Team 85 Project Progress Report

GitHub Link: https://github.gatech.edu/MGT-6203-Fall-2023-Canvas/Team-85

OVERVIEW AND PROBLEM FRAMING

The project aims to address the problem of accurately predicting the selling price of used automobiles using predictive analysis. Using the vehicle dataset from Kaggle, this project intends to understand the relationship between the selling price of used cars and various other independent variables.

The primary goal is to create a model that accurately predicts the selling price based on independent variables. The investigation will primarily utilize regression models, considering their effectiveness in the automotive industry for predictive analysis.

PLANNED APPROACH

1. Data Transformations
    - This step includes cleaning up null or zero values and investigating non-linear relationships between independent variables and selling prices through panel plots. If non-linear relationships are found, corresponding X variables will be transformed to achieve linearity. Outliers will be identified using Cook's distance and adjusted by either removal or imputation, depending on the data size. Interaction terms will be created to explore how the relationship between the target and independent variables changes based on the value of another independent variable.
2. Model Approach
    - While modeling, the focus is on predicting the selling price through multivariate linear regression. Different models will be compared, including a full linear model, log-log regression, and linear-log regression. The selection of these models is based on the nature of the data and the interpretation of coefficients. The full linear model serves as the baseline, while the log-log regression allows for interpreting coefficients regarding percent changes. The linear-log regression is chosen for lognormal distributed data, where the dependent and independent variables have such distributions.
3. Training, Optimizing, and Comparing models
    - The final step involves training, optimizing, and comparing models. This will be done using Monte Carlo K-fold cross-validation to account for important data points. The data will be split randomly into training and test sets, such as an 80/20 split. Models will be fitted to the training set and predictions will be made against the test set to calculate the mean squared prediction error (MSPE) and adjusted R-squared for each K. Adjusted R-squared and other relevant metrics will be used to assess the extent to which independent variables explain variability in selling price. After all runs, the average MSPE will be calculated for each model to determine the average root mean squared error (RMSE) and compare the models to identify the one with the lowest prediction error. Finally, the hypotheses will be evaluated based on the model performance.

REMAINING ANALYSIS

A few components remain in completing the analysis. The next steps involve validating, interpreting, and iterating as necessary, which involves the validation of each hypothesis based on model performance. The final phase involves crafting a comprehensive analysis summary, to articulate the findings and insights derived from the analysis process.

DATA TRANSFORMATIONS (COMPLETED)

The following is the comprehensive end-to-end data cleaning process.

First, we checked for NA values, of which there were none in the data frame. We looked for NAs and ran a visual to check for any missing data (see Figure 1).

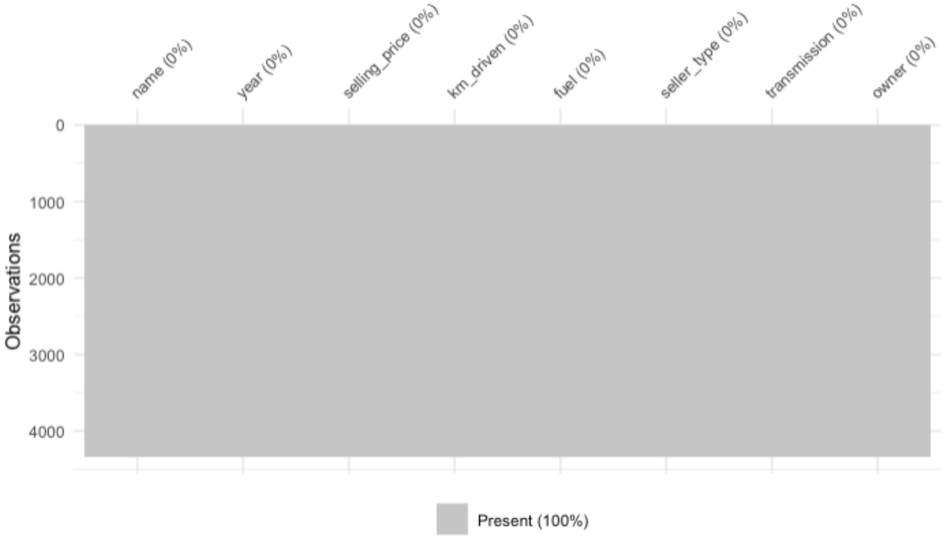
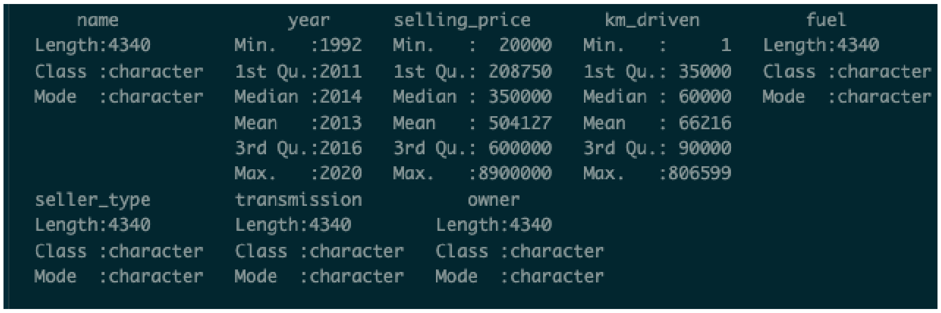


Figure 1: Vis_miss plot displaying details regarding the missing data within the data frame, where the x-axis represents our columns and their respective percentage of missing data.

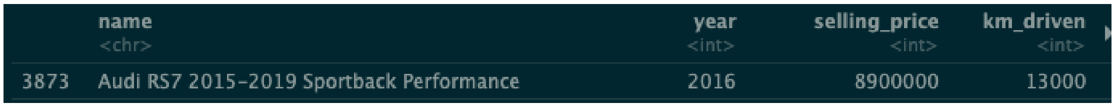
After reviewing the dataframe summary, it became evident that nonsensical outliers were present, such as an instance of a selling price of 8.9 million for a 2016 Audi, highlighting potential data irregularities (see Figure 2).



```
     name                year      selling_price       km_driven         fuel
 Length:4340        Min.   :1992   Min.   :  20000   Min.   :     1   Length:4340
 Class :character   1st Qu.:2011   1st Qu.: 208750   1st Qu.: 35000   Class :character
 Mode  :character   Median :2014   Median : 350000   Median : 60000   Mode  :character
                    Mean   :2013   Mean   : 504127   Mean   : 66216
                    3rd Qu.:2016   3rd Qu.: 600000   3rd Qu.: 90000
                    Max.   :2020   Max.   :8900000   Max.   :806599
 seller_type        transmission          owner
 Length:4340        Length:4340        Length:4340
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character
```

Figure 2: Dataframe summary captured to provide an overview of the dataset statistics and key information.

We then examined the car dataset to investigate the sale of a car priced at 8.9 million dollars. While acknowledging the potential for this to represent a legitimate data point, such as a collector's item (e.g., used in a movie), it seems more probable that this selling price is an incorrect entry. Consequently, we chose to eliminate these outliers as they do not appear to accurately represent the broader car market, which we aim to reflect in our modeling efforts.



	name	year	selling_price	km_driven
	<chr>	<int>	<int>	<int>
3873	Audi RS7 2015-2019 Sportback Performance	2016	8900000	13000

Figure 3: A detailed view focusing on the row that contains the maximum selling_price value (8.9 million) within the dataset.

After examining histograms (Figure 4,5), we observed the distribution of selling prices. To exclude extreme values, we established a threshold, opting to remove cars priced above $1 million. Implementing this threshold resulted in the removal of 7.8% of the dataset. We also removed cars that had over 200,000 miles as most buyers don't want a car that has excessive miles.

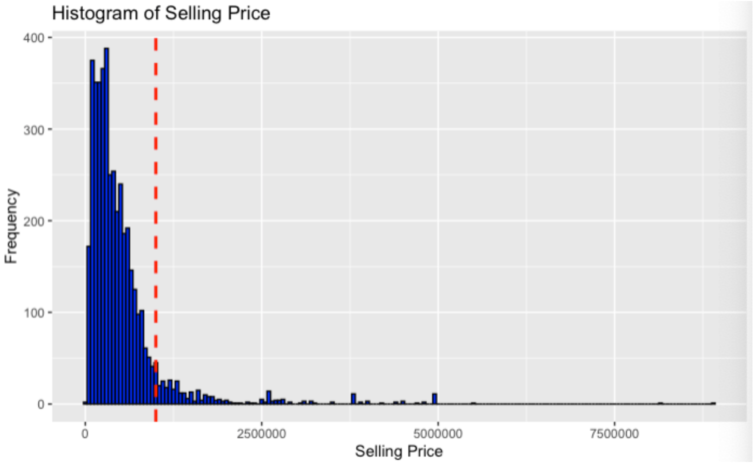


Figure 4: Histogram displaying the distribution of selling prices, where the y-axis represents the frequency of occurrence of different selling price ranges within the dataset.

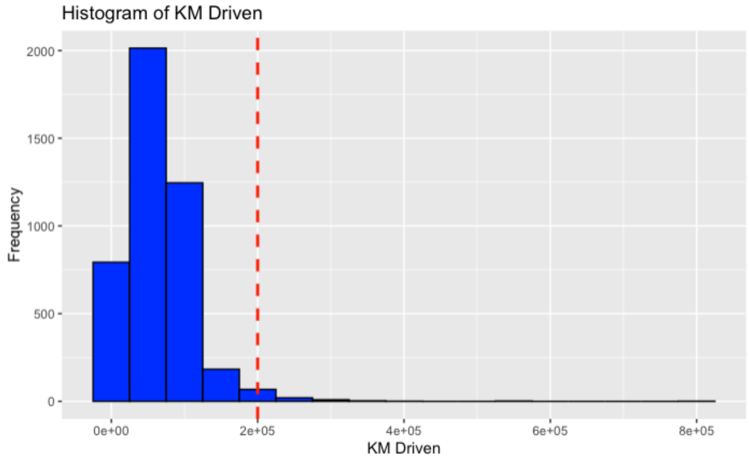


Figure 5: Histogram displaying the distribution of kilometers driven, illustrating the frequency of different ranges of distances covered by vehicles in the dataset.

Following the data transformations, we visualized relationships between variables. Below are instances demonstrating relationships between variables, considering both numeric vs. categorical and numeric vs. numeric scenarios.

We plotted the selling price against kilometers driven (Figure 6), revealing an anticipated negative correlation between these variables. The findings suggest that as the mileage of a car increases, it notably impacts the sale price, indicating a lower selling price for vehicles with higher mileage.

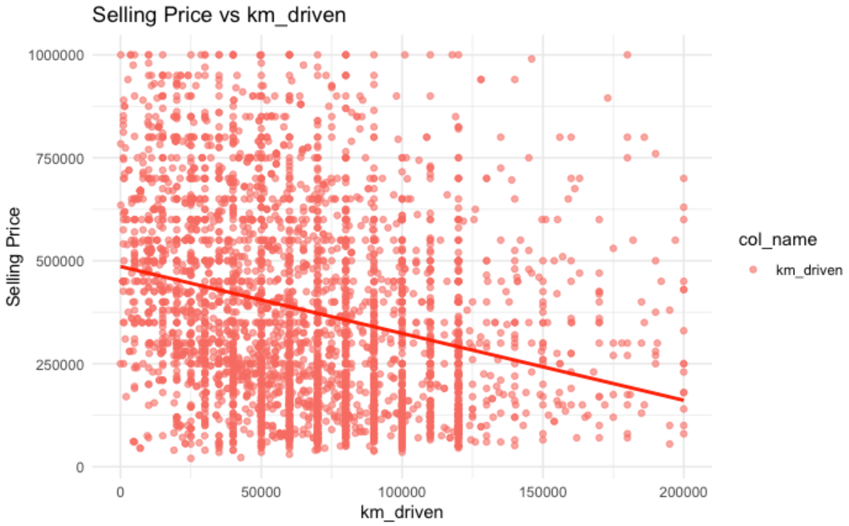


Figure 6: Displays the plot of the selling price vs kilometers driven. The x-axis captures the kilometers driven, while the y-axis captures the selling price. The red line showing a negative relationship is the regression line.

Additionally, we constructed a box plot to visualize the relationship between the selling price and seller type (Figure 7). Another notable observation is the tendency for cars sold by individual sellers to exhibit lower average prices, likely attributed to reduced overhead costs, including expenses related to employees, real estate, and advertising.

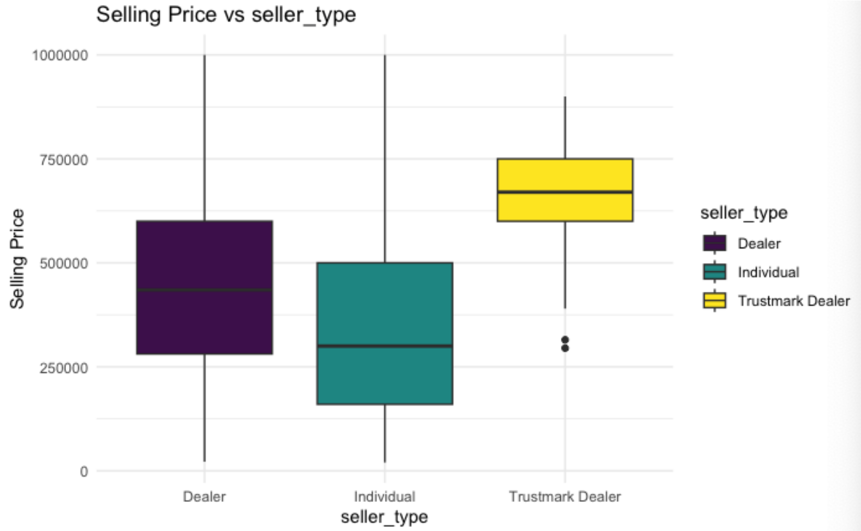


Figure 7: The box plot displays the relationship between selling price and seller type, highlighting the distinct differences in prices between various seller categories within the dataset.

We further visualized a box plot comparing the selling price and fuel type (Figure 8). The observation suggests that diesel and petrol cars tend to command higher prices, suggesting a need for potential data refinement. On average, diesel cars emerge as the costliest among the fuel types analyzed.

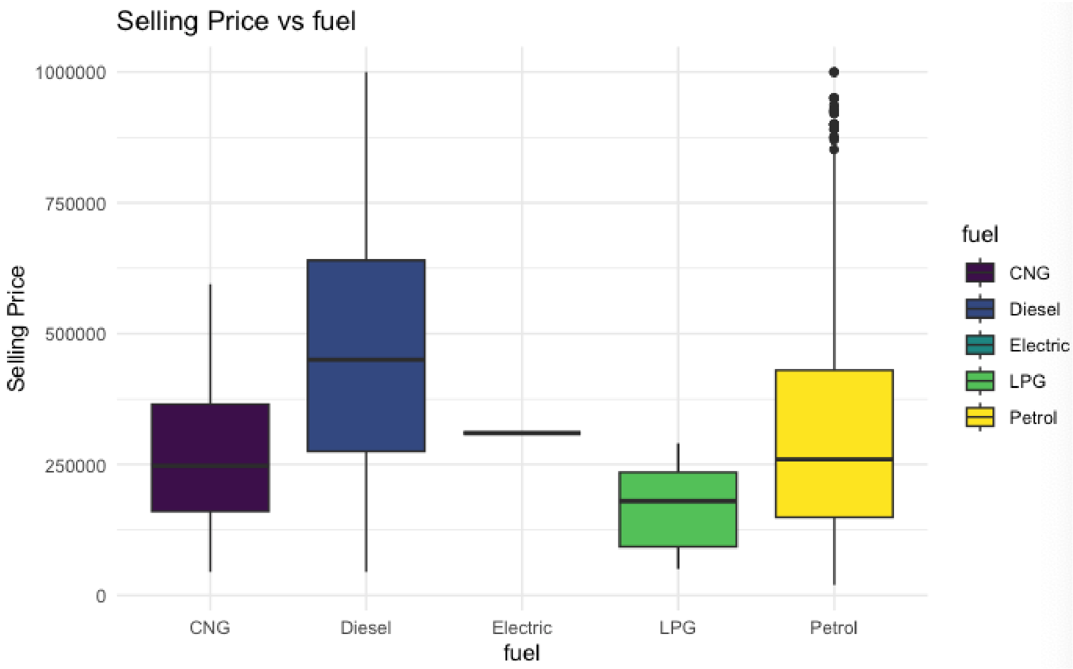


Figure 8: The box plot displays the relationship between selling price and fuel type, highlighting the distinct differences in prices between the various types of fuel within the dataset.

Figure 9 implies that it is unsurprising that newer cars generally command higher prices.

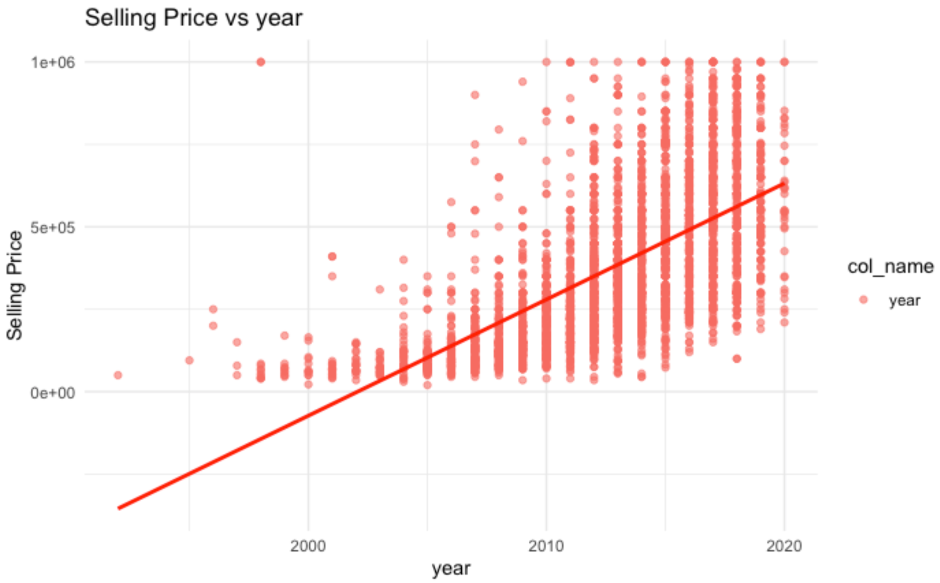


Figure 9: Plot displaying selling price (y-axis) vs year (x-axis). The red line is the regression line.

Finally, we created another box plot to compare the Selling Price against Owner status (Figure 10). The analysis indicates that test drive cars have higher selling rates, although this category represents one of the smallest segments. Notably, there are instances of second-owner cars listed for $1 million. As a team, we must address and determine an approach to ascertain the legitimacy of these data points or decide whether they should be excluded from the dataset.

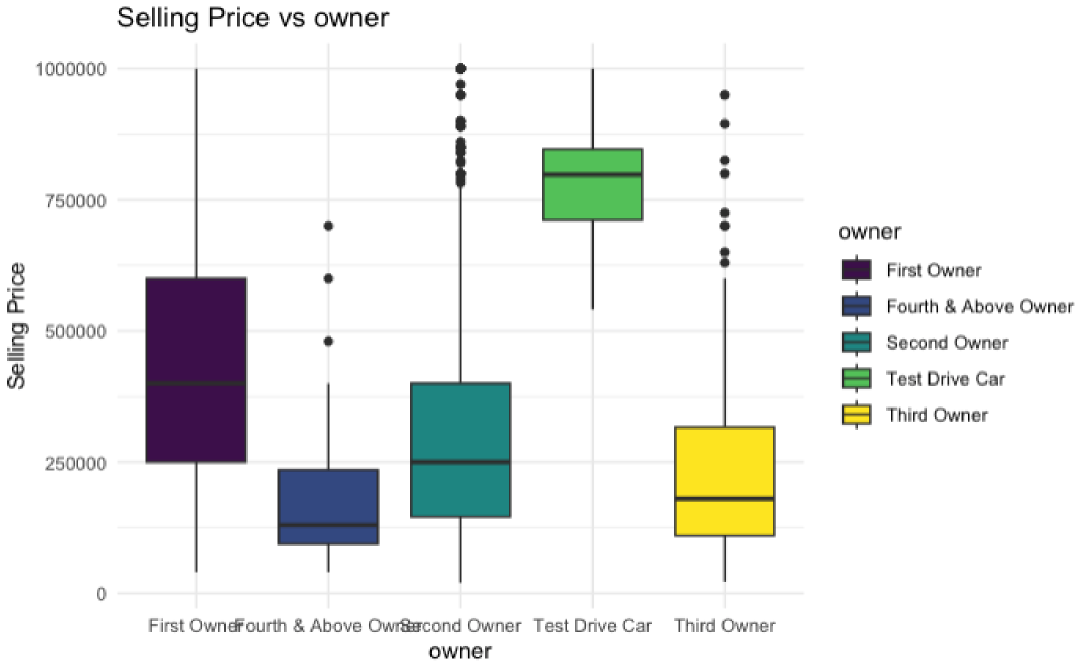


Figure 10: The box plot displays the relationship between selling price and owner status, highlighting the distinct differences in prices between the various owner statuses within the dataset.

MODELING (COMPLETED)

The following models were used thus far in our analysis:

1. **Full linear model**: This model was primarily used because most of the variables showed a linear relationship with the selling price, except for 'km driven' and 'year'.
2. **Log-log model with log transformation**: Utilized for the selling price, 'km driven', and 'year'. This transformation was applied because 'km driven' and 'year' did not exhibit a linear relationship with the selling price. The hypothesis was that the year and kilometers driven are significant factors in determining the selling price for a used car.
3. **Stepwise regression**: Used for feature selection to assess whether the full linear model could be enhanced or improved by including or excluding certain variables.

The linear models and stepwise regression across 100 runs displayed an r-squared value of .68, the average mean absolute prediction error was approximately $100,000, indicating a significant discrepancy in predicting selling prices. This discrepancy is attributed to the selling price's non-linear relationship with factors like kilometers driven and year, leading to poor predictive performance by the linear model.

Additionally, the root mean absolute error (331) was notably high, further emphasizing the model's inability to accurately predict prices. While the log root mean absolute error was lower at around .48, it's important to note that this metric is in log units due to data transformation, indicating the limitations and challenges in accurately predicting selling prices using these methods.

REFERENCES

1) Kuiper, S. (2008). Introduction to Multiple Regression: How Much Is Your Car Worth?. Journal of Statistics Education, 16(3). https://doi.org/10.1080/10691898.2008.11889579

2) Puteri, C. K., & Safitri, L. N. (2020). Analysis of linear regression on used car sales in Indonesia. Journal of Physics: Conference Series, 1469, International Conference on Innovation in Research, 28–29 August 2018, Bali, Indonesia, 012143. https://doi.org/10.1088/1742-6596/1469/1/012143

3) Monburinon, N., Chertchom, P., Kaewkiriya, T., Rungpheung, S., Buya, S., & Boonpou, P. (2018). Prediction of prices for used car by using regression models. In Proceedings of the 2018 5th International Conference on Business and Industrial Research (ICBIR). https://iopscience.iop.org/article/10.1088/1742-6596/1469/1/012143/pdf

4) Mehra, R. (2023, May 11). The untapped potential of the used car market in India. The Times of India. https://timesofindia.indiatimes.com/blogs/voices/the-untapped-potential-of-the-used-car-market-in-india/

5) Pal, N., Arora, P., Sundararaman, D., Kohli, P., & Palakurthy, S. S. (2018). How much is my car worth? A methodology for predicting used cars prices using Random Forest. Paper presented at the Future of Information and Communications Conference (FICC) 2018, Retrieved from https://arxiv.org/ftp/arxiv/papers/1711/1711.06970.pdf

DATA SOURCE

https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho