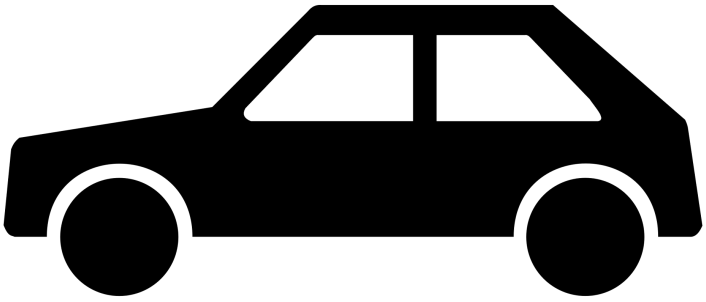

Driving Success: Comprehensive Analysis of Selling Prices For Vehicles

Team 85

Akshay Bahl
Kenji Hagiwara
Rutvij Rana
Samuel Robert Schreer
Shakil Aryal

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

TABLE OF CONTENTS

INTRODUCTION

BACKGROUND

In the ever-evolving landscape of the used automotive industry, predictive analysis emerges as a linchpin, irrespective of a business's role as a buyer or seller. The ability to accurately price cars is the key to maximizing profitability, particularly in a market that is exploding globally. In India, the used car market is expected to more than double over the next 5 years as buyers begin to forego going to a dealership for their new vehicle. This report focuses on an exploration of key variables influencing vehicle pricing, leveraging a data set of used cars sold on Car Dekho, which is an online marketplace for used cars. Highlighting the efficacy of regression models in the automotive sector, the study aims to investigate the interplay of variables for accurate predictive analysis.

PROBLEM STATEMENT

Amidst the complexities of the automotive market, the relationship between the selling price of used cars and critical vehicle attributes remains ambiguous. Historically, prices were set by industry experts trying to follow market trends. This analysis, centers on discovering the relationship between variables using predictive analytics, with a primary focus on understanding the correlation between the selling price of used automobiles and independent variables. The primary objective is to use this information to precisely predict used car prices based on the provided independent variables, which will be explored through various regression analyses. Our primary hypothesis states years and kilometers driven would be the most significant variables influencing the selling price. Our additional hypotheses can be found later in this report.

BUSINESS JUSTIFICATION

From a strategic business perspective, the accurate prediction of used vehicle sale prices holds immense value. For buyers, it ensures cost-effective decisions, fostering confidence and yielding substantial savings. Sellers stand to benefit from setting competitive prices, facilitating quicker inventory turnover and heightened profit margins. As hypothesized above, if a seller understands the significant influence of kilometers driven and year, they can accurately and quickly determine the selling prices of certain cars and add profit margins. This, in turn, can contribute to overall revenue growth. Beyond transactions, the streamlined process reduces operational costs for dealerships and online marketplaces, fostering efficiency. Such data-driven precision builds trust among consumers and elevates market share and customer loyalty, thereby influencing the financial health and competitiveness of the entire automotive industry.

METHODOLOGY

DATA & VARIABLE DESCRIPTION

For this project, the data was sourced from the vehicle dataset from Kaggle. The dataset comprises detailed information about used cars. The variables with their metadata used in the analysis are listed in the figure below.

Variable Name	Type of Variable	Variable Description
selling_price	Dependent	The price at which the car is being sold
name	Independent	Name of the cars
year	Independent	Year of the car when it was bought
km_driven	Independent	Number of Kilometers the car is driven
fuel	Independent	Fuel type of car (petrol / diesel / CNG / LPG / electric)
seller_type	Independent	Tells if a Seller is an Individual or a Dealer
transmission	Independent	Gear transmission of the car (Automatic/Manual)
Owner	Independent	Number of previous owners of the car
interaction_fuel	Interaction Term	Interaction between year and fuel
interaction_seller_type	Interaction Term	Interaction between year and seller_type
interaction_transmission	Interaction Term	Interaction between year and transmission

Figure 1: Table includes variable names with their respective type and descriptions information.

Furthermore, we generated dummy variables for categorical attributes like fuel, seller_type, transmission, and owner, contributing to the analysis.

DATA TRANSFORMATION & INSIGHTS

The data cleaning process involved a comprehensive end-to-end approach. Initially, we checked for NA values, finding none in the data frame. Visual inspection through a Vis_miss plot provided insights into missing data percentages across columns. The missing values were imputed using an average of values for that variable.

Following this, a review of the data frame summary highlighted unusual outliers, such as a 2016 Audi with a staggering listing of 8.9 million rupees. This could have potentially been a limited edition trophy car, explaining its high cost. Recognizing the likelihood of data irregularities, especially within the automotive market context, we opted to remove these outliers to improve the precision of our modeling endeavors.

Creating histograms (Figure 2) helped us understand the distribution of selling prices and kilometers driven. To manage extreme values, we set a threshold, removing cars priced above INR 1 million and those with over 200,000 miles.

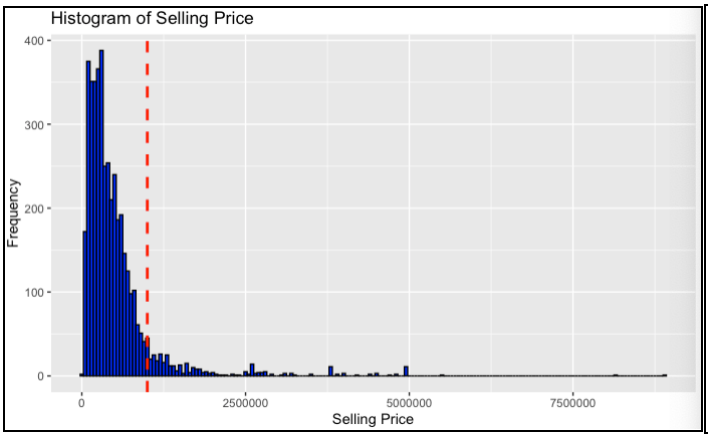

Figure 2: Histogram displaying the distribution of selling prices, where the y-axis represents the frequency of occurrence of different selling price ranges within the dataset.

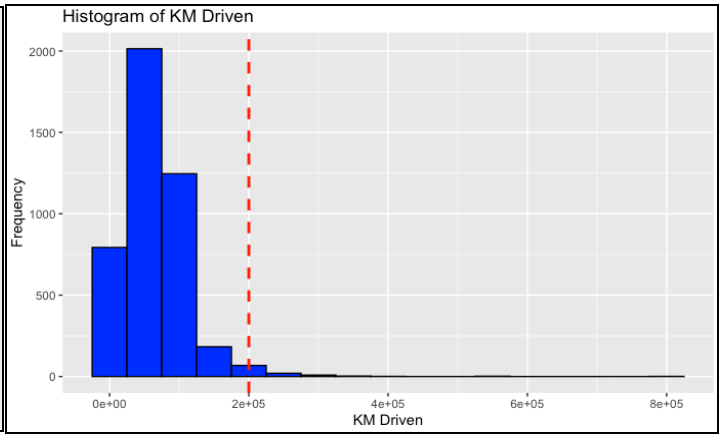

Figure 3: Histogram displaying the distribution of kilometers driven, illustrating the frequency of different ranges of distances covered by vehicles in the dataset.

This process resulted in the removal of 7.8% of the dataset. The subsequent data transformations allowed us to visualize relationships between variables, such as the negative correlation between selling price and kilometers driven (Figure 4).

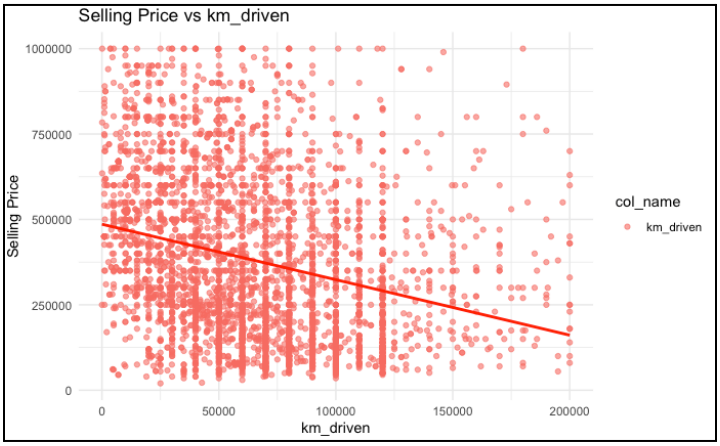

Figure 4: Displays the plot of the selling price vs kilometers driven. The x-axis captures the kilometers driven, while the y-axis captures the selling price. The red line showing a negative relationship is the regression line.

We further explored relationships through box plots, uncovering distinctions in selling prices based on seller type, fuel type, year of manufacture, and owner status. We also plotted a

correlation matrix (Figure 5) to provide some insights into relationships between variables prior to the analysis. The correlation coefficients within the matrix indicated both the strength and direction of linear associations. Beyond gauging relationships, the matrix aids in detecting multicollinearity in the regression analysis and serves as a diagnostic tool for checking assumptions.

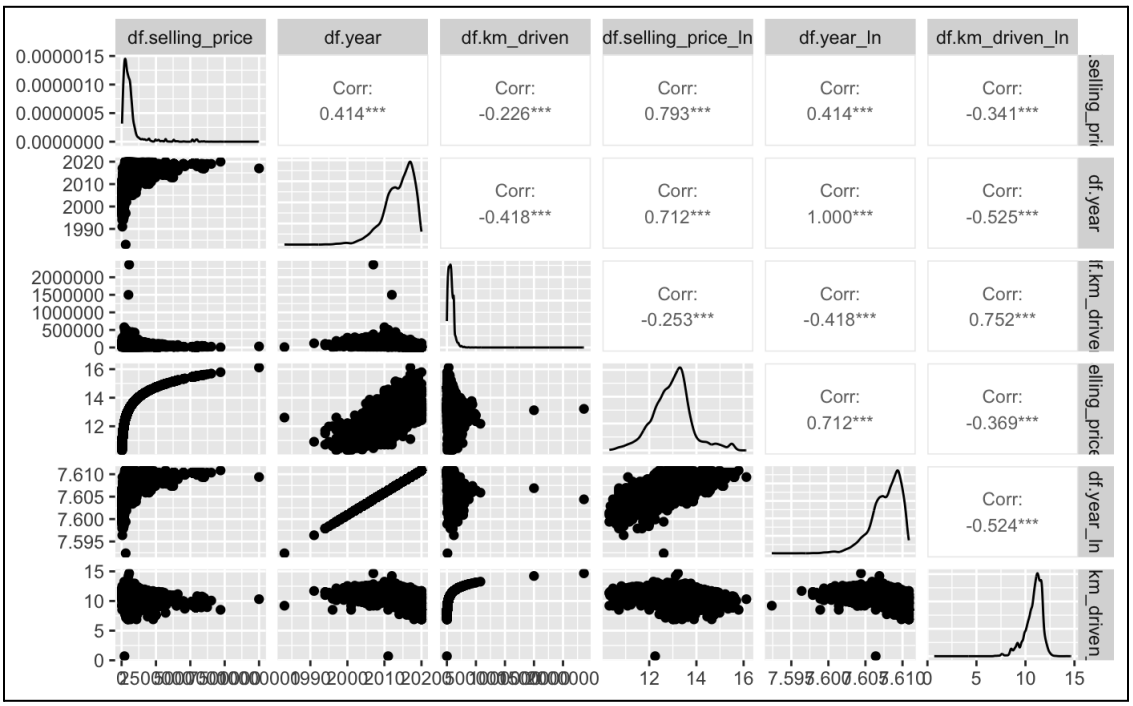

Figure 5: Correlation matrix illustrating the relationships between variables. Each cell displays the correlation coefficient, ranging from -1 to 1, providing insights into the strength and direction of associations between variables.

Notably, the analysis prompted a need for potential data refinement, especially in addressing instances of second-owner cars listed for INR 1 million. These findings informed our understanding of the dataset, guiding us in preparing a cleaner, more focused dataset for subsequent modeling. Overall, the data cleaning process was intricate, involving the identification and removal of outliers, establishment of thresholds, and insightful visualizations to refine the dataset for meaningful analysis.

APPROACH

In the process of predicting selling prices for cars, the choice of models plays a crucial role in capturing the underlying relationships within the data. The models selected for comparison are multivariate linear regression variants: a full linear model, log-log regression, and linear-log regression. The full linear model serves as a baseline, providing a standard against which other models can be compared. The log-log regression is chosen to enable the interpretation of coefficients in terms of percent changes, which is particularly valuable in understanding the magnitude of the impact of independent variables on the selling price. Additionally, linear-log regression is employed when dealing with lognormal distributed data, a choice influenced by the nature of the distributions of both dependent and independent variables in the dataset.

The training of these models involves the use of Monte Carlo cross-validation, a robust technique for assessing model performance by accounting for variations in important data points. The dataset is randomly split into training and test sets in an 80/20 ratio. The models are then fitted to the training set, and predictions are made on the test set to calculate metrics such as mean squared prediction error (MSPE) and adjusted R-squared for each iteration of the cross-validation process. The selection of adjusted R-squared and other relevant metrics is motivated by the need to understand the extent to which independent variables explain the variability in selling prices. The average MSPE is then computed for each model, yielding the average root mean squared error (RMSE), which serves as a basis for model comparison. The model with the lowest prediction error is considered the most suitable for predicting selling prices.

The initial hypothesis centers around key variables, with a focus on the year and km_driven. The assumption is that newer cars and those with lower mileage are likely to command higher selling prices. This aligns with common expectations in the automotive market, where newer models and those with less usage often hold greater value. The evaluation of these hypotheses is an integral part of the final step, where model performance is assessed based on the chosen metrics. The goal is to determine which model provides the most accurate predictions and to validate or refine the initial assumptions about the impact of variables such as year and km_driven on selling prices.

MODELING

The following models were utilized in our analysis:

1. **Full linear model**: This model was primarily used because most variables showed a linear relationship with the selling price, except for 'km driven' and 'year'.
2. **Log-log model with log transformation**: This model was employed 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 (linear-log)**: This model was used for variable 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 model across 100 runs displayed an r-squared value of 0.68, and the average mean absolute prediction error was approximately INR 100,000, indicating a significant discrepancy in predicting selling prices. This discrepancy was 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 predict prices accurately. While the log root mean absolute error was lower at

around 0.48, 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. The log-log model was selected for the analysis.

ANALYSIS

LOG-LOG MODEL SUMMARY

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.64714 -0.16491  0.02714  0.18601  1.69970

Coefficients:
                              Estimate    Std. Error t value            Pr(>|t|)
(Intercept)                -1574.12721435   22.51896609 -69.902 < 0.0000000000000002 ***
km_driven_ln                  -0.03549651    0.00623179  -5.696    0.000000012891554 ***
year_ln                      208.43981187    2.95981410  70.423 < 0.0000000000000002 ***
mileage_num                    0.01539233    0.00171223   8.990 < 0.0000000000000002 ***
engine_num                     0.00023870    0.00001866  12.789 < 0.0000000000000002 ***
maxpower_num                   0.00878015    0.00021946  40.007 < 0.0000000000000002 ***
seats                          0.04406619    0.00606030   7.271    0.000000000000405 ***
fuelDiesel                     0.20928139    0.04323094   4.841    0.000001326760637 ***
fuelLPG                        0.19284799    0.06840873   2.819             0.00483 **
fuelPetrol                     0.07508105    0.04341471   1.729             0.08379 .
seller_typeIndividual         -0.06926040    0.01266075  -5.470    0.000000046842832 ***
seller_typeTrustmark_Dealer    0.07662869    0.02513718   3.048             0.00231 **
transmission_manual           -0.07503537    0.01570081  -4.779    0.000001806344465 ***
OwnerFourthAndAbove           -0.16542034    0.02662464  -6.213    0.000000000557384 ***
ownerSecond_Owner             -0.07133136    0.00951783  -7.494    0.000000000000077 ***
ownerThird_Owner              -0.11470785    0.01623492  -7.066    0.000000000001797 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.281 on 5560 degrees of freedom
Multiple R-squared:  0.8115,    Adjusted R-squared:  0.8109
F-statistic:  1595 on 15 and 5560 DF,  p-value: < 0.00000000000000022
```

Figure 6: Output of log-log model summary.

Our final model was trained using log transformations of selling price, kilometers driven, and year. It demonstrated an Adjusted R-squared of .8109 on the training set, which was a promising start. We then performed Monte Carlo cross-validation across 100 samples, which resulted in a mean Adjusted R-squared of .8081, nearly identical to our initial assessment. This means that our model is able to explain over 80% of the variance in selling price for a used car.

COEFFICIENTS

When interpreting the coefficients of our model, it is important to keep in mind that the selling price is log-transformed. We also had to log transform the data for year and kilometers driven, so those coefficients are the most difficult to interpret (and also most influential). Year is positively correlated with price, so newer cars command a higher premium. For every 1% increase in year, the price of the car is expected to increase by 208.44%. Similarly, for every 1% increase in kilometers driven, car price is expected to decrease by -.035%. For the remaining variables, the coefficients represented the expected percentage change in price if that variable

is increased or selected. Consider the seller type as an example. Our base case was “dealer”, and we had categorical variables for “individual” and “trustmark dealer”. Based on our coefficients, if a car is sold by an individual, the selling price is expected to be 6.9% lower than a dealer-sold car. Similarly, cars sold by a trustmark dealer command a 7.5% premium on average over a regular dealer-sold car.

DIAGNOSTICS

Reviewing diagnostics summaries of our model highlights some of the problems we had working with our data set. Ultimately, the used car market can be very volatile and is subject to many factors that were not included in our data set. Notably, the only usage-related data point we had was kilometers driven. Necessary or completed repairs were not directly included in our report, though you could assume some amount of correlation with age and kilometers driven for each of those variables. In addition, car models were not included, so high-end vehicles are not easily identified in our data set. Combining both of these factors means we have trouble with the tails. This can be clearly seen in our Residuals vs Fitted graph.

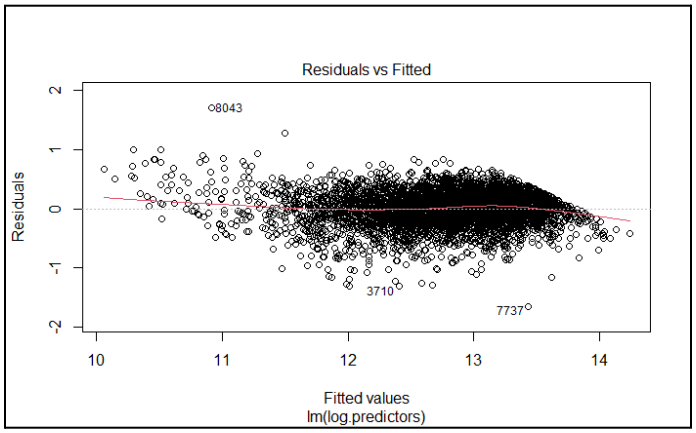

Figure 7: Displays the Residuals vs Fitted graph of our final model, demonstrating heteroscedasticity at the tails.

A studentized Breush-Pagan test supports the evidence of heteroscedasticity by returning a BP of 388.95 with a p-value of near 0. Higher BP values indicate greater heteroscedasticity.
studentized Breusch-Pagan test.

```
data:  log.log
BP = 388.95, df = 15, p-value < 0.00000000000000022
```

Figure 8: Output of Breush-Pagan test.

A histogram and Q-Q plot of our residuals shows that they are close to normally distributed, but again there is some skewing, particularly when it comes to underprediction.

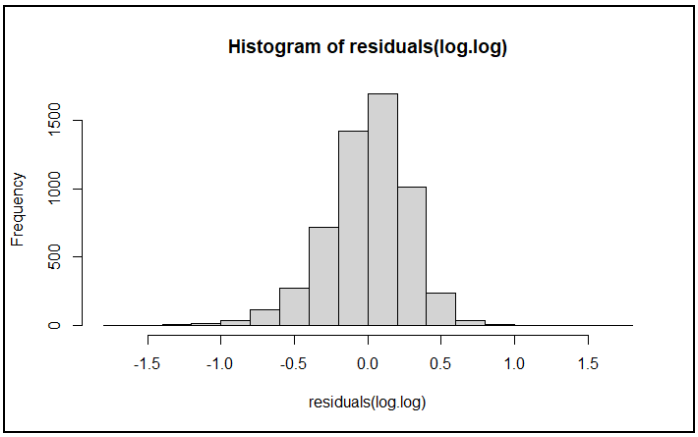

Figure 9: Histogram illustrating the distribution of residuals, indicating proximity to a normal distribution with a slight skew towards underestimation.

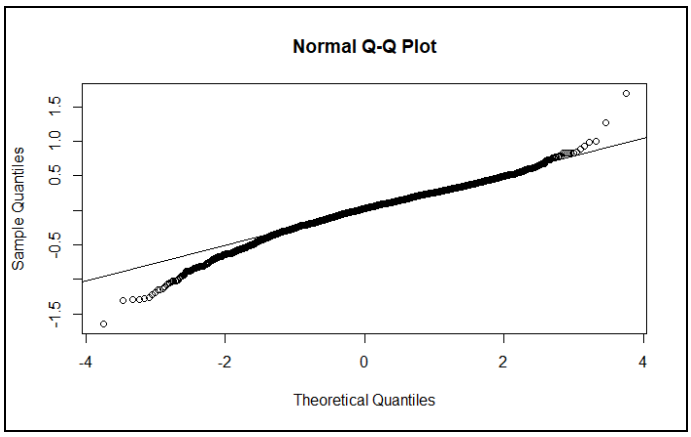

Figure 10: Q-Q plot comparing sample quantiles to theoretical quantiles. The model shows good alignment in the middle but deviates at the extremes.

Finally, we reviewed the VIF of each variable to check for collinearity. Our model showed a high value for two of our categorical fuel type variables, namely Diesel and Petrol. All other variables were below the ideal max.

```
> vif(log.log)
          km_driven_ln                year_ln           mileage_num            engine_num          maxpower_num
              1.834494               2.322577               3.186207               4.778364               1.967420
                 seats             fuelDiesel                fuelLPG             fuelPetrol  seller_typeIndividual
              2.224512              32.991461               1.651144              33.233715               1.334035
seller_typeTrustmark_Dealer  transmission_manual  OwnerFourthAndAbove  ownerSecond_Owner       ownerThird_Owner
              1.288861               1.177911               1.105610               1.277296               1.202161
```

Figure 11: Variance Inflation Factor (VIF) output depicting the results for the final model.

HYPOTHESES

We proposed 4 hypotheses before starting our research. Here we will explore each of them based on the model summary:

1. **Year and kilometers driven will be the most significant variables, year being positively correlated with price and km driven negatively correlated.**

   Our model supported these hypotheses as our coefficient for year was 208.44, with a p-value of 2e-16. This means that our model predicts for each 1% increase in year, car price increases by 208.44%, and this is a statistically significant relationship. Similarly, the coefficient for KM driven was -.035 with a p-value of 1.3e-8 at the p<.05 level. This means for every 1% increase in kilometers driven, car price decreases by -.035%, and again this estimate was statistically significant. The large coefficient for year demonstrates the large impact a small change in year can make. While the coefficient for km driven is lower, given the range of possible values, this results in a large impact on price as opposed to other variables like seats. Even though each increase in seats results in a 4.4% increase in price, our seats value ranged from 3-14, which is a very small range.

2. **Cars sold by individual owners are more likely to be used by only one owner and hence may be better maintained as compared to cars that may have switched ownership multiple times. This leads to the "seller_type" independent variable having a significant effect on the selling price of the car.**

   Our model supports, as each increase in ownership is associated with an increasing negative impact on selling price. A second owner is expected to get about 7.1% less than a first owner, a third owner is expected to get 11.5% less than a first owner, and subsequent owners are expected to receive 16.5% less than a first owner. The p-values for each of these relationships were 5.6e-10, 7.7e-14, and 1.8e-12 respectively, indicating they were all statistically significant at the p<.05 level.

3. **With fewer moving parts, a manual shift has the advantage of being easy to maintain, and hence, people perceive that from among two cars of the same make and year, a manual transmission vehicle will have a higher value and hence a higher selling price**

   Our model does not support this hypothesis as the coefficients indicate that having a manual transmission car is associated with a 7.5% price decrease over an automatic vehicle with similar characteristics. This relationship was statistically significant at the P,.05 level, with a p-value of 1.8e-6.

4. **Cars sold by dealers are more expensive than cars sold by an individual.**

   Our model supports this hypothesis, as individual sellers have a price about 6.9% lower than dealers. In addition, trustmark dealers which bear a higher level of scrutiny are associated with a 7.5% premium over regular dealers. The p values for these relationships were 4.7e-8 and 2.3e-3, indicating they were both significant at the p<.05 level.

LIMITATIONS

Our data set was exclusively populated by cars from a single online marketplace in India. While it is likely that this is a good representation of the used car market there, it cannot be applied to data from other markets. In addition, our analysis focused on the most quantifiable aspects of a vehicle such as year and mileage, but did not include other more qualitative points such as macroeconomic conditions and internal/external conditions. Future research could be enhanced by data sets with more variables.

In addition, our model exhibits some small amounts of heteroscedasticity at extremes, so it is most useful in predicting the prices of cars with attributes that fall in the inner quartiles of our data set's range. This was likely exacerbated by the size of our data set, which had less than

10000 entries. More entries would have potentially allowed our model to interpret edge cases with greater precision.

SUMMARY & CONCLUSION

Our research was an initial look into the dynamics of used car pricing in the Indian market. By employing regression analysis on a dataset from Car Dekho, key insights were revealed into how various factors such as the age of the car, mileage, type of seller, and transmission influence the selling price. The log-log model, in particular, proved effective in capturing the nuances of these relationships, explaining over 80% of the variability in selling prices.

One of the most notable findings is the strong influence of a car's year and kilometers driven on its selling price, with newer, less-driven cars commanding higher prices. This reinforces the commonly held belief in the automotive industry about the depreciation of car value over time and usage. The analysis also sheds light on the impact of seller type, revealing that cars sold by individual owners generally fetch lower prices than those sold by dealers, and trustmark dealers command a premium over regular dealers.

These findings have significant implications for various stakeholders in the used car market. For sellers, including dealerships and individual owners, this analysis provides a data-driven basis for setting competitive prices. For buyers, it offers a guideline to assess the fair market value of a used car, potentially leading to more informed purchasing decisions. Moreover, online marketplaces like Car Dekho can leverage these insights to refine their pricing algorithms, enhancing customer experience and trust.

The limitations within the analysis primarily stem from the dataset's scope and representativeness. The data, being limited to a single online marketplace in India, may not fully capture the diversity and complexity of the broader used car market. Key variables like specific car models, internal and external conditions of the cars, and macroeconomic factors were not included in the dataset, potentially affecting the comprehensiveness of the analysis.

Future research should aim to include these additional variables, which could provide a more holistic view of the factors affecting used car prices. A larger and more diverse dataset, possibly encompassing multiple marketplaces and geographical regions, would enhance the generalizability of the findings. Additionally, exploring advanced analytical techniques and machine learning algorithms could further refine predictive accuracy, especially for outlier cases where the current model shows limitations.

In conclusion, this study underscores the potential of data-driven approaches in revolutionizing the used car market, particularly in emerging markets like India. While none of the insights from our modeling were a surprise to market participants, providing them with data paves the way for more sophisticated pricing strategies and models in the future. As the automotive industry continues to evolve, such analytical approaches will become increasingly important in driving business decisions and shaping market trends.

REFERENCES

- 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

- 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

- 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

- 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/

- 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

- Verma, N., & Kushwaha, N. (2023). *Vehicle dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho