Vehicle Sales and Market Trends Analysis

Brandon Nava
March 16, 2024

Contents

1 Objective

The automobile industry is known to be highly dynamic due to the multitude of factors that determine the price of a vehicle. By understanding what determines the price can be crucial for dealers, buyers, and manufacturers in order to set competitive prices and understand the market trends. This project aims to analyze how specific factors such as vehicle condition, manufacturing year, mileage, and market trends influence the overall selling price. Through assessing a comprehensive dataset which includes information on sales transactions, this analysis seeks to identify significant predictors in order to develop a model that can predict selling prices in correlation with those specific variables.

2 Overview

The dataset utilized in this project, *Vehicle Sales and Market Trends Dataset,* consists of the following categories: year, make, model, trim, body type, transmission type, Vehicle Identification Number, state of registration, vehicle condition rating, odometer reading, colors, seller, Manheim Market Report value, selling price, and sale date. Using R Studio, we can use this dataset which provides the grounds to analyze specific factors on vehicle prices.

3 Methods & Procedure

3.1 Data Preparation

This involved loading the dataset and ensuring it was suitable for analysis. I handled missing values by omitting them according to each perspective category. I broke down the condition category into broader categories in order to simplify the analysis and make it more interpretable when dealing with a large dataset.

3.2 Model Building

I used a linear regression model to predict the selling price of vehicles based on the predictors: year, odometer, condition, and mmr. I also split the dataset into training and testing sets to ensure a random split of data.

4 Data Analysis

4.1 Type of Statistical Analysis

Using the multiple linear regression analysis helped to understand the relationship between the selling price of vehicles (the dependent variable in this case) and other predictor variables (serving as independent variables). This type of regression model fits a linear equation to observed data in order to understand how the dependent variable changes as the independent variables change as well, helping us to ultimately make predictions based on the manufacturing year, odometer reading, condition category, and Manheim Market Report (MMR) values.

5 Results

5.1 ANOVA Analysis

Predictor	DF	Sum Sq	Mean Sq	F value	P value
year	1	1.8270e+13	1.8270e+13	6435246	< 2.2e-16 ***
condition category	3	1.3412e+12	4.4707e+11	44.01	< 2.2e-16 ***
odometer	1	1.7533e+12	1.7533e+12	617569	< 2.2e-16 ***
mmr	1	3.0167e+13	3.0167e+13	10625667	< 2.2e-16 ***
residuals	558830	1.5865e+12	2.8390e+06		

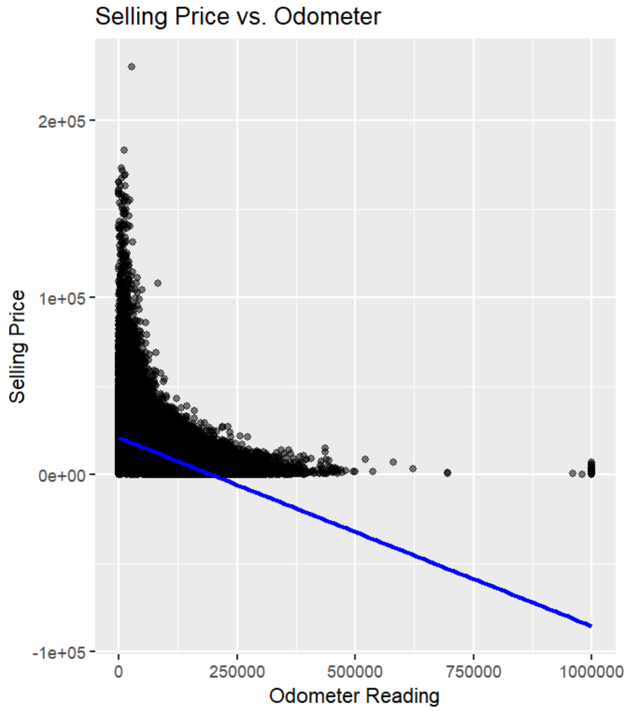
The ANOVA results indicate that the manufacturing year is a significant predictor in the selling price. The high F value and extremely low p-value indicate that newer vehicles tend to have different selling prices compared to older. The condition of a vehicle, the mileage, and Manheim Market Report value also have high F values and low p-values. This supports the assumption that these predictors contribute to the overall variability in vehicle selling prices.

5.2 VIF Analysis

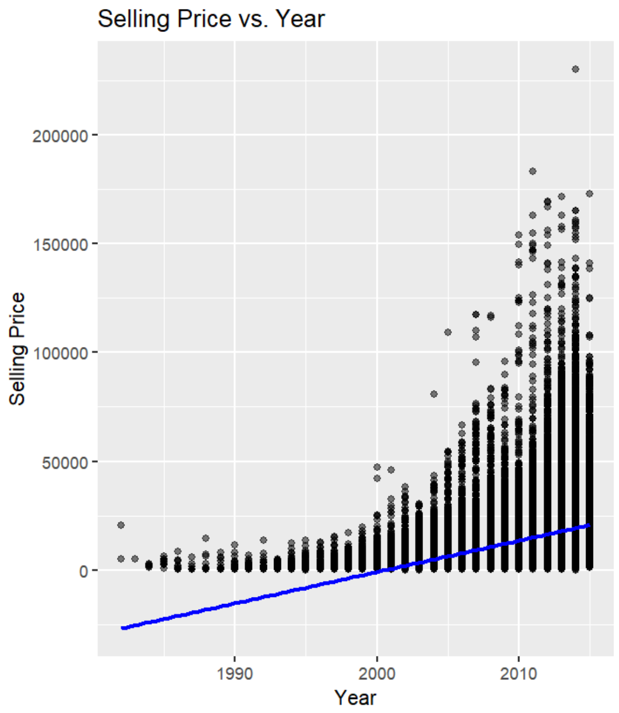
Predictors	**DF**	**GVIF**	**GVIF^(1/(2*Df))**
year	1	2.741483	1.655742
condition category	3	1.243537	1.036995
odometer	1	2.685019	1.638603
mmr	1	1.676095	1.294641

The VIF values indicate whether there is multicollinearity among the predictors in the regression model. This can affect the standard errors of the coefficients and make the model less reliable. Values under 5 indicate that there is no multicollinearity concern, while values ranging from 5 to 10 suggest moderate levels, and 10+ indicating high multicollinearity. All our values have low VIF values, indicating that the predictors do not have a high correlation with each other. This makes our estimates more reliable.

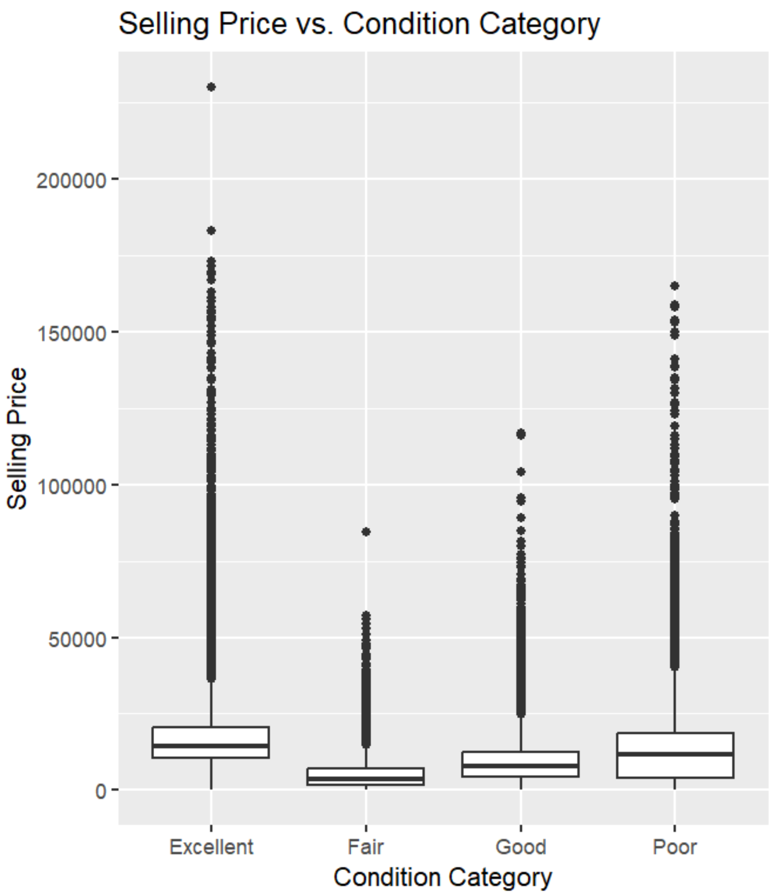
5.3 Predictor Plots



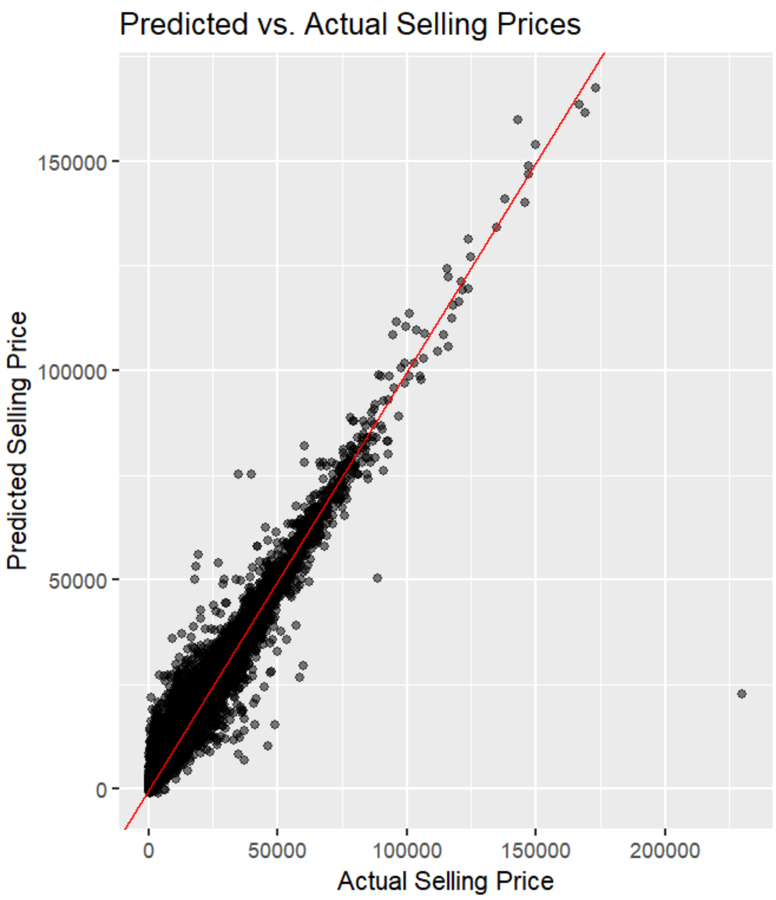
This scatterplot shows there's a downward slope of the linear regression line. This indicates a negative correlation between odometer readings and selling prices. This means that as the odometer reading (mileage) increases, the selling price of the vehicle tends to decrease. The wide spread of data points around the regression line indicate variability in selling prices for vehicles with similar odometer readings. The outliers could represent vehicles with unusual conditions or market circumstances affecting their selling prices.



The scatter plot analysis of selling price vs. year reveals a strong correlation between the age of a vehicle and its selling price. Newer vehicles tend to have higher selling prices, while older vehicles tend to have lower prices. This trend is consistent with the expectation that newer vehicles are more valuable. However, the variability in selling prices for vehicles of similar ages indicates that other factors also play a significant role in determining vehicle prices.

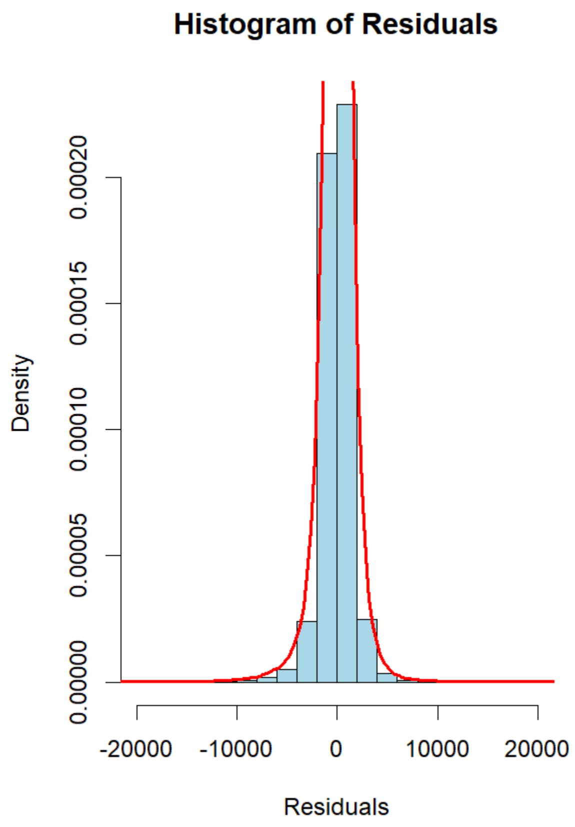


The box plot of selling price vs. condition category provides a clear visualization of how vehicle condition influences the selling price. Here we see that the median price for fair condition vehicles seems to be the lowest, with the highest for excellent condition. Poor seems to have the most variability in price, with good condition vehicles falling in between fair and poor.

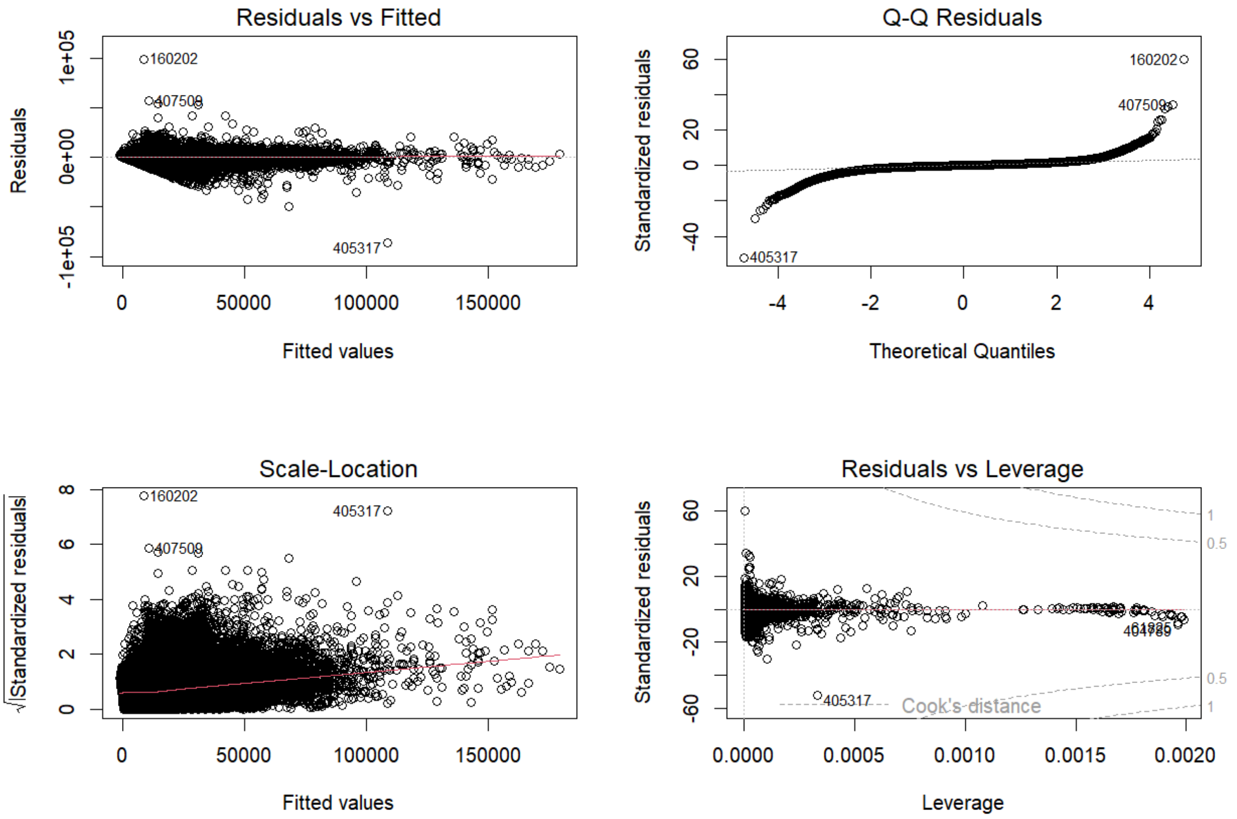


The plot comparing predicted selling prices against actual selling prices shows the quality of the model's predictions as seen in the proximity of the points to the reference line. This is further supported by the symmetrical distribution of these same points. There are a few outliers but overall the model seems to be a good fit for the data presented.

5.4 Residual Diagnostics



The histogram has a bell-shaped curve which indicates that the residuals follow a normal distribution. This is an important assumption for the validity of the linear regression model. This is further supported by the way it's centered around zero. There seems to be less variability as seen in the narrowness of the model. Overall, the histogram also supports the validity of this model for the data presented.



The residuals vs fitted plot indicates that the model generally meets the linearity assumption despite having outliers. The Q-Q plot suggests that the residuals have a normal distribution overall, with a few deviations at the tails. The scale-location plot tells us that there is heteroscedasticity as seen by the spread of residuals as the values increase. Lastly, the residuals vs leverage plot identifies a few influential outliers that might have an effect on the model, however the model seems to fit the data pretty well.

6 Discussion

Through this study, we analyzed how various factors play a role in affecting the selling price of vehicles. By analyzing a comprehensive dataset, we concluded that predictors such as year, odometer reading, condition category, and MMR are significant in helping make informed decisions on purchasing vehicles and finding reasonable prices.

7 Sources

1. Vehicle Sales Data, Vehicle Sales Data (kaggle.com)