**Statistical Analysis of the Price of Used Cars**

BAIS:3250 Data Wrangling: Final Project Report

Sylvia E. Jerden | May 9th, 2025

**Introduction**

The automotive market is a constantly shifting landscape, influenced by factors such as economic trends, supply chain disruptions, technological advancements, and consumer preferences. While new car prices are often dictated by the manufacturer pricing and dealership incentives, used cars present a far more complex pricing structure. One that can vary significantly for the same make and model based on individual factors such as mileage, year of manufacturing, and more.

For buyers and sellers alike, determining a fair price for a used car can be challenging. Unlike new cars’ standardized price, used vehicles can fluctuate due to many factors, such as depreciation, market demand, and public opinion. The emergence of online car marketplaces, such as Cars.com, has made it easier to compare listings, but the sheer variety of pricing can still leave consumers wondering: What truly determines the cost of a used car?

In this study, I aim to analyze data from Cars.com for vehicle listings across both 2023 and 2025, assessing how key attributes correlate with pricing trends. By using structured data collection and analysis, I seek to provide insights into how used car prices are determined and what buyers should expect to pay in today’s market. This project will not only highlight common pricing patterns but also explore whether certain vehicle characteristics lead to consistent price differences, helping customers make informed purchasing decisions.

**Data**

The data for this analysis comes from two sources. One is a Kaggle dataset, that claims to be scraped from Cars.com. The dataset was uploaded to Kaggle[[1]](#footnote-14421) 2 years ago, so all the data that comes from that dataset will be labeled as 2023. The dataset is 4009 rows long, with some outliers omitted. These outliers are either too old, too expensive, or have too many miles to be considered.

Additionally, I personally scraped more data from Cars.com. These rows will be labeled as 2025. The data was scraped from the website on May 9th of 2025, if the web scraping were to be run again, the results would be slightly different due to constant updates. This dataset is 280 entries long.

Combining the datasets was relatively easy, as the data from the Kaggle datasets was scraped from the same website. However, some columns were omitted from the Kaggle set I wanted to include, and some columns I was unable to scrape that were in the Kaggle set. The analysis for these columns, therefore, must be considered with that in mind.

There are several columns of data in the dataset that are not necessarily used for any specific statistical testing. These are left in for two reasons. For one, I wanted to leave them in when searching for specific cars within the set so that I would be able to find more specific details about the cars. Secondly, if I wanted to do more with this dataset in the future involving said columns, I would not need to edit the current dataset.

**Data Dictionary:**

|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | Description |
| brand | object | The brand that makes the vehicle, examples may include Ford, Honda, or Mitsubishi. Sometimes referred to as “Make”. |
| model | object | The model of the vehicle. The specific name or initials given to the vehicle to differentiate it from others made by the same company. Examples may include Civic, Model Y, or F-150. |
| model\_year | int64 | The year the vehicle was manufactured. |
| mileage | int64 | The miles that the vehicle has already been driven. |
| fuel\_type | object | The type of fuel the vehicle takes. |
| transmission | object | The type of transmission the vehicle has. Can typically be divided by Automatic and Manual, but these both have many variants. |
| ext\_col | object | The exterior color of the vehicle. |
| int\_col | object | The color of the vehicle's interior. |
| data\_year | int64 | The year in which the data was collected. This column has either a value of 2023 or 2025. 2023 indicates that the data comes from the Kaggle dataset, and 2025 indicates the data was scraped. |
| Exclusive to Data Year = 2023 | | |
| accident | object | If the vehicle has been involved in an accident or experienced damage. |
| clean\_title | object | If the title of the vehicle is clean. A clean title means that a vehicle has not been significantly damaged, rebuilt, has had no issues and so on. A clean title also means ownership is more straightforward. |
| Exclusive to Data Year = 2025 | | |
| Drivetrain | object | Drivetrain of the vehicle (Front-Wheel, AWD, etc.) |
| MPG | object | Miles per gallon the vehicle receives. |

**Analysis**:

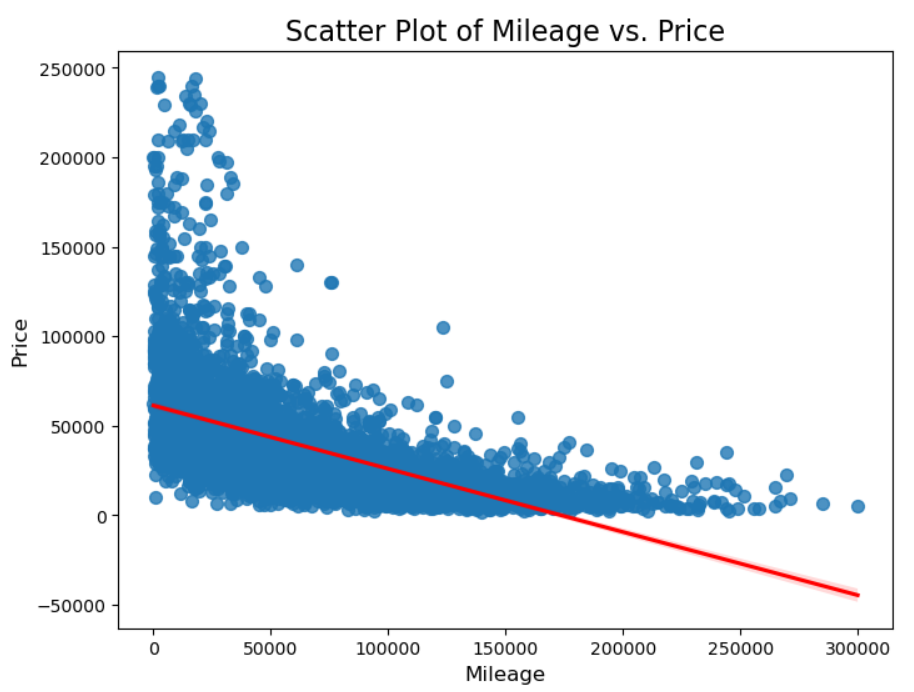
**General Analysis**

I ran a general analysis of my target variable of price. In the dataset, the average price of a vehicle is $38,694 and a median of $30,599, with a standard deviation of $32,565. This shows me that my price column is skewed to the right even after tuning the data to remove outliers. Additionally, the range of prices is $2,000 to $244,968.   
 I also ran some additional statistics on mileage, as this is one of the features I use the most in my analysis. The vehicles in my dataset had an average mileage of 64,156 miles and a median of 52,500, which shows me that compared to the price, mileage is skewed to the right significantly more. Additionally, the range of mileage is 100 miles to 300,000.

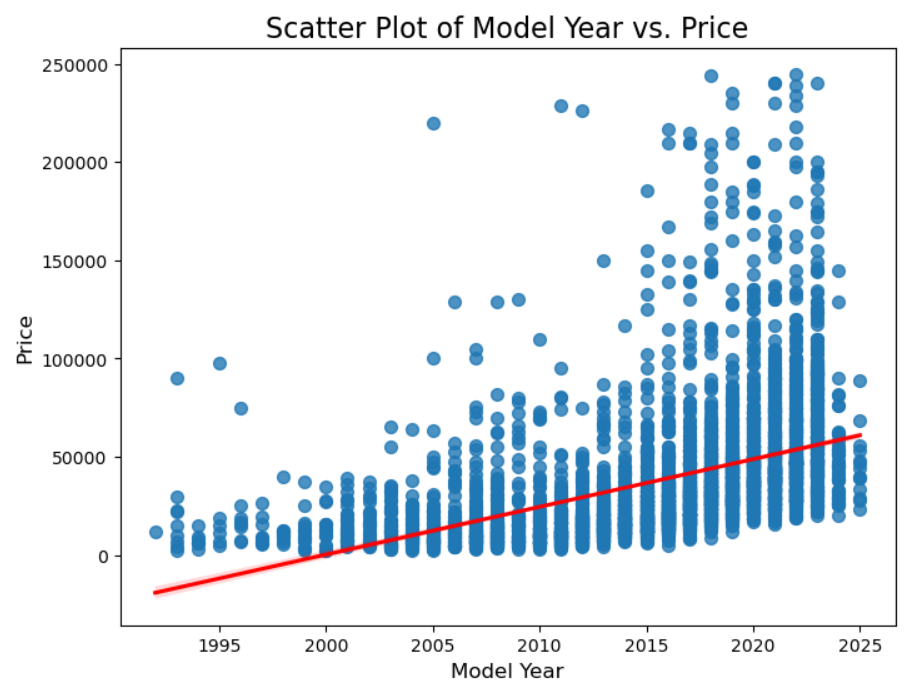
**Pearson Correlation Tests**

To begin, I decided to run two Pearson correlation tests on the primary columns I was considering for my analysis later. These were done with my target variable of price, and my considered features of mileage and model year. I figured that these would likely be determining factors in the price of a vehicle, but if I ran correlations on them, it would give me stronger evidence for this for later.

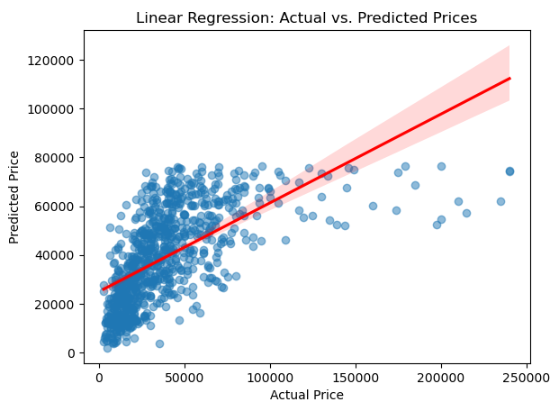
The first correlation test I ran was between the price of a vehicle and the mileage on the vehicle. I predicted that the two features would be negatively correlated, as typically the more that something has been used, the less that someone is willing to purchase it. This is especially true for vehicles, because as they age, they tend to have costly mechanical issues. What my correlation test produced was a negative correlation coefficient of -0.5493. This means that there is a moderate negative correlation between the price and mileage of a vehicle. I also found a p-value of 0, finding that a significant relationship exists between the two features.

 The second correlation test I decided to run was between the price of a vehicle and the year it was manufactured. This required a bit finer tuning to run compared to the previous correlation, since a lot of the cars in both datasets tended to be newer with the vast majority being manufactured between now and 2015. I also needed to focus on removing one outlier, as there was one vehicle that was manufactured around 1975, which would’ve not only skewed this correlation, but also other tests as well. Another reason specific older vehicles need to be removed is because even though older cars do not have the desired features and (sometimes) quality of newer cars, as certain cars age the demand increases. This would throw off testing, as what I am seeking out is a car that could be used as a daily vehicle, not an antique.

Since newer cars tend to be the most expensive due to many factors such as them being considered less ‘used’, I predicted that there would be a positive correlation between model year and price. The correlation my test produced was 0.4526, which implies that there is a moderate positive correlation. This is likely closely related to my previous correlation, as newer cars tend to have less miles on them. In my testing, I found that this test had a p-value of 0 as well, meaning that a significant relationship was found.

**Regression Analysis**

For my regression analysis, I chose my target variable to be the price of a vehicle, and for my features I chose mileage, price, and if the car is Japanese. I first ran a Linear Regression, which yielded an r^2 score of 0.4789 and coefficients of mileage = -1.217e-05, model\_year = 3.373e-04, and Japanese = -3.029e-01. This also produces a regression shown below.



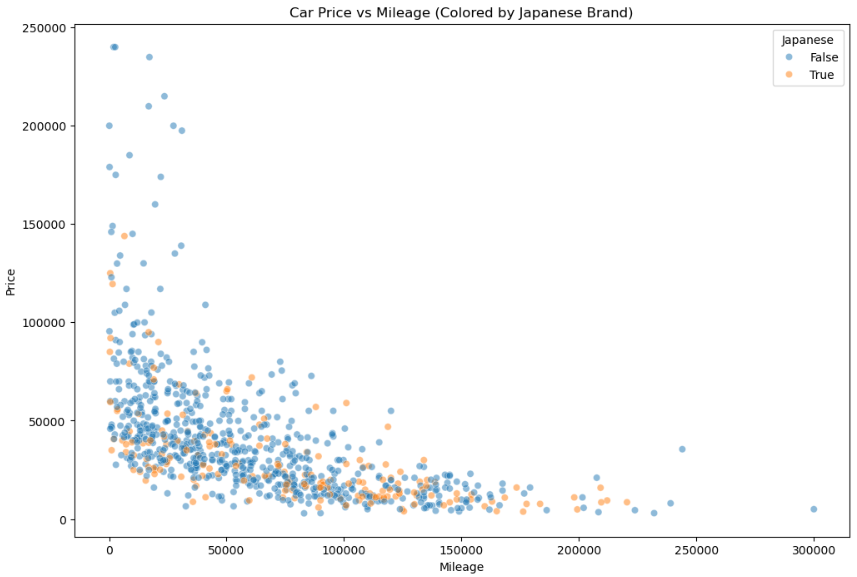
What this shows is that the model is not the best at predicting the actual price of the car, and this is likely due to several factors. One would be that I didn’t give it enough features, but on the other hand adding more features could lead to overfitting, as the rest of the columns within my dataset are categorical and would require a large number of dummy variables. Additionally, my two features of mileage and year may be causing collinearity problems, as an older car likely has more miles. Additional to all of this, the location of the points doesn’t necessarily have a strong positive linear relation to begin with.

In addition to a linear regression, I ran two additional models: Ridge and Lasso. Both models produced the same r^2 score of 0.33. This is probably since my features given for both models are very limited, so my lasso model is not reducing features meaning they are both running the same features. Therefore, nothing is changing between models, and they are running the exact same.

Overall, I think that regression is not the best fit for this dataset. Most of my features are not numeric or have too many unique values to effectively be transformed into dummy variables. In combination with this, a lot of my nonnumeric features do not have the same phrasing for similar things, such as colors or transmissions which are manufacturer names and not necessarily standard names.

**Classification Analysis**

For my classification analysis, I am unable to use my target variable from previous sections, as it would not make sense to use classification. Instead of my target variable being price, I chose to try and predict if a car is made by a Japanese manufacturer. Japanese cars are considered some of the best cars on the market in general by some, and this is for numerous reasons, but reasons that are important to consider when buying used. For one, they are common in the United States, which makes it easier to find a mechanic to work on one for routine maintenance as well as the occasional problem. Additionally, they tend to be reliable, with the typical vehicle making it to 200,000 miles with minimal maintenance[[2]](#footnote-31789). However, I’d like to see just how true these assumptions are. In my dataset, the 10 car manufacturers that are Japanese are: Lexus, INFINITI, Acura, Toyota, Nissan, Honda, Subaru, Scion, Mitsubishi, Mazda, and Suzuki (some of these are owned by others, but these are all that are found individually). For all of the classification models, the features considered are Price and Mileage. They follow a distribution within the dataset as shown:

 I did four total classification analyses: Logistic Regression, K-Nearest Neighbor, Support Vector Machine Analysis, and a Decision Tree.

I started with a logistic regression, which yielded an accuracy of 0.7887. The regression had similar coefficients to my linear regression in terms of size. Because of this, it was very difficult to form a logistic regression curve. I think in the future for better visualization, the features may need to be scaled. This is especially true if other features are added that are not as large as the average price and mileage.

In the three other models I ran, Support Vector Machine analysis yielded the best accuracy out of the four models, with an accuracy of 0.7863. This was followed by K-Nearest Neighbors with an accuracy of 0.7391 and Decision Tree with only 0.6800. This shows me several things. Firstly, I believe that the decision tree performed the worst simply because I did not include many features, leading to not many splits being performed in the first place. KNN performing second worst implies to me that there isn’t much clustering occurring, which would be backed up by the scatter chart I made previously of these datapoints. Japanese cars do appear to have a relatively even split across the dataset in terms of the features provided. However, given the fact that only two features were considered for Logistic Regression, it succeeded due to it does not doing as much additional analysis as other classification models, causing it not overcomplicating the predictions as much.

**Conclusion:**

One of the things that I initially set out to uncover with this project was the relationship between certain features of a car and how they relate to price, which related to my regression analysis. While a low r^2 score is not necessarily a bad thing, it shows me that my model has room to improve. I think that certain features could be uncovered in this dataset, such as drivetrain, brand reputation, and the condition of the car. This would require more intensive analysis of features currently in the dataset. However, in its current state, the model does show me that there is a significant relationship between the features already. Alongside this, I believe that adding these features could enhance the ridge and lasso models, letting them function the way they are intended.

My classification models enhance this point. My logistic regression model outperforming my other models implies that they would have increased accuracy if more features were added. However, what the logsitic regression does prove is that there is a moderate linear separation between Japanese and Non-Japanese cars based on price and mileage.

Overall, the best thing that can be done in the future with this dataset is more refinement of current features in a way that allows them to be turned into dummy variables without virtually making a new column for each individual row. Alongside this, more data scraped from cars.com without the Kaggle set with every feature possible off the page would enable deeper analysis. I think that all of these could create a better model in terms of regression and classification.

1. https://www.kaggle.com/datasets/taeefnajib/used-car-price-prediction-dataset/data [↑](#footnote-ref-14421)
2. https://www.next-drive.co.jp/blog/trust-japanese-vehicles [↑](#footnote-ref-31789)