**Domain:**

Automobile Industry

**About:**

There is an automobile company XYZ from India which aspires to enter the Indian used-car market by setting up their company locally to give competition to their counterparts.

**Challenges:**

They want to understand the factors affecting the pricing of cars in the market, since those may be very different from the new car market. Essentially, the company wants to know:

* Which variables are significant in predicting the price of a used car?
* How well those variables describe the price of a car

Based on various market surveys, the consulting firm has gathered a large dataset of different types of used cars across the market.

**What is Expected?**

Being a data analyst, you must come up a first step document that lists output of your exploratory analysis, any issues or problems you may see with data that need follow up, and some basic descriptive analysis that you think highlights important outcomes/findings from the data. Based on your findings, the next level of analysis will be charted out.

Here are some indicative types of analysis you can perform. Please note that this is not an exhaustive list, you may add more

**Data Preparation, Data Manipulation, Visualization and Exploratory Data Analysis and Statistics:**

* Come up with appropriate results and visuals for the following:
  + Which variables are significant in predicting the price of a used car?
  + How well those variables describe the price of a car?
  + Which brands are selling most?
  + Are there specific locations selling more?
  + Which factors are more important in deciding cars' selling price? Ex. kms driven or type of owner or fuel type?
  + Descriptive statistics for both numerical and categorical and draw few insights from them.
  + Perform relevant hypothesis testing
* Prepare the data by handling missing values, outlier analysis, data transformation and normalization.

**Model Building:**

* Build appropriate ML model/s on the data.
* Compare various ML models with appropriate regularization and/or hyper-parameter tuning.
* Evaluate the performance of the models.
* Identify the right metric to evaluate the performance of the model.
* Identify issues and concerns on the given data and suggest the best technique/s to overcome the issues.

**NOTE:** Results and graphs must be backed with appropriate inferences and insights.

.**Data Dictionary:**

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Sales\_ID | Sales ID |
| name | Name of the used car |
| year | Year of the car purchase |
| selling\_price | Current sellling price for used car |
| km\_driven | Total km driven |
| Region | Region where it is used |
| State or Province | State or Province where it is used |
| City | City where it is used |
| fuel | Fuel type |
| seller\_type | Who is selling the car |
| transmission | Transmission type of the car |
| owner | Owner type |
| mileage | Mileage of the car |
| engine | engine power |
| max\_power | max power |
| seats | Number of seats |
| sold | used car sold or not |

**Questions I Want Answered:**

1. How many used cars were sold in a particular year?
2. What is the average price of used cars sold in a particular location?
3. What is the average number of miles on the used cars being sold in a particular location?
4. What is the most common fuel type of the used cars being sold in a particular location?
5. What is the average selling price of used cars with different seat numbers?
6. What is the average trade-in value for used cars in a particular location?
7. What is the most popular model of used car being sold in a particular location?
8. What is the total value for used cars in a particular location?
9. What is the total revenue generated from used car sales in a given year for central region?
10. How many used cars were sold in each city in a given year (top 10 with under 50K miles)?
11. How many used cars were sold in each price range (e.g. under $10,000, $10,000-$20,000, etc.) in a given year?
12. What is the total number of used cars sold by make and model in a given year?
13. What is the total number of used cars sold by fuel type (e.g. gasoline, diesel, electric) in a given year?
14. Find count of cars sold and group by make (find the top 5 cars makes that have been sold)
15. Find total sum of cars sold and group by year in decreasing order, limit 5
16. Find count of cars not sold and group by year and make, order by desc, limit 5
17. Categorize the horsepower as low, average, medium, high in a new column called HP Ratings (case statement)
18. Find top car make, then use case statement to filter for count of cars based on miles driven