Introduction

With the development of society, the daily activity range of a person becomes much larger than a decade before. With such a change of society, the transportation tool becomes more significant and various in modern society. Compared to other transportation tools, such as the subway, the cars are the majority of the transportation tools for daily life. Therefore, the market related to cars becomes thriving and active. Based on the report of used car market analysis by Grand View Research, “The global used car market size was valued at USD 1.57 trillion in 2021 and is expected to expand at a compound annual growth rate (CAGR) of 6.1% from 2022 to 2030.” (GVR, 1) Hence, the used car market becomes the main market for the trade of cars. More and more people would like to buy a used car when they need a transportation tool because of the lower price and attractive finance options. Additionally, when people want to upgrade their cars, selling the used car is the first choice for them. Our project is about this used car market. and try to figure out the factors that would affect the price of used cars in the market. We found that there are various platforms that would provide the service of selling and buying used cars. Additionally, the prices of the cars are close to each other when they have similar characteristics, such as mileage or years. However, the sellers have to look up the price of cars which have the similar characteristics with their cars and decide the retail price of their own in the same market, which is really trivial and unreliable. Therefore, we would like to collect the data from the real platform of the used car market and analyze the different variables that would affect the retail price of a car. As a result, the users could determine the retail price of their cars based on this analysis.

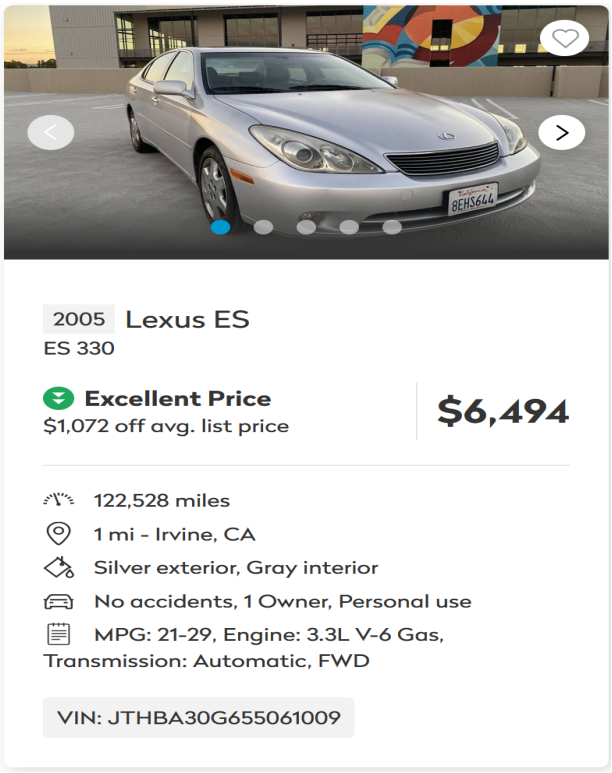
Related Work

Everyone in our group is passionate about automobile performance and the market. There is an ebay project sample provided in class which inspired us to do studies on automobile pricing data. However, many sources on ebay are actually not verified. There is also not enough information

about each car's conditions. To make our data analysis more reliable, we decided to choose a platform with verified inventories from dealers and private sellers.

Data scraping and cleaning

For this project, the data source is truecars.com. truecar.com is an automobile pricing and information platform. The data is about information of all used car inventories in Irvine, CA. Here is a used car's pricing information from truecar.



This figure shows the make, model, mileage, color, VIN, and other information about this car. Different kinds of information can be provided as significant factors in our data analysis. The data range is all used car inventories in Irvine, CA with a radius of 10 miles.

To scrape all data from the website, we used a data science software called Octoparse. This software can help scrape the data automatically with just a URL. To explore all inventories and scrape data from them, we set the step guidelines for the program. The program opens the webpage first. Then it starts scraping data from the page and proceeds to the next page to repeat

scraping. The scraped data was exported as a csv file. We captured more than 3000 rows of data without duplicates.

In order to analyze the data more efficiently and effectively, we have also cleaned and processed the result from the web scraping. We use pandas which is a library of Python to clean the data. We first convert the raw CSV to data frames. Then, since there are massive unclear and combined columns like this:

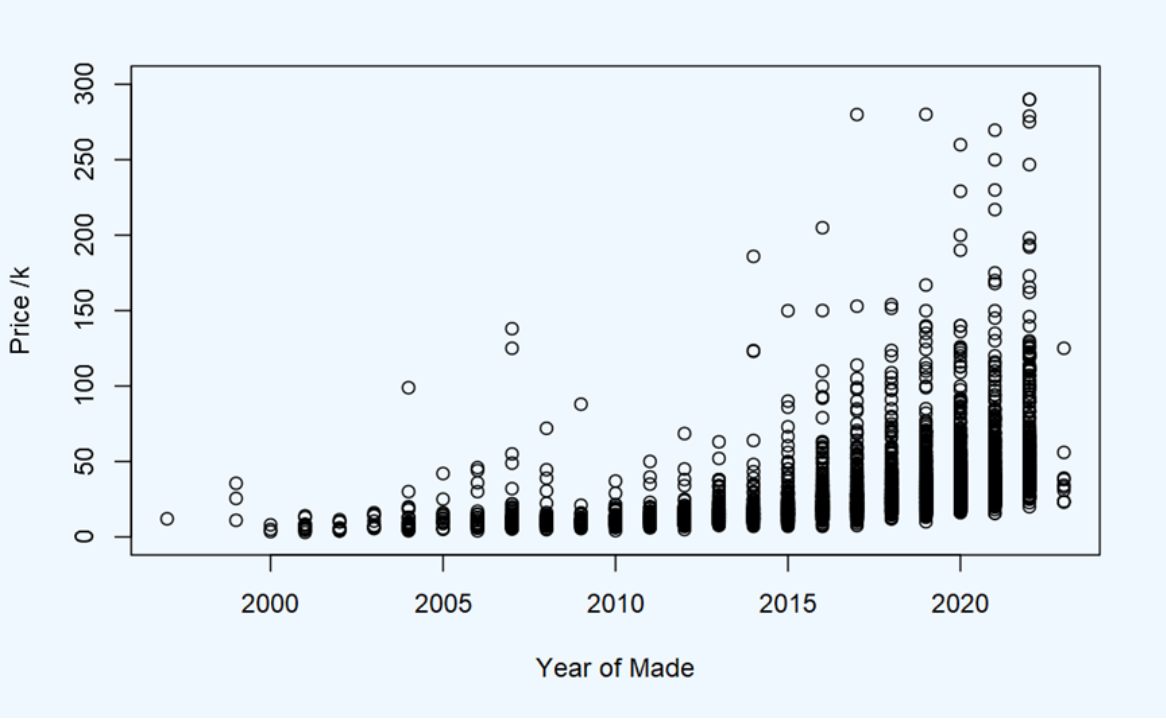
	A	B	C
1	truncate	Year	Price_Level
2	Ford Fusion	2018	Fair Price
3	Hyundai Santa Fe	2020	Excellent Price
4	Lexus RX	2019	Excellent Price
5	Hyundai Elantra	2020	Great Price
6	Honda Accord	2021	Excellent Price
7	Chevrolet Silverado 1500	2019	Excellent Price
8	Mercedes-Benz C-Class	2019	Great Price
9	Hyundai Santa Fe	2020	Excellent Price
10	Mercedes-Benz GLC	2022	Great Price
11	Honda Accord	2019	Great Price
12	BMW i3	2017	Excellent Price
13	Mercedes-Benz GLC	2022	Great Price
14	Hyundai Santa Fe Sport	2015	Excellent Price
15	Chrysler 200	2015	Excellent Price
16	Toyota Mirai	2019	Excellent Price
17	Hyundai Sonata	2020	Great Price
18	Hyundai Elantra	2020	Great Price
19	Chevrolet Equinox	2020	Excellent Price
20	Toyota Prius	2013	Excellent Price
21	Toyota Camry	2012	Excellent Price
22	Mercedes-Benz GLC	2022	Great Price
23	Honda CR-V	2016	Excellent Price
24	BMW 3 Series	2011	Excellent Price
25	Mercedes-Benz GLC	2022	Great Price
26	BMW 3 Series	2019	Excellent Price
27	Mazda Mazda6	2017	Excellent Price
28	Lexus ES	2012	Great Price
29	Tesla Model 3	2020	Great Price
30	Mercedes-Benz E-Class	2019	Great Price

We have separated those columns into several clearer columns by the split method of string. Furthermore, we have filled the null value in order to make it easier and safer to analyze. For example, we have filled the “No Discount” to the null value of the column “Discount”. After all the valuable information was clearly classified and separated, we convert the data frames back to a CSV which can be used for the following analysis.

Analysis/Visualization

For this project, we crawled the truecar website through the web and integrated the data to form a csv file. in the data analysis part, we used different methods for visualization. The data set used in this case study is a crawler tool that crawls the vehicle information of the used car trading

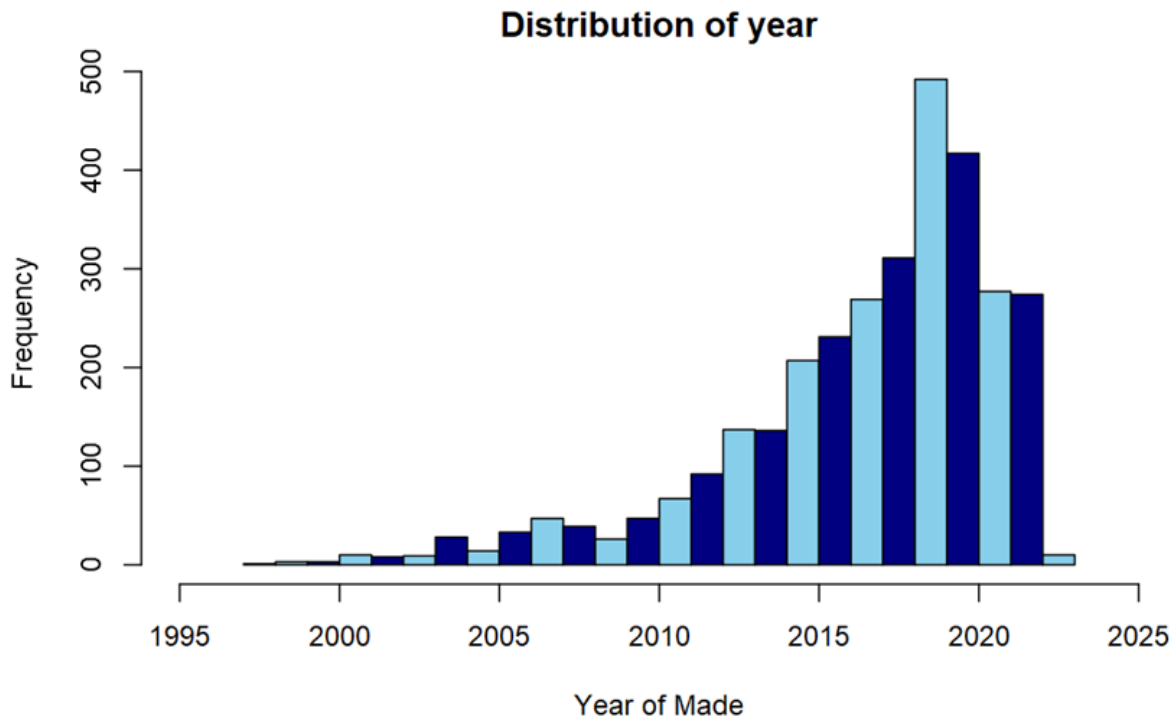
platform (Truecar) from the year 2000 to June 2022. The data is stored in a csv file, with mileage in miles and price in dollars. The original csv file contains pictures of used vehicles, exterior colors, interior pictures, etc. After consolidation and cleaning, the file only contains the information needed for this analysis. Then, we first wanted to find the relation between the different variables of the cars. Hence, the first variable we noticed was the year of make. Though there are huge stocks of the same model and same make, even some of them have the same color, the prices change rapidly with the change of year of make Hence, we use the R studio to plot the relationship between the price and the year of make.



Relationship Between Year and Price (P1)

As the relation shown above, there is an obvious trend between the price and year of make. With the year closer to the current time, 2022, the price increases exponentially. Additionally, we noticed that the amount of price is mainly concentrated on the closer year to the current time.

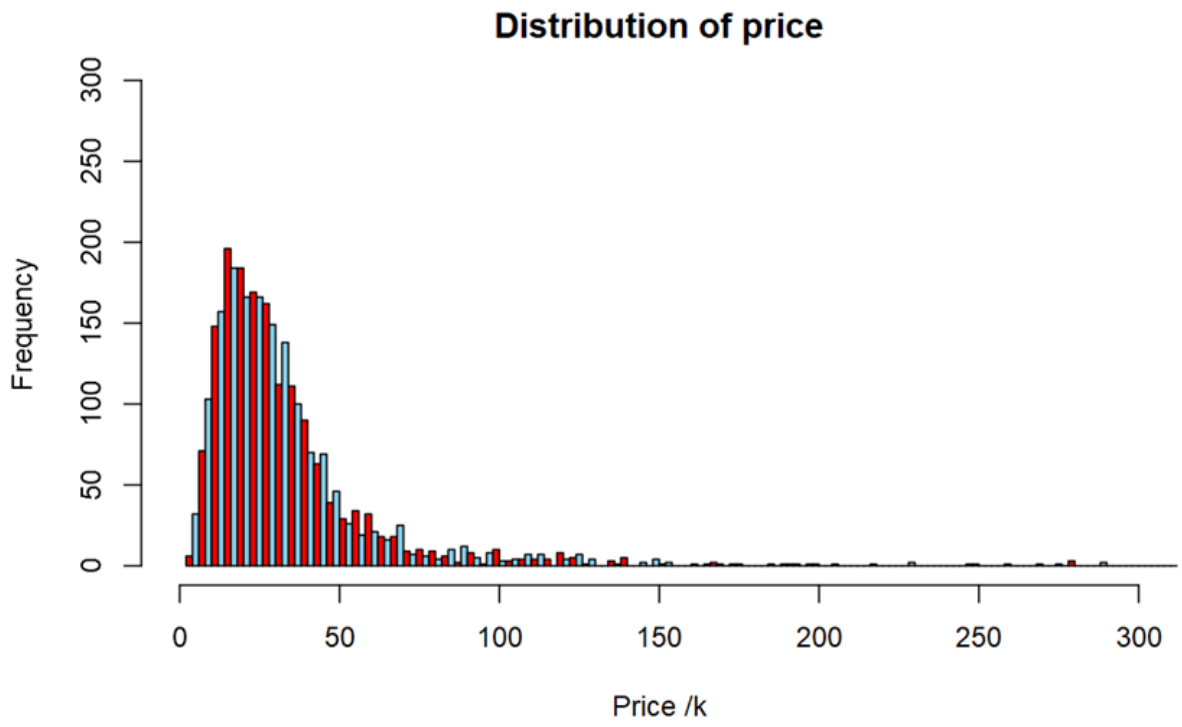
Hence, we assumed that the closer the year of make, the more popular the market. In order to support or against this assumption, we make the analysis about the distribution of year of make based on the data we collect.



Distribution of Year(P2)

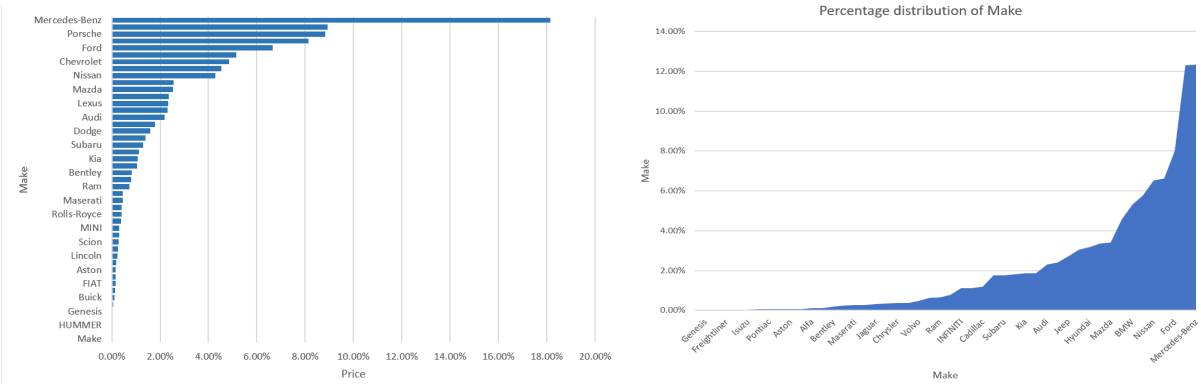
From the frequency (y-axis), this plot supports our assumption that most of the used cars in the market are around the newer year of make, from 2015 to 2021. Hence, the newer cars are more popular in the market. Additionally, we also find that the amount of cars made in 2022 are few. This also indicates that people would not like to sell the cars which have not been used for one year. Then, we wanted to find what prices are most acceptable and popular within the market. Despite other characteristics of one car, the price is the most important factor that would affect people to buy a car because they have to spend the money within the range that they can afford.

Therefore, the sellers who set the reasonable retail price could sell the car efficiently. Hence, we make the following plot to visualize the distribution of the price in the data we collect.



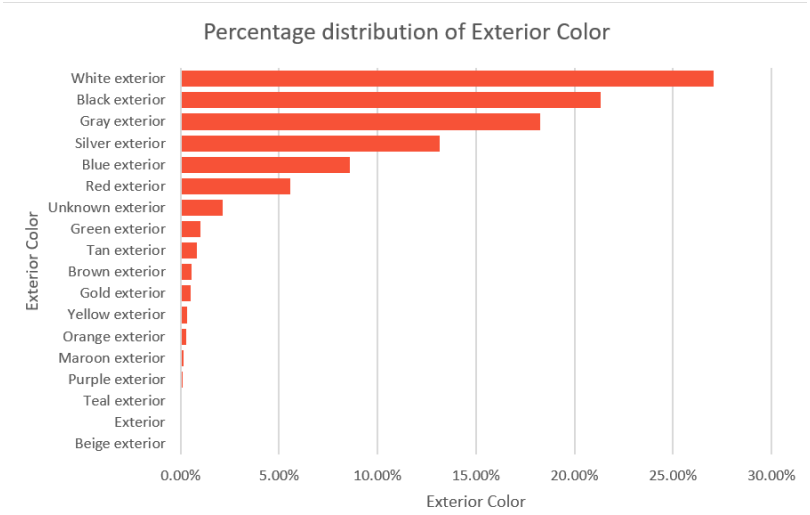
Distribution of Price(P3)

From this plot, we find that most of the prices are set between the range from $0 to $50,000. This should be the range that people could accept as the buyer. Then, We used data visualization (R) to visually present the percentage of different makes and models in the overall data.



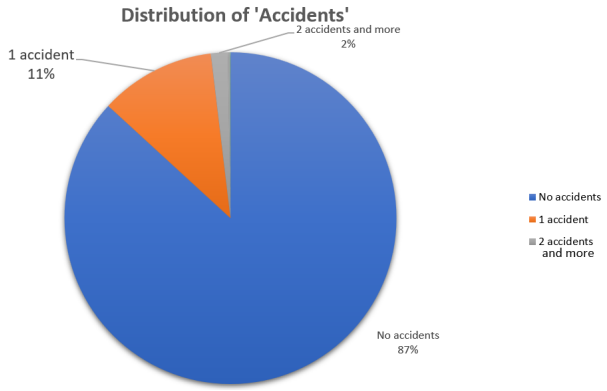
Distribution of Make(P4)

In addition, the price and quantity of used vehicles were visualized as a percentage of the data set, comparing the two visualized charts. We find that Mercedes Benz has the largest share of used vehicle quantity and price, and it is worth noting that the price of used Porsche vehicles occupies the second position, indicating the high value retention rate of Porsche.



Distribution of Color(P5)

To explore the research for this analysis report, R was used to create a comparison about the value of used cars and mileage, accidents, and exterior paint colors. We can clearly see from the chart below that used cars with a white exterior tend to sell for a higher price than used cars with a black exterior, and in third place is the gray exterior. In addition, a key factor affecting the price of a used car is the condition of the used car and whether the vehicle has experienced a major traffic accident. We visualized and analyzed the used car accident data collected, and we can clearly see from the graph that 87% of the vehicles sold on the truecar platform have not experienced structural damage or engine damage.



Distribution of Accidents Cars(P6)

Conclusion:

After our group analyzed the truecars model, we found that significant attributes are Year, Price_level, Price, Mileage, Discount, Model, Certification, brand, Number of Accident, Exterior Color, Interior Color, Number of Owners.

Among these attributes, the number of accidents and the year they were made are two most important factors that affect the price. The newer the cars are, the higher the prices are. The less accidents the cars have had, the higher the prices are. In addition, people prefer cars that have black or white exterior color.

Therefore, we would recommend truecars to add more important attributes such as recall history, etc and create categorical variables to distinguish cars, such as engine displacement, transmission, etc. in order to increase the sales.

Work Cited

Grand View Research. “Used Car Market Size & Share Report, 2022-2030.” *Grand View Research*, https://www.grandviewresearch.com/industry-analysis/used-car-market. Accessed 8 December 2022.

Used cars for sale in Irvine, CA. TrueCar. (n.d.). https://www.truecar.com/used-cars-for-sale/listings/location-irvine-ca/?excludeExpandedDelivery=true&searchRadius=10 Accessed December 8, 2022.