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

**Introduction**

(copy+paste)

**Exploratory Analysis**

* Plots , summary
* Discuss a bit outliers
* Data preprocessing: missing values, split the “name” into “brand”, based on the median of brand, create “brand\_level”

NOTE: plots here are messy, might place them in panel by ggplot2

**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** (Appendix 2.03). 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** (Appendix 2.04), 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.

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 2.05), 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 2.06). 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 2.07) 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 2.08), 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 2.09). 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 (Appendix 2.10) and generated a Q-Q plot (Appendix 2.11). Both diagnostics confirmed that the model exhibited no strong violations of homoscedasticity or normality assumptions, further validating our choice.

**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. **Statistical Analysis**
3. **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. **Plot of Residuals for Interaction Model**

A graph showing a number of dots

Description automatically generated

1. **Q-Q Plot for Interaction Model**

**A graph of a normal q-q plot

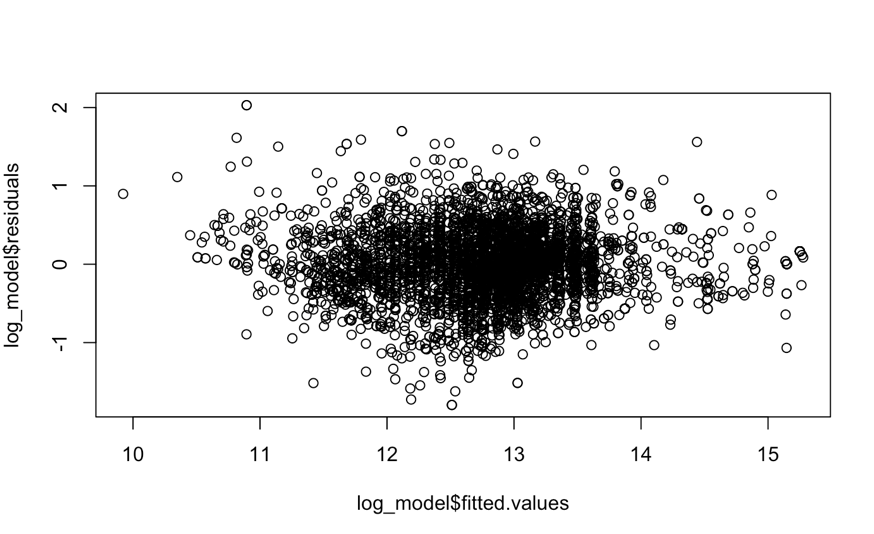
Description automatically generated**

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**

1. **Plot of Residuals for Categorical Interactions**

**A black and white diagram with numbers

Description automatically generated**

1. **Q-Q Plot for Categorical Interactions**

A graph of a normal q-q plot

Description automatically generated