CarDekho Dataset Analysis Report

1. Dataset Overview

- The CarDekho dataset consists of **8,148 rows** and **13 columns**. The dataset includes information on various vehicles, including their specifications and market attributes such as price, brand, transmission type, fuel type, and owner details.

```
Data columns (total 13 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Name           8148 non-null   object
 1   year           8148 non-null   int64
 2   selling_price  8148 non-null   int64
 3   km_driven      8148 non-null   int64
 4   fuel           8148 non-null   object
 5   seller_type    8148 non-null   object
 6   transmission   8148 non-null   object
 7   owner          8148 non-null   object
 8   mileage        7927 non-null   object
 9   engine         7927 non-null   object
 10  max_power      7933 non-null   object
 11  torque         7926 non-null   object
 12  seats          7927 non-null   float64
dtypes: float64(1), int64(3), object(9)
```

2. Data Cleaning and Preprocessing

To ensure the dataset is clean and suitable for analysis, the following steps were performed:

- **Dropping Null and Duplicate Values**: Null and duplicate values were removed to eliminate inconsistencies and avoid biases in the analysis.
- **Data Type Conversion**: Columns such as mileage, torque, engine, and max power were converted to numerical (float) values to facilitate further analysis.
- **Creation of New Features**:
    - **Vehicle Age**: This feature was created by subtracting the year of manufacture from the current year.
    - **Mileage Condition**: Vehicles were categorized as high mileage, normal mileage, and low mileage based on their recorded mileage to assess their usage levels.

	year	selling_price	km_driven	mileage	engine	max_power	torque	seats
count	6260.000000	6.260000e+03	6260.000000	6260.000000	6260.000000	6260.000000	6260.000000	6260.00000
mean	2013.572684	4.450674e+05	69524.590575	19.765917	1375.531789	83.686846	151.562685	5.38099
std	3.898454	2.534658e+05	39686.904876	3.922068	445.078133	25.613021	78.814351	0.93999
min	1994.000000	2.999900e+04	1.000000	0.000000	624.000000	32.800000	4.800000	2.00000
25%	2011.000000	2.500000e+05	39000.000000	17.100000	1196.000000	67.100000	91.000000	5.00000
50%	2014.000000	4.000000e+05	68000.000000	19.700000	1248.000000	81.800000	140.000000	5.00000
75%	2017.000000	6.000000e+05	100000.000000	22.702500	1497.000000	97.700000	200.000000	5.00000
max	2023.000000	1.250000e+06	190000.000000	33.440000	3498.000000	272.000000	789.000000	14.00000

3. Exploratory Data Analysis (EDA)

3.1. Univariate Analysis

This analysis focused on understanding the distribution and frequency of individual columns, particularly those with categorical data.

1. **Vehicle Name**:
    - There are **1,809 unique vehicle names** in the dataset.
    - The most frequent vehicle names include:
        - **Maruti Swift Dzire VDI**: 118 counts
        - **Maruti Alto 800 LXI**: 76 counts
        - **Maruti Alto LXi**: 69 counts
    - These results indicate that Maruti is a dominant brand in the dataset.
2. **Fuel Type**:
    - The dataset contains vehicles with different fuel types:
        - **Diesel**: 3,250 vehicles
        - **Petrol**: 2,926 vehicles
        - **CNG**: 50 vehicles
        - **LPG**: 34 vehicles
    - Diesel and petrol vehicles dominate the dataset, suggesting that these fuel types are most common in the market.
3. **Seller Type**:
    - **Individual sellers** account for the majority of listings (**5,686**), followed by **dealers** (**548**), and **Trustmark dealers** (**26**).
    - This indicates that the platform is mainly used by individual sellers rather than professional dealers.
4. **Transmission**:

- Vehicles with **manual transmission** are significantly more common (**5,883**), compared to those with **automatic transmission** (**377**).
- This finding highlights a preference for manual vehicles in the dataset.

5. **Owner Type**:
   - **First Owner** vehicles are the most common, making up **3,869** entries, while **Second Owner** vehicles account for **1,776** entries.
   - **Third Owner** and **Fourth & Above Owner** vehicles are less frequent, with **471** and **144** entries respectively.
   - This shows that the majority of the vehicles listed are still in their first or second ownership.
6. **Company**:
   - The dataset contains vehicles from various companies:
     - **Maruti**: 2,075 vehicles
     - **Hyundai**: 1,193 vehicles
     - **Mahindra**: 648 vehicles
     - **Tata**: 602 vehicles
   - Maruti, Hyundai, Mahindra, and Tata are the top brands, indicating their strong market presence.
7. **Vehicle Age**:
   - The vehicles were categorized based on age:
     - **Old** vehicles: 6,160 entries
     - **Latest** vehicles: 65 entries
     - **Very old** vehicles: 35 entries
   - Most vehicles fall into the "old" category, indicating a higher prevalence of older vehicles in the dataset.
8. **Mileage Condition**:
   - Vehicles were grouped based on mileage condition:
     - **High mileage**: 3,608 vehicles
     - **Normal mileage**: 2,638 vehicles
     - **Low mileage**: 14 vehicles
   - A majority of vehicles have either high or normal mileage, while only a few have low mileage.

3.2. Bivariate Analysis

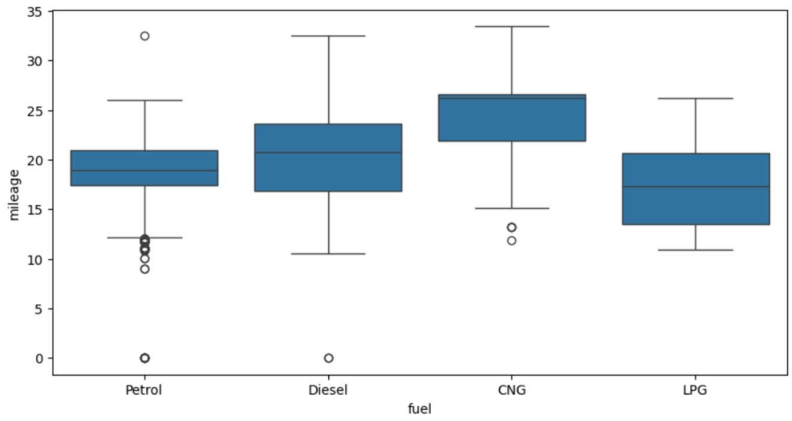
The bivariate analysis aimed to explore relationships between two variables, providing insights into how different features interact with each other:

1. **Price vs. Vehicle Age**:
    - A trend showing that older vehicles tend to have lower prices, while newer vehicles (categorized as "Latest") have higher prices.
    - This relationship aligns with the typical market behavior where depreciation impacts vehicle prices over time.
2. **Transmission Type vs. Fuel Type**:
    - Diesel vehicles are more prevalent in manual transmission, while automatic transmission options are more often associated with petrol.
    - This suggests that manual diesel vehicles are favored in the used car market.
3. **Owner Type vs. Seller Type**:
    - First-owner vehicles are predominantly sold by individual sellers, while multiple-owner vehicles have a balanced distribution among individual sellers and dealers.
    - This observation highlights that first-owner cars might be considered more valuable and are often sold directly by individuals rather than dealerships.

3.3. Visualization

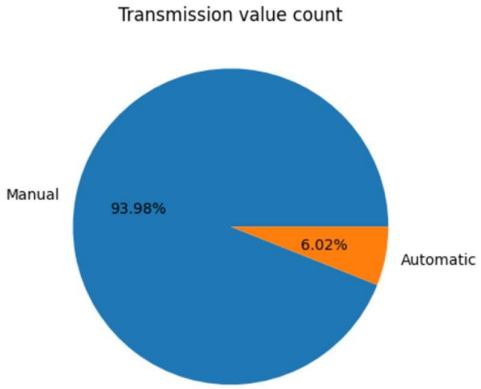
Several visualizations were created to better understand the dataset:

1. **Boxplots**:
    - Boxplots were used to analyze the distribution and outliers in the price and mileage columns, showing how these features vary across different brands and vehicle ages.

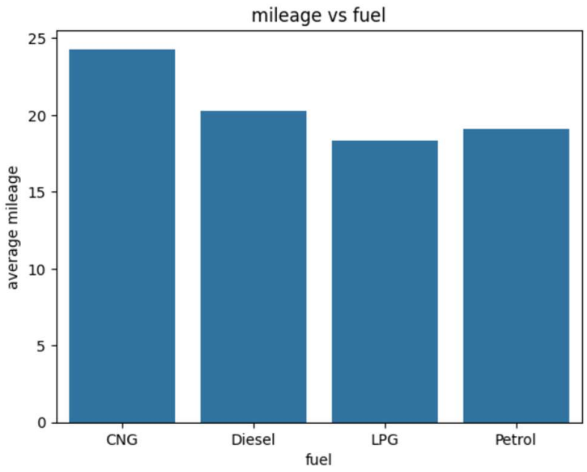


2. **Univariate Plots**:
    - Bar charts were generated for categorical features like fuel type, transmission, owner type, and company to visualize their distribution.
    - These charts clearly showed the dominance of certain categories, such as manual transmission and individual sellers.
    - **Pie Plot**: In this analysis, a pie plot is used to visually represent the proportions of different categories within a feature. It helps to illustrate the distribution of categorical variables, such as fuel type, transmission type, or seller type, showing each category's percentage of the total. By displaying these proportions as slices of a circle, pie plots

offer a clear and intuitive way to understand the composition and dominance of each category in the dataset, making it easier to identify the most and least common groups at a glance.

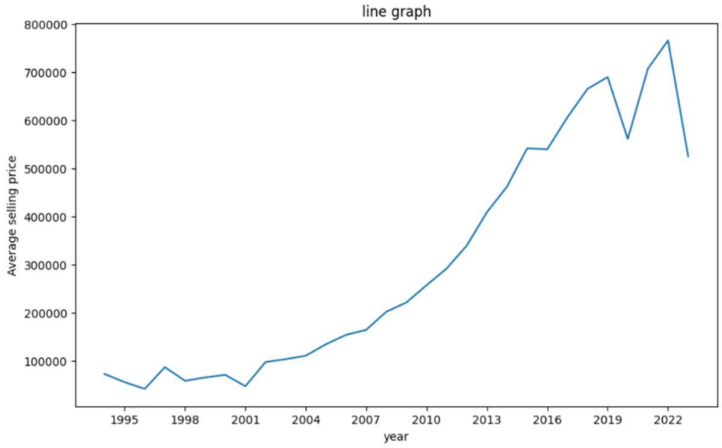


3. **Grouped Analysis**:
    - Using the groupby function, detailed statistics were computed for both numerical and categorical columns, providing summaries like average prices per company or average mileage per vehicle age category.
    - **Bar plot:** A bar plot is used in this analysis to visually represent the distribution of categorical data. It displays data using rectangular bars, where the length of each bar is proportional to the count or value of that category. Bar plots are particularly effective for comparing different categories within a feature, such as the number of cars from each company or the distribution of fuel types. By providing a clear visual comparison, bar plots make it easy to identify the most and least frequent categories, trends, or patterns in the dataset.



    - **Line Plot**: A line plot is used to visualize the relationship between two variables, typically when one is continuous, such as time or age. In this analysis, line plots can be useful for showing how a numerical feature like vehicle age or mileage changes across different observations. Line plots are especially effective for illustrating trends or patterns over a continuous interval, helping to identify increases, decreases, or other

shifts in the data over time. They provide a simple and intuitive way to track changes and identify trends.



---

4. Conclusion and Recommendations

- The CarDekho dataset reveals a strong preference for manual transmission, diesel vehicles, and older cars in the used car market. Maruti emerges as the most dominant brand in terms of the number of listings.
- Based on the analysis, it is recommended to explore price trends further by considering additional factors such as vehicle features, location of sellers, and specific brand popularity.
- Future analysis could also include a predictive model to estimate the price of a vehicle based on its attributes, which could be beneficial for both buyers and sellers on the platform.