BAIS:3250 Data Wrangling Project Proposal

Introduction:

While it is a near necessity in our current society to own a vehicle, shopping for a car can be a daunting task for many, including myself. Nowhere is that more obvious than the used car market. However, due to the advent of the internet, it is much easier to compare vehicles than on a lot while talking to a salesman.

Due to many factors, such as the chip shortage of 2020, we’ve seen an increase in the cost of not only new cars but used cars as well. This is partially due to the demand for cars not being met by the production of new cars, leading to many turning to buying used cars instead. With COVID-19 restrictions lifting, the demand for motor vehicles increased due to increased travel was met with a car industry unable to meet the demands of customers. This has led to a sharp increase in the price of used cars, that due to the tariffs placed recently, is unlikely to go down anytime soon.

I would really like to upgrade my car sometime in the not distant future. I currently drive a 2005 Chrysler PT Cruiser GT Coupe in Navy Blue with a Tan soft-top. It was purchased in 2018, before the chip shortage and during the month of February, for $4000. I don’t think that is a conceivable price for a vehicle. If I were to try and sell it, I could likely get it for the exact price it was bought, despite the miles put onto it and its several issues. There are aspects that I like about my car, such as the color, which was done with a reflective paint. However, there are other aspects that I do not like, such as the amount of maintenance I have had to have done to it. Because of the experiences with my own car, as well as things I have heard from the people around me and those on the internet, I have begun to form opinions on what I’d like in a new car. Some of these things include:

* Blue Preferably
* A car from a Japanese brand, such as:
  + Toyota
  + Honda
  + And more
* Something made more recently (2015 or later)
* Less than 70,000 miles
* And many more things

There are many aspects to consider when buying a car from a general perspective as well. Buying a car with a clean title and no accident or damage reports is good practice, but some things such as minor scrapes can be ignored depending on your budget. While color preferences aren’t a bad thing, sometimes meeting other criteria may lead you to picking your second or third favorite color. It is also not a good idea to buy a manual car if you do not know how to drive manually, unless you are willing to learn if the car checks off all your other boxes.

I want to have a better scope of the current used car market, and the trends with pricing compared to the make and model as well as the mileage. I would also like to more statistical trends that may be impacting the market near me. Besides this, I would like to compare this to data that is from 2022, to see how the current market compares, and how soon I should be looking at possibly buying a new car.

Data:

Currently, the [Kaggle dataset](https://www.kaggle.com/datasets/taeefnajib/used-car-price-prediction-dataset/data) that I have found was scraped from cars.com in 2022. My plan is to scrape more data from cars.com and use it as a comparison tool. As of now, I have three sample entries scraped. Most data should match, although my scraped data does not have all the same columns as the 2022 data. This can possibly be scraped at a later time.

|  |  |  |
| --- | --- | --- |
| Field | 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 | int32 | The miles that the vehicle has already been driven. |
| fuel\_type | object | The type of fuel the vehicle takes. |
| engine | object | The type of engine the vehicle has. |
| 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. |
| accident | bool | 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. |
| price | int32 | The price of the vehicle. |
| Data Exclusive to Scraped Data | | |
| Drivetrain | Object | Drivetrain of the vehicle (Front-Wheel, AWD, etc.) |
| MGP | Object | Miles per gallon the vehicle receives. |
| VIN | Object | Vehicle identifying number. The serial number of the car. |

Proposed Analysis:

One research question I would like to ask is how different the datasets from both 2022 are from present. For this, I would like to perform many statistical tests on the data, such as t-tests. I’d also like to do a large number of visualizations to help see the difference between the datasets much more clearly.

I’d like to see how much certain features of the dataset correlate with each other. For instance, I want to see the relationship between price and mileage, or the relationship with transmission and other features.

For a lot of features, it could also be interesting to see what is seen most frequently. It could be interesting to see if car color trends match used cars the same when trends are forecasted now.

I can also compare luxury car brands to economy cars. I could also form many other clusters, such as by origin nation, years, or by price buckets.

I’d also like to see if there is a car that can meet my needs from a car. As stated in my introduction, I have a lot of things I’d like out of a car, and I want to know if there is one that can meet at least some of my expectations. I think it could be interesting to see what the cars I may like would’ve cost in 2022 compared to today.

Current Bugs:

As of right now, while trying to scrape larger entries, random amounts of data will not be scraped off the page, despite verifying the path being correct. This can be the name of the car, the price, as well as all of the other data all at once. It is not consistent with each time scraping, so that will need to be an issue figured out when I scrape a larger amount of data. Either that, or entries that are missing data get removed right off the bat.