Car Price Prediction Model: A MATH 189 Project

Mu Yang Anthony Liu, Pallavi Gaikwad, Catherine Pang, Boyu Tian, Matthew Ware

March 18, 2024

Abstract

The automobile industry holds significant importance in the realm of transportation, with car ownership being a prevalent choice among individuals in the United States. According to recent data, 56% of college students in the United States own a car, underscoring the widespread adoption of car ownership among various demographic groups. Our analysis delves into the car dealing industry to predict prices of Petrol, Hybrid, and Plug-in Hybrid vehicles, aiming to assist future consumers in making informed decisions when purchasing vehicles. Hybrid vehicles currently compose approximately 5.5% of the car market, reflecting a growing interest in eco-friendly transportation options. Our focus on environmentally conscious options reflects the growing concern about climate change. While electric vehicles offer emission-free driving, concerns persist regarding charging infrastructure and limited range. Hybrid vehicles emerge as a compromise, offering eco-friendliness alongside the convenience of traditional fuel sources for longer journeys. By examining industry data, our analysis seeks to provide insights into pricing trends for these vehicle types, helping consumers weigh their options based on cost, environmental impact, and practicality. It's worth noting that our study specifically targets future car buyers, aiming to empower individuals to make choices aligned with their values and needs in a rapidly evolving automotive landscape.

1 Introduction

Our research delves into the complex interplay between various factors and car prices within the U.S. automotive market. Employing two hypothesis tests and diverse regression models, our aim is to elucidate the determinants influencing the pricing dynamics of Petrol, Hybrid, and Plug-in Hybrid vehicles. This investigation is particularly relevant amidst escalating environmental concerns, notably regarding climate change, highlighting the significance of understanding and promoting sustainable transportation choices. Through rigorous statistical analysis, our study endeavors to equip prospective car buyers with the insights necessary to navigate the evolving landscape of eco-friendly options. By shedding light on the intricate relationships between factors such as car category, fuel type, and pricing, we strive to empower individuals to make informed decisions that align with both their values and practical considerations. Ultimately, our research seeks to contribute to a more sustainable and efficient transportation ecosystem, wherein consumers are empowered to choose vehicles that not only meet their needs but also minimize environmental impact.

2 Data Source and Description

The data gathering and cleaning process for this analysis involved several steps to ensure the dataset, sourced from Kaggle's Car Price Prediction Dataset, was well-prepared for hypothesis testing and regression modeling.

The dataset initially includes the following columns: ID, Price, Levy, Manufacturer, Model, Prod. year, Category, Leather interior, Fuel type, Engine volume, Mileage, Cylinders, Gearbox type, Drive wheels, Doors, Wheel, Color, and Airbags.

Initially, we extracted relevant features from the dataset, dropping the 'ID' of the vehicle as it's unrelated to our analysis, and we dropped the 'Levy' column as it was deemed non-predictive due to being based on the taxation and sales system employed by the seller. We then focused on the top 20 manufacturers, representing 95% of the industry, to keep the dataset focused on major manufacturers

in the United States market and prevent significant data loss. We chose to one-hot encode them to make it feasible to conduct regression modeling. Only vehicles produced after 2010 were retained because hybrid cars were massively introduced and produced in the market after 2010, and old cars with ultra-low salvage value are considered outliers in our prediction model. The 'Prod. year' was used to calculate the age of each vehicle. Next, we dropped buses and trucks since they are not commonly bought by college students. Then, we categorized vehicles into specific types, including Sedan, Jeep, Hatchback, Minivan, Coupe, and Universal, and converted categorical variables into numerical representations. This one-hot encoding helped us differentiate the type of vehicle in the dataset and conduct regression models. The leather interior was encoded as binary (0 for No, 1 for Yes) as it remains one of the most important interior factors that might impact the price of the vehicle. The dataset was further refined to include only Petrol, Hybrid, and Plug-in Hybrid vehicles, sourced from the Kaggle dataset as they are the main three types of energy sources we are going to explore. 'Turbo' was extracted as a new feature from the 'Engine volume' column as it is one of the key factors in describing engine performance, which impacts the price significantly. Mileage was converted to float, and only cars with automatic gearboxes were considered because most of the cars in production in the United States are automatic and all hybrid cars are automatic. Only left-wheel cars were retained because we are analyzing the car market in the United States, which is a left-wheel driving country, aligning with the focus on the USA market. Finally, 'Drive wheels' was encoded as 0 for Front, 1 for 4x4, and 2 for Rear, and irrelevant columns such as 'Doors', 'Color', and 'Airbags' were dropped.

This rigorous data-cleaning process ensures that the dataset is quantifiable and well-suited for regression modeling, enabling accurate analysis of pricing trends in the automobile industry.

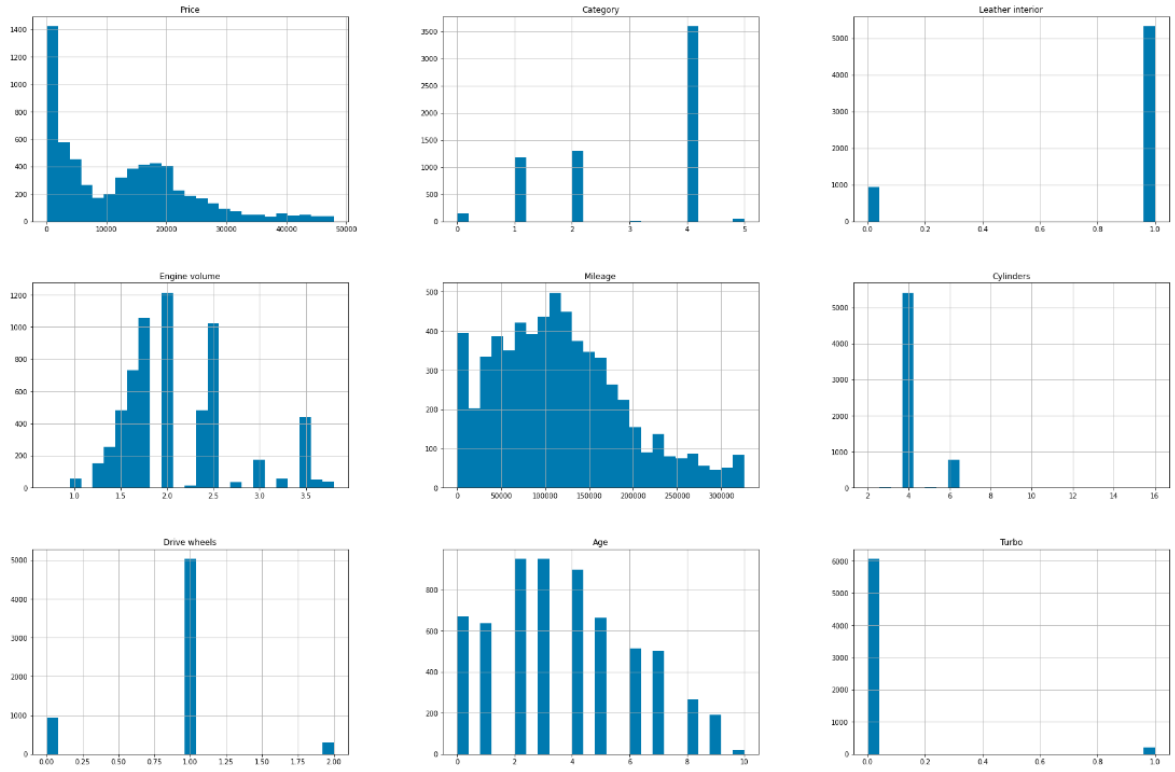

Figure 1: General View of Dataset after Cleaning.

3 Exploratory Data Analysis

After cleaning and filtering the dataset, we performed several data analyses to get a better understanding of the data. Bar graphs and a correlation matrix were computed to help us find patterns and new information that are relevant to predicting car prices.

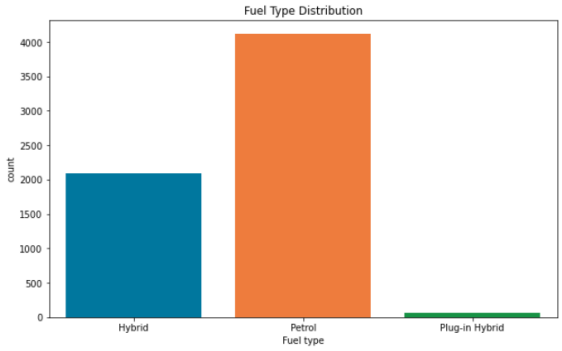

Figure 2: Bar Graph of Counts for Each Fuel Type.

From Figure 2, we see that that about 67% of the cars in our filtered dataset rely are petrol-powered cars, while only about 33% of cars are hybrid or plug-in hybrid. Assuming that the cars in this dataset were randomly chosen, there is a larger market for petrol-powered cars than there is for hybrid cars. Also, gas-fueled cars are dominant fuel-type in the automobile.

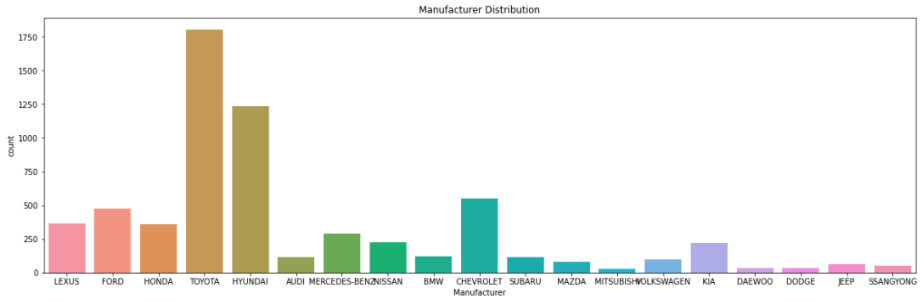

Figure 3: Bar Graph of Counts for Each Manufacturer.

In Figure 3, we see that the that Toyota, Hyundai and Chevrolet are the most common car manufactures in the dataset, with Toyota cars making up about 29% of the dataset, Hyundai making up 20.8% dataset, and Chevrolet making up about 11% of the dataset. We can infer that Toyota cars, Hyundai cars and Chevrolet cars are the most common cars bought and sold in the market. The counts of all other car manufacturers were relatively the same.

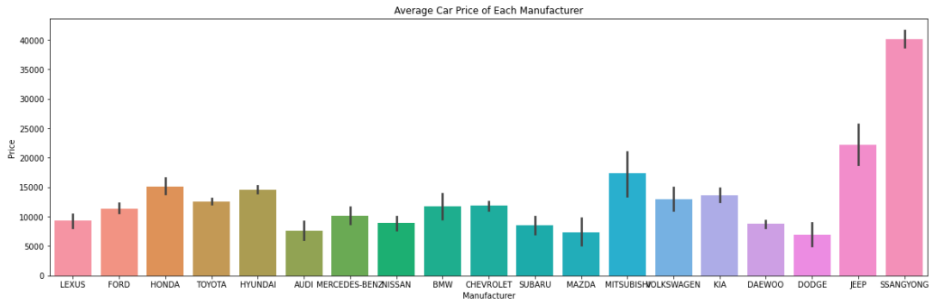

Figure 4: Bar Graph of Average Price for Each Manufacturer.

In Figure 4, we looked at the average price of a car given its manufacturer for each manufacturer in our dataset. The average price for a car given its manufacturer were all relatively the same with the average price being around $10,000 to $15,000. The notable exceptions were cars manufactured

by Jeep, which had an average price of $22,500, and cars manufactured by Ssangyong which had an average price of $40,000.

In Figure 5, we calculated a correlation heatmap for all the attributes in the dataset, where each value in the grid represents the correlation between the attributes in each corresponding row and column. We observe that the two largest numbers in the heatmap are the relationship Cylinders and Engine Volume (r = 0.67) and the relationship between Age and Mileage (r = -0.52). The first correlation value tells us that the number of Cylinders and Engine Volume are highly positively correlated with each other, which makes sense because cars with more cylinders are more powerful than cars with less cylinders, and the extra power requires an increased engine volume to make the car go. The second correlation tells us that Age and Mileage are strongly and negatively correlated with each other. This result makes sense because when a car has a higher 'Age', then the year that the car was manufactured was more in a more recent year. If a car was manufactured recently, then the car is less likely to have a lot of miles driven on it.

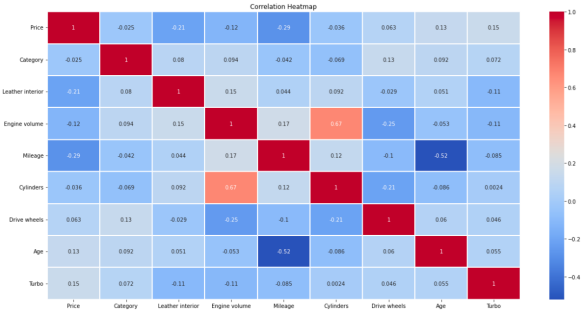

Figure 5: Correlation Heatmap for Each Attribute.

From Figure 5, we can also observe that the attribute most strongly correlated with Price is Mileage (r = -0.29). This negative value makes sense because we usually expect cars with higher mileage to be sold for a cheaper price due to depreciation. All other attributes had a weaker relationship with price.

After performing these exploratory data analyses, we have a better understanding of our dataset. We learned that the fuel-powered cars dominate the car market, Toyota is the largest manufacturer of the cars in our dataset, Ssangyong has the most expensive average car price, and that a car's Mileage is the strongest predictor of a car's price. Because of this better understanding, we can start performing more advanced analysis.

4 Analyses Performed for the Project

4.1 Hypothesis Testing: Is there a statistical difference between the prices of cars that run on standard petrol versus cars that are hybrid?

We sought to answer the above question because the rising popularity of hybrid and electric cars as alternatives to standard gas powered vehicles has continued steadily in the past 2 decades. On the same strain, policy limiting the sale of new gas powered vehicles is set to be introduced in California and rising gas costs made us question whether the prices of gas powered vehicles is similar to hybrid vehicles. Choosing between purchasing a gas powered vehicle versus a hybrid vehicle is likely to be a critical decision many consumers make when looking to purchase a car, so answering this question could better inform consumers when making that decision.

We chose to test this hypothesis using a permutation t-test because we want to know if our observed prices of petrol cars are drawn from the same distribution of observed prices of hybrid cars. Also, permutation tests are convenient because we don't need to meet any underlying assumptions about the distribution of the data. I plotted the distribution of both hybrid (combined hybrid and plug-in hybrid) and petrol powered vehicles separately to get an overall view of the data.

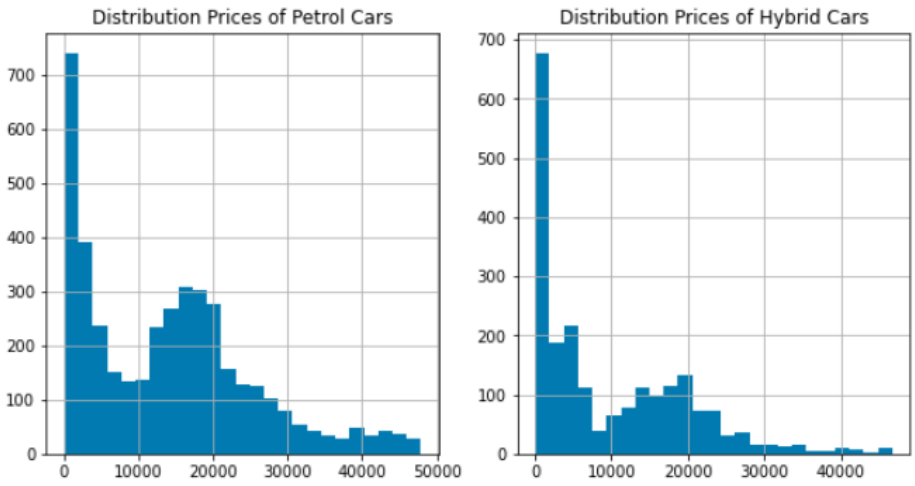

Figure 6: Distribution of Petrol Car Prices and Hybrid Car Prices

In Figure 6, we can see that both distributions have the same shape being right-skewed. This skew makes sense because we would anticipate that most cars are around the same price around $10,000 or $20,000, and then luxury cars tend to be more expensive.
After plotting the distributions, we followed the standard procedure of running a permutation test:

1. Defining a null and alternative hypothesis
2. Defining a test statistic and calculating the observed test statistic
3. Permuting the data and on each iteration calculating the test statistic
4. Plotting our distribution of sample test statistics
5. Calculating the p-value and making some conclusion

4.1.1 Null and Alternative Hypotheses

H_0: Prices of petrol fueled cars and hybrid cars come from the same distribution, and any difference is due to random chance.
H_1: Prices of hybrid fueled cars are lower than petrol cars, on average. The observed differences in our sample cannot be due to random chance alone.

4.1.2 Test Statistic

Our test statistic is defined as the following:

Mean of Petrol Car Prices - Mean of Hybrid Car Prices

Observed test statistic: $4213.23

4.1.3 Permuting the Data

We ran our permutation test, running 1000 iterations to permute the data. We shuffled the data by the 'Fuel type' column to randomly reassign the new labels to cars. After each iteration we calculated the sample test statistic, using our newly permuted data. We felt 1000 iterations was a sufficient number of trials for generating sample test statistics, without causing significant computational costs for running the simulation.

4.1.4 Sample Test Statistics Plotted

As we can see in Figure 7, we see the distribution of the sample test statistics that we generated, and the red bar denotes our observed test statistic. Our observed test statistic lies very far outside the sampled test statistics we generated.

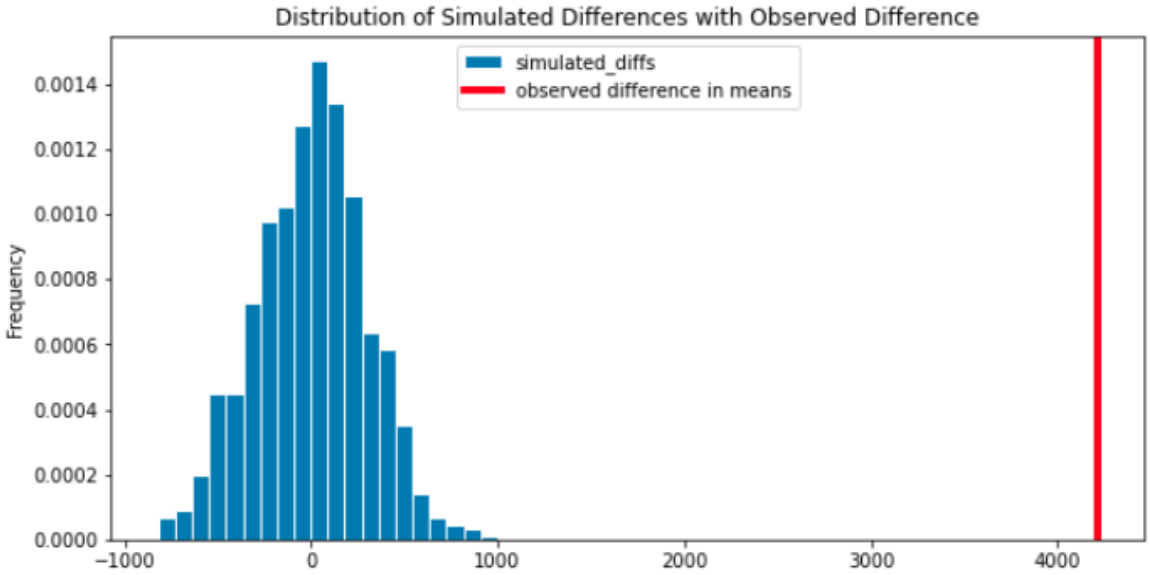

Figure 7: Distribution of Sample Test Statistics and Observed Test Statistic

4.1.5 p-value and Interpretation

After looking at Figure 7 and calculating the p-value to be 0.0 and $\alpha = 0.01$, we can reject the H_0 hypothesis. This essentially tells us that the mean price for petrol cars is significantly higher than hybrid cars, and that the difference in price means cannot be explained by chance. We cannot conclude that the reason for the difference of price is caused by fuel type, but there is a clear association between price differences and fuel types.

4.2 Hypothesis Testing: Is there a significant difference between car prices for cars of different drive wheel orientations?

After finding that hybrid cars on average are less expensive than petrol fueled cars, we wanted to explore relationships within the hybrid car subset of data. We wanted to explore the relationship between drive wheel orientations because we did not have any preconceived notions about how drive wheel orientation can affect car prices. We understood that different drive wheel orientations can affect whether a car is able to drive in different terrains, for example a 4x4 drive wheel orientation can drive through snow, while a front wheel drive car requires chains. For this reason, we wanted to see if specific drive wheel orientations would have significantly different car prices.

4.2.1 ANOVA and Assumptions

We selected ANOVA to run this test because we had three drive wheel categories (front wheel, rear wheel, 4x4) and wanted to explore the difference between each categorical mean. Before running the hypothesis test, we had to ensure that the data satisfied the following assumptions:

- each group drawn from a normal distribution
- all populations have a common variance
- all samples are drawn independently of each other
- with samples, observations are selected randomly and are independent

Looking at our dataset, we can assume that samples were drawn independently and randomly. However to check the first assumption we generated a graph to show the distribution of the data, and to check the second assumption we generated box plots to show variances in the data.

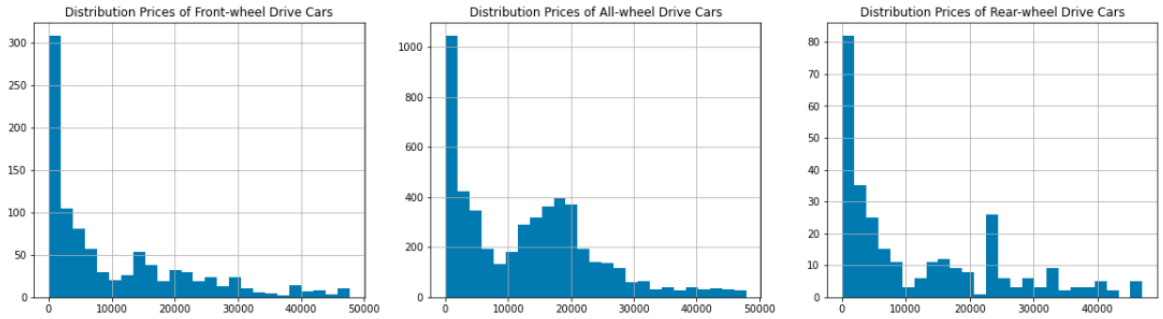

Figure 8: Distribution of Sample Test Statistics and Observed Test Statistic

In Figure 8, we see that the plots are not normally distributed, violating the first assumption. You can ignore this violation if the distributions look to follow the same distribution pattern (in this case, all three are right-skewed), and if the samples are approximately equal and sufficiently large. While all three samples are sufficiently large, they are not equal, so this assumption fails. We still proceeded with this hypothesis test, however we understand that our findings may not be accurate because we fail this assumption.

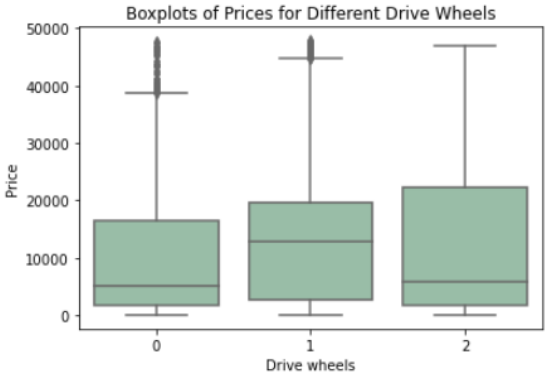

Figure 9: Distribution of Sample Test Statistics and Observed Test Statistic

Looking at Figure 9, we see that the variance of the different groups is roughly equal with majority of the points being concentrated at the bottom of the graph and having long tails. In this case, we pass the second assumption.

4.2.2 Hypotheses and Test Statistic

$H_0 : \mu_0 = \mu_1 = \mu_2$ (where 0 denotes front wheel, 1 denotes 4x4, 2 denotes rear wheel)
H_1: At least one μ is not equal.
Test Statistic: F statistic

4.2.3 ANOVA Test Results and Conclusion

We ran the one-way ANOVA test using the the scipy.stats import and the f_oneway function. After running the test, our F-stat was approximately 22.23 and our p-value 0.239. Looking at our F-stat initially, we may think that we can reject our H_0, but with a p-value as large as the one that we have, it may not necessarily be unlikely to get the following F-stat. This discrepancy is likely due to the fact that we do not meet the assumption for normality. We cannot reject the H_0 with these results.

4.3 Regression Modeling

4.3.1 Linear Regression

We first attempted to perform a linear regression on the data set with the price being the response variable and our explanatory variables being car category, leather interior, engine volume, mileage, number of cylinders, age, turbo presence, fuel type, and car manufacturer. We chose not to include 'Drive wheels' as an explanatory variable in our model due to the conclusion we made from our hypothesis test in section 4.2.

The calculated RMSE (root-mean-square error) of the linear regression model was found to be about 9,488.66. This means that the typical difference between our model's prediction and the actual price is \$9,488.66 and consequently, our model has an average error rate of 19.83%. The R^2 was calculated to be about 23%. Because the RMSE is relatively large and R^2 is relatively low, the linear regression model is not a good fit, and we should attempt to use a different model.

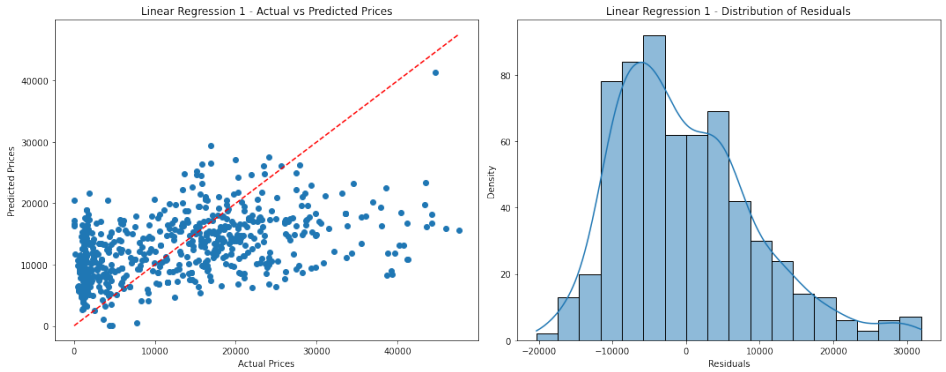

Figure 10: Linear Regression 1: Actual vs Predicted Prices and Distribution of Residuals

4.3.2 Polynomial Regression

Our second attempt to fit a regression model to the dataset was by doing polynomial regression. We used the same explanatory variables and response variable as the ones used in the linear regression model. We calculated polynomial regression models of degree 2, degree 3, and degree 4.

The calculated RMSE of the polynomial regression model of degree 2 was found to be about 9,098.94. This means that the typical difference between our model's prediction and the actual price is \$9,098.94 and our model has an average error rate of 19.02%. The R^2 was calculated to be about 29.2%. Although the R^2 and RMSE values for this model were better than the values from our linear regression model, we recognized that the RMSE was still relatively too large, and the R^2 was relatively too low for our purposes.

For the polynomial regression model of degree 3, we observed an even higher RMSE (11,039.27) and a negative value for R^2 (-0.04). Not only were these worse values than the polynomial regression model of degree 2, but they were worse than the values observed under the linear regression model. The case was the same for the polynomial regression model of degree 4, except we observed even more extreme values for RMSE and R^2. We concluded that we have an over-fitting issue when the degree of the model is greater than 2.

Because the polynomial regression model of degree 2 did not meet our standard of performance, and the other polynomial regression models performed worse than the linear model, we decided to proceed with a different model.

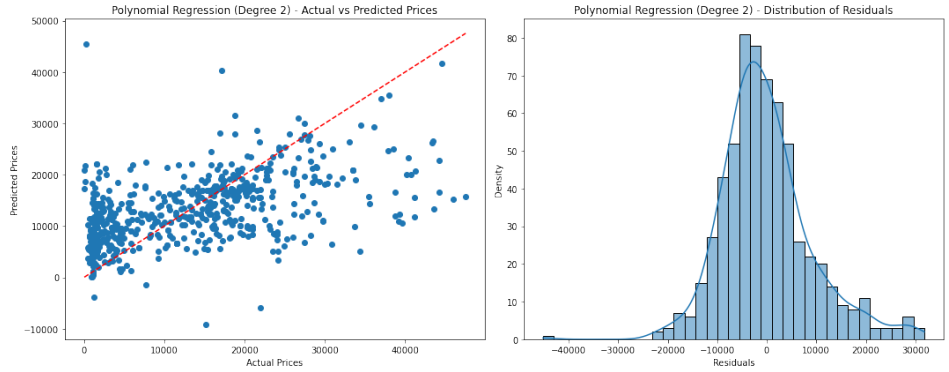

Figure 11: Polynomial Regression 1: Actual vs Predicted Prices and Distribution of Residuals

4.3.3 Decision Tree Regression

Our third attempt to fit a regression model to the dataset was by performing a decision tree regression. Decision tree regression was chosen for its ability to capture non-linear relationships and interactions between variables without requiring feature scaling or transformation. We used the same explanatory variables and response variable as the ones used in the linear model.

The calculated RMSE of the decision tree regression model was found to be about 8,712.83. This means that the typical difference between our model's prediction and the actual price is \$8,712.83 and has a corresponding average error rate of 18.21%. The R^2 was calculated to be about 35.09%. Although this model performs better than the polynomial regression model of degree 2, the increased performance was significant enough for meeting our expectations.

Figure 12: Decision Tree Regression 1: Actual vs Predicted Prices and Distribution of Residuals

4.3.4 Random Forest Regression

For our fourth attempt, we performed a random forest regression. We chose to proceed with a random forest regression model because of its high accuracy, robustness to overfitting, and ability to handle non-linear data effectively. By aggregating the predictions of numerous trees, random forest regression improves prediction stability and accuracy, making it well-suited for complex datasets with a large number of features and potential interactions. This method is particularly beneficial for capturing the intricate relationships within car data, providing a deeper understanding of factors affecting car prices while mitigating the variance and bias associated with individual decision trees. We continue to use the same explanatory and response variables as before.

The calculated RMSE of the decision tree regression model was found to be about 7,115.98. This means that the typical difference between our model's prediction and the actual price is \$7,115.98 which corresponds to an average error rate of 14.87 %. The R^2 was calculated to be about 56.7 %.

Because this model achieved a promising RMSE value and a large enough R^2, this model indicates a moderate level of accuracy and explanatory power in predicting car prices based on their characteristics. As such, this regression model can be trusted and used for the benefit of consumers when buying a car, or for the benefit of firms when selling a car.

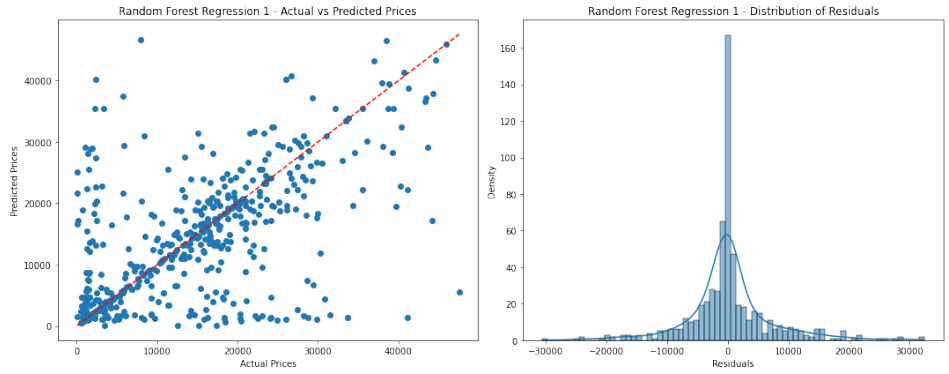

Figure 13: Random Forest Regression 1: Actual vs Predicted Prices and Distribution of Residuals

4.3.5 Moving Forward with Regression: Regression models on one manufacturer's cars

In our quest to refine the accuracy and applicability of our predictive models, we embarked on a strategic pivot towards analyzing a subset of the original dataset, focusing on vehicles from a singular manufacturer. This decision was informed by the understanding that consumers often have a predetermined preference for a specific car brand before they consider other purchase parameters. By tailoring our regression models to datasets concentrated on a single manufacturer, we aim to capture the nuanced variations and unique attributes inherent to that brand's vehicle offerings. This specialized approach allows for a deeper dive into the data, enabling the development of more precise and relevant predictive models. Such models are expected to outperform our previous, more generalized versions by providing insights that are closely aligned with the specific preferences and expectations of consumers loyal to a particular brand. This focused strategy not only enhances the model's predictive accuracy but also its practical utility for potential car buyers, offering them tailored recommendations that resonate with their brand-specific considerations.

In this case, we choose the manufacturer with the most instances of data in the dataset, which is Toyota. Then we perform the same 4 regression models as we did before.

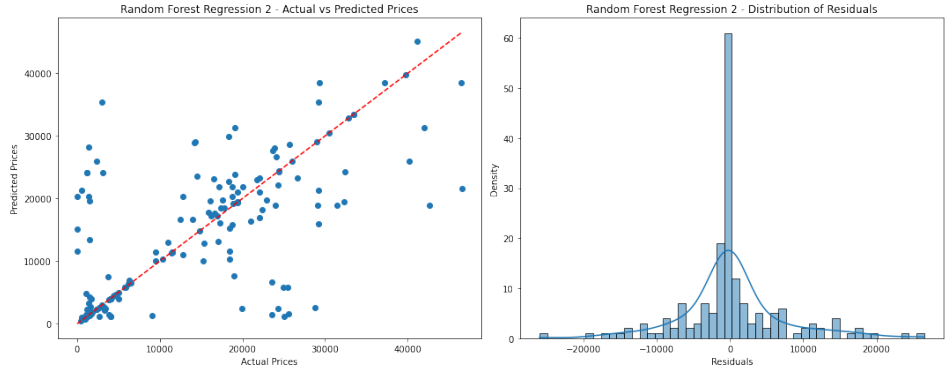

Figure 14: Random Forest Regression 2: Actual vs Predicted Prices and Distribution of Residuals

As we can see from the RMSEs, R^2 scores, and graphs in the notebook, focusing on one manufacturer indeed improves the models' performance in predicting the prices of a car given all other attributes. For example, from Figure 14, we can see that the distribution of points on the Actual vs Predicted

Prices graphs are more concentrated to the red line, indicating a higher prediction accuracy. Also, the distribution of residuals, which is still a normal distribution, is more concentrated to the 0 point and less outliers compares to the graphs in previous models. Therefore, we can draw to the conclusion that if we select one specific manufacturer and want to predict the price for its cars, the performance of the models are noticeably better than before. This result can help people better understand and foresee the future prices of the cars they intend to buy if they have decided which manufacturer they want to buy the car from.

5 Analyses Performed on Similar Dataset

Upon review, it has come to our attention that dome analyses have been conducted on a dataset similar to the one chosen for our project.

In the initial analysis, the relationship between selling price and kilometers driven was investigated. However, the findings did not conclusively establish a direct correlation between higher kilometers and lower prices based solely on the presented plot. This ambiguity may stem from the distribution of kilometers, particularly the prevalence of values within the 0–100k range. Further exploration is necessary to determine the precise impact of kilometers driven on pricing.

In the subsequent analysis, attention was directed toward the relationship between the selling price and the present price. The results indicated a positive correlation, suggesting that higher present prices correspond with higher selling prices. This observation aligns with common intuition, as it is reasonable to expect that more expensive cars would generally command higher resale prices.

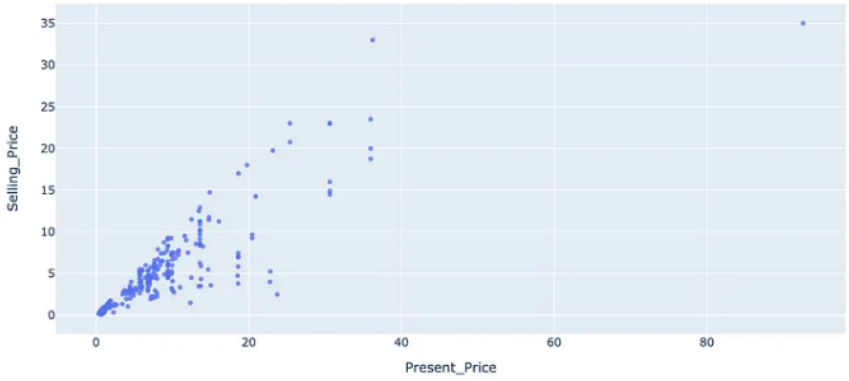

Figure 15: Relation between selling price and the present price.

The third analysis centered on the relationship between the selling price and the year of the car. The graphical representation displayed a discernible trend: newer cars typically fetch higher selling prices. This finding reinforces the conventional understanding that vehicles with more recent manufacturing years typically retain greater market value.

Furthermore, a correlation matrix was utilized to identify relationships between variables. The analysis revealed a strong correlation between the Present Price (the price at which the car was purchased when new) and the Selling Price. This significant correlation suggests that the feature "Present Price" could serve as a reliable predictor for our desired output, the selling price.

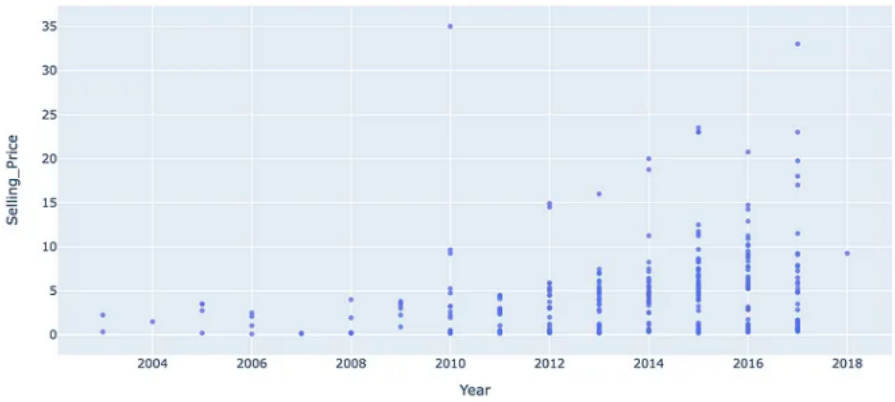

Figure 16: Relation between the selling price and the year of the car.

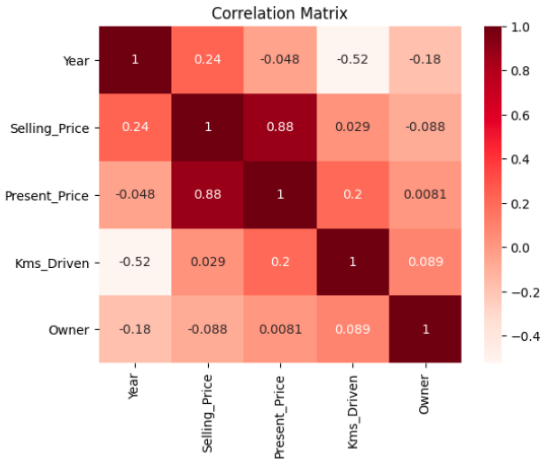

Figure 17: Correlation Matrix

Moreover, comparisons were conducted regarding selling price relative to fuel type, seller type, and transmission. Examination of the provided graphs indicated that cars fueled by Diesel tend to command higher prices. Additionally, vehicles equipped with Automatic Transmissions appear to be priced higher than those with Manual Transmissions. Lastly, it was observed that cars sold by individual sellers tend to be priced lower.

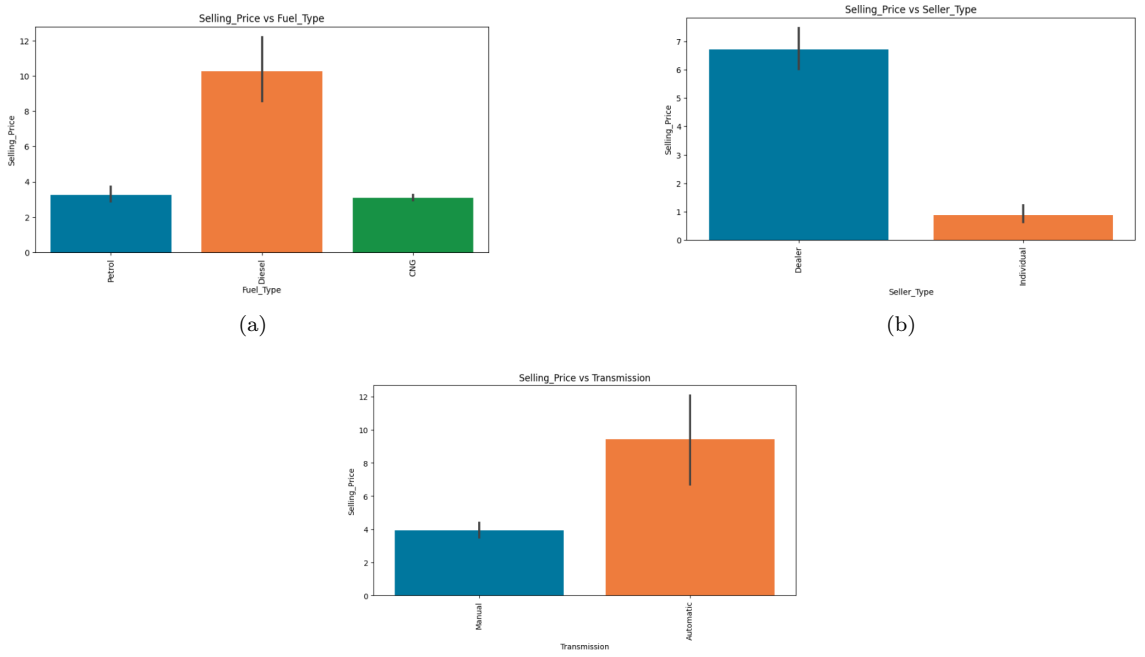

Figure 18: Fuel Type, Seller Type, and Transmission Type Comparison

6 Conclusions and Future Work

This study delves into the intricate dynamics influencing the market prices of automobiles, shedding light on the impact of specific variables. Through rigorous hypothesis testing, it was determined that the distribution of hybrid cars differs significantly from that of petrol cars. Conversely, the average prices among different wheel types were found to exhibit no significant differences. Additionally, various regression models were explored for predicting car prices based on explanatory variables such as car category, leather interior, engine volume, mileage, number of cylinders, age, and turbo presence.

Among the models examined, the random forest regression emerged as the most effective. This model was further utilized to forecast the price of Toyota-manufactured cars, leveraging the aforementioned explanatory variables. The integration of both hypothesis testing and regression modeling equips both consumers and car manufacturers with valuable insights into the automotive market landscape.

Our pursuit of refining car price predictions has unveiled promising avenues for future research. Feature engineering stands out as a critical area for improvement, holding the potential to uncover deeper relationships within the data through sophisticated variables and interaction terms. Furthermore, the optimization of model performance via techniques like hyperparameter tuning, encompassing grid search, and Bayesian optimization across various regression models, presents strategic opportunities.

Collaboration with industry experts is paramount for enriching our understanding of the automotive industry, consumer behavior analysis, and economic factors. This interdisciplinary approach will enhance our models, ensuring their alignment with real-world scenarios. Additionally, exploring alternative modeling approaches such as deep learning holds promise for capturing complex, nonlinear patterns within expansive datasets.

As we navigate this trajectory, our models will not only enhance predictive accuracy but also become more relevant and user-centric, in tune with evolving consumer demands and market trends. This holistic approach signifies a significant stride toward intuitive predictions that resonate with the complexities of the automotive landscape.

References

1. Dataset (https://github.com/diellor/machine-learning/tree/main/liner_regression_car_price_prediction)
2. Website link (https://medium.com/@diellorhoxhaj/linear-regression-car-price-prediction-and-data-analysis-112883cdd39b)