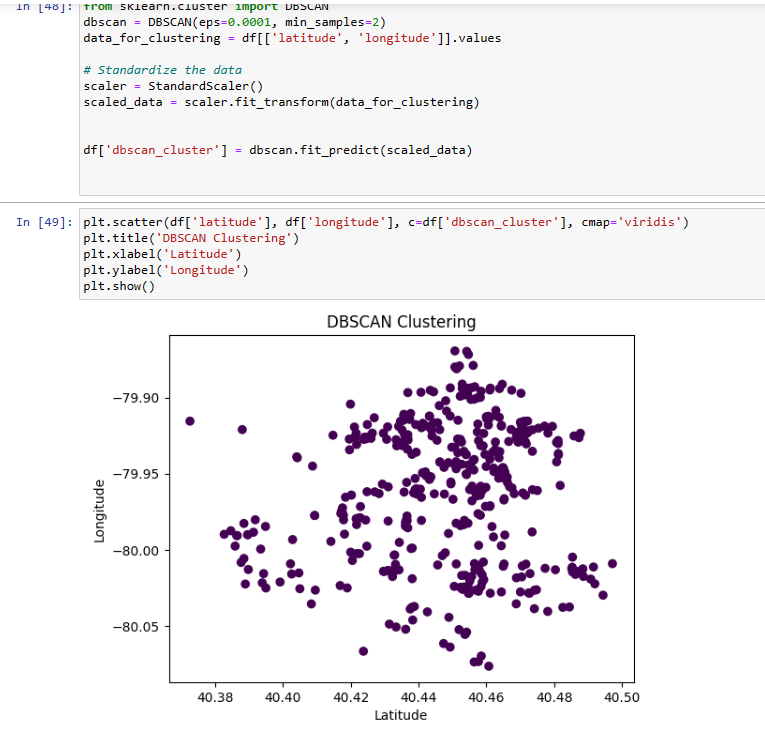
Hierarchical Clustering:

1. **Dendrogram Analysis:**
   * Constructed a dendrogram to visualize the hierarchical relationships between data points.
   * Identified three clusters based on the dendrogram's structure.
2. **Hierarchical Clustering:**
   * Implemented hierarchical clustering with 3 clusters.
   * Hierarchical Silhouette Score: 0.146, suggesting a moderate degree of separation.
   * Cluster distribution: Cluster 1 (252), Cluster 2 (59), Cluster 3 (109).

­

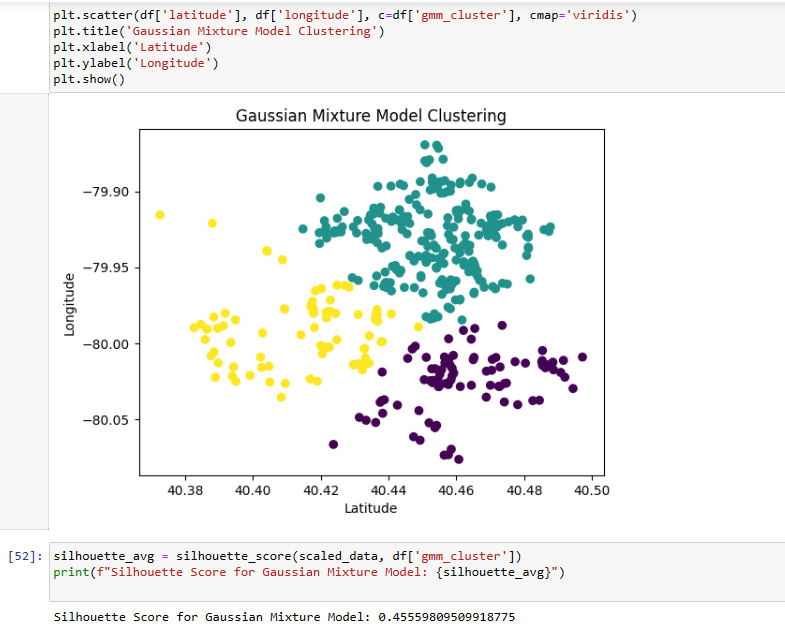
DBSCAN:

1. **DBSCAN Clustering:**
   * Applied DBSCAN with specified parameters to detect dense regions.
   * Visualized clusters using a scatter plot.
   * **Issue:** DBSCAN assigned only one label, indicating potential challenges in identifying distinct clusters.



Gaussian Mixture Model (GMM):

1. **GMM Clustering:**
   * Applied Gaussian Mixture Model (GMM) with 3 components.
   * Visualized clusters using a scatter plot.
   * Silhouette Score for GMM: 0.456, indicating a good separation between clusters.



Conclusion:

* **K-Means and GMM:**
  + Both K-Means and GMM demonstrated meaningful cluster separation with silhouette scores, suggesting robust results.
* **Hierarchical Clustering:**
  + Hierarchical clustering showed a lower silhouette score, indicating less distinct separation compared to K-Means and GMM.
* **DBSCAN:**
  + DBSCAN faced issues, possibly due to the nature of the data or parameter settings. Further investigation is needed.

Recommendations:

* **Optimal Clustering Method:**
  + K-Means and GMM appear suitable for this dataset. Consider the application requirements and the characteristics of the clusters.
* **DBSCAN Adjustment:**
  + Revisit DBSCAN parameters for better cluster identification.
* **Further Analysis:**
  + Explore additional features or advanced techniques for improved clustering performance.