Group 2
MIST 7770: Business Intelligence & Analytics
Dr. Rios
11/27/2019

Group 2: Punjab Used Car Sales Regression & Analysis

Context and Question

The given report investigates the key determinants of pricing for the used car market in the southeastern region of Asia. Growing at a staggering value of about $62.31 billion, the knowledge of the car market must evolve with the exponential growth. Due to the COVID-19 pandemic, supply chain constraints caused backlogs in manufacturing and materials, resulting in an uptick in the demand for used cars. Industries that work closely in the automobile industry include industrials and technology. We have gathered a variety of different variables that act as key determinants when dictating the pricing of a vehicle. By employing regression models, the study seeks to discern the significance of key features, such as mileage, model year, and brand in determining the price of used cars.

We're using data from Kaggle, including information on 46,000 used car sales. We're particularly interested in variables that deal with the wear and tear of the vehicle due to usage (mileage), the age of the vehicle (model year), and the brand of the car (manufacturer). The age and brand of the vehicle will be closely evaluated, as the age of the vehicle entails the technology, amenities, and engineering of the vehicle. The brand frequently reflects the prestige and goodwill a manufacturer possesses. The conflict we seek to remedy involves the efficiency at which a dealership can operate, as well as the clients looking to purchase a used vehicle. With the great influx of buyers, dealerships must be aware of how they can maximize revenues and improve inventory turnover. Knowledge of what features hold a vehicle's value can allow dealerships to locate and purchase vehicles that will provide greater profits. Dealerships also can better identify the optimal features to sell when targeting clients, and therefore they can hold vehicles in their inventory that are in high demand. Clients of Punjab also stand to benefit from the study. Those wishing to sell or purchase a vehicle from a dealership frequently lack the knowledge to be rightfully rewarded for the amount they are paying or receiving upon their interaction with a dealership. Dealerships are known for either lowballing when purchasing a vehicle or are frequently found guilty of having clients pay more upon negotiations. If individuals had knowledge of what features provide value to a vehicle, they could better formulate a plan of action upon arrival at the dealership, leaving both sides of the transaction feeling better.

Our plan is to use regression models to analyze all the datasets to come up with a course of action to address our current conflict. This will allow us to figure out which features of a car play a significant role when it comes to pricing. In the end, we hope to create a way to predict

used car prices that's easy for everyone to understand and use, whether they're selling or buying a used car.

Data and Variables

Where and How We Obtained the Data:

We sourced our dataset from Kaggle, a widely recognized platform for data science and machine learning. This extensive dataset encapsulates 46,000 transactions of used car sales across the Middle East and Asia, rendering it a robust and comprehensive foundation for our regression and models.

Adequacy of the Data:

Our selection of this dataset aligns seamlessly with the requirements of our project, as it delivers a thorough perspective on the used car market within the pertinent geographical region. The volume of observations, totaling 46,000, guarantees a statistically significant foundation for our analysis. The dataset's variable richness enables exploration of the factors influencing a used car's pricing in the selected region. Offering a comprehensive coverage of the automobile industry in our region, it encompasses a diverse array of brands, models, and vehicle ages, providing a holistic view for our analytical models

Description of Variables:

Our dataset includes ten useful variables:

- Mileage: The total distance the car has traveled since its creation.
- Manufacturer: The brand or company that produced the car.
- Model: Specific model of the car.
- Model year: The year each vehicle was manufactured
- Color: The color of the vehicle.
- Location: Geographical location where the car is being sold.
- Engine Type: Type of engine (e.g. petrol, diesel, hybrid).
- Body Type: The style or body shape of the car (SUV, Sedan, etc.).
- Transmission: Type of transmission (manual, automatic).
- Condition: Overall condition of the car (new, used, etc.).
- Registration Status: Evaluate whether the vehicle is registered or not registered.

Target and Predictors:

In our research, we are examining the intricacies of the used car market, with a particular focus on understanding the determinants of sale prices. Our objective is to explain the influence of key variables, such as mileage, model year, and manufacturer, on the final transaction price of these vehicles.

Variable Types:

Within our dataset, we encounter a diverse array of data types. Numeric variables, like mileage and model year, coexist with categorical variables such as manufacturer, color, and engine type. It's noteworthy that the absence of date-related variables simplifies the dataset's structure, allowing us to concentrate on the essential elements in our analysis.

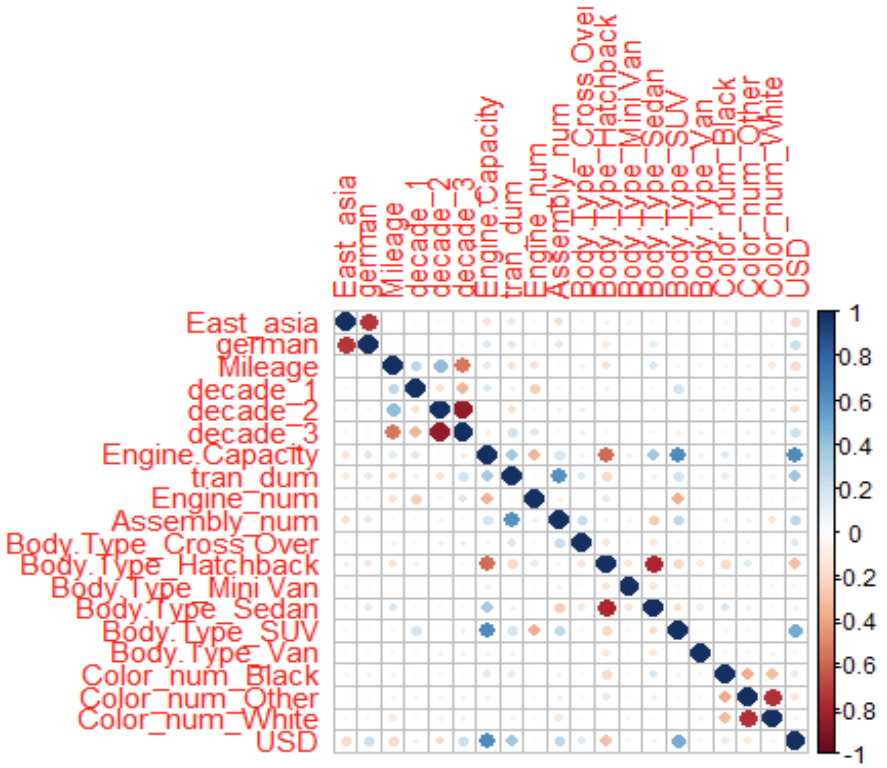
Stats Overview:

A brief exploration of statistics enables us to gain insights into the underlying patterns of each variable. Utilizing metrics like averages, medians, and measures of dispersion, we aim to offer a nuanced understanding of central tendencies and variability. However, it's crucial to acknowledge that the ensuing statistical insights are illustrative, given our limitations in direct data manipulation.

Observations:

Our dataset contains a substantial 46,000 entries, significantly surpassing the conventional threshold of 1,000 for robust statistical analysis. This abundance of observations provides a solid foundation for our comprehensive exploration of the dynamics in the used car market.

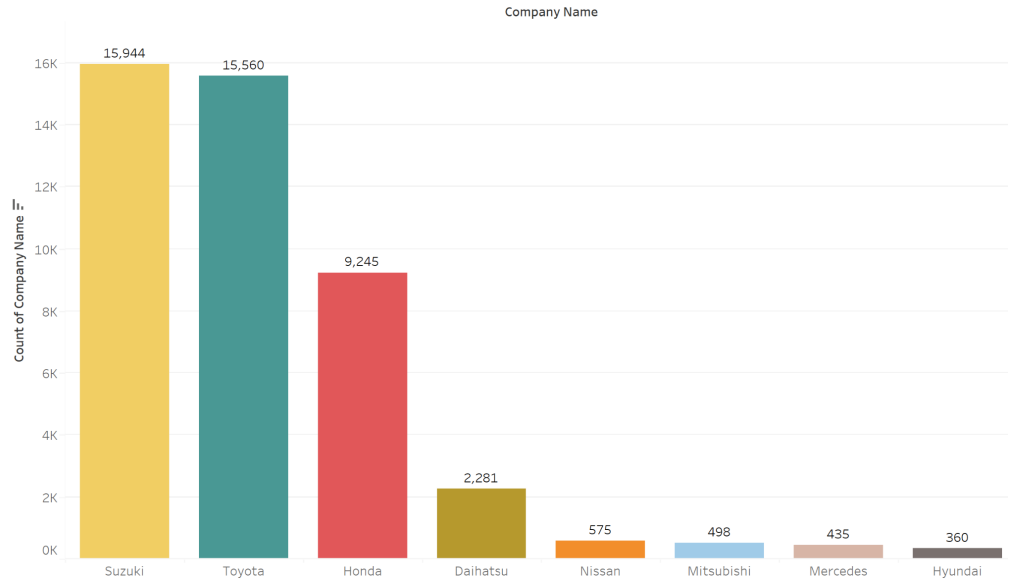
Correlation Matrix:



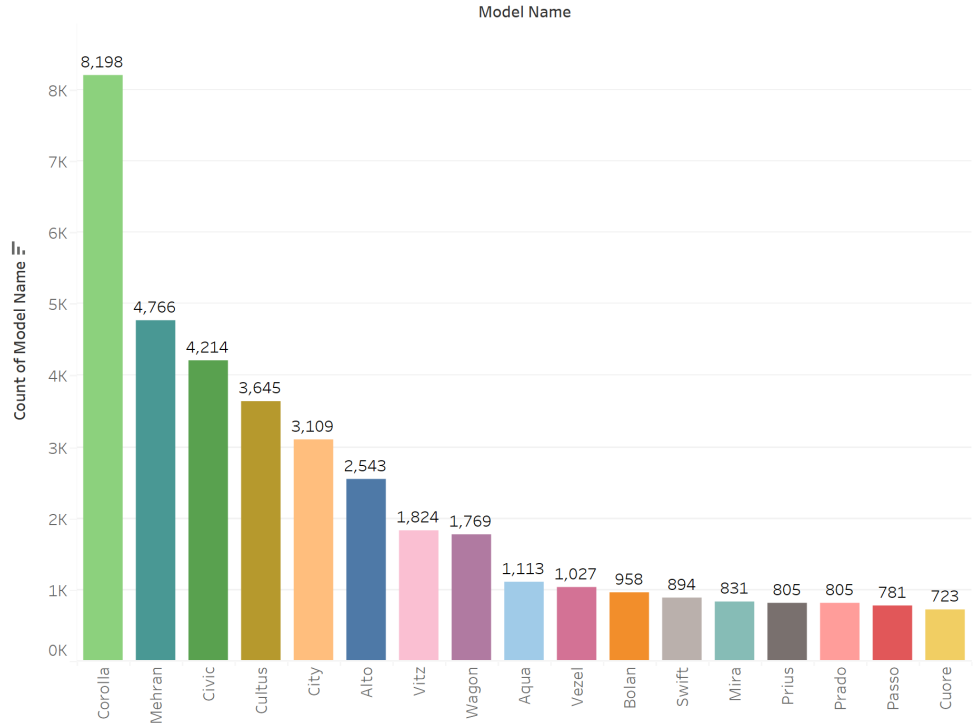
Descriptive Visualizations

*Descriptions are under the “Descriptive Visualizations” section in Appendix

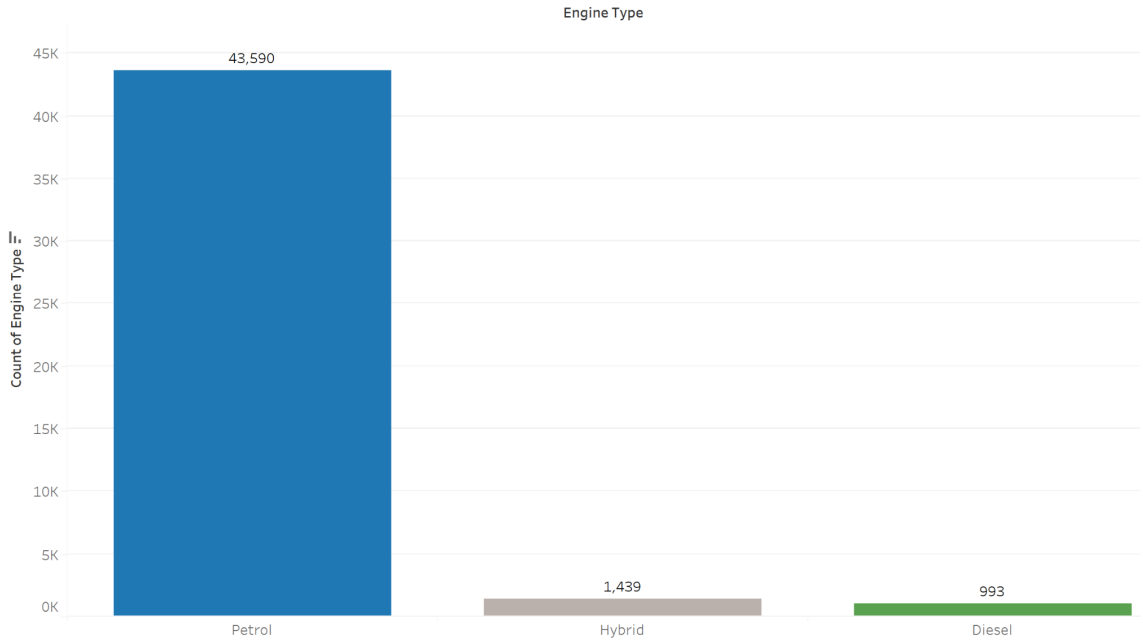
1. Count Make



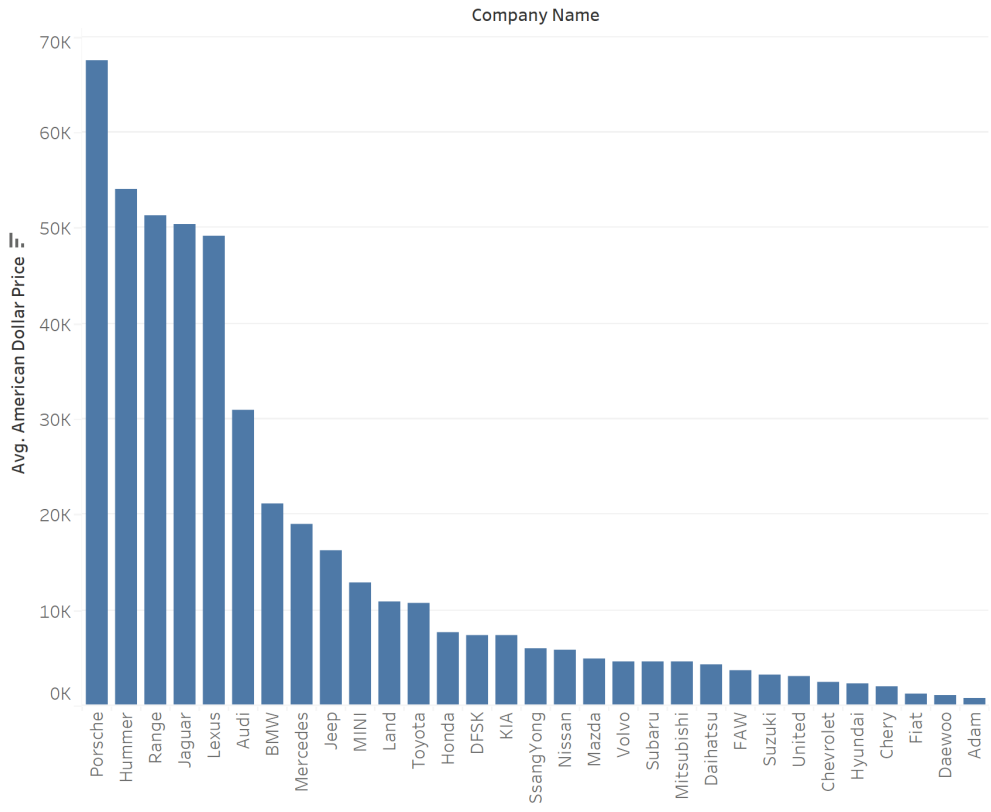
2. Count Model



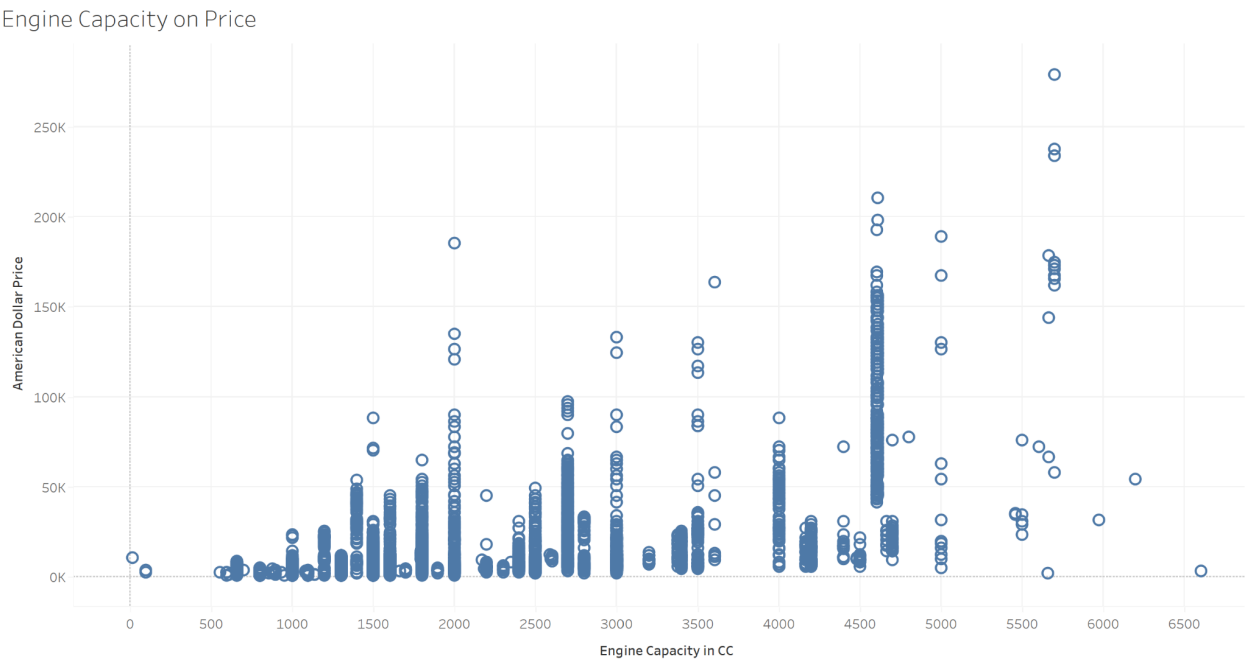
3. Count Engine Type



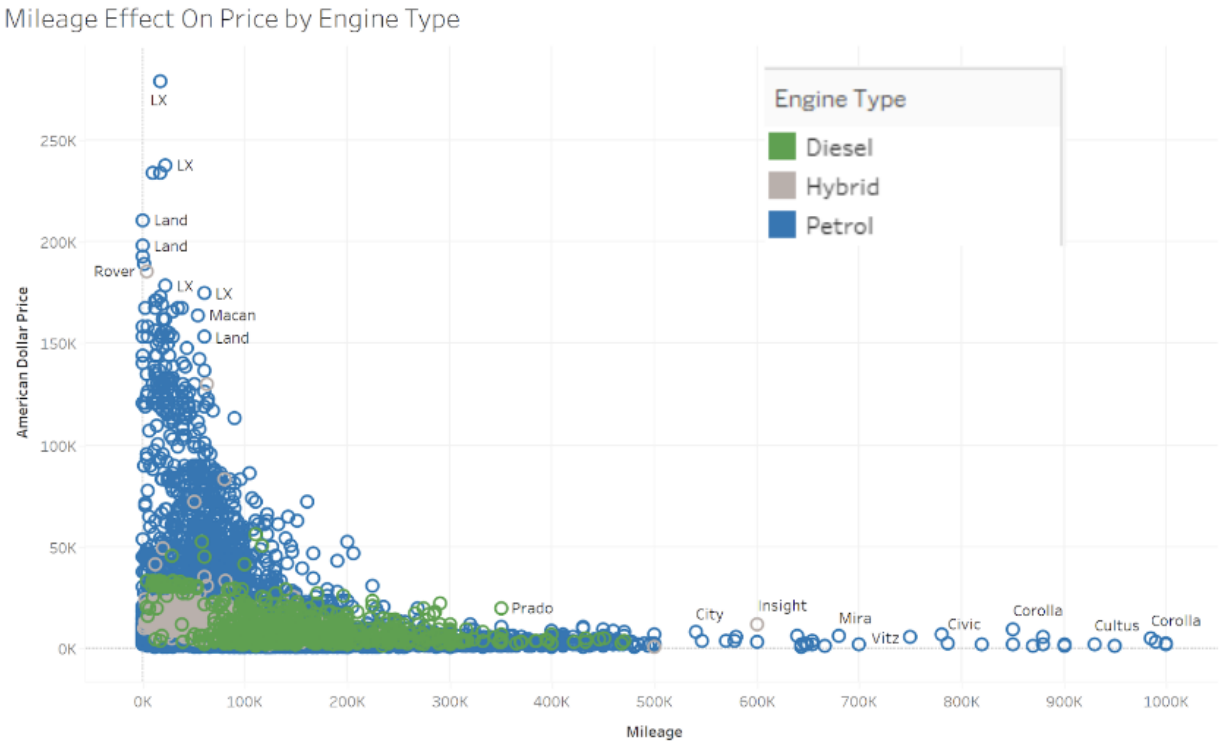
4. Make Effect On Price



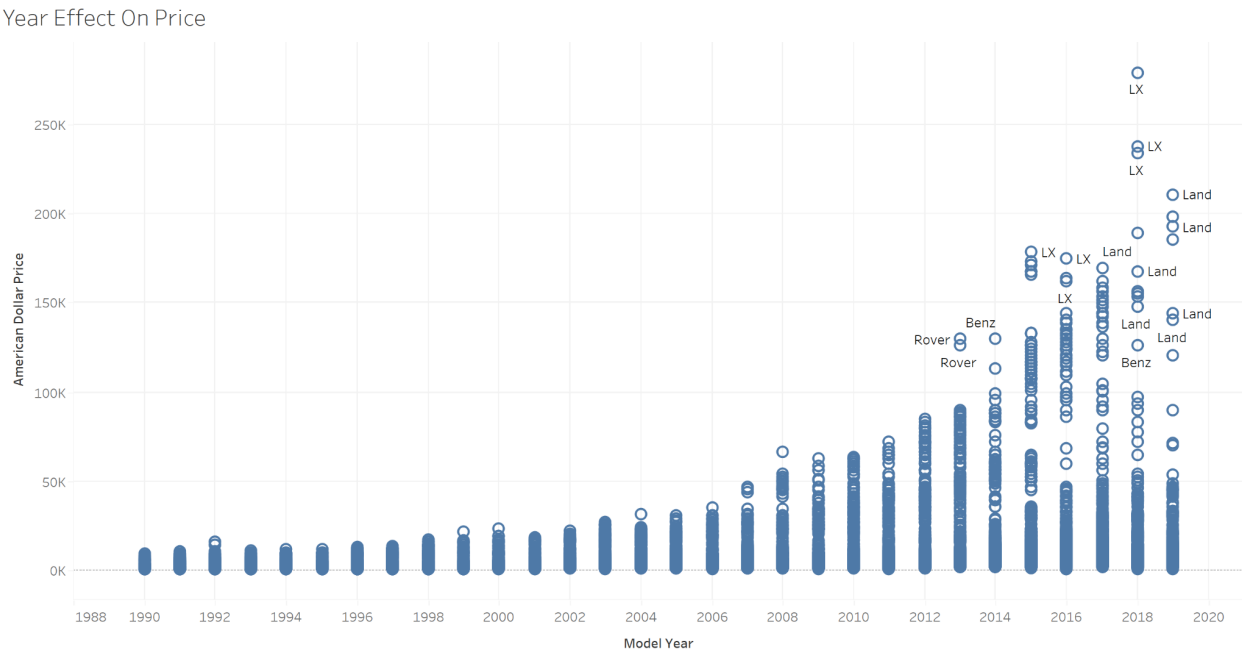
5.


Engine Capacity on Price
250K
200K
150K
100K
50K
0K
American Dollar Price
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
Engine Capacity in CC

6.


Mileage Effect On Price by Engine Type
Engine Type
Diesel
Hybrid
Petrol
LX
LX
Land
Land
Rover
LX
LX
Macan
Land
Prado
City
Insight
Mira
Vitz
Civic
Corolla
Cultus
Corolla
250K
200K
150K
100K
50K
0K
American Dollar Price
0K
100K
200K
300K
400K
500K
600K
700K
800K
900K
1000K
Mileage

7.


Year Effect On Price
American Dollar Price
250K
200K
150K
100K
50K
0K
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
Model Year
LX
Land
Benz
Rover

Method

Our method is an iterative process starting with identifying the key features contributing to car prices. The first round included all features except for the necessary omission to give the model baseline features came from assessing the correlation of many features to the target, price. We chose variables that have a high correlation with the target while not having a high correlation with the other independent variables. From here, we conducted a multiple regression. This will serve as a good baseline model to see the most important features, and to test our prediction. By trying different combinations of variables, we will be able to directly see which To assess this portion, we will look at variables that contribute most to the linear model while also statistically significant at 95% confidence. For each of these models, we will measure their success based on Root Mean Squared Error (RMSE). Since we are predicting price, RMSE will give us the average distance between the predicted values and the actual values, and it will be the best method to assess how our models perform.

For another prediction attempt, we use a lasso regression. The benefit of this model is the automatic feature selection. This model will assign higher weights to variables that contribute more to the model, and it will skip over variables that don't contribute much. We used this to reduce the dimensions of the model, considering we have 13 features in our first regression. This will also be assessed by RMSE.

For our final model, we will utilize a random forest with 500 decision trees in order to find an optimal model. The random forest utilizes the power of decision trees but produces them en masse to address the propensity to overfit decision trees. One of the primary methods used to inject randomness into the many different decision trees of a random forest is Bootstrapping. This is a resampling method with replacement. This means that each tree will have a different sample from the original dataset. These samples can have variations such as missing/repeated values, but each sample will be the same size as the original data. This randomly generated dataset will help with generalization. An individual decision tree might have an overfitting problem but these issues will average out in the results. This will serve as the primary model, and we anticipate it to yield the best results of the three.

We create a training and testing split of 80-20 to ensure that the best-performing model can be generalized to new data. We did this for each model, to ensure that whichever performed best could be generalized and handle new data. The training/testing split will help with any overfitting problems. In addition to this, the Lasso regression will cause any variables that are not contributing to have a coefficient of zero. This is called an L1 regularization. While we do not have an unreasonable amount of variables (i.e. more variables than observations), it will still be interesting to see how the model performs. Lastly, the random forest is a method to reduce overfitting. Any overfitting will be averaged out.

Results

Our results will be split up into two categories, features and prediction. Regarding features, we see that many of our variables were statistically significant. We also look at the coefficient to signify economic significance and get a more holistic view of the impact of the features. From our multiple regression, we see that engine capacity, van style vehicles, imported assembly vehicles, petrol/hybrid engines, black & white color (as opposed to the "other" category), and high mileage contribute greatly to car price in this dataset. All of these features are statistically significant at 95% confidence. Notably, color, engine capacity, imported assembly, and petrol engines have the most significant positive impact on the price of a vehicle.

As for predictions from other models, they can be viewed in our appendix. This paper will only detail the results from our best model.

In our final model, we used multiple random forest models. For each model, we ran predictions and evaluated them using RMSE and R-squared.

Random Forest - 500 Trees, all Variables:
RMSE: 7358
R-Squared: 0.9329

Random Forest - 500 Trees, all Variables - Cars below $8000 excluded:
RMSE: 10033
R-Squared: 0.9087

Random Forest - 500 Trees, all Variables - Cars above $60000 excluded:
RMSE: 3358
R-Squared: 0.9076

Random Forest - 128 Trees, all Variables:
RMSE: 8251
R-Squared: 0.9394

Random Forest - 128 Trees, Engine Capacity + Mileage + Transmission Type:
RMSE: 14442
R-Squared: 0.774

After running these variations of the Random Forest Model, we found that running it with the upper range of cars excluded yielded the best results. The usefulness of this model is that it can be used to predict the price that a car can be listed based on the characteristics that it has.

Limitations

Limitations

1.) A limitation experienced in our study included the geographic location in which the dataset was collected. The market for vehicles varies depending on the region or country in which it operates. The strength of the economy, consumer demand, and the supply of cars greatly influence the pricing of a vehicle. Many of these variables failed to be included within the dataset and our model. Cars in stronger economies tend to resell for greater value, and the consumer demand for a particular vehicle also fluctuates. A specific model or brand in Europe may be more sought after than another brand or model in the Middle East. Therefore, in scope, our model is specifically catered to dealerships and clients solely in Punjab.

Additional Useful Data

- A useful variable needed to benefit our model included whether the vehicle was sold by an unorganized dealer or an organized dealer. Over the last few years, there has been a transition in the Middle East and Southeast Asia towards unorganized players. Organized vendors had become completely consolidated into very few retailers and distributors selling vehicles at a larger scale. Many small used car dealerships have created strategic alliances to sell used cars on a grander scale, allowing for buying and selling to customers to be done at a greater rate. The implications of the type of dealership can impact the price of a vehicle and provide valuable insight into the composition of our data. As unorganized vendors continue to grow, it would have been helpful to see the impact this variable has on pricing.
- Manufacturer Suggested Retail Price (MSRP) would have been incredibly useful for the dataset by providing insight into whether users are receiving a fair price for their vehicles. An MSRP variable would have also provided valuable insight by acting as a comparison to the actual price the car sold for. Testing the correlation between MSRP and other variables could have provided insight into what manufacturers deem as “valuable” when pricing a car.
- Data further describing whether the vehicle is of “luxury” or not would have also been incredibly useful for greater recognition of whether or not the prestige of the vehicle is a key determinant of price. A car brand or a specific model can be identified as luxurious, but quantifying what is considered as “luxury” is subjective to that of the buyer and seller. For the region of Punjab to come to a greater understanding of what luxury may look like, it would be helpful to know what is classified as a luxury vehicle in that region.

Causality constraints:

There are a variety of causality constraints that prevent a firm inference from our findings. Constraints applicable to the study include ethical constraints and confounding variables. As for ethical constraints, conducting business with an organized dealership versus an unorganized dealership can lead to drastically different outcomes in pricing. Unorganized dealerships have a greater chance of unethically acting in favor of the dealership's pricing.

Financial records, accessible information, and overall operations impact a customer's purchasing power at an unorganized dealership. The lack of accessible information and well-maintained financial records make it difficult to interpret whether the selling price is justified. Due to the dataset not possessing further information on the location of the sale, justification of the sale becomes increasingly challenging. If a dealer were to act unethically, the price of a vehicle could be inflated or the value could be undercut. Therefore, this results in an error in the accuracy of many of the variables measured by the model. Confounding variables can also impact the inference about causality in our experiment. Variables such as the condition of the car can alter pricing, regardless of the other determinants. Damage to the interior and exterior of a car drives down the price of a vehicle, and without knowledge of the car's condition, it makes justification for some of the sales difficult to make.

Recommendations

Our recommendations coincide with the nature of our findings. Regarding the features of vehicles, we suggest that the automotive company continues to invest in vehicles that contribute to higher sale prices (see referenced features in the results section). In addition to this, monitor the features of the vehicles that sell at an optimal price and receive high demand. Understanding the price and demand is essential for selecting inventory. Our visualizations show that the Toyota Corolla is the vehicle purchased the most. With this fact and analysis of the contribution of certain features, there can be more insightful planning around the inventory of cars with high demand but also the right features to provide a high selling price for the dealership. This list of features can also help floor employees with their sales strategies and their understanding of what will increase the bottom line.

As for our prediction tool, continue to gather data on vehicles, and delve into more sales data to better inform the model. Elements of the vehicle such as luxury status will help the model perform better. Our model, with limited data, was able to predict car prices with substantial accuracy. The use of a random forest model is essential in this pursuit, and more data will only make the accuracy more robust. Our model can increase the knowledge of salespeople to find an optimal price to sell a vehicle. This optimized model can set a solid foundation for selecting inventory, projecting sales, and having a solid outlook on the financial future of the company.

Reproducibility and Submission

All work is located in a GitHub repository which is attached to the appendix as a hyperlink.

https://github.com/charlmc/MIST7770-BI-Project/tree/main

Appendix

Other Models

Our other models and algorithm outputs will be included in the project GitHub repository linked here (https://github.com/charlmc/MIST7770-BI-Project/tree/main). In the HTML file or PDF file is the output which does not require the computation of the models, and also includes any output or visualizations for our project.

Descriptive Visualizations Explanation

https://github.com/charlmc/MIST7770-BI-Project

1. This visualization shows the count of each make of a car in the data set. From the bar chart, we can deduce that Suzuki (15,944), Toyota (15,560), and Honda (9,245) are the most common makes of cars in our used car market. These makes have many different models spanning many different categories of body, engine, miles, and luxury which will be taken into account in the regression and explanatory models. It is important to also note that some highly luxury brands (i.e. Mercedes, BMW, etc) are low in the count but make up a high section of the luxury market.
2. This visualization shows the count of individual models in the data set. The visualization is cut off to show the top 17 models by number in the market (count). The most common car is the Toyota Corolla (8,198), Suzuki Mehran (4,766) which is a rebadged version of an Alto, and the Honda Civic (4,214). These are also very common cars in the world (not just in Pakistan and Asia) which is also obvious in our data set in the Pakistan and Middle Eastern / Asian markets.
3. The count of engine types shows that in our data set, the most used cars are petrol engines at 43,590 which leads to hybrid and diesel engines by a very large margin. The count is not determined in finding the price but for the purpose of knowing how our data set looks. There is a correlation between price and engine type which can be observed in visualization 6, which shows that petrol engine cars often have a greater range of price (most likely because more luxury cars are not hybrid or diesel).
4. For visualization 4 we are showing the average price of the cars for each individual make. In this visualization, we can see that the common luxury brands have a higher average price for their used cars sold. We can see that the top most expensive makes are Porsche, Hummer, Range Rover, Jaguar, Lexus, Audi, BMW, and Mercedes which is common among other markets as well due to scarcity, build quality, and demand for the cars of these makes.

5. For visualization 5, we are showing the look of engine capacity in CC (cubic centimeters) versus the price of the car. Looking at the graph, it is clear that the higher capacity of the engine will most likely mean the car will have a higher price. Higher engine capacity is often related to the speed, luxury, performance, or work power of the vehicle. An example of this would be a Porsche 911 GT3 would have a much higher engine capacity than a Toyota Prius which will also be reflected in the price of the vehicle. You can also observe spikes in the price at 2000cc which can be associated with boxer engines or smaller v6 engines which are often in smaller sports cars and will also produce a higher price compared to other vehicles like an old truck with a larger engine capacity.
6. Mileage of the car compared to the price of the car is an interesting visualization that shows a clear correlation between the newer cars (mileage-wise) and older more driven cars. Usually, in almost all cases, cars with lower miles will sell for more money. Also, it should be noted that more luxury cars or performance cars will be driven less by their owners to preserve the quality and price of their vehicles, which can also be seen in the visualization. We can also see that diesel and hybrid cars do not have the same range of price that petrol vehicles do, and often will not have as many miles as petrol cars like the Corolla or Civic.
7. Year-on-price visualization shows the newness of the vehicle on price. In the visualization, we can see that newer cars hold a higher sales price, which is often the same in other markets. For our regression and explanatory models we have changed the vehicle year variable to bins for each decade to make the regression and interpretation of results easier for the client.