**Assignment 2**

**Venkatesh Thirunagiri - 16342355**

1. **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.**

A screenshot of a computer

Description automatically generated

**Mileage**: there are only two missing values, we’re replacing them with the median value of the existing data.

**Engine, Power, Seats**: the number of missing values in these columns is relatively small compared to the total dataset, dropping rows with missing values might not significantly affect the analysis. We’re replacing these missing values with the median values of their respective columns.

New\_**Price**: This column has a significant number of missing values (5032 out of 6019 entries). Dropping this column, as it's missing in the majority of entries.

A screenshot of a computer

Description automatically generated

1. **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”).**

A screenshot of a computer

Description automatically generated

Removing units from text using str.extract() method in pandas with the help of regular expressions

**C)** **Change the categorical variables (“Fuel\_Type” and “Transmission”) into numerical one hot encoded value. (4 points).**

using the get\_dummies() function in pandas to convert categorical variables (“Fuel\_Type” and “Transmission”) into numerical one hot encoded value

A screenshot of a computer

Description automatically generated

1. **Create one more feature and add this column to the dataset (you can use mutate function in R for this). For example, you can calculate the current age of the car by subtracting “Year” value from the current year**

**A screenshot of a computer

Description automatically generated**

1. **Perform select, filter, rename, mutate, arrange and summarize with group by operations (or their equivalent operations in python) on this dataset**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**



A screenshot of a computer

Description automatically generated



A screenshot of a computer

Description automatically generated

1. Evaluation of car prices

A graph of a car price

Description automatically generated

A graph of a graph showing the price of cars

Description automatically generated