**# project1-group2**

This repository contains an exploratory data analysis of used car sales in the US between the years 2018 and 2020. As a group, our objective was to find answers to the following three questions:

1. What specifications of a vehicle appear to have the largest impact on price?

2. What is the most commonly seen price point for certain vehicles?

3. How does the value of vehicle change over time? And are there any statistically significant differences between the trends of different makes and models?

**Data Selection and Clean Up**

To find the answers to these questions, we used data from a US used car sales CSV file found on Kaggle (the original CSV as well as a CSV containing the cleaned version of the data set can be found within the Resources folder of this repository entitled "used\_car\_sales.csv" and "cleaned\_used\_car\_data.csv" respectively) This CSV file showed used car sales on eBay between 2018 and 2020.



The original data set (pictured above) had some issues that we cleaned up in the final data set including non-numerical zip codes and unusual price and mileage values. We eliminated zip codes that contained non-numerical characters and ensured that the remaining codes matched a record of US zip codes using a database maintained by the USPS (can be found at - https://postalpro.usps.com/ZIP\_Locale\_Detail or in the Resources folder of this repository). We also eliminated records for vehicles that were sold for less than $100 or had more than 400,000 miles. The resulting data set is pictured below.



**Linear Regression based on Vehicle Specification**

The results of our exploratory data analysis showed that after controlling for the make and model of the vehicle, age and mileage explained a respectable amount of the variation in the prices of different vehicles. One model that showed a particularly strong correlation was the Subaru Legacy. In that case roughly 61% of the price variance could be explained by mileage and 75% by age. Below is a scatter plot showing the correlation for age for Subaru vehicles sold on eBay.

Chart, scatter chart

Description automatically generated

**Common Pricing**

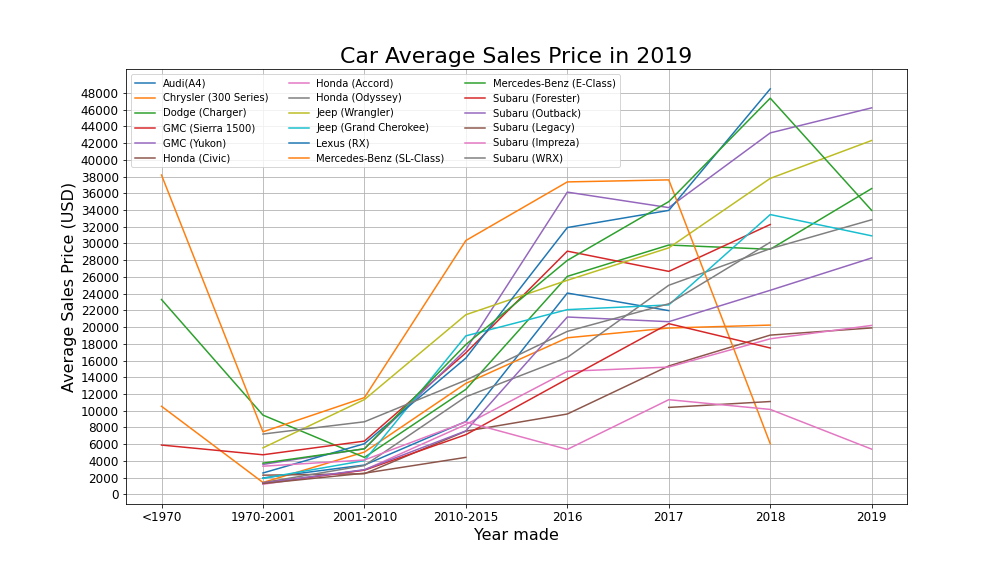
The most common price point for vehicles varied wildly across different makes and models (as shown by the long tail of the histogram depicted below). Some of the most budget friendly options on average included Honda models (Civic, Accord) and Subaru models (Odyssey, Forester, Outback, Impreza) at =< $5000 on average, while Porsche and Mercedes-Benz held some of the most expensive used vehicles (X) on average.

Chart

Description automatically generated

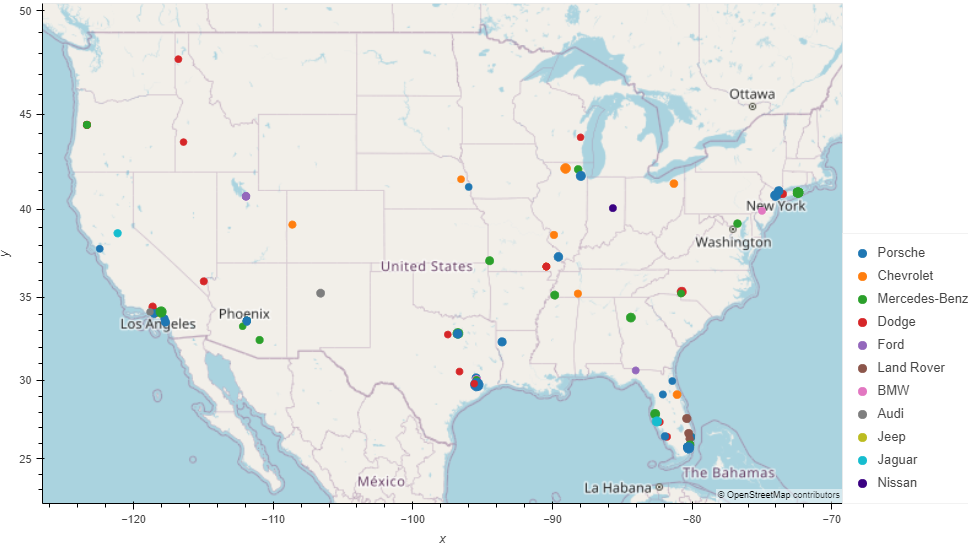
**Vehicle Value over Time**

Based on this exploratory data analysis we were unable to find any consistently explainable differences between the price trends of different makes and models over time. With the exception of some cars that are considered collectors’ items/vintage, most vehicles saw their value dwindle over time at a fairly similar rate as shown in the line graph below.

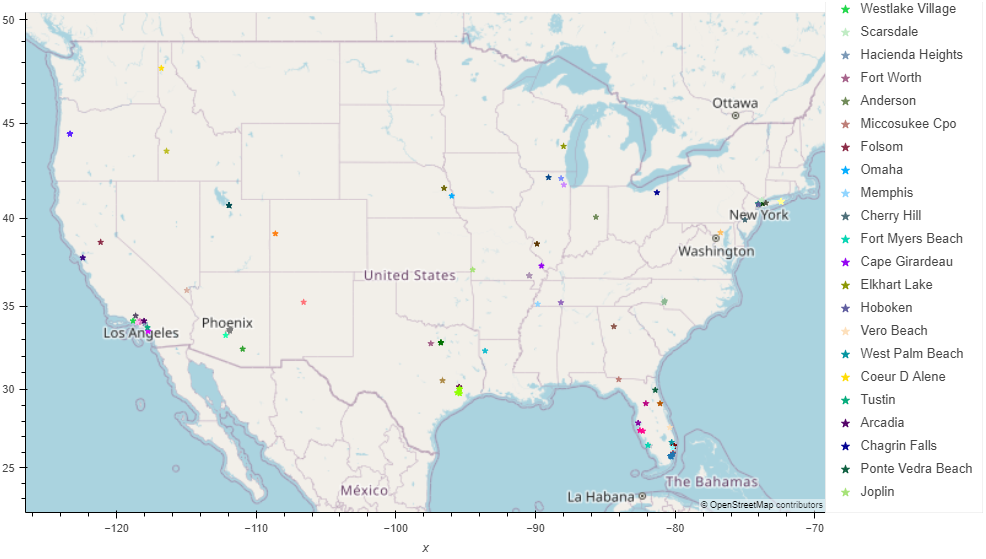


**Luxury Vehicle Map Plotting (API Bonus)**

A part of the data we found particularly interesting was that it was not only low to moderately priced used vehicles for sale on eBay, but luxury and collector vehicles as well. In total there were 86 vehicles that sold between 2018 and 2020 that exceeded $100,000. Using the matplotlib plotting library we were able to show the location in which these sales occurred, sorting by color for make and sizing the plot points based on sale price.



One of the questions that new owners of these high vehicles my start to ask themselves is, “Where is the nearest repair shop to maintain the condition of my new luxury vehicle?”. Using Geoapify’s API category list, we were able to determine the nearest repair shop to all of these sale locations within a 15-mile radius.



**Conclusion**

In conclusion, we were able to answer the three questions we asked going into this analysis:

**Q1:** What specifications of a vehicle appear to have the largest impact on price?

**A1:** While at the highest level there seems to be no explanation of variability due to a car’s mileage or age, as you dive deeper into the make and model of a vehicle type, there is a fairly strong correlation between the two factors and price, with age being slightly higher than mileage

**Q2:** What is the most commonly seen price point for certain vehicles?

**A2:** The median price for all cars sold on eBay between 2018 and 2020 was $6390 and the common price overall was less than $2500. However, knowing that most consumers do have some idea of make or body-type, we took a closer look at certain models and created buckets based on median price. What this showed was that most Honda and Subaru models on average ranged anywhere from $2000-5000 for their median price, whereas models like the Jeep Wrangler, Dodge Charger, and Subaru WRX have a median price of greater than $10,000.

**Q3:** How does the value of vehicle change over time? And are there any statistically significant differences between the trends of different makes and models?

**A3:** We concluded that there was not statistically significant evidence at the make/model level to explain or predict price trends. High level conclusions we were able to gleam from the current dataset were that, in general, cars do see some level of depreciation over time and for collection type vehicles, the value of those cars actually appreciate over time. A deeper analysis with a more advance dataset would needed make conclusions based on specific make/models.