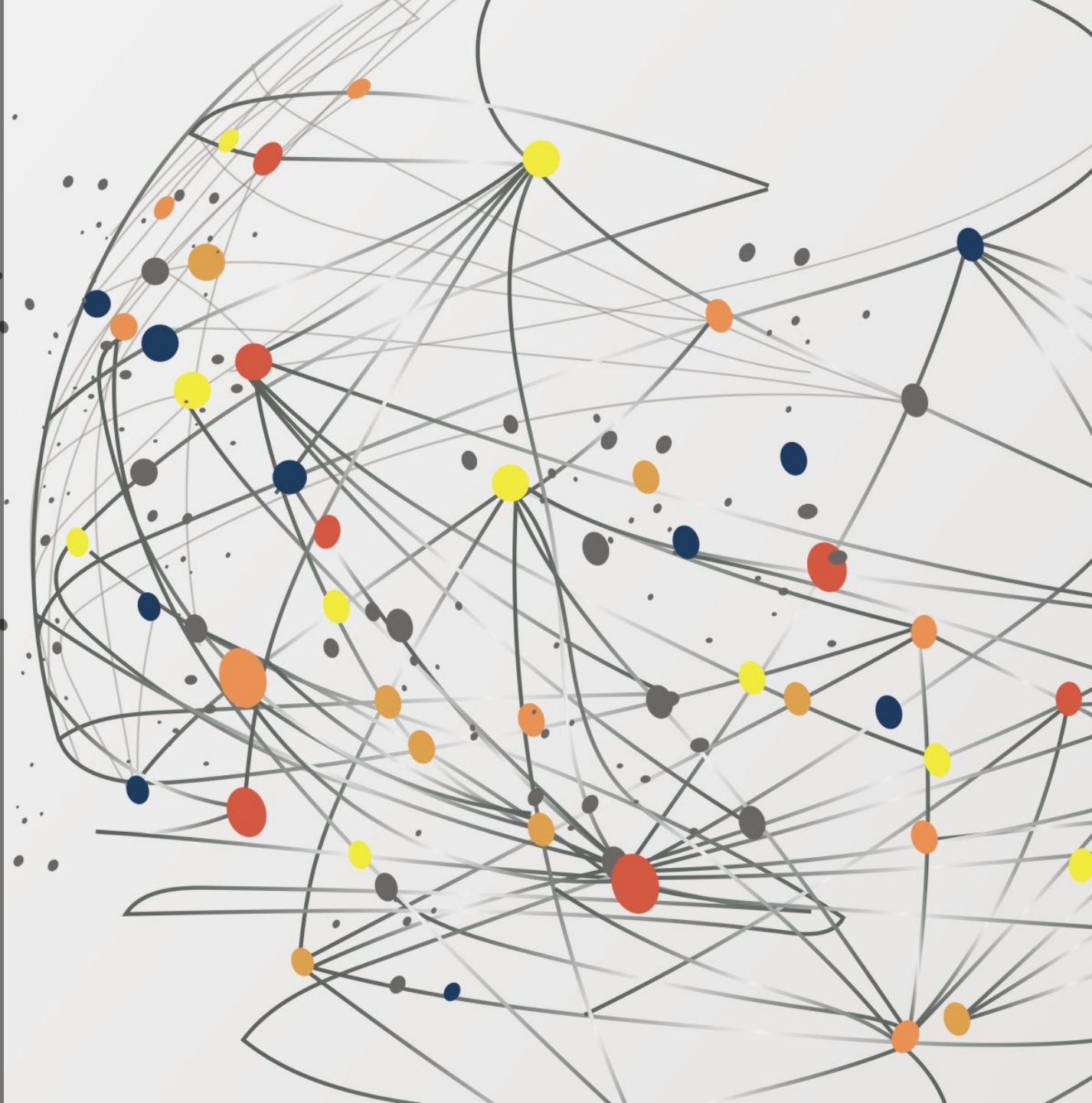
Building a Used Car Price Prediction System

-Jyothi H



Business Question

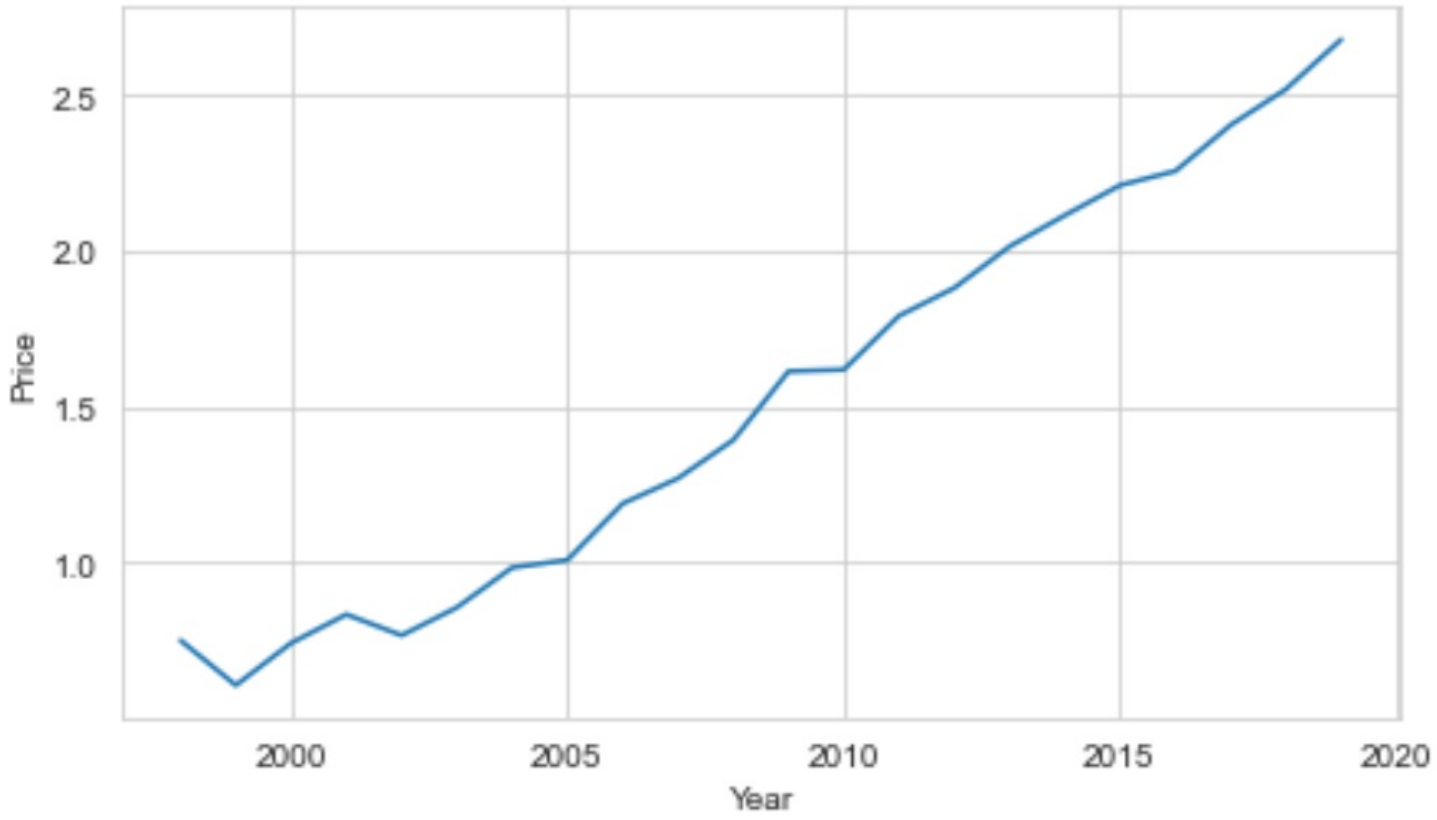
How do we accurately build a used car prediction system to effectively determine the market value of a used car ?

Key Questions To Address

The following questions need to be answered in order to effectively predict the price of used cars

- Does various predicting factors really affect the price of used cars?
- What are the predicting variables affecting the price of used cars?
- Does the year of the car model really effect the price?
- Does the number of kilometers driven really effect the price of used cars?
- What is the impact of engine power on the price?
- Is automatic transmission lined to higher price than manual transmission?
- What is the impact of fuel type on the price?
- Does changing multiple ownership have any effect on the price of the car?
- Does mileage have an impact on the car?

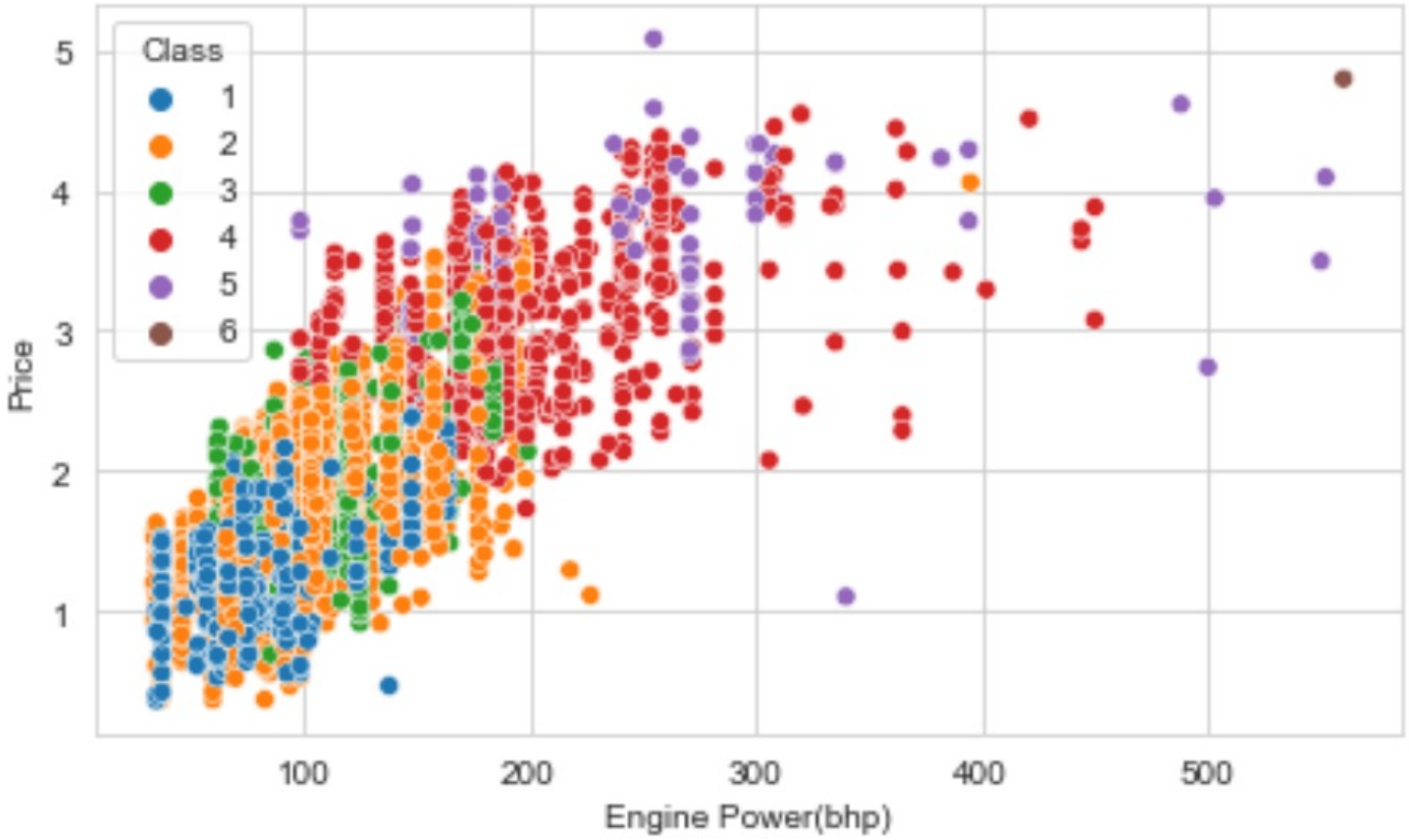
Correlation - Price Vs Year



As depicted in the graph, the price of the used cars tends to be positively correlated with the year of the car model.

Therefore, the year of the car needs to be accounted for in the pricing model of the used car

Correlation - Price Vs Engine Power Vs Car Class

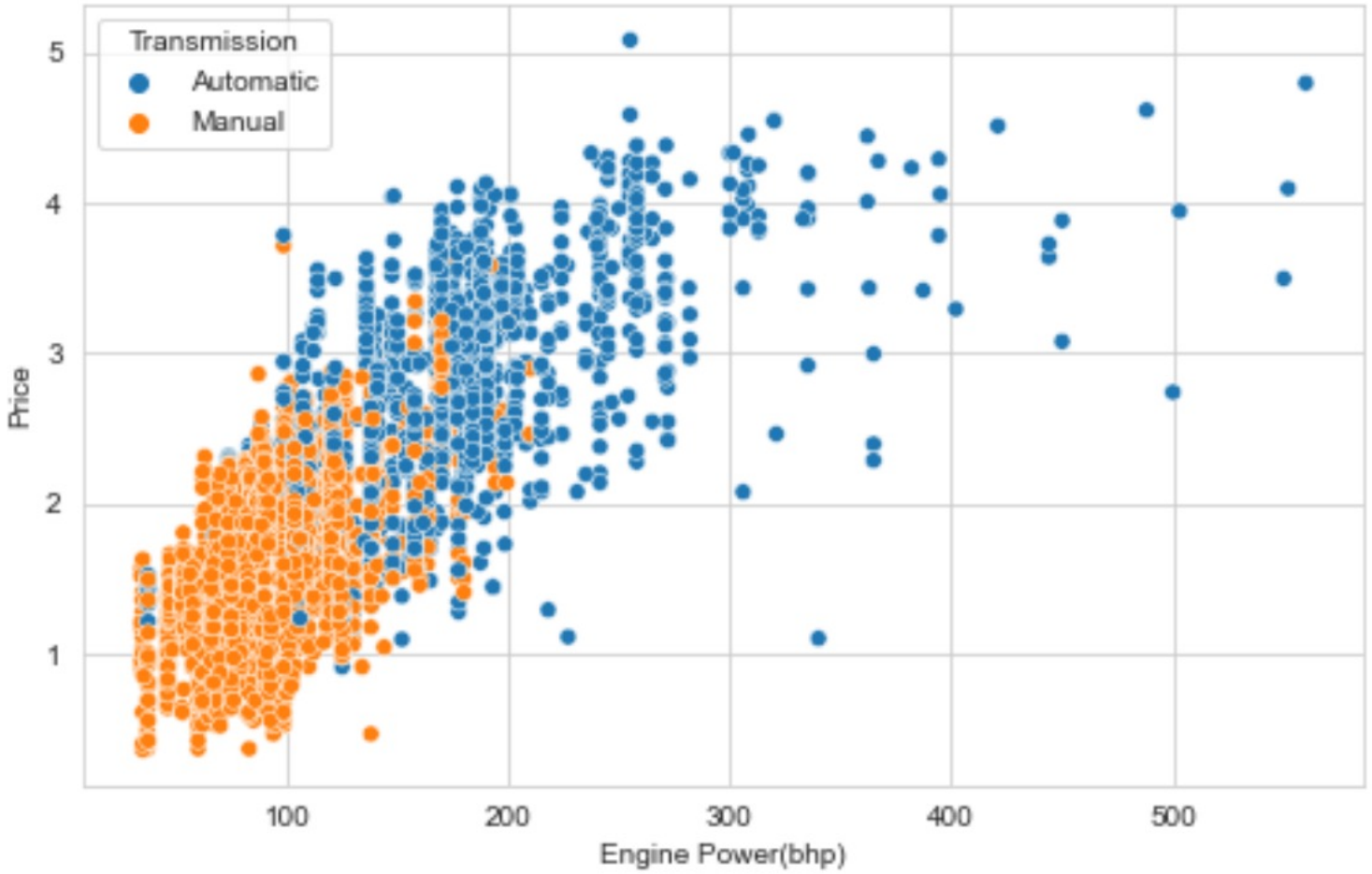


The graph shows that any increase in the engine power is associated with positive correlation with the price of the car.

Also, high-end cars (Class 6) tend to be priced higher than lower end of the cars.

Therefore, the engine power and class attributes need to be incorporated in the pricing model.

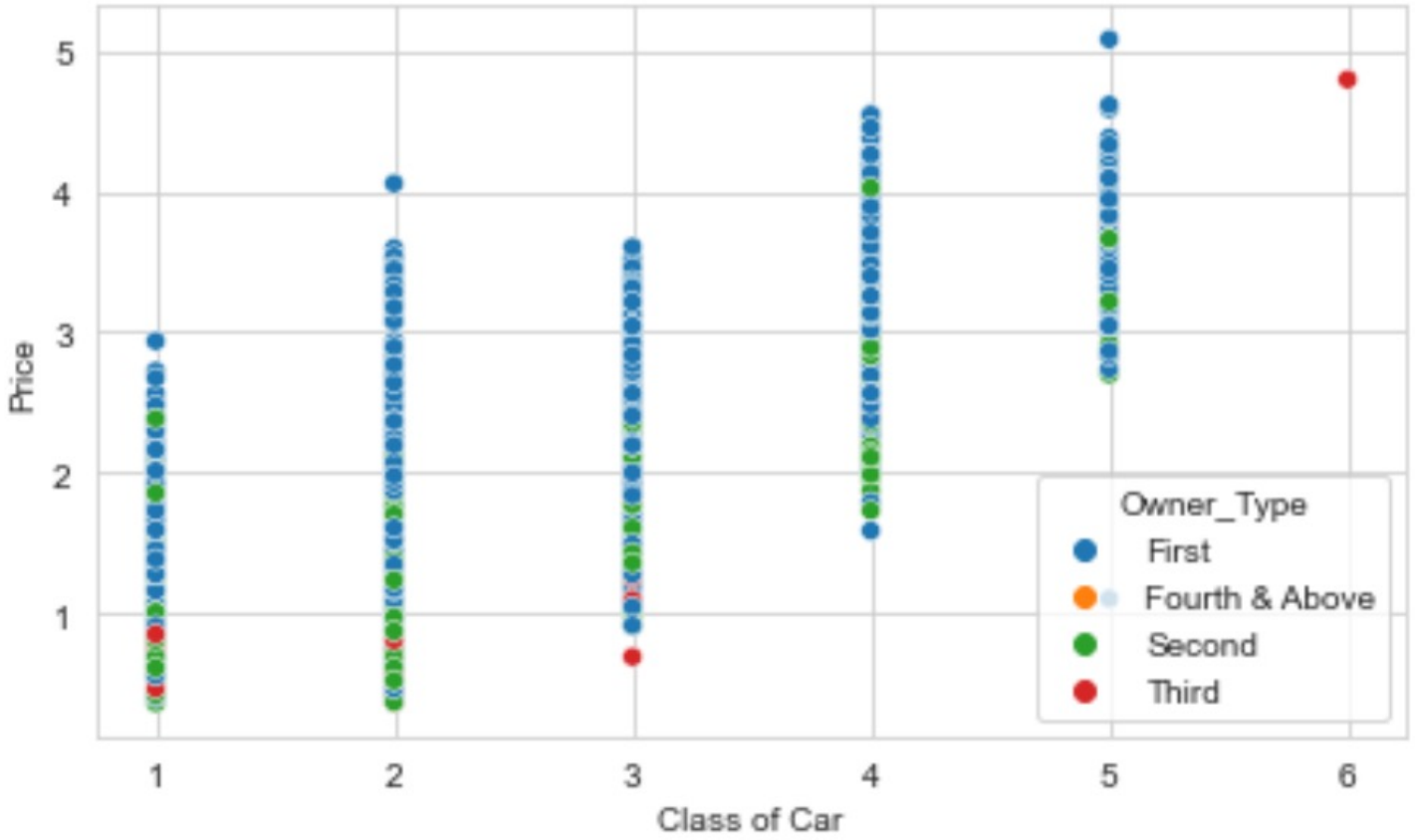
Correlation - Price Vs Engine Power Vs Transmission



Cars with manual transmission is associated with lower price in comparison to cars with automatic transmission.

Therefore, transmission should be a part of pricing strategy since the type of transmission is linked to the price.

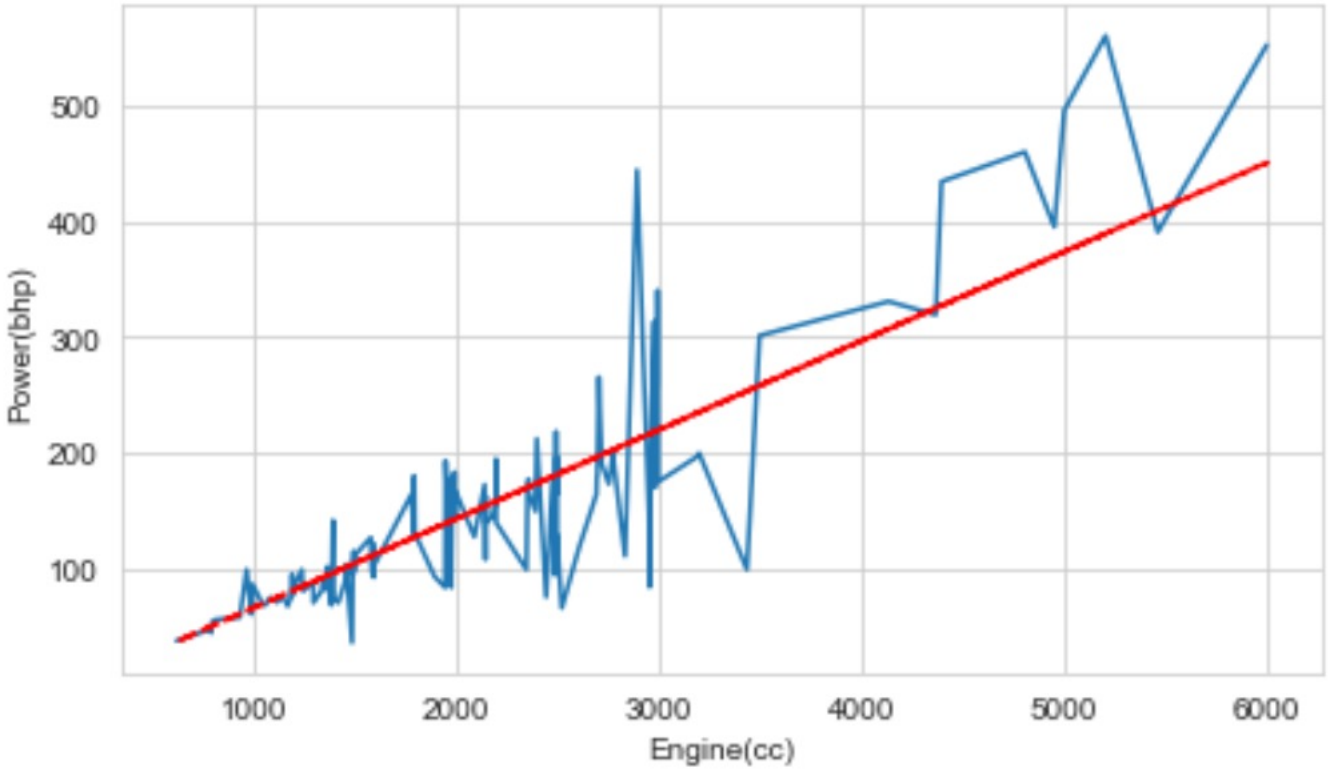
Correlation - Price Vs Car Class Vs Owner Type



On an average, first-hand owned cars tends to fetch a higher price than the second and henceforth, given a car of same class and all else equal.

Therefore, the owner type is associated with the price and needs to be a part of the pricing model.

Correlation – Engine(cc) Vs Power(bhp)



The graph here shows that engine power "Power(bhp)" and engine capacity "Engine(cc)" are highly correlated with each other. Therefore, we will only include one of these attributes in our pricing model since both are conveying same information.

VIF[1] for Engine is higher than VIF for Power. Therefore, we will not include Engine in the pricing model.

[1] Variance Inflation Factor (VIF) estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model

Linear Regression Model – Performance Metrics

Regression Model Measures	**Mean Absolute Error *(MAE)***	**Mean Absolute Percentage Error *(MAPE)***	**Root Mean Squared Error *(RMSE)***	**Adj. R-Squared**
Training Data	0.159277	8.884288	0.210376	0.917069
Testing Data	0.156532	8.836784	0.2208	0.918239

- In both the training and testing data, about 91% of the variance in the price is explained collectively by the independent variables (Year of the car model, kilometers driven, mileage, seats, engine power, transmission mode, owner type, and location). This makes the model is good fit.
- The MAPE on the testing data set suggests we can predict within 8.8% of the price of used cars.
- The RMSE, an indication of how close the observed data points are to the model's predicted values, of testing and training are comparable (with 21% for training data and 22% for testing data) .

Linear Regression Model – Magnitude of Impact (Coefficients)

Variables	Coefficients
Const	-192.6323
Year	0.0967
Seats	0.0185
Mileage(kmpl)	-0.0211
Power(bhp)	0.0055
Kilometers_Driven_log	-0.0565
Location_Bangalore	0.1028
Location_Coimbatore	0.0593
Location_Delhi	-0.0652
Location_Hyderabad	0.0798
Location_Jaipur	-0.0395
Location_Kochi	-0.0393

Variables	Coefficients
Location_Kolkata	-0.1993
Location_Mumbai	-0.0616
Location_Pune	-0.0475
Fuel_Type_Petrol	-0.2532
Transmission_Manual	-0.1105
Owner_Type_Second	-0.0498
Class_2	0.3034
Class_3	0.4208
Class_4	0.8285
Class_5	1.0107
Class_6	0.7119

Regression Equation :
Log(Price) = -192.6323 + 0.0967 * Year + 0.0185*Seats - 0.0211 * mileage + 0.0055*Power(bhp) - 0.0565*llog(Kilometers_Driven) + 0.1028 * Location_Bangalore + 0.0593*Location_Coimbatore - 0.0652*Location_Delhi + 0.0798*Location_Hyderabad - 0.0395 * Location_Jaipur - 0.0393*Location_Cochi - 0.1993*Location_Kolkata - 0.0616*Location_Mumbai - 0.0475*Location_Pune - 0.2532*Fuel_Type_Petrol - 0.1105*Transmission_Manual - 0.0498*Owner_Type_Second + 0.3034*Class_2 + 0.4208*Class_3 + 0.8285 * Class_4 + 1.0107*Class_5 + 0.7119 * Class_6

Note: All the above-mentioned coefficients are significant

Insights Summary - Primary Features Effecting Car Price

Features Impacting Price	Relevancy Rank	Magnitude (co-efficient)	Comments
Year	1	0.0967	One year increase in the year of the car model results in **9.67%** increase in price of the car, given all else remains the same.
Kilometers Driven (log)	2	-0.0565	A percentage increase in kilometers driven decreases the price by **0.56%**, provided all else remains the same.
Mileage	3	-0.0211	When mileage of the car increases by 1 kmpl, the price of the car decreases by **2%**, given everything else remains the same. This is because increased mileage is often associated with bigger cars, which tends to give a lower mileage.
Seats	4	0.0185	When a car seat increased by 1, the price of the car increases by **1.8%**, given all else remains the same. Often, an increased seat count is associated with increased car size, which in turn is associated with increased price.
Power	5	0.0055	When power increased by 1bhp, the price of the car increases by **0.55%**

Note: The above insights hold true only for petrol and diesel-based cars

Insights Summary - Additional Factors Effecting Car Price

Other attributes to consider when setting the price of the used cars

- Fuel Type – Fuel type attribute for petrol has a negative coefficient implying that petrol-based cars are associated with lower car price than diesel-based cars.
- Transmission Type - Manual cars have a negative coefficient compared to automatic cars, implying that manual cars are associated with lower car price than automatic cars.
- Owner Type – Owner type for second owned cars has negative correlation compared to cars owned by only one owner historically. This implies that a pre-owned cars with 2 past owner is linked to lower price compared to pre-owned car with only 1 owner, historically.
- Class of the Car [2]– Class 2 tends to have a positive correlation compared to class 1 , and this magnitude of correlation for classes higher up continues to increase as we move to higher class of car. This implies that cars of high-end brand tends to be associated with higher price than low end cars, which Is obvious.

Note: *The above insights hold true only for petrol and diesel-based cars*

2 – Class 1 car is associated with low end car brands and higher number class is associated with high-end car brands

Potential Benefits Of Accurate Pricing Model

- Risk Mitigation – An accurate pricing model allows the company to set appropriate buying and selling price for a particular car, thereby protecting them from any loss on sale arising from incorrect prices.

- Accurate Profit Margin Predictability - Clarity in the right market price allows the company to set a predetermined buying and selling price, thereby helping them in turn set the margin in a predictable way.

- Improved Accuracy in Financial Forecasting - Reduces the variability in financial forecasting and related financial metrics.

- Ease in Securing Investors – A lower volatility in prediction and forecasting, lowers the company's perceived risk by capital markets, thereby helping the company secure funding easily.

- Lower Cost of Capital – A lower volatility in financial forecasting and a positive profit margin , eventually translates to lower cost of conducting the business (arising from lower interest rates) , eventually resulting in an increase in market capitalization and growth.

Recommendation

The model described in the previous sections can predict with a 91% accuracy, the price of used cars, given a set of features that are described by the regression model. Therefore, I recommend leveraging this model using the below mentioned key factors , along with their corresponding coefficients described earlier) to determine its pricing

- Year of model
- Engine power (bhp)
- Mileage
- Kilometers Driven
- Transmission Type
- Car Class
- Fuel_Type
- Seat Count
- Location