Used car sell Analysis

1. Business Understanding

A. Business Overview

Finding a good car can sometimes be a hustle especially for first time car owners. This can be attributed to the fact that it is very easy to get conned into buying a vehicle that does not serve your purpose for the period you intend to own the vehicle. It might also be equally hard to resell your vehicle once you are done with it and you may end up selling it at a throw away price.
It is equally hard for car dealers, especially second hand car dealers whose whole business is to buy and sell used cars as they face almost similar problems in selling second hand vehicles.
It is therefore really important to understand how the second hand car market works before committing to purchase the car of your dreams.

Main Objective

We aim to create a model that predicts if a dealer will manage to sell a used car or not

3. Specific Objectives

1. To identify the best performing model for our car sell prediction
2. Determine which was the most common car brand sold.
3. Determine which was the most common fuel type
4. To identify the main features that will be used in our model.
5. To Investigate the relationship between the engine size and the horsepower of the vehicle.
6. Build a model that can predict if a used car will be sold.

B. Assessing the situation

1. Data Inventory

The dataset is from kaggle and can be accessed here.The dataset is based on various market surveys, the consulting firm has gathered a large dataset of different types of used cars across the market.

2. Assumptions

- The data provided is correct and up to date
- The data has minimal anomalies
-

3. Constraints

- Since the data has labels we can only conduct supervised learning on the same

Data Mining Goals

- To predict if a dealer will manage to sell a used car or not using different models
- To determine which was the most common used car brand put up for sale.
- To establish which was the most common fuel type
- To identify from which country do most purchases are made

Data Understanding

The dataset is from kaggle and can be accessed here.The dataset is based on various market surveys, the consulting firm has gathered a large dataset of different types of used cars across the market.

Data Description
It consists of 7906 rows and 18 columns

1. Sales_ID (Sales ID)
2. name (Name of the used car)
3. year (Year of the car purchase)
4. selling_price (Current selling price for used car)
5. km_driven (Total km driven)
6. Region (Region where it is used)
7. State or Province (State or Province where it is used)
8. City (City where it is used)
9. fuel (Fuel type)
10. seller_type (Who is selling the car)
11. transmission (Transmission type of the car)
12. owner (Owner type)
13. mileage (Mileage of the car)

14. engine (engine power)
15. max_power (max power)
16. seats (Number of seats)
17. sold (used car sold or not

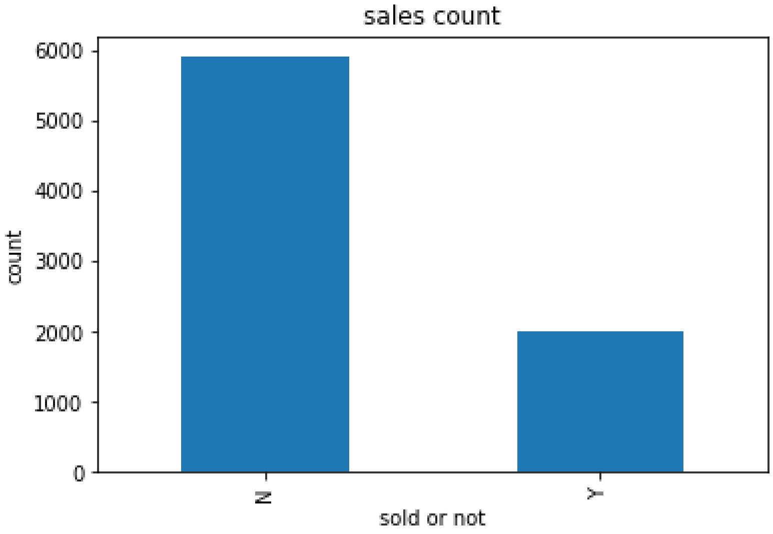
Data Preparation

1. Importing the necessary libraries into our colab notebook for analysis
2. Loading our csv file to our colab notebookComputing data description in rows and column description - The data has 7906 rows and 18 columns
3. Displaying the head and tail of the dataset
4. Checking for null values in our data- There are no null values in the dataset.
5. Checking for anomalies and outliers in our dataset
6. Checking for unique values in our dataset

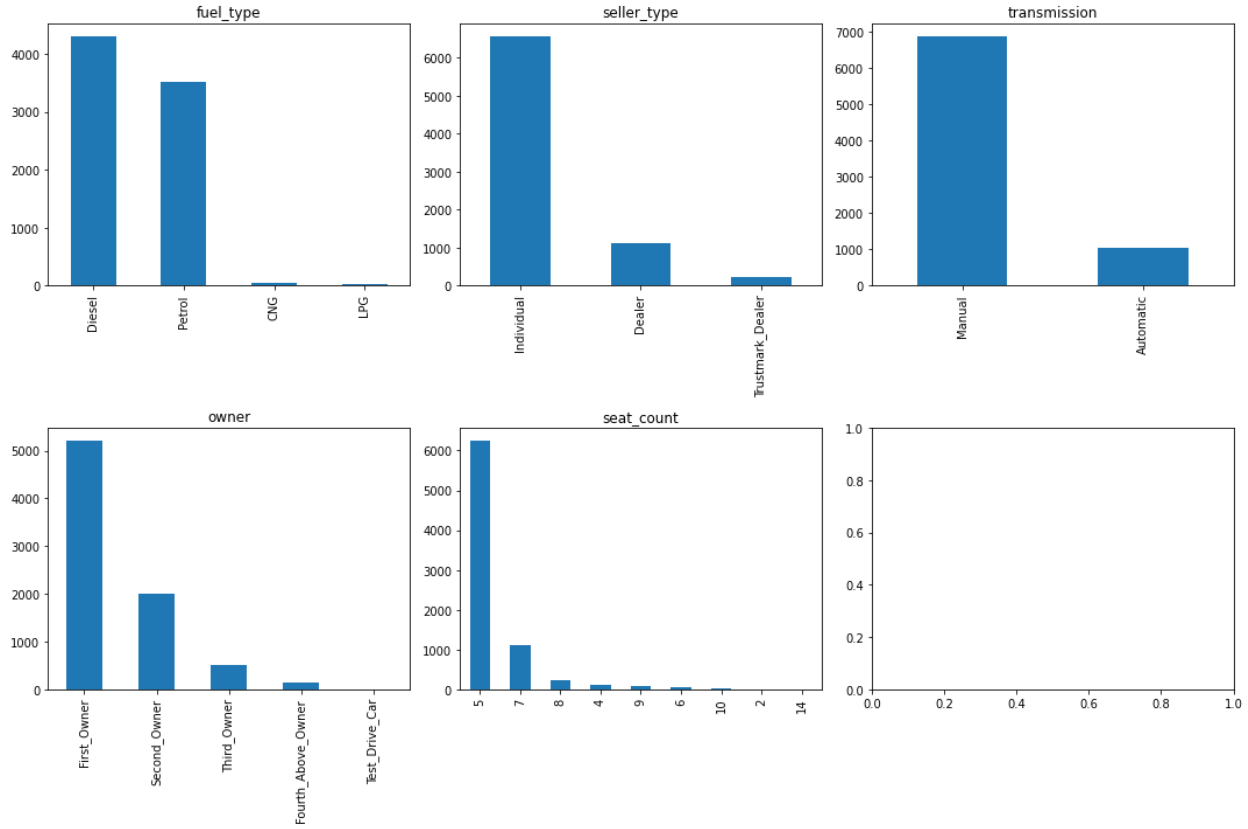
Exploratory Data Analysis

Univariate Analysis

1.Number of sold/unsold cars

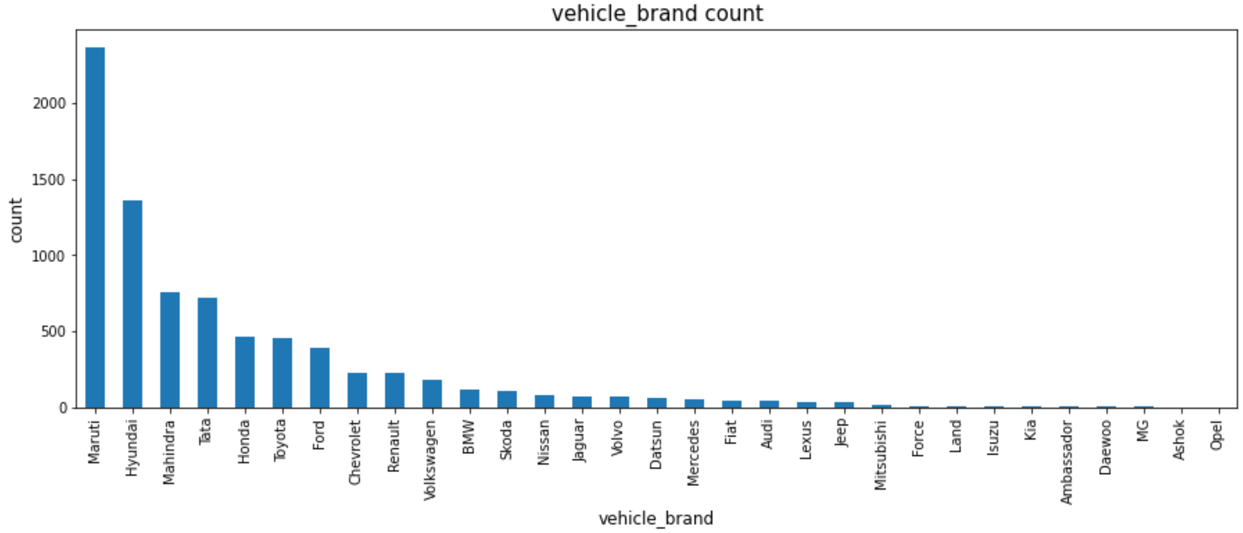


2.



- **Most vehicles were diesel engine, manual vehicles owned by first time individual owners with 5 seats**

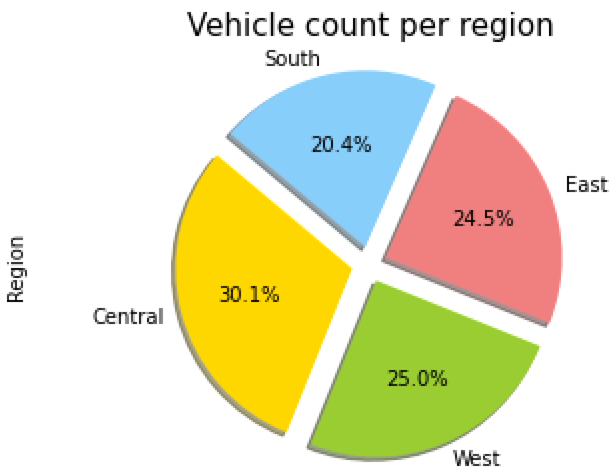
3.The most popular car brand/name



The most popular car brand is maruti with the least popular being opel

4.Region with the highest number of vehicles

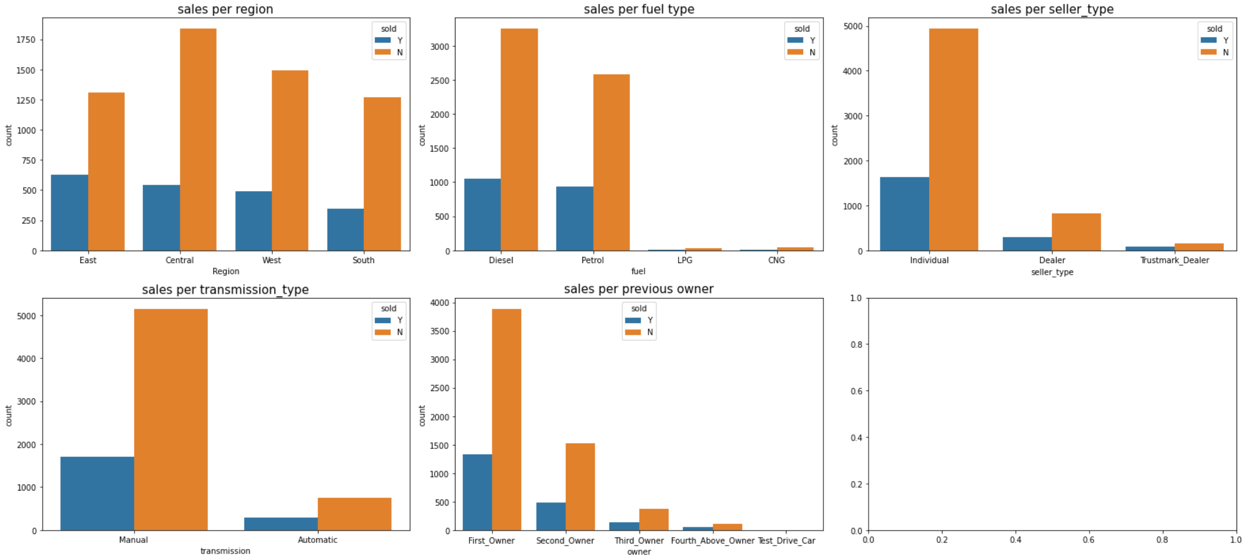
```
Central    0.300531
West       0.249937
East       0.245257
South      0.204275
Name: Region, dtype: float64
```



- **central region had the highest number of vehicles while the south region had the least**

Bivariate Analysis

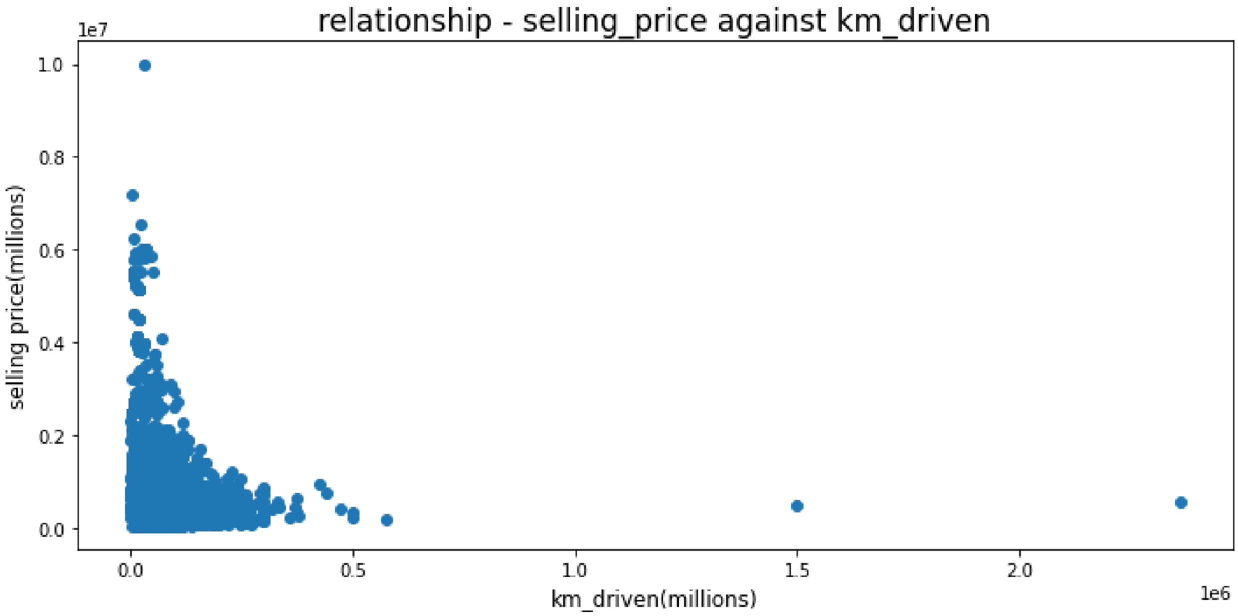
1.

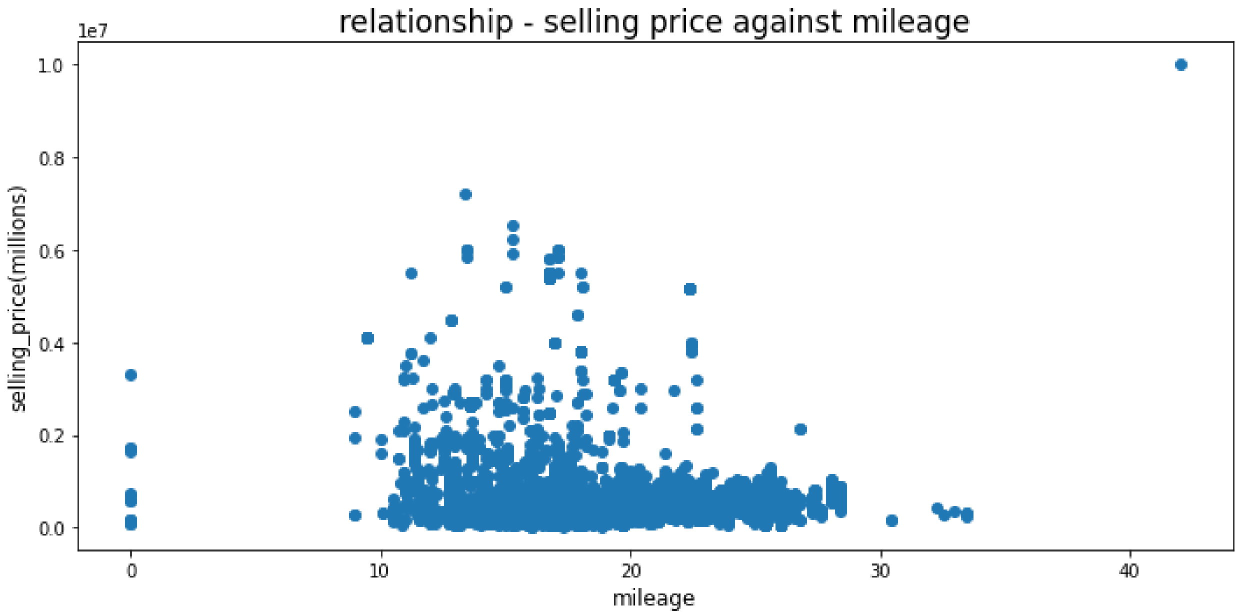


- **From all the charts we can see that cars that were not sold were the most.**

- **The Eastern region had the highest count of cars both sold and not sold with cars not sold being the most**
- **Most vehicles were diesel vehicles and most of them were not sold**
- **Most vehicles were owned by individuals, also, most of them were not sold.**
- **Most vehicles were vehicles with manual transmission and most of them were also not sold.**
- **First time owners of vehicles were the most, but even with this, most of them did not sell their vehicles**

2.





- **Most vehicles have travelled for few kilometers and have a low price as well.**
- **Most vehicles have a mileage of between 10-30 while the selling price is low as well**

Modeling

Our task was to build a model able to predict if a manager would be able to sell a used car or not. In such a problem, prioritizing minimizing the number of false positives is more important Therefore, in looking at the different evaluation metrics eg recall, accuracy etc, precision was the most important metric to optimize.

The models tested and optimized were built applying 5 types of classifiers:

- Random forests
- Naive bayes
- Support vector machines
- K-Nearest-Neighbors
- Logistic regressors

1.Logistic regression

The **precision** using **Logistic regression** approach for cars not sold :-

Before hyper-parameter tuning: **0.75**(not sold) and **0.25**(sold)

After hyper-Parameter tuning: **0.75**(not sold) and **0.27**(sold)

- The precision score for the first remained the same for group 0.
- However, there was an improvement in the precision for group 1.
- This is the baseline model.

2.Naive Bayes

The *precision* using *Naive Bayes* approach for cars not sold :-

Before hyper-parameter tuning: 0.75(not sold) and 0.25(sold)

After hyper-Parameter tuning: 0.75(not sold) and 0.25(sold)

- There was however an increase in the accuracy after tuning.
- This model did however not perform better than baseline model.

3.Random Forest Approach

The ***precision*** using ***Random Forest*** approach for cars not sold :-

Before hyper-parameter tuning: **0.75**(not sold) and **0.25**(sold)

After hyper-Parameter tuning: **0.76**(not sold) and **0.26**(sold)

1. There was an improvement after hyper-parameter tuning.
2. This model also performed better than the base line model and it is better than Naive Bayes.

4.KNN Approach

The ***precision*** using ***KNN*** approach for both categories (sold and not sold cars) :-

Before hyper-parameter tuning: ***0.75***(not sold) and ***0.28***(sold)

After hyper-Parameter tuning: ***0.76***(not sold) and ***0.25***(sold)

- There was an increase after hyper_parameter tuning.
- It however did not peform better than random forest

5.SVM Approach

The ***precision*** using ***SVM*** approach for cars not sold :-

Before hyper-parameter tuning: ***0.75***(not sold) and ***0.24***(sold)

After hyper-Parameter tuning: ***0.75***(not sold) and ***0.25***(sold)

There was an improvement after hyper-parameter tuning.

- There was an improvement after hyper-parameter tuning.
- However, this model did not perform better than the baseline model.

5.0 Discussion and Conclusion

The 18 features in our dataset included -Sales_ID ,name ,year ,Selling_price,km_driven ,Region,State or Province ,City,fuel ,seller_type ,transmission ,owner,mileage,engine ,max_power ,seats .Sold was the target column.

During modeling, the different models tested and optimized were based on 5 types of classifiers:

- Random forests
- Naive bayes classifier
- Support vector machines
- K-Nearest-Neighbors
- Logistic regressors

Given that the primary task was to predict a car will be sold or not, precision of the positive class was the most important metric to optimize because minimizing the number of false positives is essential.

The main objective of 'Creating a model that predicts if a car dealer will be able to sell a car or not' was achieved.

- The best model was the Random Forest after tuning with parameters {'max_depth': 4, 'min_samples_split': 5, 'n_estimators': 50}. It had the best precision score of 0.76 of the 'not sold' class.
- The main features used to model the data are 'fuel','seller_type' and 'transmission', which are the main features after doing feature selection.
- The relationship between the sale of a car and other predictor variables investigated is shown in the data exploratory analysis.