Title: (Proposal: Understanding Used Car Prices: Insights from CarDekho)

**Introduction**

**Motivation**

Nowadays, as technological advancements revolutionize industries, digital car trading platforms have emerged as vital marketplaces, transforming the automotive market and the way people buy and sell second-hand vehicles. Notably, CarDekho.com is a well-known car trading platform based in India. The primary motivation for this analysis stems from the growing importance of platforms like CarDekho.com in connecting buyers and sellers. In particular, for buyers, understanding the relationship between car traits and pricing can facilitate purchasing decisions, and help assess whether a car is priced fairly. Meanwhile, for sellers, the insights into pricing trends can help attract more potential customers while still setting competitive selling prices, determining the optimal price for their listings. Thus, this study aims to investigate a deeper understanding of these relationships by addressing the following research questions:

* What are the relationships between the selling price and continuous covariates (year, km\_driven)?
* What categorical variables are associated with the selling price, and how do these variables influence pricing?
* What regression model can effectively explain and interpret the relationships in the data?

By answering these questions, we aim to develop a statistical model that contributes to the car trading marked by offering data-driven insights.

**Dataset**

The dataset analyzed in this report contains information about second-hand cars provided by CarDekho.com (<https://www.cardekho.com/>). The dataset was published on Kaggle (<https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho/data>) on June 26, 2020.

The following is a description of the variables found in the dataset:

|  |  |  |
| --- | --- | --- |
| **Variables** | **Type** | **Descriptions** |
| name | categorical | the name of the car; start with brand name which is a possible factor |
| year | continuous | the year of the car when it was bought (in years) |
| selling\_price | continuous | the price of car when it was sold (in Indian Rupee (INR)) |
| km\_driven | continuous | the distance that the car is driven (in kilometers) |
| fuel | categorical | the fuel type of the car; 5 levels (CNG/Diesel/Electric/LPG/Petrol) |
| seller\_type | categorical | the type of the seller; 3 levels (Dealer, Individual, Trustmark Dealer) |
| transmission | categorical | the type of gear transmission; 2 levels (Manual, Automatic) |
| owner | categorical | the number of previous owners; 5 levels |

By setting selling\_price as our response, we have a total of 5 categorical variables and 2 continuous variables.

**Exploratory Analysis**

**Preliminary analysis:**

To answer our research question, we commenced with an exploratory data analysis (EDA). From this, we aim to understand the structure and characteristics of the dataset while identifying underlying patterns, potential outliers, and relationships among the variables.

This analysis used the following R packages: dplyr, ggplot2, tidyr, tidyverse, readr, corrplot, and patchwork.

As previously determined, the dataset contains 5 categorical and 3 numerical variables. A summary of the dataset provides a preliminary overview of the variables in the data (Appendix 1.01).

From the summary, we observe several key findings in the continuous variables:

• The variable **year** ranges from 1992 to 2020, with a median of 2014, indicating that most vehicles in the dataset are relatively recent (modern).

• The variable **selling\_price** ranges from 20,000 to 8,900,000, with a median of 350,000 and a mean of 504,127, suggesting the presence of some high-priced outliers

• The variable **km\_driven** ranges from 1 km to 806,599 km, with a median of 60,000 km and a mean of 66,216 km, again indicating a large amount of variance in vehicle usage.

In terms of the categorical variables, we determined that fuel has 5 categorical types, transmission has 2 categorical types, owner has 5 categorical types, and seller type has 3 categorical types.

Furthermore, we identified that the brand of vehicles might also be an important factor. Therefore, we created a variable called **brand** by extracting the car manufacturers’ names from the name column in the dataset. This variable allows us to identify trends or patterns that are specific to certain manufacturers. In addition, this simplifies and reduces the redundancy in the dataset, making it more interpretable and easier to work with in the statistical analysis phase.

**Note:** No missing values were found in the data. Thus, it was not necessary to perform any additional changes to the data to handle NaN values.

**Visualizations:**

To begin, we first created histograms to visualize the data, check for normality, and understand the distributions of the continuous variables.

A graph of a car

Description automatically generated

From these plots, we observe that all distributions are not normal. The distribution of year is left-skewed, with a sharp increase in counts around 2005 and peaks around 2017. Meanwhile, the distribution of selling price is right skewed with the majority of vehicles sold under 625,000 and only a small number of vehicles price above that price. Similarly, km driven is also right skewed with most of the vehicles reaching a mileage of under 200,000 km, and a couple over 200,000 km. This skewness can indicate outliers in higher values, something we will need to consider when performing further analysis

To analyze the distribution of categorical variables, we used boxplots.

A group of colored boxes

Description automatically generated with medium confidence

Intuitively, from the distribution of owners, most of the cars offered just have one owner, while the proportion decreases by approximately half for each additional owner. Additionally, we see that petrol and diesel are the most popular in terms of fuel type while manual is the most frequent in terms of transmission. Also, we seem to have more individual sellers, followed by dealers, and finally, Trustmark dealers.

**Relationships with Selling Price:**

To investigate the relationship between the continuous variables and selling price, we created scatterplots that plotted the variables against selling price (Appendix 1.02).

For the year vs selling price plot, even though we cannot observe a straight away linear regression, we can see that the overall values seem to increase as the years go by, which is intuitive since older cars always tend to have lower prices. There isn’t an obvious relationship between selling price and km driven, likely because the axis needs to be scaled since there are several outliers in the data. To solve this, we created a scatterplot of selling price vs. year by transmission type.

A graph with red lines

Description automatically generated with medium confidence

After adding the transmission variable to the plot, we can see that there is more of a linear relationship for manual vehicles. As for automatic vehicles, there seems to be more variance and potentially a higher slope.

To investigate the relationship between the categorical variables and selling price, we created boxplots that plotted the variables against selling price (Appendix 1.03).

The outliers seemed to make the interpretation of the difference between each categorical type against selling price hard. For example, there seems to be a lot of outliers for Diesel and Petrol cars, but it could be due to the size of data for each type. However, it is still discernable that cars with less previous owners, cars that operate on diesel, automatic cars, and cars sold at Trustmark dealers have higher selling prices.

To mitigate this problem with the outliers, we plotted the boxplots once again without outliers (filtering by rejecting any value outside of the 5th quantile to the 95th quantile range). The results are consistent with our previous findings when we kept outliers in the plot.

A group of colored rectangular shapes

Description automatically generated with medium confidence

Regarding the brand variable, we created a preliminary boxplot to assess the difference in selling prices among different brands

A graph with different colored squares

Description automatically generated

There is an obvious difference between the selling prices among different brands. In particular, Audi, BMW, Land Rover, Benz, and Volvo have higher selling prices. However, the number of cars is significantly different among these brands. Thus, for model construction, we might merge brand into two levels: luxury and economical. This is determined by the median selling price of each brand. According to the plot (Appendix 1.04), we determined 1250000 to be the threshold that divides luxury and economical brands in terms of selling price.

According to this threshold, the following brands are luxury vehicles: Audi, BMW, Isuzu, Jaguar, Jeep, Kia, Land, MG, Mercedes-Benz, and Volvo. The rest of the brands are considered economical. The box plot comparing selling price by brand level shows consistent results with luxury vehicles having a higher selling price than economical vehicles.

Finally, we created a correlation matrix to evaluate the correlation strength between each continuous variable in the dataset (Appendix 1.05).

This correlation matrix supports our hypothesis that newer cars (cars with higher year value) tend to be more expensive. This is reflected in a moderate positive correlation with selling price (+0.41), and is consistent with our previous findings, though it is not as strong as expected. The kilometers driven display a weak negative correlation with selling price (-0.19). This indicates that the fewer kilometers driven, the higher the price of the car. This aligns with the logic that a car with lower mileage is closer to being new, and therefore, more valuable. Finally, the correlation between year and kilometers driven is negative and moderate (-0.42), implying that newer cars are less likely to have high mileage since older cars have been used longer, which is intuitive as well. However, this correlation between covariates may lead to multicollinearity which we would need to consider in the statistical analysis phase.

**Statistical Analysis**

After exploring the data and gaining an initial understanding of its characteristics, the next step was to construct a regression model to better interpret the data and quantify the relationship between the selling price and various covariates. This modeling process aims to provide actionable insights into how these factors influence second-hand car prices.

Before fitting the regression model, it was crucial to assess the predictors for any potential concerns, particularly multicollinearity, which could undermine the stability and interpretability of the model. Multicollinearity occurs when predictors are highly correlated with one another, potentially inflating standard errors and obscuring the true relationship between the response variable and covariates.

To diagnose multicollinearity, we calculated the Variance Inflation Factor (**VIF**) for each numerical variable. The results revealed that both year (**VIF = 1.57**) and km\_driven (**VIF = 1.48**) exhibited expected correlations without crossing the critical threshold of 10, which typically signals problematic multicollinearity. These values confirmed that the numerical predictors could be confidently included in the regression model without risking redundancy or instability in parameter estimates.

|  |  |  |
| --- | --- | --- |
| **0 < 5** | **5 < 10** | **10+** |

**VIF**

**Year**

**1.57**

**Km Driven**

**1.48**

With these results, we moved forward in our analysis by fitting a linear regression model using all covariates. To enhance the model, we explored the potential interaction between the continuous covariates km\_driven and year. Two models were compared: one with additive effects (Appendix 2.01) and another including the interaction term (Appendix 2.02). The comparison revealed that the interaction term significantly influenced the selling price, aligning with the objectives of Research Question 1.

Further observations indicated that among fuel types, only diesel cars were significantly different from the baseline (CNG), suggesting that the remaining fuel types could potentially be merged into a “non\_diesel” category. For seller type, only Trustmark dealers showed significant differences compared to the baseline (dealer). Transmission type and brand level also exhibited notable effects, with significant differences in their coefficients. Regarding ownership, the “Second Owner” category was the only one to show statistical significance, which raised the possibility of merging the other ownership levels.

Given these findings, and the significance of the interaction term as suggested by its p-value, we decided to include this interaction in our model. We then conducted diagnostic tests to evaluate whether the model met the assumptions required for linear regression.

The first diagnostic test involved **plotting residuals against fitted values** (Figure 1). This revealed a distinct “fan” shape, indicating a violation of the constant variance assumption, or homoscedasticity. This finding suggested the presence of heteroscedasticity and the need for a transformation of the response variable. To further validate this, we generated **a Q-Q plot** (Figure 2), which showed deviations from the Q-Q line, particularly in the tails. These deviations pointed to a heavy-tailed distribution of residuals, indicating a potential violation of the normality assumption.

**A graph of a normal q-q plot

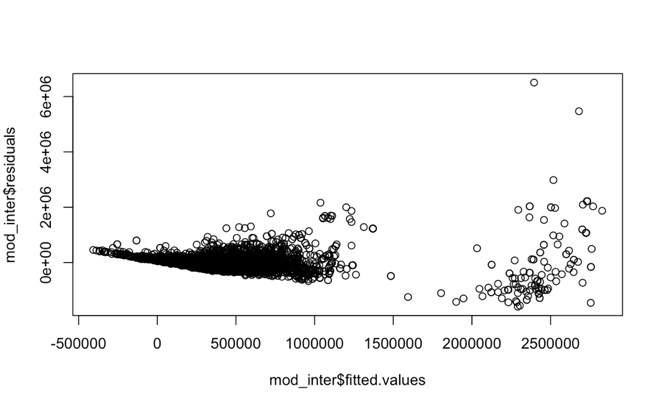
Description automatically generated**

Figure 2

Figure 1

**Residual vs. Fitted**

Both diagnostics provided sufficient evidence that transforming the response variable was necessary. Following this, we implemented the transformation, resulting in a new proposed model that better addressed these issues.

By running our fitted log-transformed model (Appendix 2.03), we observed significant improvements compared to the previous model. The transformation resulted in more coefficients of the owner variable becoming significant, while all coefficients of the **seller\_type** variable were now significant. For the fuel variable, only one level (diesel) showed significance, suggesting that the remaining levels could potentially be merged. Additionally, the R-squared value increased from **0.6288** to **0.7252**, indicating that the log-transformed model explained a greater proportion of the variation in the selling price.

To evaluate whether the transformation improved the model’s assumptions, we revisited diagnostic plots. First, we examined the residuals against fitted values (Appendix 2.04). Unlike the original model, the residuals plot showed no obvious patterns, suggesting that the issue of heteroscedasticity had been mitigated. Next, we generated a Q-Q plot (Appendix 2.05) to assess the normality of residuals. While some data points still deviated from the Q-Q line, particularly in the tails, the deviations were less pronounced compared to the previous model. This indicated that the assumption of normality, though not perfectly satisfied, could now be reasonably approximated with slight heavy tails.

Overall, the log transformation significantly improved the model’s adherence to linear regression assumptions. Given these results, the log-transformed model was deemed suitable for addressing Research Question 3, offering a better interpretation of the relationships between the covariates and the selling price.

Once our model appeared more stable, we shifted our focus to improving its accuracy. From the transformed model, we observed that some categories were not significant, prompting us to merge certain levels to simplify the analysis. For instance, we consolidated fuel types into **Diesel and Non-diesel**, based not only on their lack of significance but also to enhance interpretability. These modifications reduced the model’s complexity while retaining its explanatory power. The revised model was then re-fitted (Appendix 2.06), and the results indicated that all coefficients, except for the coefficient of Test Drive Car, were now significant. This confirmed that further variable removal through model selection was unnecessary.

One of our key research goals was to achieve an effective interpretation of the model. While the log-transformed model was a better fit than its predecessor, the presence of categorical covariates with multiple levels posed challenges in determining which interactions to include. To address this, we employed a **backward selection** procedure focused on interactions between categorical and continuous variables, specifically:

• **year:owner** to assess how selling price changes with the year of purchase across ownership levels.

• **km\_driven:owner** to examine how kilometers driven impact selling price across ownership levels.

• **year:transmission** to determine whether the relationship between year and selling price differs by transmission type.

• **km\_driven:transmission** to explore how kilometers driven influence selling price by transmission type.

• **km\_driven:year**, as the continuous interaction term, to evaluate how the combined effect of kilometers driven and year influences selling price.

Starting with a full model that included all these interactions, we iteratively removed terms with the lowest significance until all p-values were below the threshold **alpha = 0.05**.

|  |  |  |
| --- | --- | --- |
| Iteration | Removed Term | p-value |
| 1 | **km\_driven : transmission** | 0.665 |
| 2 | **km\_driven : year** | 0.436 |
| 3 | **year : transmission** | 0.149 |

The backward selection process resulted in a model with interactions between owner and each continuous variable (year and km\_driven). We referred to this as the **Model with Categorical Interactions** (Appendix 2.07). The final equation for this model was:

With this model finalized, we compared it to two others: the **Full Model**, which included all interactions (categorical and continuous), and the **Log-Transformed Model**, which featured only continuous covariates. To assess model performance and select the best fit, we used Mallow’s cp statistic, which measures how well a model balances fit and complexity. The results were as follows:

|  |  |  |
| --- | --- | --- |
| Model | Expected p+1 | Mallow’s cp |
| Full Model | 23 | **23.00** |
| Log – Transformed | 13 | **50.56** |
| Categorical - Interactions | 20 | **19.87** |

Although the Full Model achieved a cp value equal to its expected p+1, its complexity made it less desirable due to potential overfitting and reduced interpretability. The **Categorical Interactions Model** emerged as the preferred choice, as its cp value closely aligned with the expected p+1, demonstrating an optimal balance between simplicity and accuracy.

To further validate our findings and ensure consistency with the cp analysis, we conducted a comparison of R-squared and Adjusted R-squared values across the three models. These metrics provided insight into the variance explained by each model while accounting for model complexity. The results were as follows:

|  |  |  |
| --- | --- | --- |
| Model | R-squared | Adjusted R-squared |
| Full Model | **0.73** | **0.73** |
| Log – Transformed | **0.72** | **0.72** |
| Categorical - Interactions | **0.73** | **0.73** |

The results aligned with the cp analysis, reinforcing the robustness of the **Categorical Interactions Model**. While both the Full Model and the Categorical Interactions Model had similar R-squared and Adjusted R-squared values, the latter offered slightly better performance when analyzed to more decimal places. Given its reduced complexity and interpretability, the Categorical Interactions Model was preferred over the Full Model.

Confident in these results, we conducted additional diagnostics to ensure the Categorical Interactions Model adhered to linear regression assumptions. We plotted the residuals against the fitted values (Figure 3) and generated a Q-Q plot (Figure 4). Both diagnostics confirmed that the model exhibited no strong violations of homoscedasticity or normality assumptions, further validating our choice.

A graph of a normal q-q plot

Description automatically generated**A black and white diagram with numbers

Description automatically generated**

Figure 4

Figure 3

**Residual vs. Fitted (Categorical Interactions)**

**Final Model Selection**

In conclusion, based on the cp statistic, R-squared, Adjusted R-squared, and diagnostic evaluations, the **Categorical Interactions Model** emerged as the best fit. Its balance between simplicity and accuracy, combined with its interpretability, made it the optimal choice for addressing our research questions. This model effectively captures the interaction between categorical and continuous variables, providing meaningful insights into the factors influencing selling price.

**Discussion**

Ideas:

* Based on the final model, answer research question 1,2,3
* Summarize/interpret the new found model
* Extra assumption of normal distribution of error term, since Q-Q plot is NOT perfectly normal
* limitation: the outliers/large value observations are hard to handle, though we use log transformation to reduce the effects.
* etc.

**Appendix**

1. **Exploratory Data Analysis**
2. **Summary of Dataset**

A white background with black text

Description automatically generated

1. **A screenshot of a graph

   Description automatically generatedContinuous Variables vs. Selling Price**
2. **Categorical Variables vs. Selling Price**

**A group of graphs with different colored lines

Description automatically generated with medium confidence**

1. **Boxplot of Selling Price by Brand Level**

A graph with a blue and red rectangle

Description automatically generated

1. **Correlation Matrix of Continuous Variables**

A diagram of a number of numbers

Description automatically generated with medium confidence

1. **Statistical Analysis**
2. **Additive Regression Model with Full Covariates**

A screenshot of a computer

Description automatically generated

1. **Regression Model with Full Covariates and Interaction (Year-Km\_Driven)**

**A screenshot of a computer

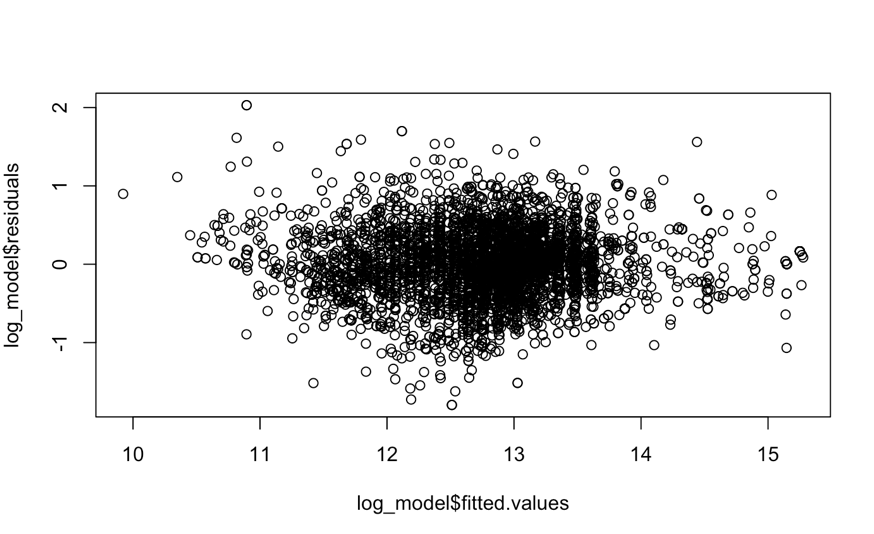
Description automatically generated**

1. **Regression Model with Log-Transformed Response**

**A screenshot of a computer

Description automatically generated**

1. **Plot of Residuals for Log-Transformed Model**

****

1. **Q-Q Plot for Log-Transformed Model**

**A graph of a normal q-q plot

Description automatically generated**

1. **Regression Model with Merged Levels**

**A screenshot of a computer

Description automatically generated**

1. **Regression Model with Categorical Interactions**

**A screenshot of a computer

Description automatically generated**