Carson Angell
DS 2500
10/5/2025

Final Project Proposal

Central Argument:

For this final project, I'm answering one of the most nagging questions I've had about the car industry. What brands have the least amount of value degradation? Everywhere, I see car listings with a huge variety of offers. High mileage - high price, low mileage - low price, and everything in between. I want to see why. And to make my inquiry more specific, I will narrow it down to the top 10 most popular car manufacturers.

There's a good deal of work that has been done on this question. Many analyses have been posted to show value degradation with more specificity into the different car models. What's interesting is that the data used to answer this question is also used to see what features consumers value. By running multiple regression between car features and listing prices on used cars, companies see what features we value most. Market analysis and price prediction are common use cases for this kind of data.

Datasets:

I will be using three datasets for my inquiry. Two from Kaggle, and an independent one I found on GitHub. All the data will be uploaded and cited to a public GitHub repository from which the notebook will make requests. Here is a list of them and what they are going to be used for:

- Car Sales:

This dataset contains metrics of specific car models, such as engine type, horsepower, etc. But more importantly, how many of each car have been sold up to when the dataset was made (2017). By grouping the data by brand and summing all the sales, we can see the brands with the most sales. This is a good indication of popularity. I will use the top 10 highest-selling brands for my analysis.

This data does contain some missing values, but none of them are in the sales or brand columns. Its structure is also well-made and organized. I will not have to do any cleaning with this dataset.

- Used Car Price Predictions

This will be the primary dataset used for my analysis. It contains listings of used cars for roughly 850,000 independent vehicles. With valuable data such as listing price, mileage, make, model, and year. I can compare the car listing prices with the model's MSRP to get a percent degradation for each car. I can use this data in later statistical testing to answer my question about

which brands have the lowest overall degradation. I can also utilize the listed mileage to indicate the car's usage in later analysis.

Very surprisingly, this data set does not contain null or empty values in the columns that will be used in the analysis. However, the listed mileage is in kilometers, and for the sake of readability, I will create a conversion column that lists the mileage in miles for each car.

- DVM-CAR

This dataset contains a handful of different tables. From which I will be using two in my analysis. The first contains MSRP data for a large number of cars, organized by make, model, and year. I will be correlating the make and models from the previous dataset to get an MSRP for each car model. I can use this to calculate the value degradation.

The second table is very similar to the Car Sales dataset; it contains different cars and how many were sold. This will be used as an external data source to check against when picking the top 10 most popular brands. It's organized by 23 columns, the first three are make, model, and an ID. While the rest are years ranging from 2001 to 2020, these columns contain integers for how many units of the car were sold that year.

In terms of cleaning, I may melt the MSRP table to make it easier to sum the sales and determine popularity. However, neither of the tables contains null or missing values, and they are structured well apart from the sale years.

Here is some extra information about each of the datasets:

Name	Last Updated	Source (where did the author get the data)
Car Sales	2017	Pulled from Analytixlabs
DVM-CAR	2023	Undisclosed collection. This is a public dataset used for AI image training, research, and analytics.
Used Car Price Predictions	2022	Unknown, based on the description, it seems like a teacher who posted the dataset for their students.

Analysis:

- When choosing the top 10 popular car brands, I will show some simple bar graphs of the car sales amounts to illustrate popularity and external data checking.
- Before starting the analysis, I will check for normality on the data I will be using, like mileage, price, and later the degradation value, with probplots and distribution graphs.
- I will show some sample scatter plots of the potential correlation between mileage and value degradation for the top 10 brands. Since mileage can be a large factor in determining a vehicle's value, seeing some sort of trend between the two variables can validate my results. I may go as far as to do some Pearson Correlation tests between these variables.
- After all this, I will be using an Analysis of Variance (ANOVA) test to determine if there is a statistically significant difference between the degradation values of each car brand. Since this test can compare multiple variables of the same unit and type, it's suitable for comparing the percent degradation between the 10 brands.
- Along with that, I will do a Tukey Honestly Significant Difference (HSD) test to show the statistical significance between each pair of car brands. This test is an elegant solution for finding out which brands are specifically different from one another after an ANOVA test. It will be shown as a table and allow me to draw conclusions about which brands retain more value than the others.

Limitations:

- Looking for a correlation between a car's mileage and the value degradation may prove to be inaccurate since mileage isn't the only factor in determining a car's value. Things like the make, model, year, and overall feature set also play a huge role.
- Trying to connect the MSRP and car listing datasets by make, model, and year in order to get the correct price may prove to be difficult. Things like string parsing, model inclusivity, and the chance that the dataset doesn't contain a certain brand can get in the way.
- The Used Car Prediction Prices dataset is large, roughly 55MB, with more than 850,000 rows. This can make analysis and programming slower and harder to interpret.