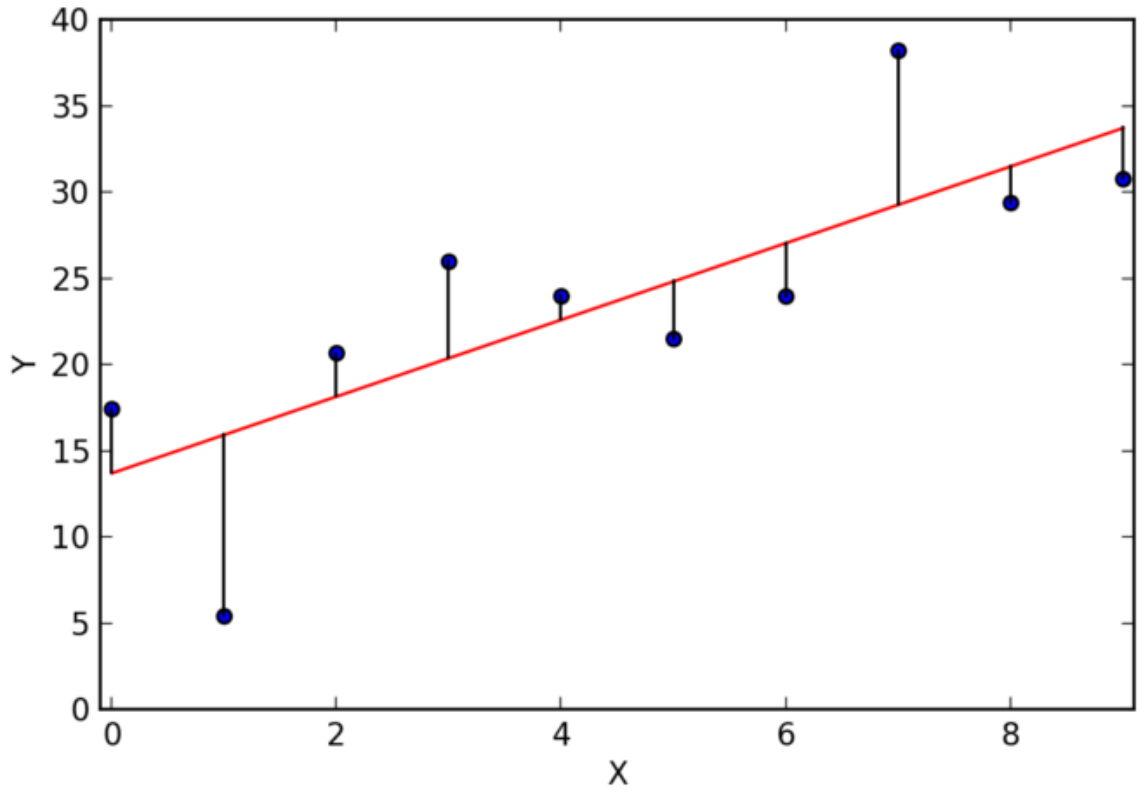
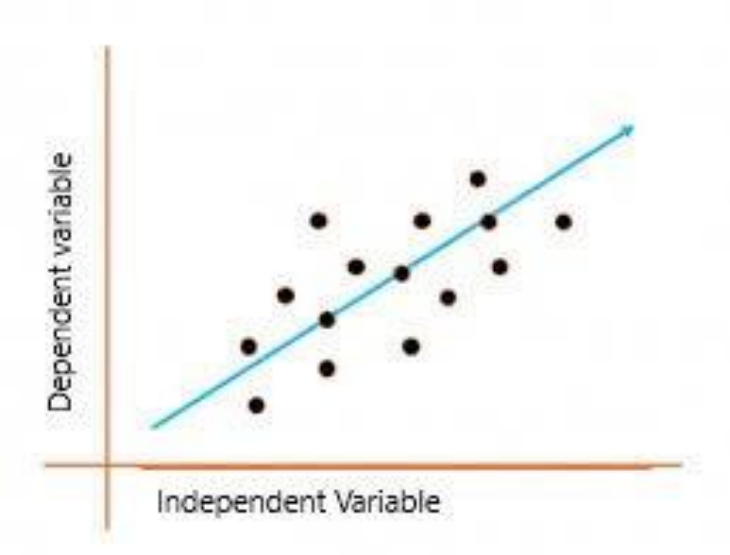
Multiple Linear Regression using Python

Team Members:
Garveet Juneja
Vansh Chugh



Linear Regression



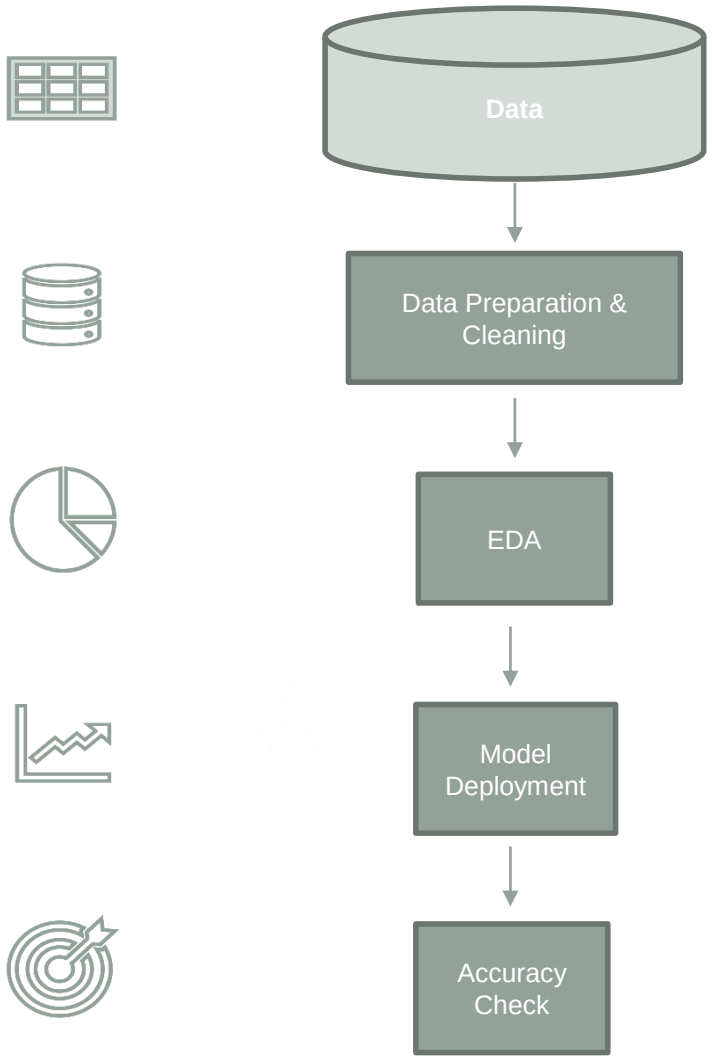
Linear regression is a quiet and the simplest statistical regression method used for predictive analysis in machine learning. Linear regression shows the linear relationship between the independent(predictor) variable i.e. X-axis and the dependent(output) variable i.e. Y-axis, called linear regression.

DATA

- The data set has been taken from a trustworthy source, consisting of the information related to pre-owned cars and its features. The input dataset contains information about used cars listed on www.cardekho.com, which we found through a dataset available on Kaggle.
- A quick glance at the data, gives us an idea of the columns and their datatypes.
- The data contains 8128 records on 13 variables.

CarDekho
BHAROSA KAR KE DEKHO

PROJECT FLOWCHART



DATA DESCRIPTION

Name	• Name of the Car Company and its model
Year	• Year of first selling of the car
Selling Price	• Selling price of the Car
Km Driven	• Kilometers driven at the time of re-selling
Fuel	• Type of fuel intake
Seller Type	• Type of Seller
Transmission	• Type of transmission
Owner	• Number of owners before re-sell
Mileage	• Mileage of the Car
Engine	• Capacity of engine in cc
Max Power	• Power of the car in bhp
Torque	• Torque of the car in rpm
Seats	• Seating capacity of the car

DATA CLEANING AND PREPARATION

(1) We started off by detecting the NA values in our columns (using the isnull() function in pandas).

There were about 220 NA values each in Mileage, Engine, Max power, Torque and seats, which are fairly important features to estimate the selling price of a car.

So we filled the NA values with the average of rest of the values using fillna() function in pandas.

(2) Now the next task was to obtain relevant information from the data available,

For example in Mileage we had data entries like “20 kmpl”

For our estimation we needed the numerical value of the entry so we obtained them by applying str.split() function.

Same was observed in engine and max_power data and thus same method was applied on them to obtain relevant information for our model.

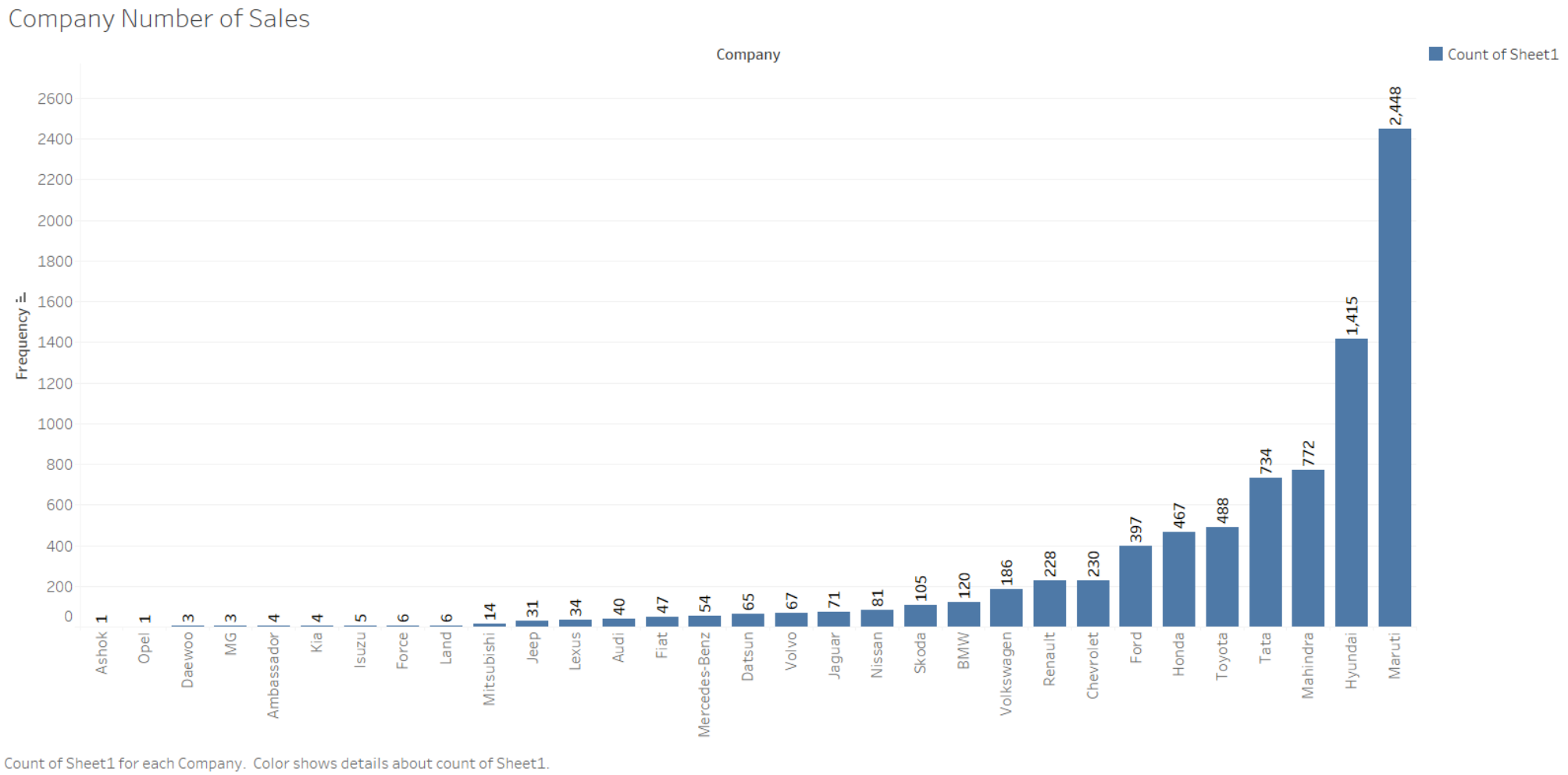
DATA CLEANING AND PREPARATION

(3) Next, we obtained the Age of the car using the current year and subtracting the year of first purchase from it, Age is an important factor to estimate selling price of the car which is primary goal of our model.

Converted values from Mileage, Engine, MaxPower, Torque to Numeric datatype to use them further in the model.

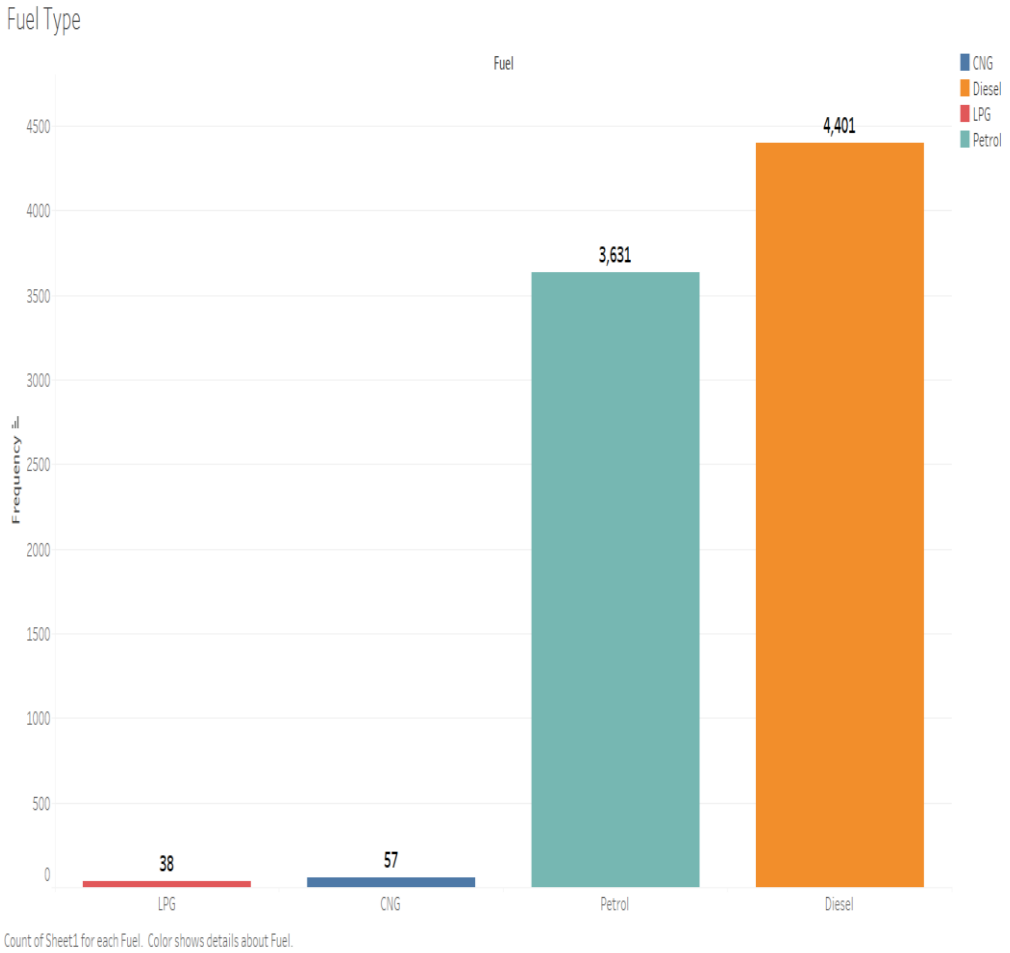
EXPLORATORY DATA ANALYSIS

Frequency of Cars of each company

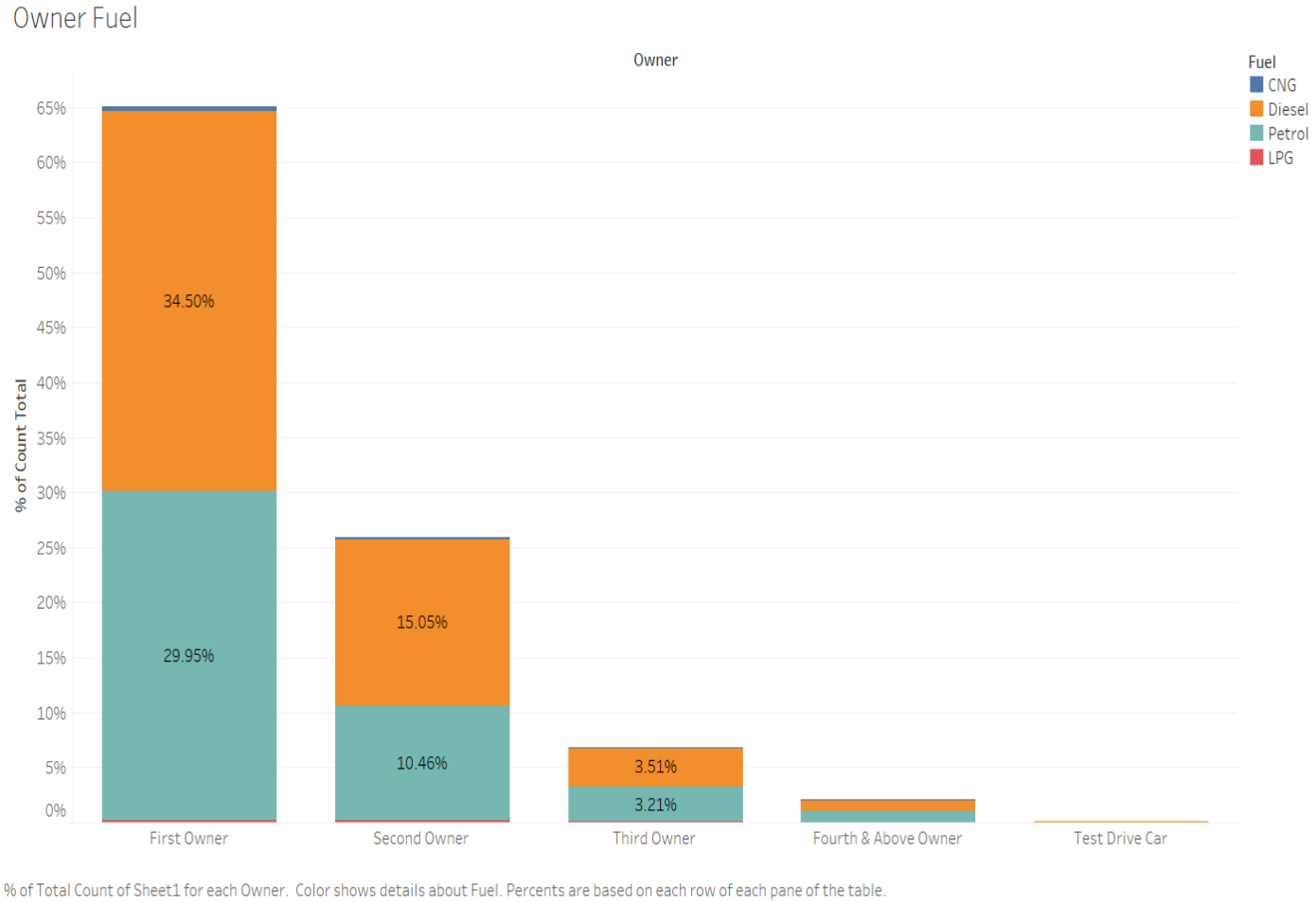


Maruti leads the race in frequency of re-selling of cars followed by Hyundai, Mahindra and others.

Frequency of cars for different type of fuel



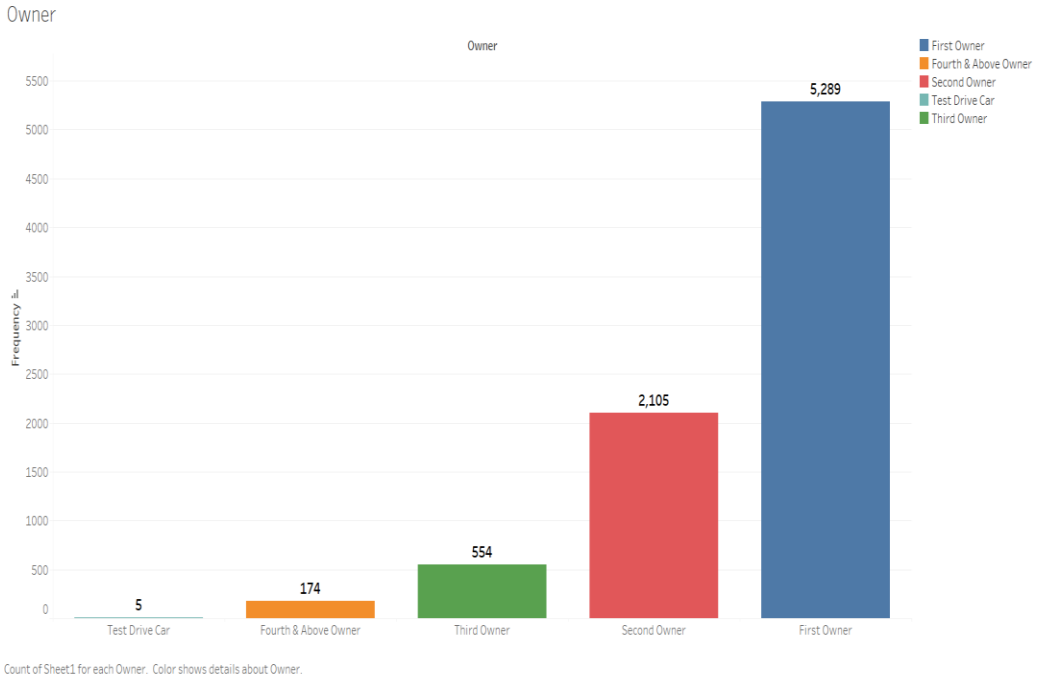
Owner and fuel



Diesel cars are most sold cars followed by petrol engine.

Diesel cars are most preferred fuel type 2nd and even 3rd owner cars ,followed by petrol engine.

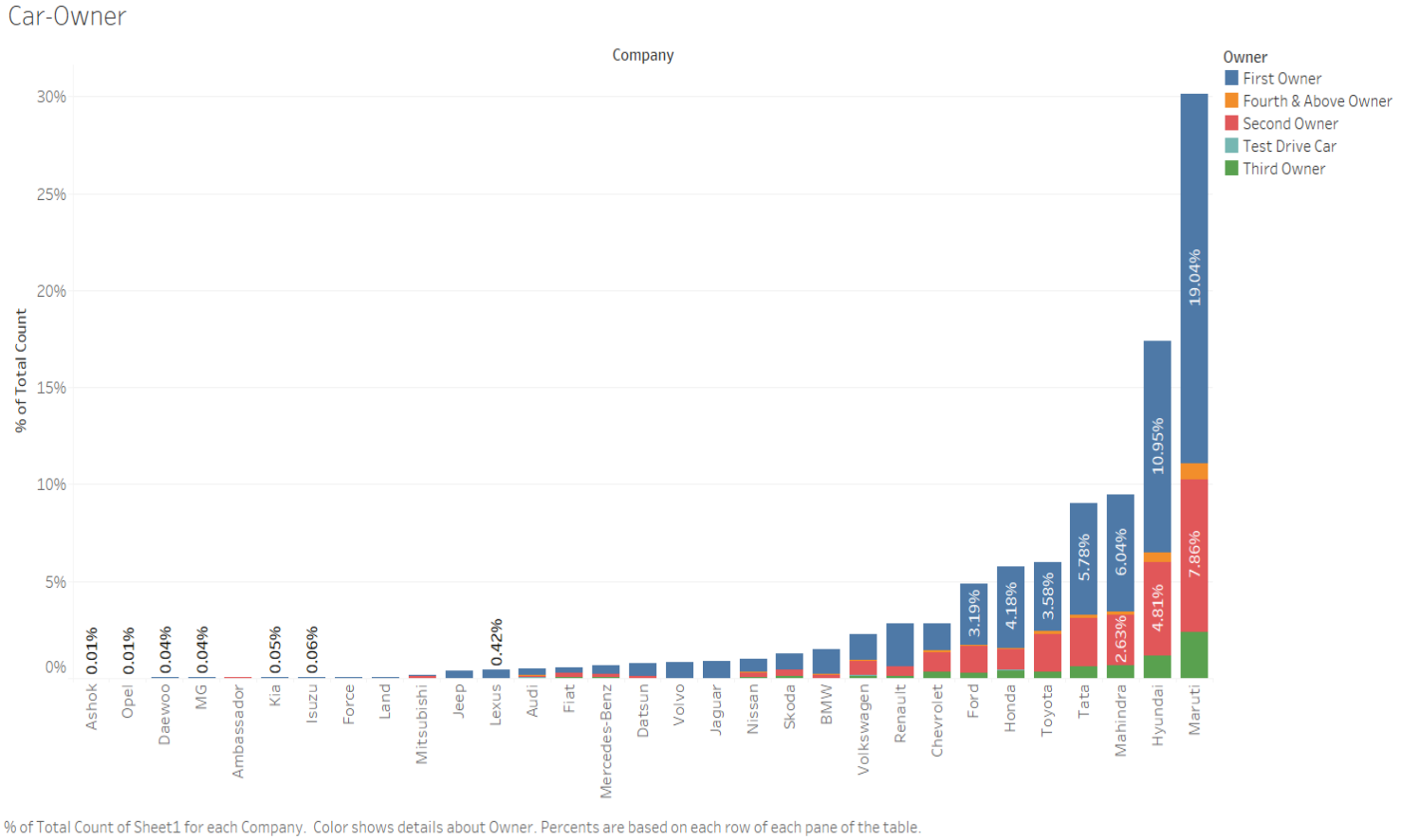
Owners of car



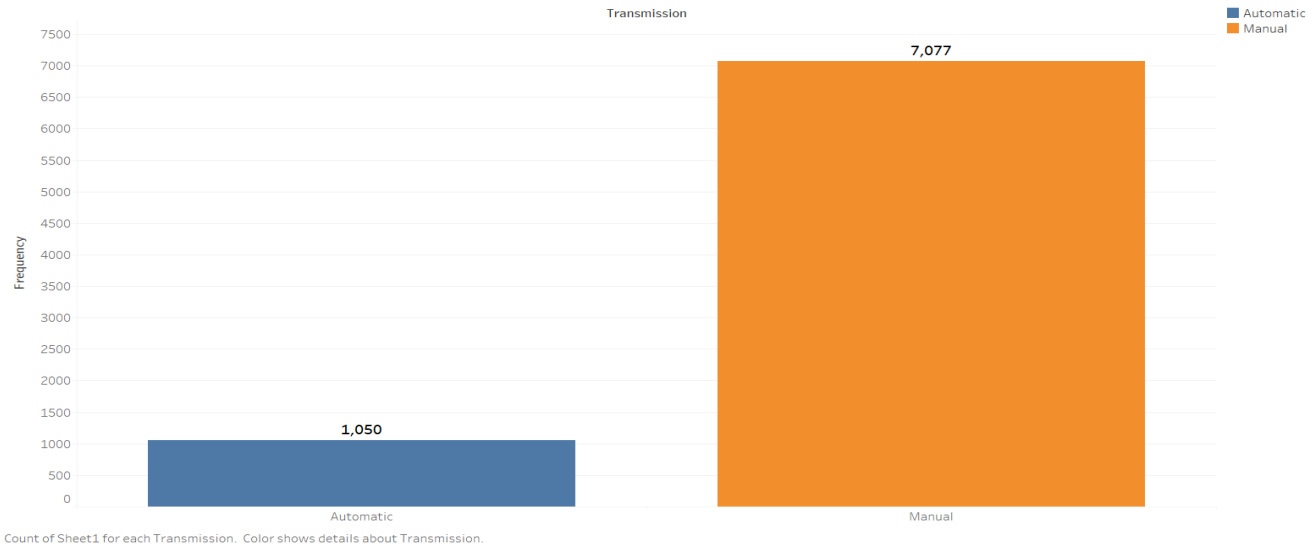
Most customers prefer 1st hand, new cars.

Maruti cars are the most preferred 2nd owner cars. Clearly, we can say Maruti is the industry leader currently.

Owner vs company



Transmission

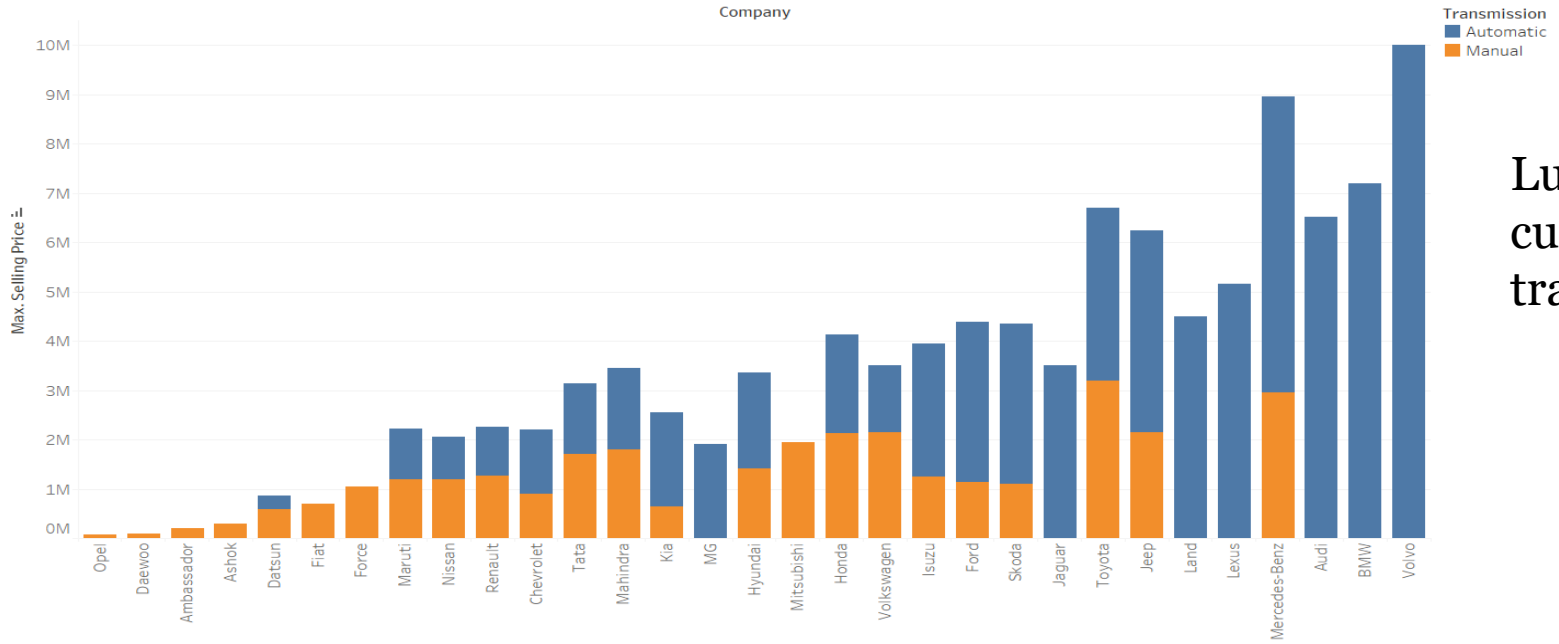


Count of Sheet1 for each Transmission. Color shows details about Transmission.

Transmission Type

Despite being more comfortable, customers prefer manual transmission over automatic, due to cost factor.

Max Selling Price

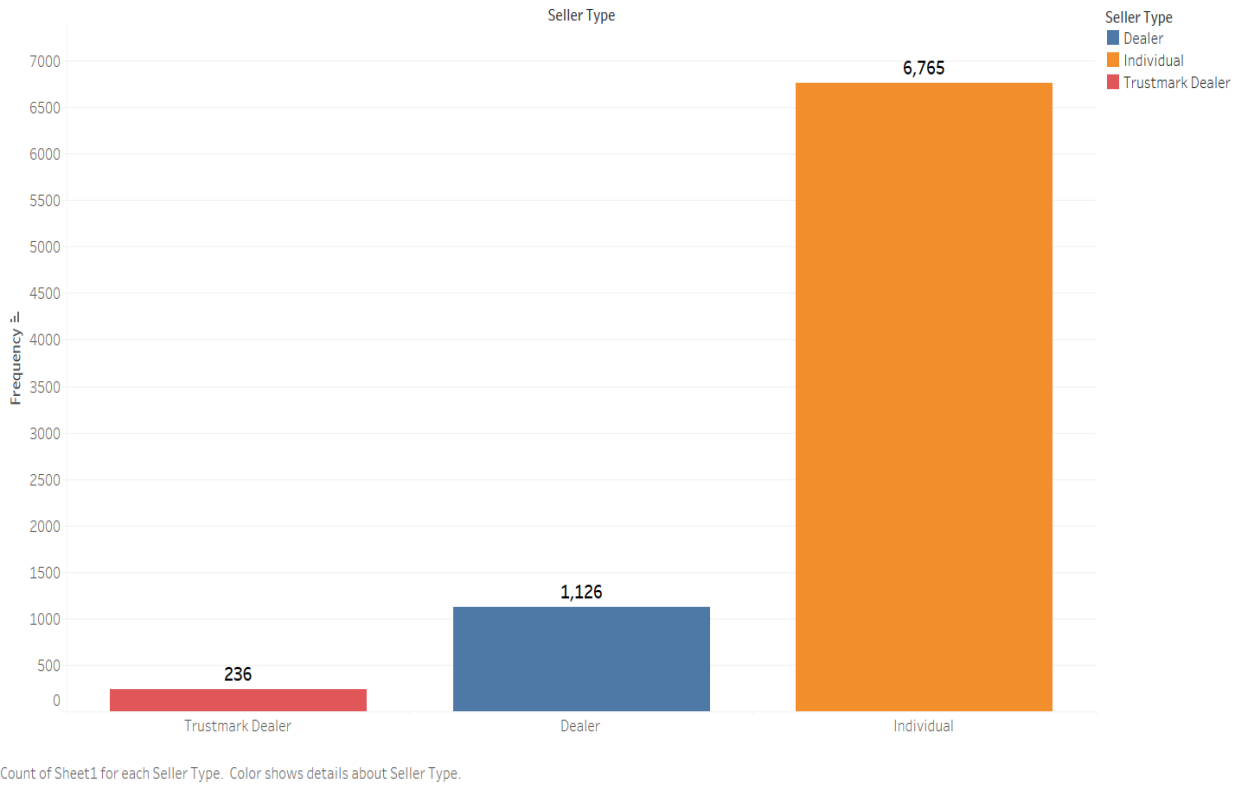


Maximum of Selling Price for each Company. Color shows details about Transmission.

Transmission Type vs company

Luxury cars have a different preference for customers. Buyers prefer automatic transmission over manual.

Seller Type



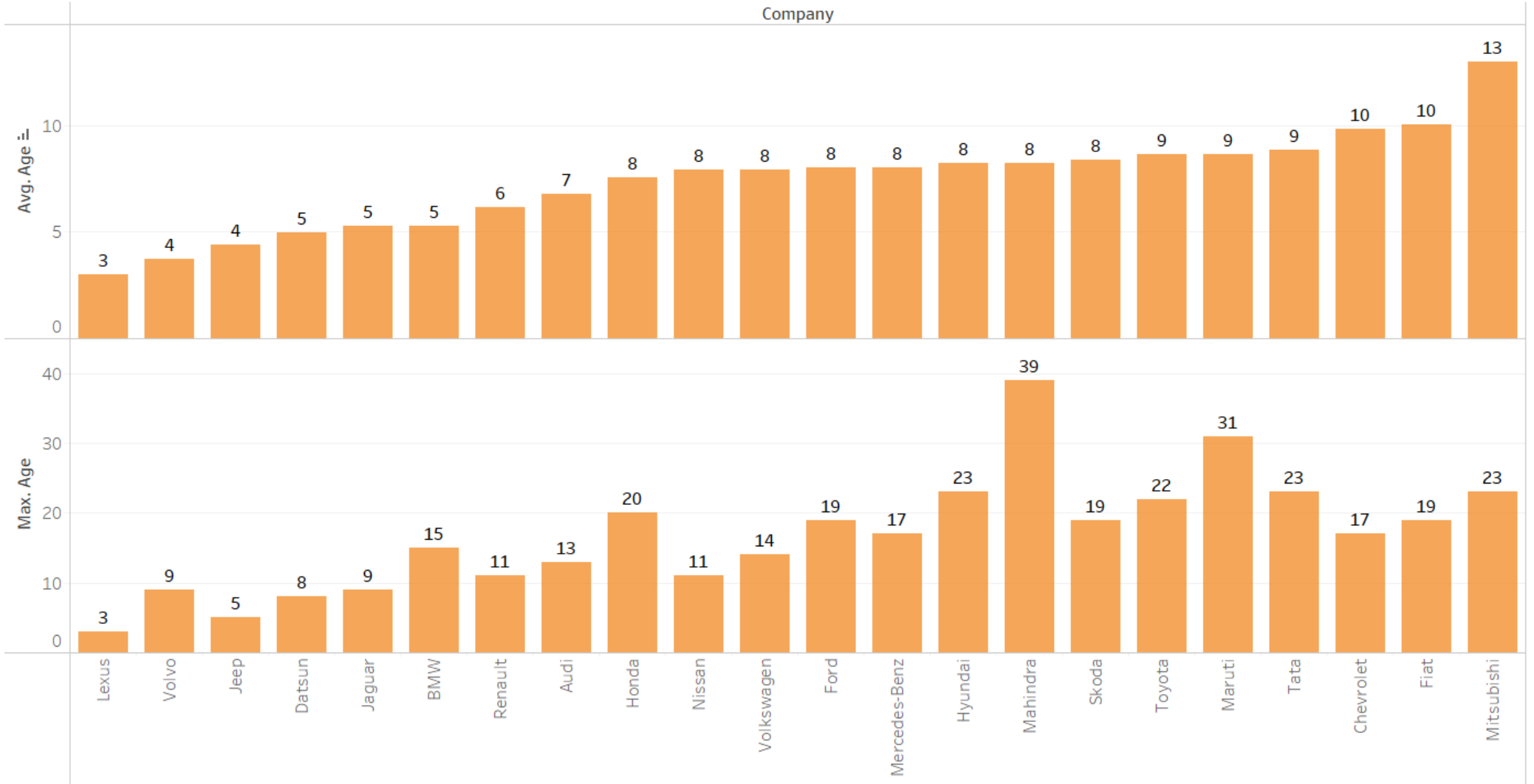
Insert graph – Seller type – Selling Price

Number of individual sellers is the highest but Trustmark dealers are selling the cars for the highest price.

Average age and maximum Age of each Company

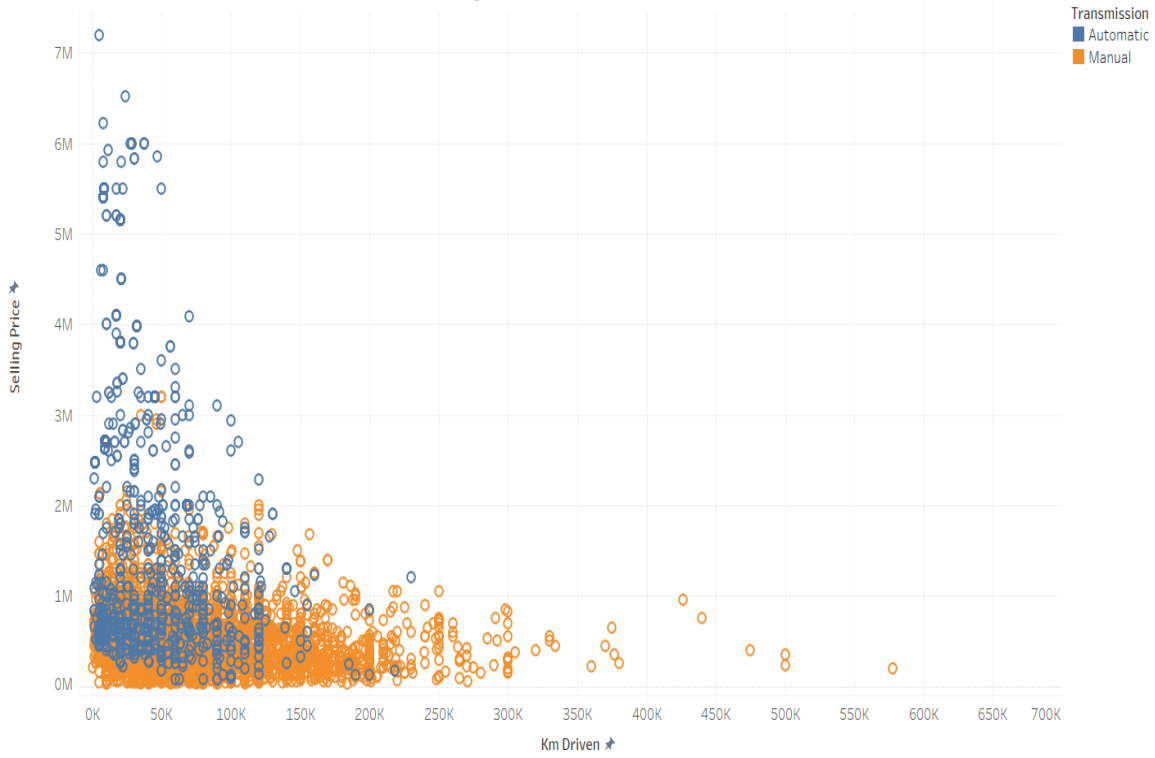
Mahindra cars have the max re selling age due to the love of customers for Mahindra jeep. However, the max average of cars is seen by mitsbuishi.

Age



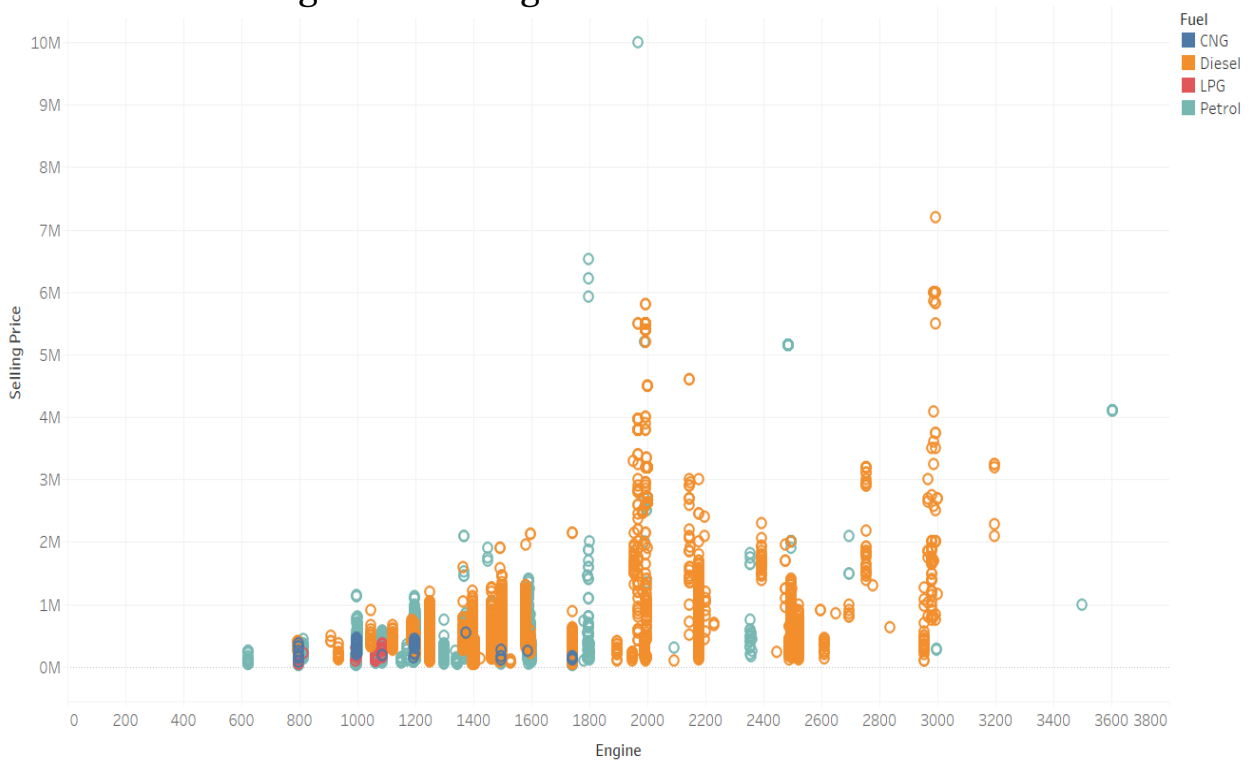
Average of Age and maximum of Age for each Company. For pane Average of Age: The marks are labeled by average of Age. For pane Maximum of Age: The marks are labeled by maximum of Age. The data is filtered on count of Sheet1, which includes values greater than or equal to 14.

Km Selling Price



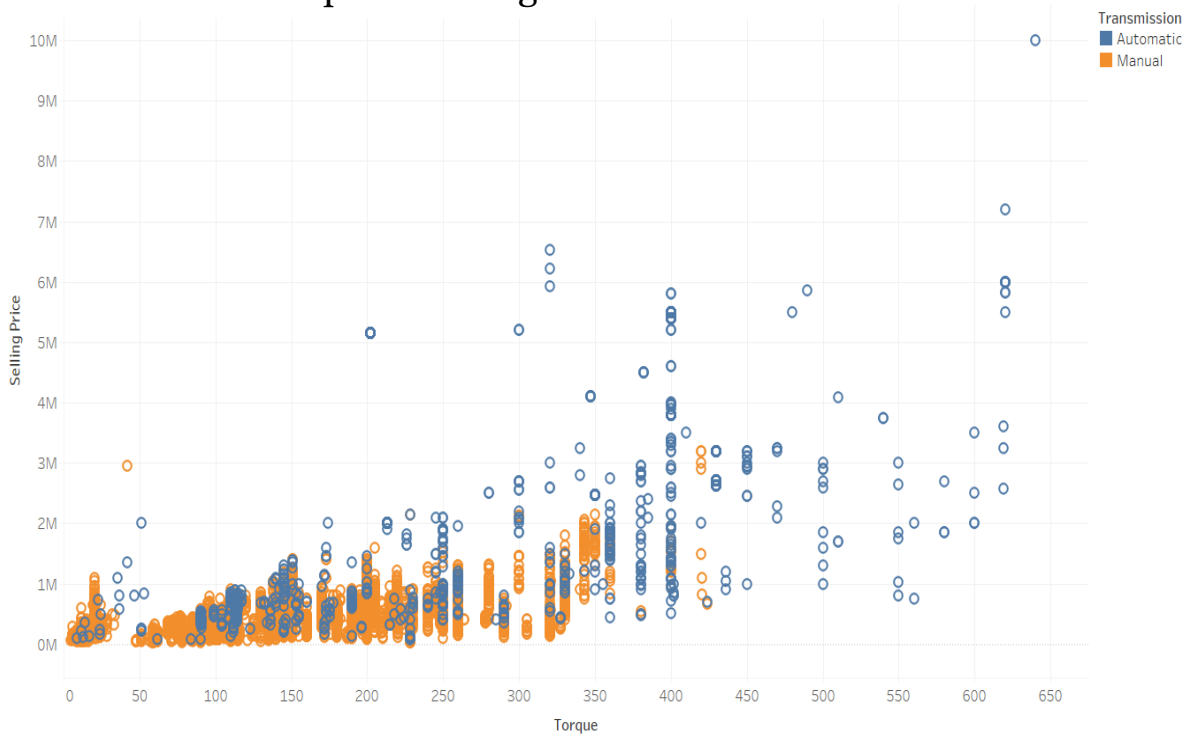
Km Driven vs. Selling Price. Color shows details about Transmission.

Engine Vs Selling Price



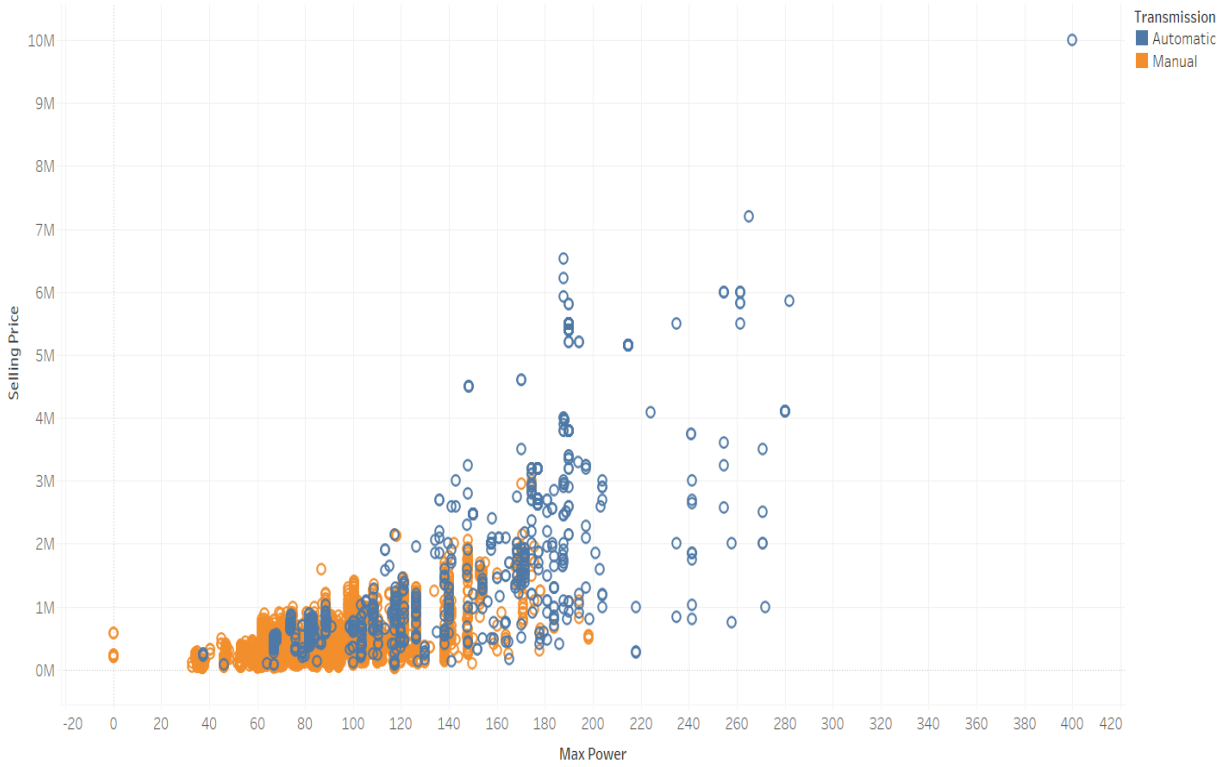
Engine vs. Selling Price. Color shows details about Fuel.

Torque vs Selling Price



Torque vs. Selling Price. Color shows details about Transmission.

Power Vs Selling Price



Max Power vs. Selling Price. Color shows details about Transmission.

LINEAR REGRESSION

Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

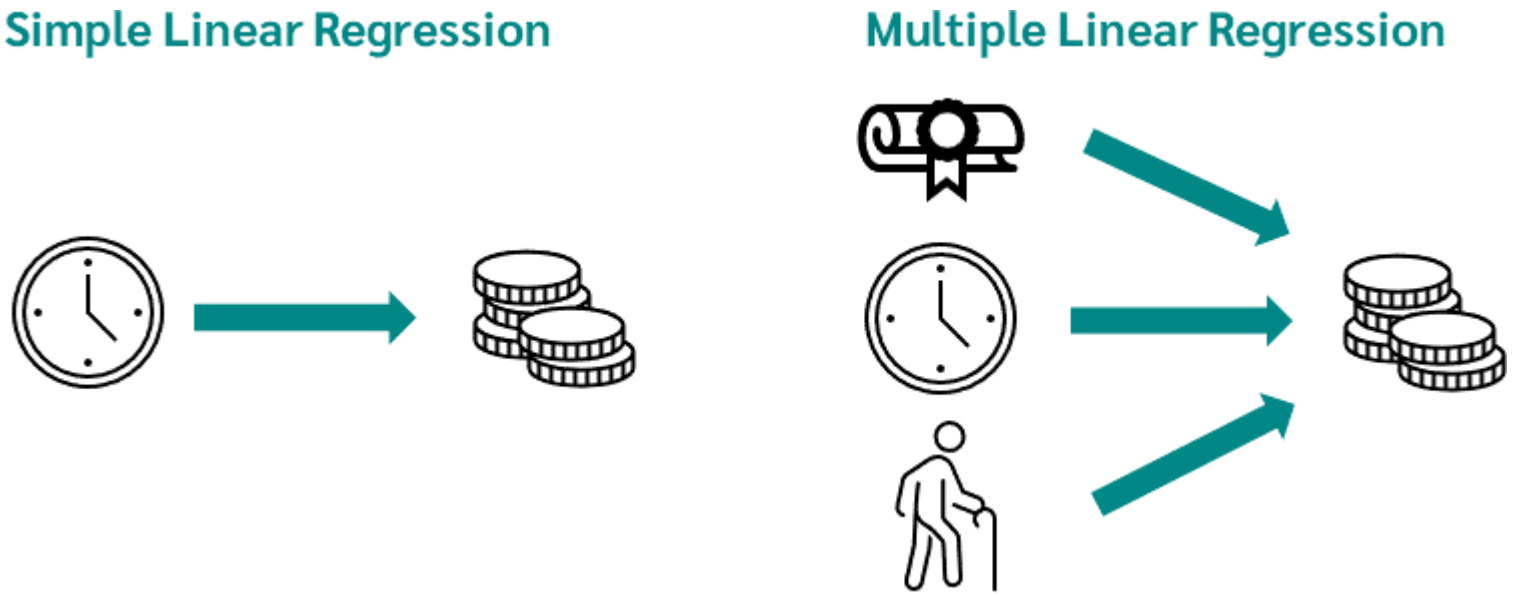
For example, If X is the indedpendent variable and we have to make a model to predict Y,

$$y = \theta_1 + \theta_2.x$$

Where theta1 and theta2 are coefficients to be estimated using normal equations (Least Square Method).

MULTIPLE LINEAR REGRESSION

Unlike simple linear regression, multiple linear regression allows more than two independent variables to be considered. The goal is to estimate a variable based on several other variables. The variable to be estimated is called the dependent variable (criterion). The variables that are used for the prediction are called independent variables (predictors).



MODEL BUILDING

As we begin with our model building

(1) Our data contains categorical variables (Fuel, Owner, Transmission, Seller type) which seems to be important for our model, so we introduced dummy variables in such variables.

Why dummy variables?

Since our variables does not have any ordinal relationship i.e there is no dependence relation between them, We cannot use Label encoding and resort to introducing Dummy variables or One-Hot Encoding.

(2) Then Imported the needed libraries such as train_test_split, MinMaxScaler, LinearRegression, metrics and accuracy score from Sklearn.

Then we normalized the data using MinMaxScaler() and we split our data to test (33%) and training data (67%).

Why normalization?

Normalization refers to rescaling real-valued numeric attributes into a 0 to 1 range to make model training less sensitive to the scale of features. This allows our model to converge to better weights and in turn leads to make accurate model.

MODEL BUILDING

(3) Fitted the linear regression model using normalized training data and checked its coefficients.

Predicted our response variable on the test data using the model built above and plotted them against the original values to explore the accuracy of our model.

Used metrics like R-squared (0.787), Adjusted R-Squared (0.78) to judge our model's accuracy.

Why used R-squared and Adjusted R-squared as metrics to judge our model?

The coefficient of determination R^2, also known as the variance explanation, indicates how large the portion of the variance is that can be explained by the independent variables. The more variance can be explained, the better the regression model is.

The coefficient of determination R^2 is influenced by the number of independent variables used. The more independent variables are included in the regression model, the greater the variance resolution R^2. To take this into account, the adjusted R^2 is used.

LIMITATIONS

Main limitation of Linear Regression is the **assumption of linearity** between the dependent variable and the independent variables. In the real world, the data is rarely linearly separable. It assumes that there is a straight-line relationship between the dependent and independent variables which is incorrect many times.

FUTURE SCOPE

Deploying model on live data

Inclusion of Features related to market sentiments

Creating a more generalized model

THANK YOU!