Authors: Ethan Freshman, Michael Klosowski, Roy (Re) Wang, Chris (Zichen) Xu

Professor/Section: Gold/B1

Code: [DS Project Final](https://colab.research.google.com/drive/1AAqe17shLvYYTNT9qIg49dTxdTKkz_Rs)

**DS110 Final Paper**

**Factors Most Essential in Predicting Car Price**

**Introduction**

For this project, our team decided to explore the predictive power of car features when attempting to predict the price, or value, that it is worth. To begin, we searched Kaggle for a dataset that contains an appropriate number of columns (variables) to predict price with, as well as a good number of rows to have a sufficient sample size. Our central question was finding if it was possible to accurately predict the price of cars based on their features. As a team, we chose to explore this because we wanted to understand which features were most important in determining the price of a car.

**Previous Work**

Used car salesmen may utilize a variety of resources to appraise their cars, such as looking at repair/accident histories and prices detailed in pricing guides published by Kelley Blue Book (Momot). Depending on the condition of the car, a car salesman may need to recondition, detail, or clean the car. Additionally, safety inspections are necessary to assure the car’s reliability. All of these services inevitably raise the price of the car if they are required. We further found some research which suggested that cars with Diesel fuel type and automatic transmissions are more expensive (D'Allegro). However, the results from our specific dataset found these variables to be inconsequential in determining the price of a car.

The importance of our central question becomes clear under the lens of the current car sales conditions. More so than ever, consumers are privy to countless details and histories about cars out on the lots (Linkov). With car salesmen being notoriously greedy, consumers should strive to know the true value of a car.

Consumer advice from the Federal Trade Commission about questions to ask before buying a car helped us decide on variables that we would later use in our regression analysis. These reports helped us hone in on registration year and miles driven as leading variables Consumer Advice). These variables, as we will discuss later, were ultimately predictive of car price.

Before starting this project we also searched the internet for any similar projects done in Python so that we could get an idea of how we would go about doing this successfully. We were able to find a helpful article from Medium detailing the steps to a linear regression model on car data, which gave us insight into the thought process behind taking certain steps in python(Diellorhoxhaj).

**Methodology - Data Preparation**

Firstly, we cleaned the dataset, eliminating any columns we deemed insignificant or that would work against our central question. We also dropped several columns including, 'genmodel\_ID' and 'adv\_ID', which did not contain meaningful information in regards to our central question.

Next, we addressed the categorical variables present in the dataset, such as 'bodytype' and 'body type'. To incorporate these variables into our regression models, we transformed them into dummy variables using one-hot encoding. This essentially makes our model treat these variables as quantitative.

Finally, we examined the presence of multicollinearity among the predictor variables. Multicollinearity occurs when two predictor variables are highly correlated to one another—the general convention is if their correlation coefficient is greater than equal to 0.8. Multicollinearity causes issues within regression models as it makes the impact of single predictor variables difficult to interpret.

To combat this, we calculated the correlation coefficients between all pairs of variables. We then visualized them using a correlation heatmap. Our heatmap, however, displayed no instances of multicollinearity.

The following provides the variables, and brief descriptions if necessary, used in our analysis:

* Price (target variable)
* Bodytype\_SUV: Binary variable indicating if the car is an SUV (1) or not (0).
* Bodytype\_Hatchback: Binary variable indicating if the car is a hatchback (1) or not (0).
* Reg\_year: The registration year of the car.
* Gearbox\_Manual: Binary variable indicating if the car has a manual transmission (1) or not (0).
* Runned\_Miles: The total mileage of the car.
* Average\_mpg: The average miles per gallon (fuel efficiency) of the car.
* Engine\_power: measured in horsepower (hp)
* Wheelbase: The distance between the front and rear axles of the car.
* Height
* Width
* Length

Cleaning our data ensured we did not encounter any unforeseen anomalies in any of our analyses.

**Methodology - Statistics**

To examine the relationships between the dependent variables and the independent variable (price), we performed a correlation analysis. Correlation coefficients were calculated between each variable and the price to identify the strength and direction of their linear associations. Variables with strong positive or negative correlations were considered potential predictors for further analysis and modeling.

Additionally, we performed a two-sample t-test to determine if there was a statistically significant difference in the mean prices between two groups of a categorical variable—convertible and non-convertible cars.

**Methodology - Machine Learning**

We employed two machine learning techniques to build predictive models for car prices:

1. Regression
   1. We started with a simple linear regression model, using registration year as the sole predictor variable to establish a baseline for predicting prices. Registration year was a variable we confidently felt would impact car price because of the previous work we had read about.
   2. To improve upon the baseline, we constructed multiple linear regression models with a combination of relevant predictor variables identified from the correlation analysis and previous research. We ran a few different multiple regressions to see the impact of several variables, and whether any variables had joint impacts.
2. K-Nearest Neighbors (KNN)
   1. We also explored the KNN algorithm. KNN models were trained using different sets of predictor variables and different numbers of neighbors (K). The performance was evaluated by measuring the accuracy scores.

In each case, the models were trained on a portion of the dataset, and their performance was evaluated on the remaining data using appropriate metrics such as R-squared (for linear regression) and accuracy scores (for KNN).

**Results - Statistics**

The correlation analysis revealed important relationships between our dependent variables and the price of cars. Registration year had a correlation value of 0.405, and engine power—0.686. These correlations suggest a strong positive correlation with price, indicating that newer cars and those with higher engine power tend to be more expensive. On the other hand, mileage had a correlation coefficient of -0.360 with price. This makes sense as its common knowledge that cars with higher mileage are generally worth less.

The categorical variables body type SUV (0.204) and coupe (0.228) exhibited a positive correlation with price, while hatchback (-0.285) showed a negative correlation. This suggests that SUVs and coupes tend to be more expensive compared to hatchbacks.

The t-test performed to compare the price means between convertible and non-convertible cars yielded a p-value of 3.105e-194, which is significantly lower than the commonly used threshold of 0.05—indicating statistical significance. We found that on average, convertible cars are more costly.

**Results - Machine Learning**

For our initial linear regression model, we used the registration year, and acquired a low R2 value of 0.164. In other words, the linear model was about 16% accurate in predicting the data. We chose the value initially because of its relatively high correlation value, however after performing more regressions, we found that engine size is actually the highest sole variable predictor in a linear regression, with a R2 value of 0.471

We ran one multiple regression with its dependent variables as registration year, gearbox type, and different bodytypes. This regression model yielded an R2 of 0.341. While this is significantly better than our linear regression, we wished to increase it further.

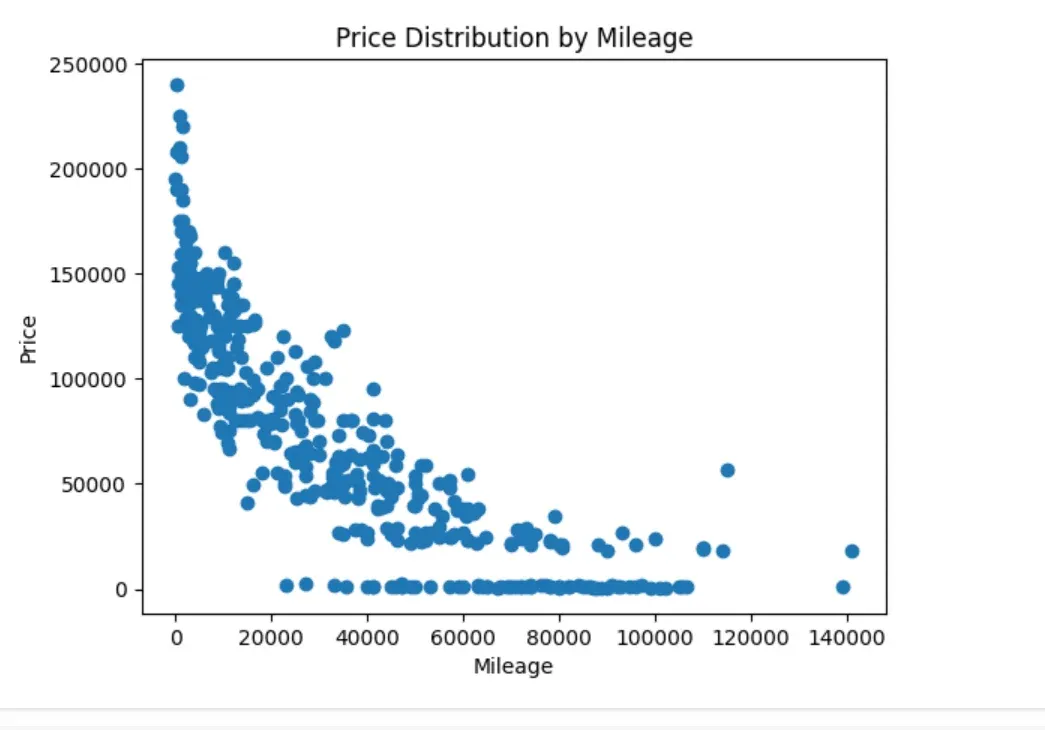
From our previous regression we determined that gearbox type and bodytype did little to add to the R2 value, so we omitted them in our second multiple regression. We also looked at the predictiveness of more variables and found that the combination of registration year, engine power, and mileage held the highest R2. We kept adding variables until the highest value was achieved, then removed variables that had minimal changes in R2 . This process resulted in a R2 of 0.604. This suggests that combining these three variables results in a multivariable regression model that is 60.4% accurate.

The coefficients of the multiple regression model revealed that registration year (999.442) and engine power (140.336) had a positive impact on the predicted price, while mileage (-0.062) had a negative impact, which aligns with the correlation findings.

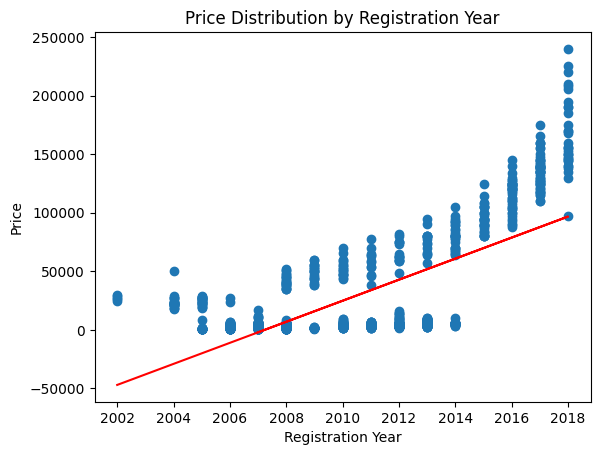
We also attempted a k-nearest neighbors (KNN) model using only the engine power, which resulted in an extremely low accuracy score of 0.005, which can also be seen as a .05% accurate model. However, when using the three most predictive features (registration year, engine power, and mileage), the KNN model's accuracy improved to 0.333, although still lower than the multiple regression model's performance. We found that the model was most accurate when k, the number of neighbors, was equal to 3.

The accuracy of either machine learning method can easily be doubted and our team believes that looking purely at features of cars does not provide enough information to make a reliable model. Intangible features, such as the perceived value of luxury, play an important role in determining the price of a car and must also be considered.

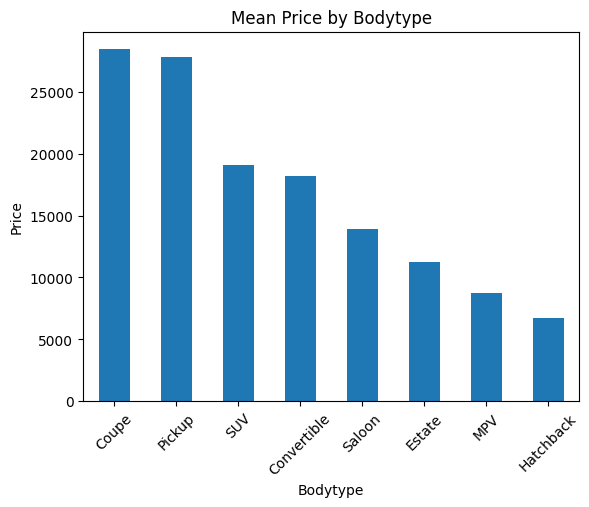
**Visualizations**



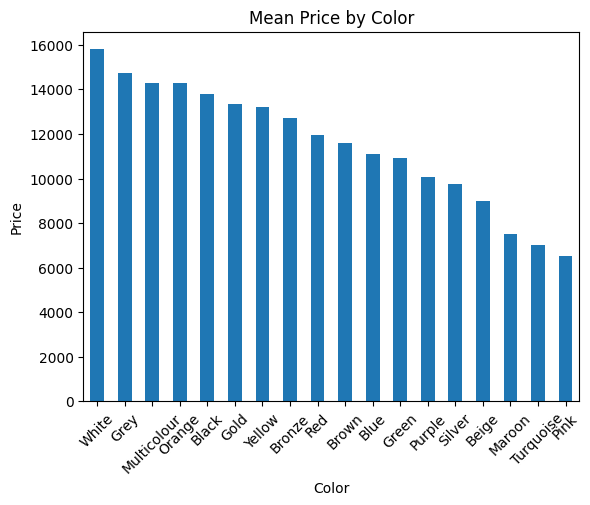
The scatter plot of mileage to price shows a negative exponential relationship, where higher mileage is associated with lower prices.



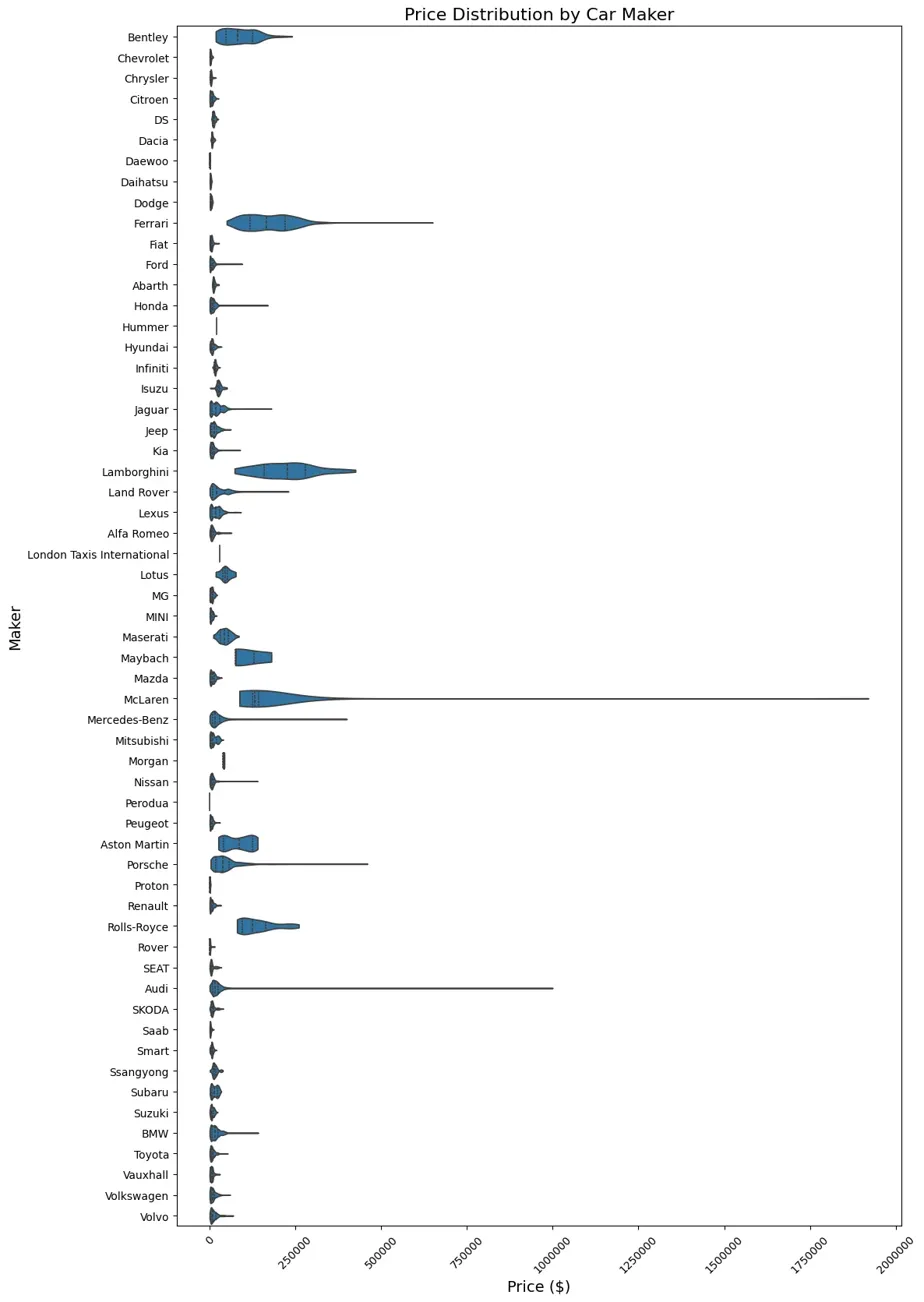
The scatter plots reg\_year vs. price shows a positive exponential relationship between registration year and price. This plot also shows a line of best fit for the data, however it would not be an accurate predictor of sale price.



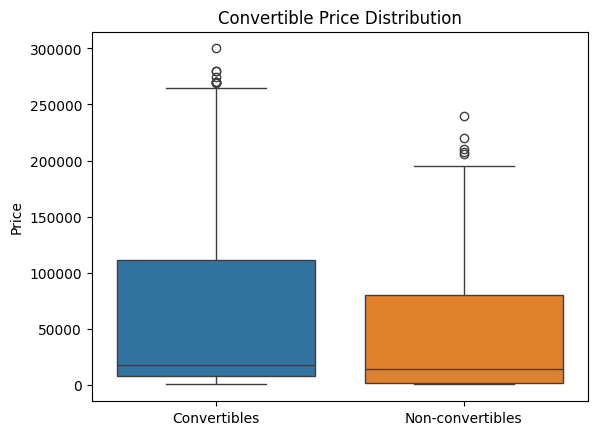
This histogram shows the mean price by car body type, with coupes and pickups being the clear leaders, while hatchbacks come in dead last.



This bar chart shows the distribution of mean car prices by exterior color. White cars are the most expensive, with pink cars being the cheapest.



The violin plot illustrates the distribution of prices across different car makers. It is evident that brands like Lamborghini, Ferrari, and McLaren have significantly higher prices compared to others like Morgan and Nissan, which occupy the lower end of the price range.



The box plot comparing the price distributions of the first 500 convertible and non-convertible cars visually confirms the results of the t-test. The boxes representing the two groups are separated, with the convertible group having a higher median price, supporting the finding that convertible cars are generally more expensive.

**Conclusions**

Based on the analysis, we conclude that registration year, mileage, and engine power are the most important predictors of car prices in this dataset. Newer cars with lower mileage and higher engine power tend to have higher prices, which aligns with common expectations.

While individual variables like registration year could not effectively predict prices on their own, combining multiple relevant variables in a multiple regression model significantly improved the predictive performance.

One limitation of this analysis is that the dataset may not be representative of the entire car market, as it likely contains a biased sample of cars available for sale. This dataset features many very high-end cars, so it is not entirely representative of the car market. High-end cars are priced more so on intangible variables, such as status or perceived value, which are nearly impossible to measure or quantify. Additionally, there could be countless other important factors not captured in the dataset that impact car prices, such as brand reputation, or the inclusion of additional features and packages.

Future work could also involve exploring more advanced machine learning techniques to further improve the predictive accuracy. Incorporating additional data sources or expanding the dataset to include a wider range of car models and conditions will also provide more comprehensive insights.

**Credits**

Ethan - He was supposed to do much of the coding for this project, including the linear regression of the multiple regressions, k-nearest neighbors, and the statistics. Ethan also helped in writing the paper. Along with Michael, he presented for the lightning talk.

Michael - He created much of the slides and paper. Additionally, he coded for the T-Test and a second multiple regression. Along with Ethan, he presented for the lightning talk.

Roy (Re) - He helped create several visualizations. He helped in conceptual areas, such as determining what variables to omit from the dataset.

Chris (Zichen) - He wrote some of the final paper. He also helped create several visualizations.

**Bibliography**

Buying a Used Car From a Dealer | Consumer Advice, https://consumer.ftc.gov/articles/buying-used-car-dealer. Accessed 28 April 2024.

D'Allegro, Joe. n.d. “Just What Factors Into The Value Of Your Used Car?” Investopedia. Accessed April 30, 2024. <https://www.investopedia.com/articles/investing/090314/just-what-factors-value-your-used-car.asp>.

Diellorhoxhaj. “Linear Regression - Car Price Prediction and Data Analysis.” Medium, Medium, 25 July 2023, medium.com/@diellorhoxhaj/linear-regression-car-price-prediction-and-data-analysis-112883cdd39b.

Linkov, Jon. “Expert Tips on How to Inspect a Used Car.” Consumer Reports, 19 October 2023, https://www.consumerreports.org/cars/how-to-inspect-a-used-car-a1377126659/. Accessed 28 April 2024.

Momot, Steve. “Demystifying Car Dealer Pricing: A Guide for Buyers.” Medium, 17 October 2023, https://medium.com/autohitch/demystifying-car-dealer-pricing-a-guide-for-buyers-c2812a681cc7. Accessed 28 April 2024.