Name: Revanth Kumar Kanapala

Student ID: 16343832

# ASSIGNMENT-2

A) Look for the missing values in all the columns and either impute them (replace with mean,

median, or mode) or drop them. Justify your action for this task

1. Firstly, we need to read the raw data ,after cleaning the raw data file we can observe that file contains 5847 observations and 14 variables

A screenshot of a computer

Description automatically generated

A screenshot of a white screen

Description automatically generated

1. Secondly, we need to find the missing values in the dataset. We can see the seats column has 38 missing values, Mileage has 2, Engine has 36, power has 36 and New\_price has 5032 missing values.

A screenshot of a computer

Description automatically generated

1. In this dataset some columns are represented as character datatypes in general they are either int or float values. Mean operation can be performed only on numeric datatype. If we try to impute the missing values without removing the unit values from Mileage, Power and Engine Column, our function will consider those columns as character values and try to impute them with mode. Therefore, before handling the missing values I am making sure my data is clean`.

A screenshot of a computer

Description automatically generated

1. Here We need to drop first column and New\_Price column because it has more than 50% missing values and also we need to check that missing values have been correctly imputed and then we need read the cleaned df\_car file to clean.csv , this is how we retrive the clean data.

A screenshot of a computer

Description automatically generated

B) Remove the units from some of the attributes and only keep the numerical values (for

example remove kmpl from “Mileage”, CC from “Engine”, bhp from “Power”, and lakh from

“New\_price”)

1. As we did in the first question, we need to follow the same steps

A screenshot of a computer

Description automatically generated

C) Changing the categorical variables (“Fuel\_Type” and “Transmission”) into numerical one hot encoded values(0,1).

A screenshot of a computer

Description automatically generated

Here the file will be saved as cars\_encoded.csv in the ‘results’ folder.

A screenshot of a data

Description automatically generated

D)Creating a new feature variable called “Current\_Age” by using the mutate function in R. Current\_Age = Year – Current Year.

A screenshot of a computer

Description automatically generated

The updated file with Current Age will be generated as “car\_results.csv” in results

A screenshot of a table

Description automatically generated

**Conclusion:**

A Used Cars dataset with 5847 observations and 14 attributes, then we need to retrieve clean data by performing cleaning ,preprocessing steps such as imputing the missing values with mean,mode,median or dropping them, removing the units from attributes like Mileage, Power and Engine. Created a new column “Current Age”.