Lavanya Ghanathey 16342393

**Assignment-2**

**Data Pre-Processing Steps for Used Cars Dataset**

1. Imputing the missing values in all columns with either mean, median, mode or dropping them.

* Loading the dataset and understanding the data and its type. The used cars dataset has 5847 observations and 14 variables.

A screenshot of a computer

Description automatically generated

* Finding 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

* Before imputing the missing values, we have to perform a proper cleaning process. 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 appropriate.

A screenshot of a computer program

Description automatically generated

* I am dropping the first column and the New\_Price column because the first column is just giving us the row number and the new\_price column has more than 50% of missing values. If we try to consider this column it will not add any weightage to our analysis so it is ideal to drop the New\_Price column from our dataset which has 5032 missing values out of 5847 observations.A screenshot of a computer program

  Description automatically generated
* As all the missing values are of numeric type, I am replacing the mean values with the Mean values and checking whether the missing values are handled or not.

A screenshot of a computer

Description automatically generated

* The cleaned dataset is stored in a new file called “clean\_data\_cars.csv”

A table with numbers and text

Description automatically generated

1. Removing the units from the attributes like Mileage, Power, Engine and New\_Price and keeping the only numeric part for analysis.

A screenshot of a computer

Description automatically generated

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

A screenshot of a computer

Description automatically generated

* Saving the results in a new file called “ c\_results\_cars\_encoded.csv”

A table with numbers and numbers

Description automatically generated

1. 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 final results are stored in a new file called “d\_results\_cars.csv”

A screenshot of a computer

Description automatically generated

***Conclusion:***

I am considering a Used Cars dataset which has 5847 observations and 14 attributes. After performing the cleaning and preprocessing steps such as imputing the missing values with mean, mode, median or dropping them, removing the units from the attributes like Mileage, Power and Engine, performing one-hot encoding for the attributes like Fuel\_Type and Transmission and using the mutate function in R to add a new column called Current\_Age of the car and saving the results in a new dataset called “d\_results\_cars.csv” which has 5847 observations and 13 attribute values.