Answering an Interview Question: The EDA Project

Thank you for the opportunity to discuss my work. I'd like to walk you through a project I completed where I performed an Exploratory Data Analysis (EDA) on a dataset of used cars.

---

1. Problem Statement & Project Goal

The core problem I aimed to solve was to gain a deep understanding of the used car market from a data-driven perspective. My goal was to explore a dataset from the Car Wala website containing information on used cars in Hyderabad, India. The primary questions I wanted to answer were:

- What are the key factors influencing the price of a used car?
- What is the typical profile of a used car available in the market in terms of mileage, fuel type, and age?
- How do different brands compare in terms of their pricing and market presence?

The insights from this analysis could be invaluable for a car dealership to optimize their pricing strategy or for a potential buyer to make an informed decision.

---

2. Data Cleaning and Preparation

Before any analysis, I started with the essential steps of data cleaning and preprocessing to ensure the data was in a usable format.

- **Initial Inspection**: I began by loading the dataset into a pandas DataFrame and used df.info() and df.describe() to check data types, identify missing values, and get a statistical summary.
- **Handling Irrelevant Data**: I identified and dropped columns that were not relevant to the analysis, such as the title and emi columns, as they contained either unstructured text or a large number of missing values.
- **Feature Engineering**: This was a crucial step. The price and mileage columns were in a string format with units like 'Lakh', 'Crore', and 'km'. I wrote a custom function to clean these columns and convert them into numerical values, which is necessary for any quantitative analysis. For instance, 1.55 Lakh was converted to 1550000. I also extracted the brand and year from the name column, which allowed me to analyze how brand and age influence price.

---

3. Exploratory Data Analysis (EDA)

This is where I applied different analytical techniques to uncover insights. I structured my analysis into three main parts: Univariate, Bivariate, and Multivariate analysis.

Univariate Analysis

This part of the analysis focused on understanding each variable individually.

- **Price Distribution**: I used a **histogram** to visualize the distribution of car prices. This showed that the data was heavily skewed towards lower-priced cars, with a long tail of a few very expensive cars, which is a common pattern in market data.
- **Fuel Type and Brand Counts**: I used **bar charts** to count the occurrences of each fuel type and brand. This revealed that Petrol was the most common fuel type, and Maruti Suzuki was the most frequently listed brand, giving a clear picture of the market composition.

Bivariate Analysis

Here, I explored the relationships between two variables.

- **Mileage vs. Price**: I used a **scatter plot** to examine the relationship between a car's mileage and its price. The plot clearly showed a strong negative correlation, indicating that as a car's mileage increases, its price tends to decrease. This is a fundamental and expected insight.
- **Fuel Type vs. Price**: I created a **box plot** to compare the price distribution across different fuel types. This highlighted significant price differences; for instance, Electric and Hybrid cars had a higher median price than Petrol and Diesel cars.

Multivariate Analysis

This goes a step further by looking at the relationships between three or more variables.

- **Price by Brand and Year**: I would use a **grouping and aggregation** approach to look at the average price of cars, grouped by both brand and year. This would allow me to see not just which brands are generally more expensive but also how a brand's price changes with the age of the car, revealing patterns like depreciation rates for different brands. For instance, a luxury brand like Mercedes-Benz might depreciate differently than a mass-market brand like Hyundai.

---

4. Conclusion and Next Steps

In conclusion, this project provided valuable insights into the used car market. The analysis confirmed that **mileage, fuel type, brand, and age are all significant factors influencing a car's price**. This understanding can be used to inform business decisions.

The next logical step for this project would be to leverage these insights to build a **predictive model**. I would use the cleaned and engineered features to train a machine learning model, such as a **Random Forest Regressor**, to predict the price of a used car based on its attributes. This would transform the descriptive EDA into a powerful predictive tool.