**![Logo, icon

Description automatically generated]()**Logo, company name

Description automatically generated

**Data Warehouse and Data Mining**

**(CS-344)**

**Open-Ended-Lab**

**Name:**

Muhammad Razi-Ur-Rehman

Zaid Ali Chandio

**Section**

BESE-26C

**Submission Date**

May 23, 2023

**Instructor**

Ma’am Laraib

**Department of Computer Software Engineering**

**Military College of Signals (MCS)**

**National University of Sciences and Technology (NUST)**

**Task:**

The automobile manufacturer is seeking to identify the closest competitors to their newly developed vehicle prototypes before launching the new model. To achieve this, they need to group existing vehicles on the market based on similarities, determine which group is the most similar to the prototypes, and use this information to identify the primary competitors for their new model.

The objective is to utilize clustering techniques to identify clusters of vehicles that possess unique characteristics. This analysis will provide an overview of the current market of vehicles and aid manufacturers in deciding on the development of new models based on the identified distinct clusters.

**Solution:**

For the data cleaning and preparation following steps are taken place:  
**Data Preparation:**

**Finding Wrong Data:**The wrong data is checked as there was date present in the model column. We changed into the null value.

**Finding Null values:**

We find the null value and then drop it using dropna function. Null value was also present in the first column.

**Finding Wrong Data like $null$**

We changed this type of wrong data and converted it into NA value. These are present in almost every column.

**Dropping row**

In this next step, we find out the row containing more than 2 nulls values and drop it. Only the 33-index row was present and it was deleted.

**Filling with Mean**

The columns like “price”, “curb\_wgt” and “mpg” have nearly one or two null values which are filled with their respective means.

**Dropping single unique value column:**

The column partition has only 1 unique value “0”. So, it will not be contributing anything so we should drop it to increase computational performance.

**Changing into respective Datatype:**

In this step, the respective column is changed into its respective data type.

**Filling With Prediction Model:**

The column like “resale” has many null values so we use prediction model Linear Regression. This is a regression model as we want to predict numerical value. To predict this, following steps are taken:

* Label Encoder to encode the string values.
* Normalization
* Splitting into training and testing data
* Model Selection
* Model Prediction

Checking duplicate value

In this step, Duplicated value is checked. As no duplicated values are found, none are removed.

**Feature Selection:**

Using the seaborn library, we have developed the heatmap of the correlation matrix to check the relation between the features. As price and resale are highly corelated, we drop resale.

**Data Modeling:**

As our data is now prepared, we must select the model. We choose the Hierarchy Clustering Agglomerative Clustering for the fitting of this data. First, we need to find the distance using Euclidean distance formula. Finding linkage using scipy library and drawing dendrogram.

**Data Visualization:**

The dendrogram is drawn to show the data. After that in the last we displaced the data in points.

**Model Evaluation:**

As it is unsupervised learning and there are not ground truth labels to evaluate this type of this model.