greatlearning
Power Ahead

Cars4U - Project 3

Contents

Business Problem Overview and Solution Approach

- There is a huge demand for used cars in the Indian Market today. As sales of new cars have slowed down in the recent past, the pre-owned car market has continued to grow over the past years and is larger than the new car market now. Cars4U is a budding tech start-up that aims to find footholes in this market.

- In 2018-19, while new car sales were recorded at 3.6 million units, around 4 million second-hand cars were bought and sold. There is a slowdown in new car sales and that could mean that the demand is shifting towards the pre-owned market. In fact, some car sellers replace their old cars with pre-owned cars instead of buying new ones. Unlike new cars, where price and supply are fairly deterministic and managed by OEMs (**O**riginal **E**quipment **M**anufacturer / except for dealership level discounts which come into play only in the last stage of the customer journey), used cars are very different beasts with huge uncertainty in both pricing and supply. Keeping this in mind, the pricing scheme of these used cars becomes important in order to grow in the market.

Objective

1. Explore and visualize the dataset.
2. Build a linear regression model to predict the prices of used cars.
3. Generate a set of insights and recommendations that will help the business.

Data Overview

Data Dictionary

1. S.No.: Serial Number
2. Name: Name of the car which includes Brand name and Model name
3. Location: The location in which the car is being sold or is available for purchase Cities
4. Year: Manufacturing year of the car
5. Kilometers_driven: The total kilometers driven in the car by the previous owner(s) in KM.
6. Fuel_Type: The type of fuel used by the car. (Petrol, Diesel, Electric, CNG, LPG)
7. Transmission: The type of transmission used by the car. (Automatic / Manual)
8. Owner: Type of ownership
9. Mileage: The standard mileage offered by the car company in kmpl or km/kg
10. Engine: The displacement volume of the engine in CC.
11. Power: The maximum power of the engine in bhp.
12. Seats: The number of seats in the car.
13. New_Price: The price of a new car of the same model in INR Lakhs.(1 Lakh = 100, 000)
14. Price: The price of the used car in INR Lakhs (1 Lakh = 100, 000)

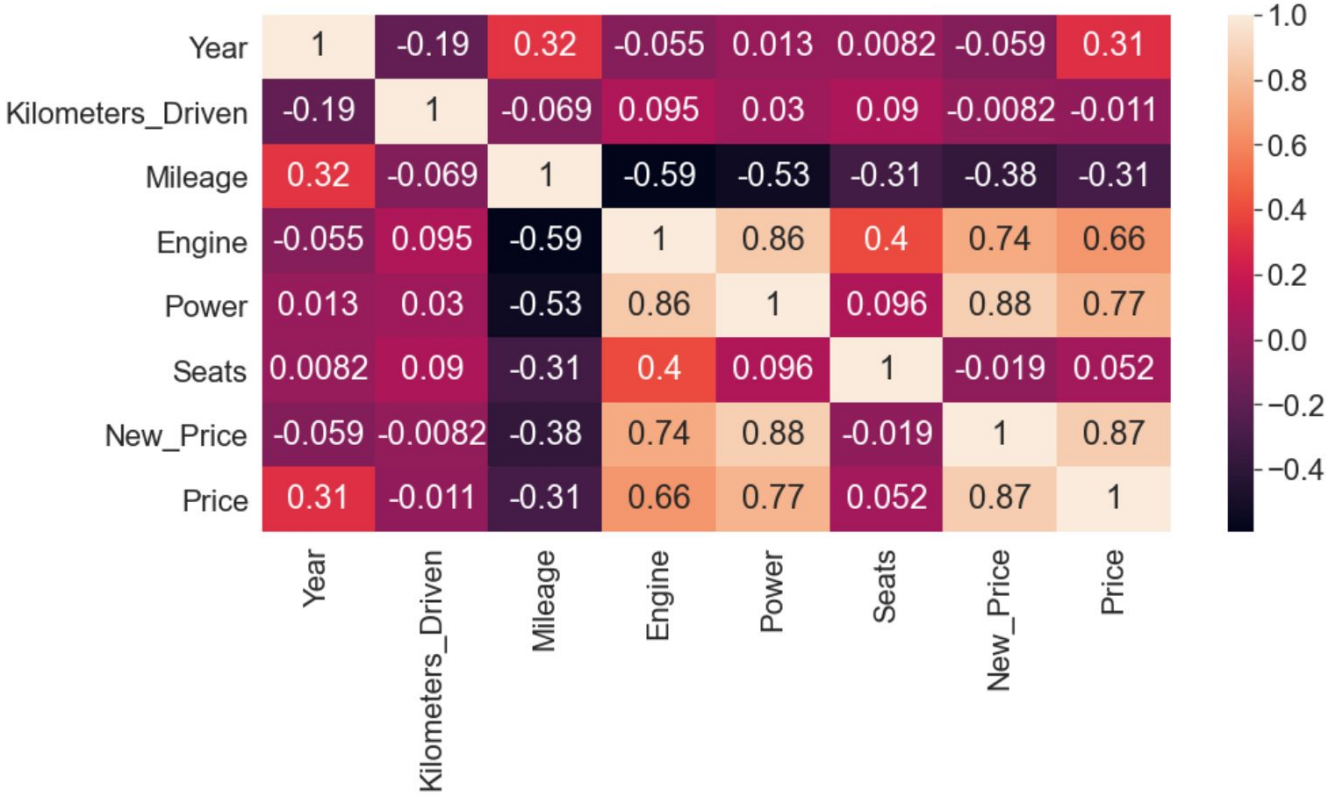
What I Changed with the Data

- First, I deleted the Serial Number because it was the same as the generated ID.
- Then, I split the Name variable into two columns, Brand and Model.
- Then, I took the labels off of Mile, Engine, and Power to turn these variables into a numeric variable.
- I also converted the Cr into Lakh so all prices were reported the same and then took off the labels and converted to a numeric variable.

Correlation

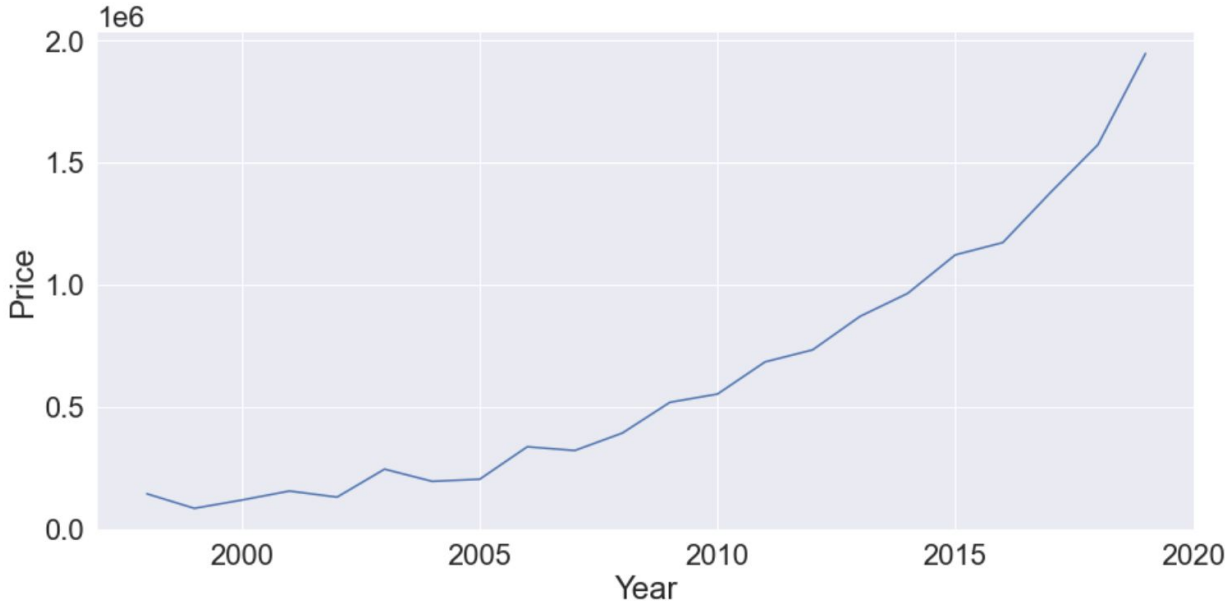
Price and New Price are highly correlated.

Price is also decently correlated with Power and Engine.



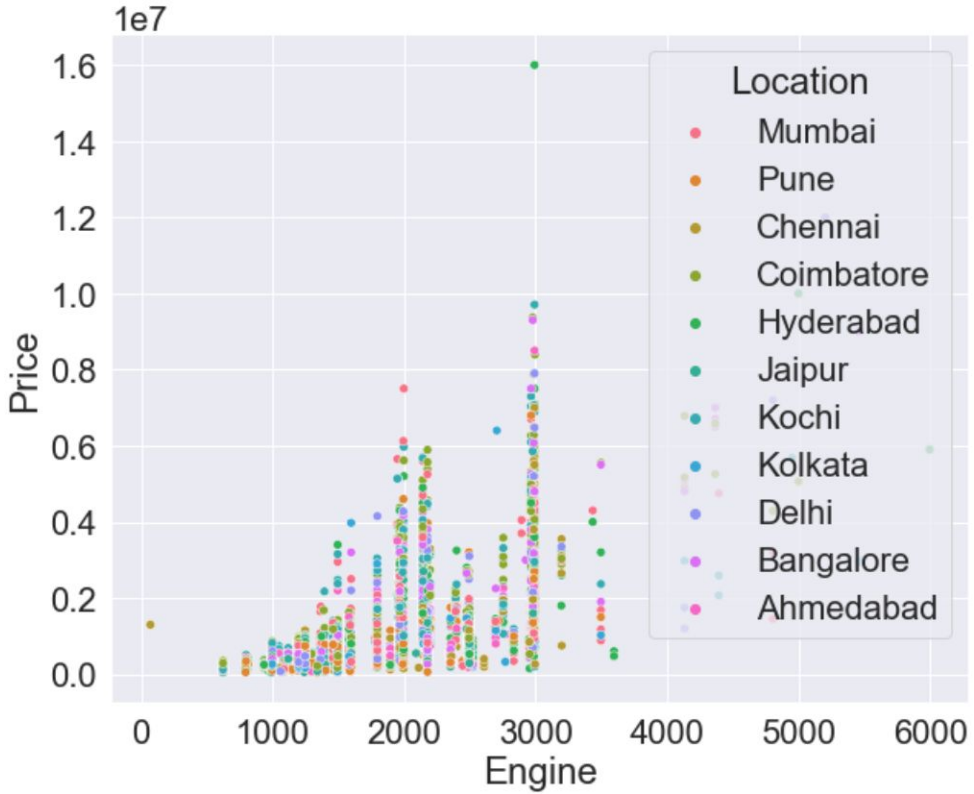
Year vs. Price

Year had a very positive correlation with Price as well.



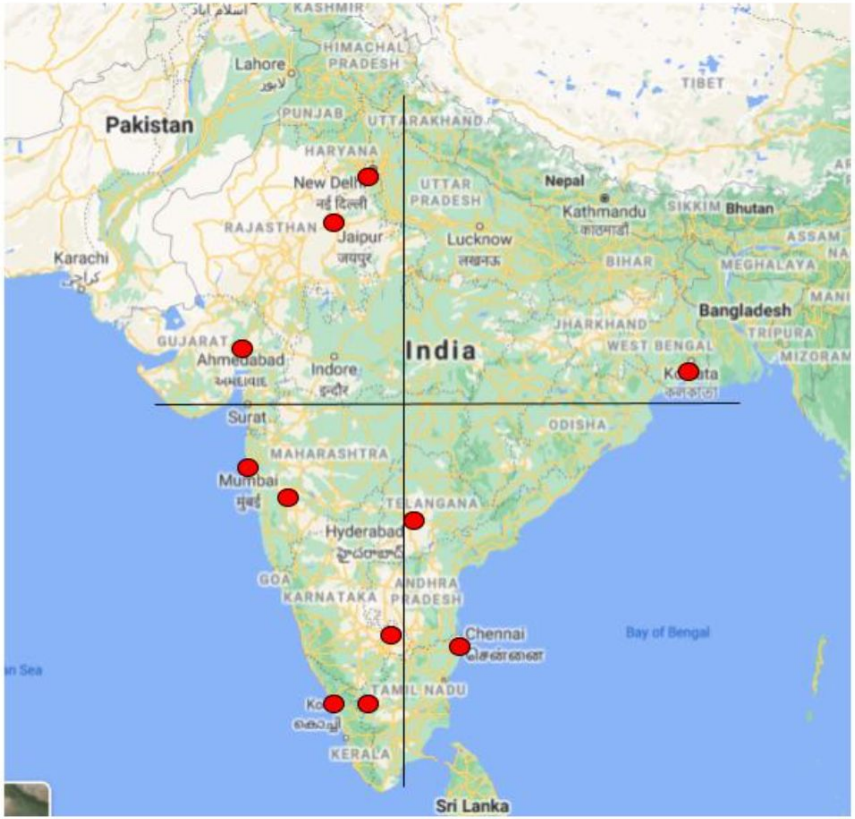
Location was an issue

It was difficult to get an idea of how location affected the price because there were a few locations that make the charts look to clustered.



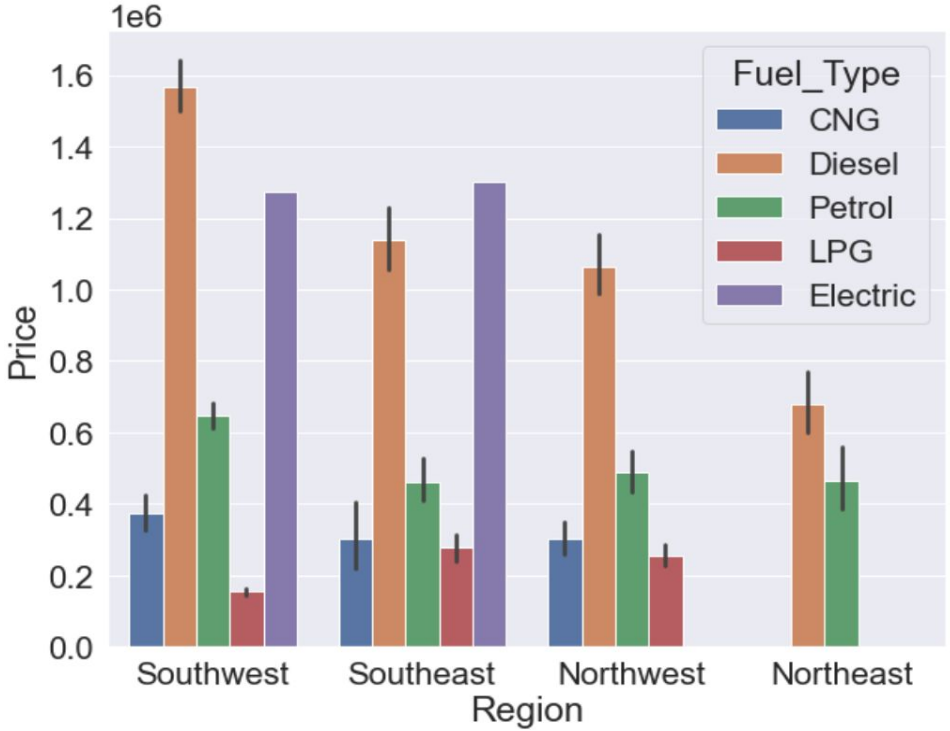
Location was an issue

Therefore I created regions to look at. I did this by creating 4 quadrants and then plotting each city using Google Maps.



Location was an issue

This allowed for a much cleaner visualization and a better understanding of the data.



Model Performance Summary

- When preparing for the model I dropped
  - Model
  - Brand
  - Region
  - Seats
- I originally had not dropped seats but my original R^2 value was lower than anticipated and every VIF value for seats was very high.

Model Performance Summary

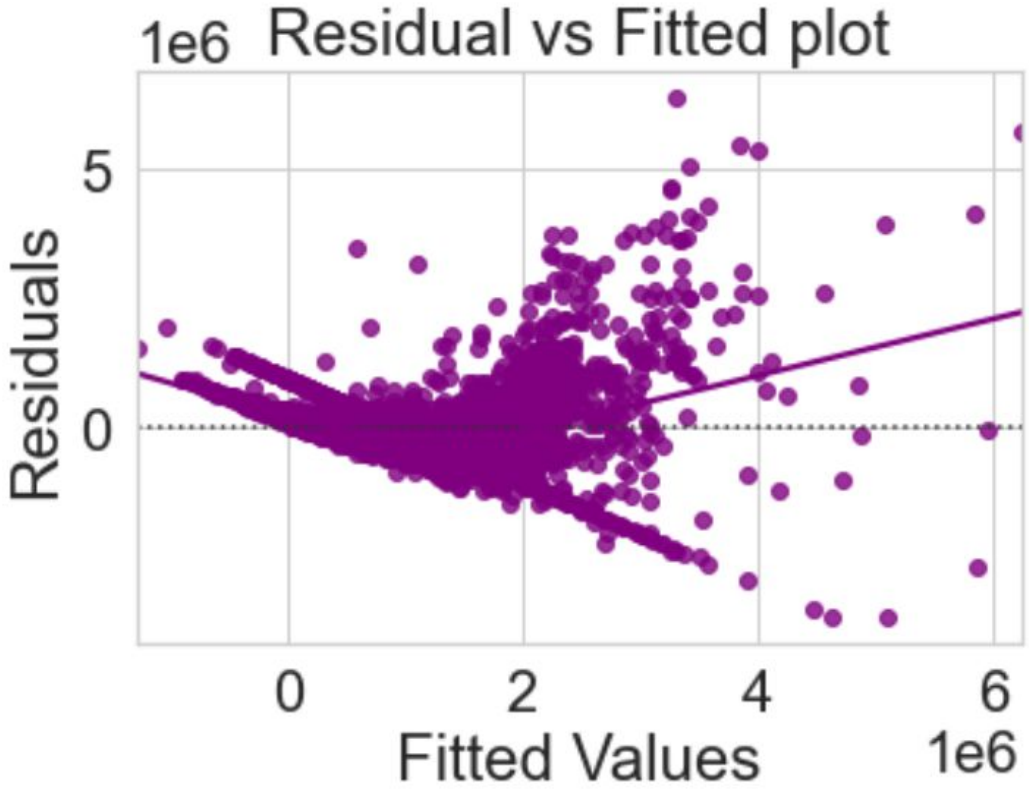
- I filled in missing values by using the mean for:
  - Mileage
  - Power
  - Price
  - Engine
- I also deleted one row from Kilometers Driven because it was an extreme outlier.

Model Performance Summary

- I made Price the dependant variable
- The R^2 value for the training data was 0.58
- While MAPE was 57.90
- The R^2 value for the testing data was 0.62
- While MAPE was 57.57

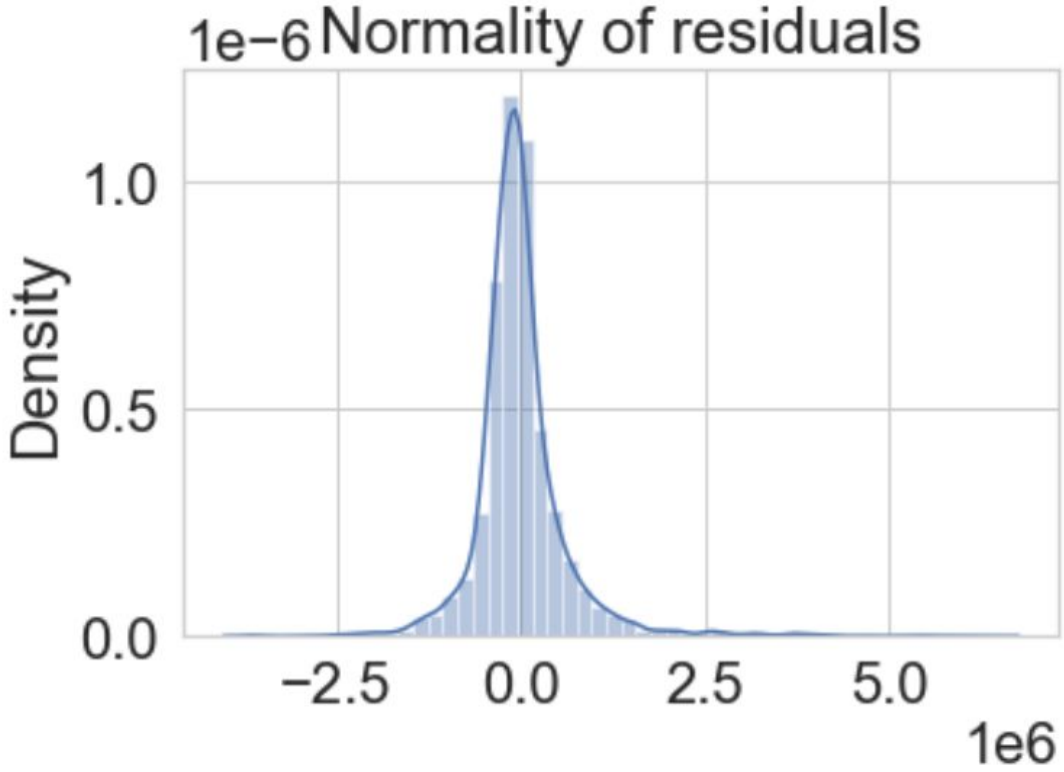
Model Performance Summary

- There is a start the the linearity but I do feel this model could use some more negotiating.



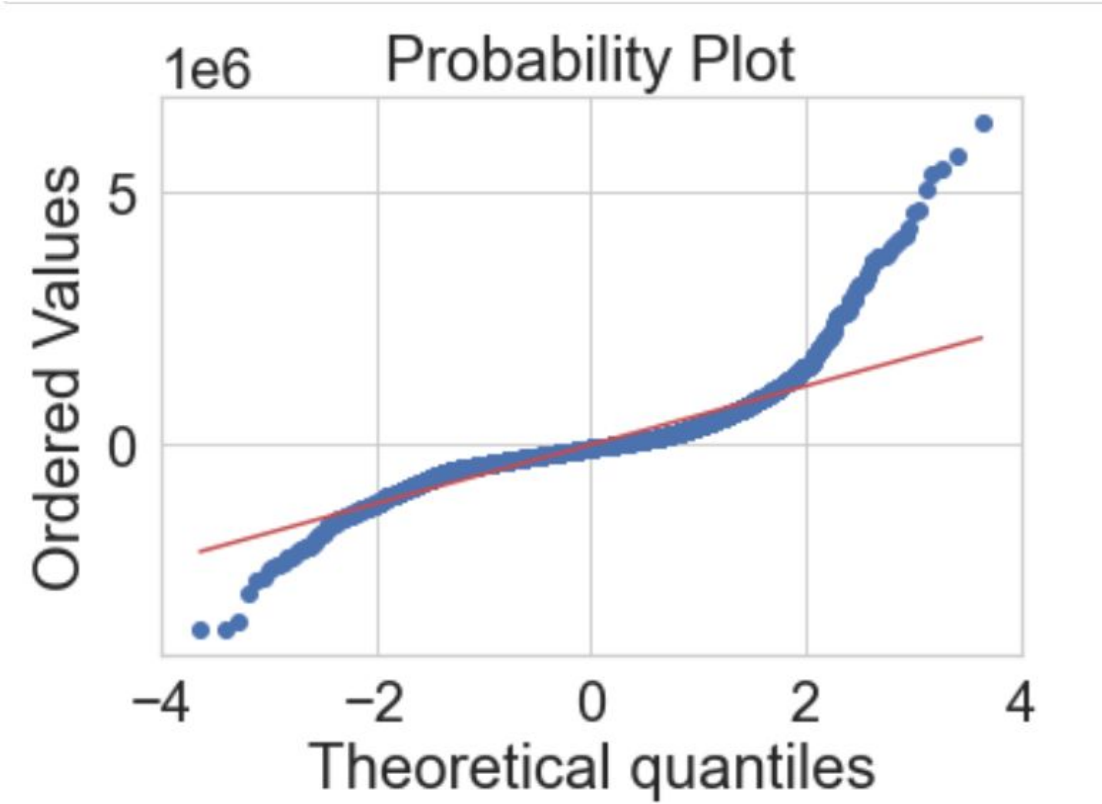
Model Performance Summary

- This model did pass the Normality of Residuals though with a normal distribution.



Model Performance Summary

- This graph tells me there are some issue with an outlier that should be looked into.



Business Insights and Recommendations

- The two variable that need special consideration when pricing used vehicles are the Kilometers Driven and the kind of Engine the car has.
- As the Kilometers Driven increases the price of the car comes down by 3.61 Lakh.
- As the engine gets bigger the price increases by 62.03 Lakh.
- Both of these do make intuitive sense because the value of a car that has been driven a lot goes down while the cars that have bigger engines generally are nicer cars.

greatlearning
Power Ahead

Happy Learning !