**Principles of Data Science**

**Saikrishna Boddula**

**Justification For Question A**

In our analysis, we encountered missing values across several columns, including Mileage, Engine, and Fuel\_Type. Instead of discarding these rows, we opted to **impute missing values** using different strategies based on the type and distribution of each attribute. Our approach maintains the integrity of the dataset by preserving the information of rows that would otherwise be lost if we simply dropped rows with missing data. Here’s how we handled each:

1. **Mileage**: We replaced missing values with the **mean mileage**. Since mileage is a continuous variable and its distribution was approximately normal, using the mean provided a reasonable estimate that minimizes bias and preserves the overall distribution of this feature.
2. **Engine**: Missing values in the engine size were imputed using the **median**. This approach was chosen because the distribution of engine sizes was slightly skewed. Using the median ensured that extreme values did not disproportionately affect our imputed values, which helps maintain accurate representation within this attribute.
3. **Fuel\_Type**: We filled missing values with the **mode** since this is a categorical variable. By filling in the most common fuel type, we minimized the impact of missing values on our analysis without distorting the dataset’s categorical balance.

By using imputation instead of deletion, we retained as much information as possible, which will contribute to a more robust analysis and predictive performance in subsequent modeling steps.