Ridma Jayawardena

1.

a.)

This dataset contains details of automobiles in terms of specifications, insurance risk ratings, and normalized losses compared to other cars. It has 205 instances (rows) and 26 attributes (columns), including 15 continuous, 1 integer, and 10 nominal attributes. Attributes include specifications like make, fuel type, body style, engine size, and price. These are used for tasks like predicting car prices or insurance risk ratings.

Variable	Data type
symboling	Integer, Categorical
normalized-losses	Float , Nominal
make	String, Nominal
fuel-type	String , Nominal
aspiration	String, Nominal
num-of-doors	String, Nominal
body-style	String, Nominal
drive-wheels	String, Nominal
engine-location	String, Nominal
wheel-base	Float, Continuous
length	Float, Continuous
width	Float, Continuous
height	Float, Continuous
curb-weight	Float, Continuous
engine-type	String, Nominal
num-of-cylinders	String, Nominal
engine-size	Integer , Continuous
fuel-system	String, Nominal
bore	Float, Continuous
stroke	Float, Continuous
compression-ratio	Float, Continuous
horsepower	Float, Continuous
peak-rpm	Integer, Continuous
city-mpg	Integer, Continuous
highway-mpg	Integer, Continuous
price	Float, Continuous

g.)

Missing values are safe to remove when the rows with missing values are sparse and do not represent a significant portion of the dataset and if the missing data is not critical to the analysis.

Missing values are not safe to remove when a significant portion of the data is lost, leading to a potential bias, or if the missing values are in key columns critical for prediction or analysis.

When we consider this dataset, it is not recommended to remove the missing values because out of 205 rows, 46 rows were removed, which is approximately 22.4% of the data. This is a significant portion. So, removing them could lead to potential bias or loss of valuable information.

i.)

One Hot Encoding is a technique for transforming categorical variables into a binary representation. Each category in the original variable is given a new binary column (0s and 1s). Every category in the original column is shown as a distinct column, with a value of 0 denoting its absence and a value of 1 denoting its existence.

Importances of one hot encoding is as follows:

- It allows the use of categorical variables in models that require numerical input.
- It can improve model performance by providing more information to the model about the categorical variable.
- One Hot Encoding eliminates the risk of mistakenly interpreting Male = 0, Female = 1 as a ranking and leading to biased predictions.
- For high-cardinality categorical data (many unique values), one-hot encoding can create sparse matrices where only one column is active (1) at a time.

2.