George Washington University
School of Engineering and Applied Science
EMSE6574

Explore and Analyze Used Cars Market in the U.S.

Rui Zhang- Data Analytics

Ran Wei- Data Analytics

Bite Xiong- Data Analytics

Fangzhou Liu- Engineering Management

Professor Maksim Tsvetovat

Submission Date: 12/06/2021

1. Introduction and Problem

Over the past year, as the pandemic disrupted supply chains and caused shortages in critical auto components, resulting in a lack of new vehicles, which caused the used car price surging.

According to the U.S. Bureau of Labor Statistics' Consumer Price Index, used car prices are up a staggering 39.8% since March of 2020, while the U.S. inflation is only up 6.3%. JPMorgan analysts say prices will continue to rise and will stay high for longer than expected. It is difficult for people to use previous data and experience to judge whether the price of a used car is a good deal. So, our team tried to create a method to determine whether the price of a used car is appropriate.

2. The Dataset

The dataset we used in this project was downloaded from Kaggle. The provider of the dataset scraped data from Craigslist every few months, it contains all relevant information that Craigslist provides on car sales. We selected data from the beginning of 2020 and the same period in 2021.

There are 426880 rows and 26 columns in the dataset of 2021 and 539759 rows and 25 columns in the dataset of 2020.We first drop some unnecessary features, like 'id', 'url', 'VIN', 'image_url', 'description','county','posting_date','lat','long', etc.

Then we found there were some outliers in numerical variables. In the dataset of 2021, the target variable price ranges between 0 to $3.7 billion. We transformed the price value into log data and removed the outliers of price by using 3-sigma rules. We finally got the price variable ranging from $320 to $105,000. The year variable ranged from 1900 to 2022, we only kept values from 1991 to 2021. For the odometer variable, the max value was about 10 million miles. According to the Federal Highway Administration, Americans drive an average of 13,476 miles per year. So, we dropped odometer values higher than 40,000.

For missing values, there were some variables missing lots of values, we didn't want to drop them in order to keep important features and enough data for our model. So, we kept all the values and dealt with them after exploratory data analysis.

We did the same cleanup (removing price outliers) for the 2020 data on Kaggle, and our next step will be to compare and analyze the data for the same period in 2020 as well as 2021.

3. Exploratory Data Analysis

In this section, we will test various characteristics of used cars and try to have a better understanding of the data. We will look at different combinations of features with the help of figures while exploring the data. This will help us to have a better understanding of data and provide us with some clues about the data patterns.

3.1 Target Variable

Price: The price of the used car, given in US dollars.
F: price distribution
Price is the feature we are predicting in this study. Before applying any model, looking at price data can give us a better understanding and judgment.

3.2 Other features of Used Car:

odometer: The distance that the car has been driven after it is bought.

year: The year in which the car was manufactured.

manufacturer: Manufacture of the car.

model: The exact model of the car.

condition: The condition of the used car, including *excellent, good, fair, like new, salvage, new*.

cylinders: The number of cylinders in the car engine.

Fuel: The fuel type of the car, including *diesel, gas, electric, hybrid, and others*.

title_status: Including *clean, lien, rebuilt, salvage, parts only, and missing.*

transmission: The transmission of the car, including *automatic, manual, and other.*

drive: Including 3 types of drive transmissions: *4WD, FWD, and RWD.*

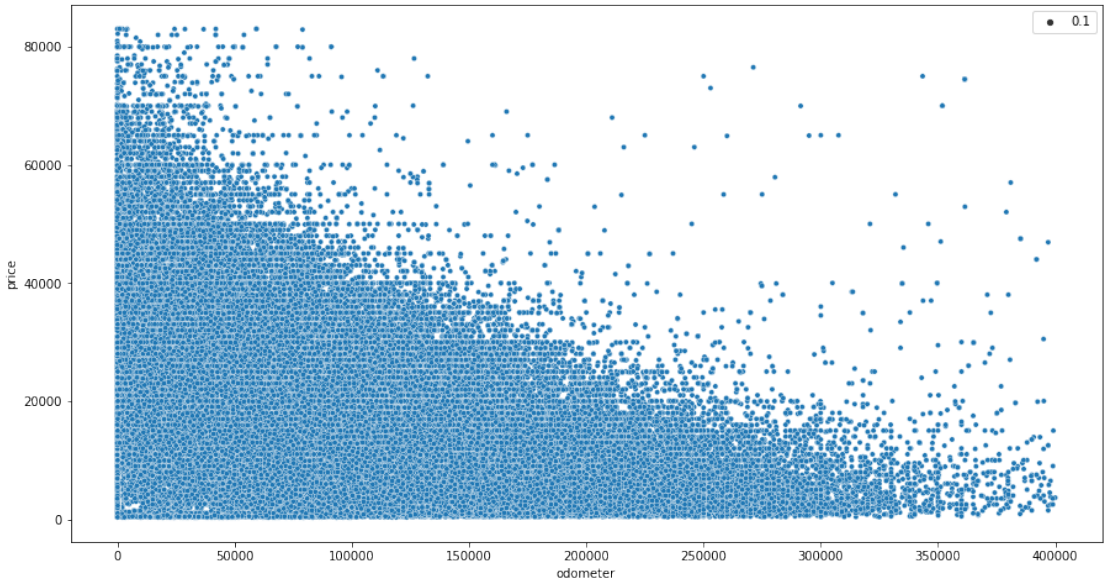
size: The size of the car, including compact, full-size, mid-size, sub-compact

type: The generic type of the car.

State: The state that the car belongs to.

Paint_color: The color of the car.

Odometer： When people buy a used car, they pay a lot of attention to the odometer value on the car. We can see that the odometer significantly changes the price of the car. On the other hand, this does not mean that only cars with low odometers are sold. Depending on the price, there are buyers for high odometer cars as well. In addition, the most popular odometer numbers for used cars under $40,000 range between 50,000 and 150,000 miles

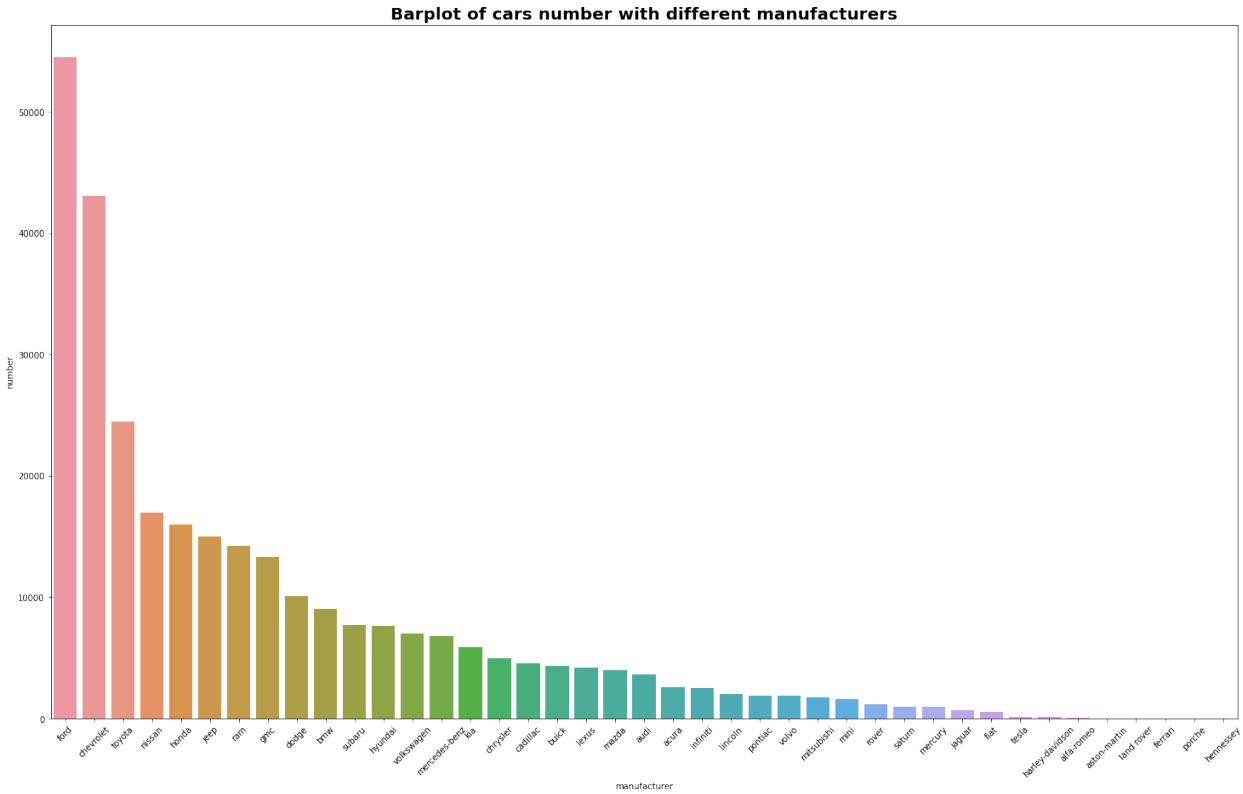


Year

There are 51,240 cars for the last 5 years, which is 16.78% of the entire market.

There are 160,951 cars for the last 10 years, which is 52.72% of the entire market.

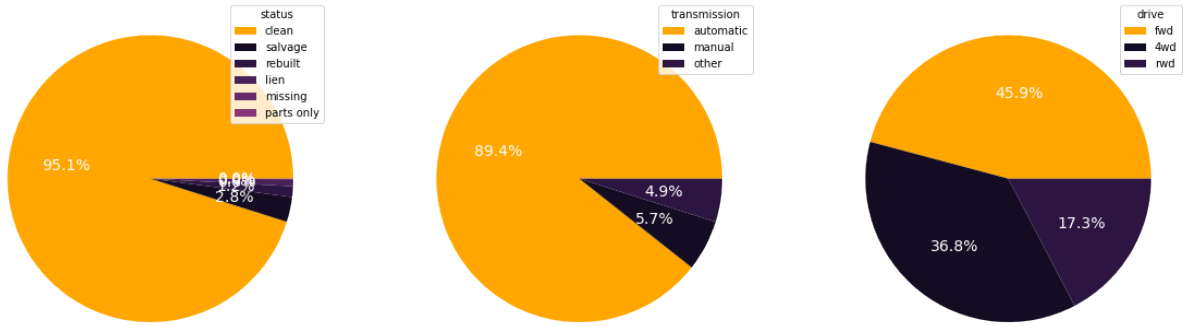
Manufacture & Models： Automobile manufacturers are another important variable in the used car market. Ford and Chevrolet are among the major manufacturers in North America. Toyota, Nissan, and Honda follow the order as big manufacturers. It can be concluded that Japanese cars have a significant share in the used car market.



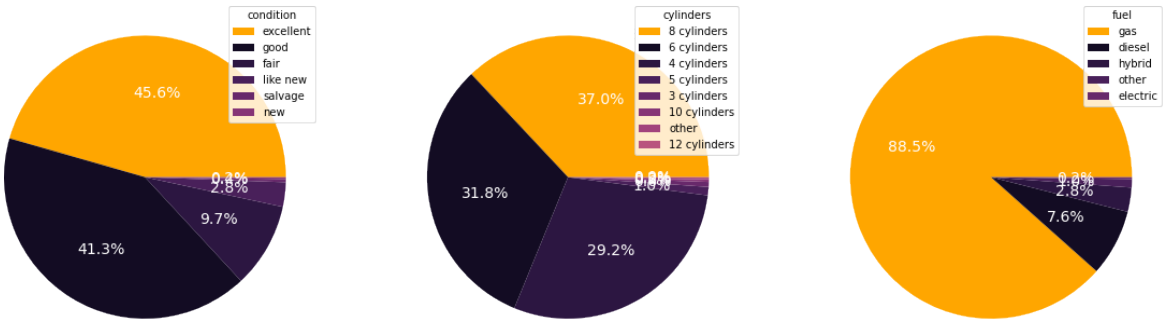
Type & Transmissions & Fuel： Transmission is another feature that dominates the used car market. Automatic transmissions have a strong influence on people's preference for cars. It can be seen that automatic transmission cars dominate across the board. covid-19 has an impact on the used car market and influences the market. Another increase in transmission types maybe the increase in Continuously Variable Transmissions (CVT). CVT is more environmentally friendly and fuel-efficient. There may be a rollout of this technology. Another possibility is that some sellers on the Craigslist website do not fill out the transmission section of the car's information.

When evaluating a car, it is important to understand the factors that affect its condition. The fwd-driven cars are more durable and reliable. As you can see, fwd cars are the most popular in terms of numbers. In the long term, they maintain better runnability compared to rwd and fwd

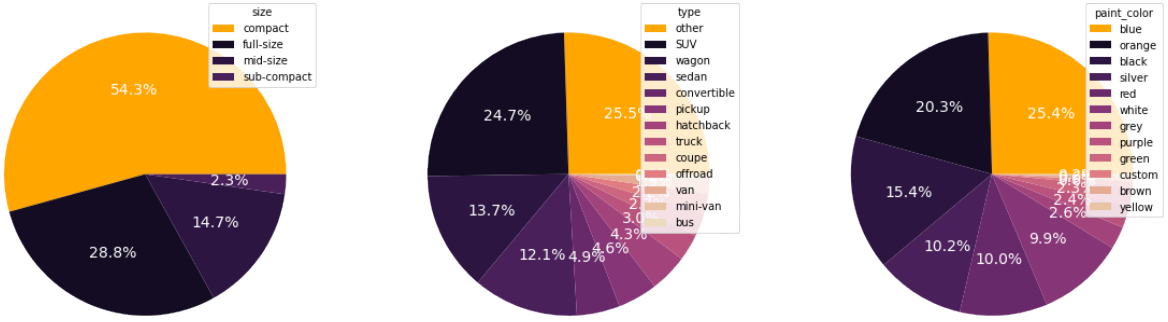
drivetrains.



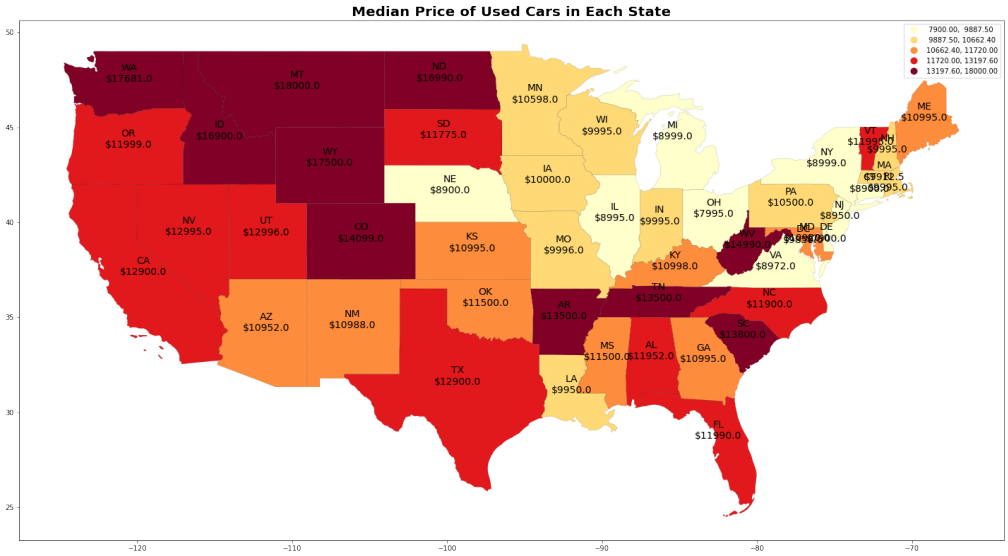
Condition： We can find that most of the used cars are below $20,000. In addition, we see that there are still a significant number of cars that cost more than $20,000. We can guess that all types of cars can be cheap or expensive. However, excellent, like-new and good condition cars are still the most popular cars on the used car market. Salvage cars follow the popularity of these three types of cars. Therefore, it is difficult to make a strong estimate of a car's price by considering only the type or condition of the car. However, cars in certain conditions are popular and have a better chance of being sold.

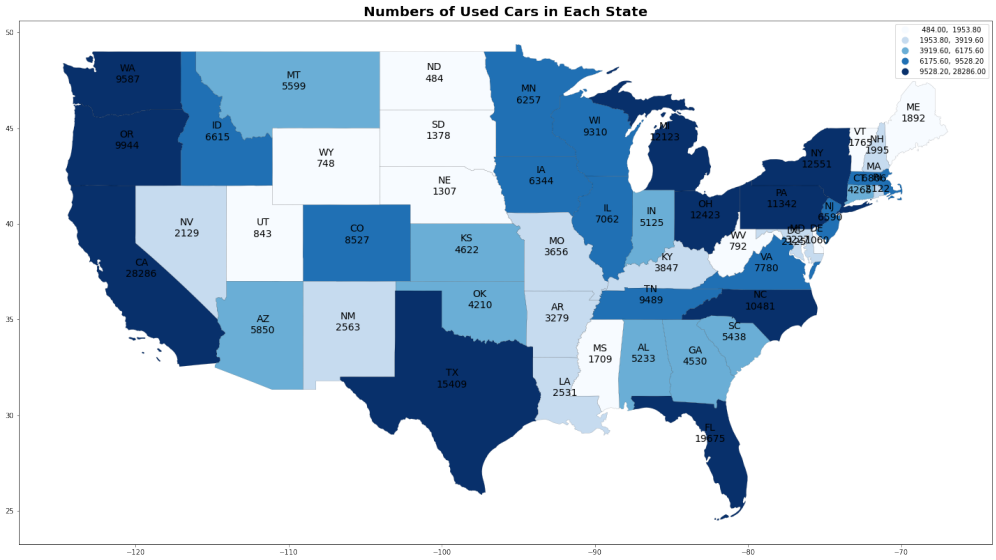


Drive & Title Status & Size & Type & Color： Nearly fifty percent of people choose compact used cars because of their more practical size, and everyone has their own preferences when it comes to choosing a car type and color, so this distribution is all relatively even.



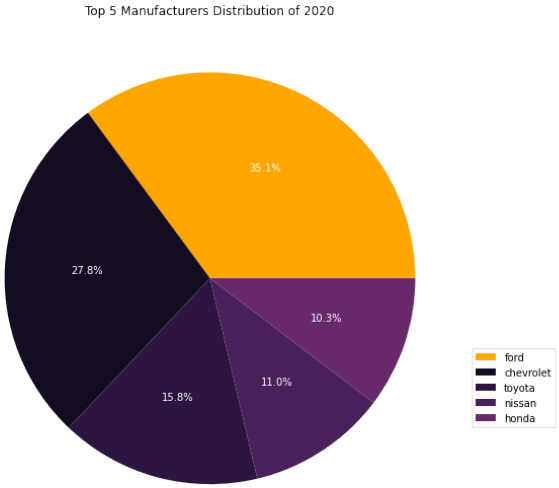
State： The median used car price by state shows that used car prices are generally higher in the West and in some cities in the Southeast than in the Northeast, with the number of used cars concentrated in large cities in the East and West.

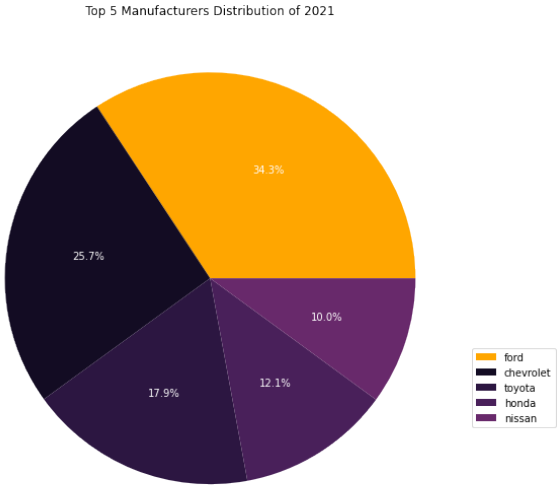




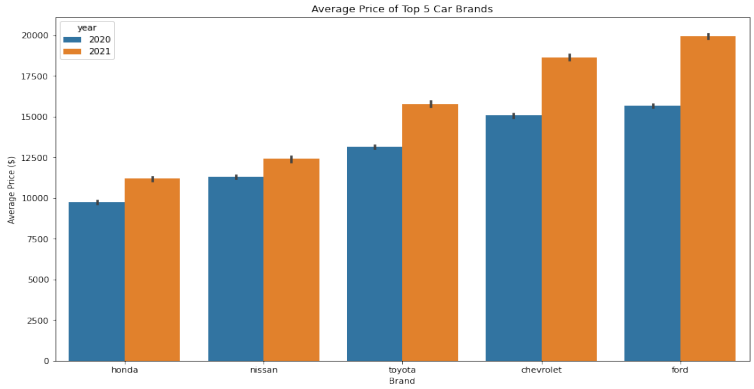
3.3 Comparison Between 2021 and 2020

After we cleaned the car data for 2020 and 2021, we compared the results we obtained. First, we extracted frequencies for the top 10 manufacturers, then visualized their manufacturer distribution by year (pie chart), and finally compared the top 5 manufacturer distributions for 2020 and 2021. The Top 3 car manufacturers remain the same, whereas Honda beat Nissan in the year 2021. Although Ford, Chevrolet, and Toyota remain to be the Top 3, there is a dip in the market share of Ford and Chevrolet while there is an increase in Toyota's market share.

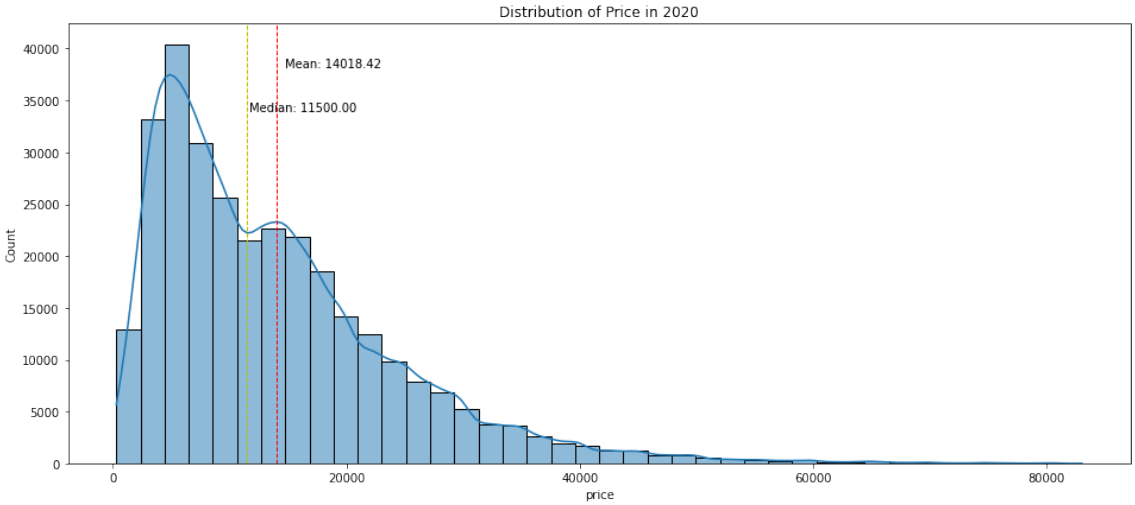


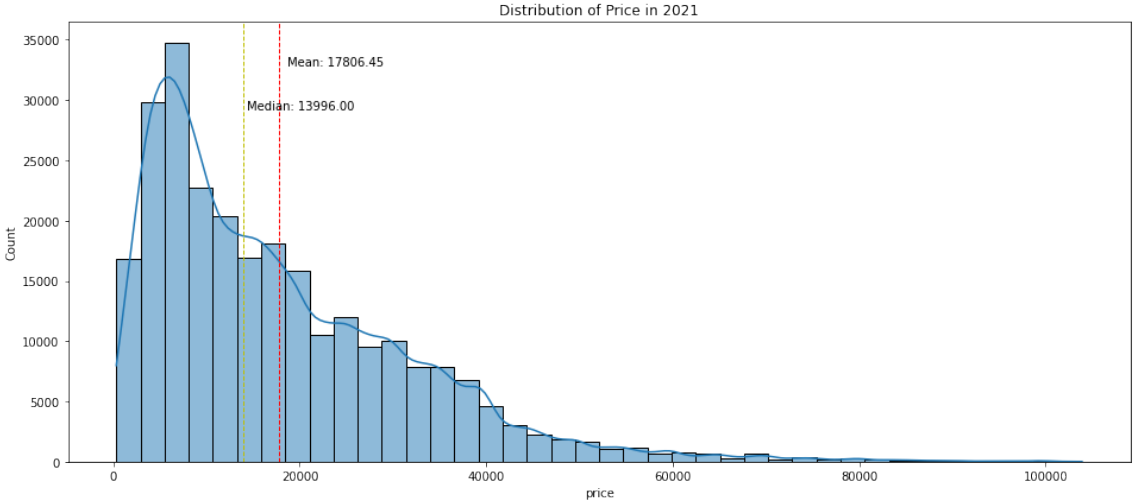


Then we extract the data of the top 5 car brands (Honda, Nissan, Toyota, Chevrolet, Ford) and visualize the average price of manufacturers for two years, and finally show the average price of the top 5 car brands. It can be seen that the average price of the car for the Top 5 brands has increased. We can see a drastic rise in average car prices of Ford and Chevrolet, which could be the reason behind the dip in market share in 2021 for Ford and Chevrolet.

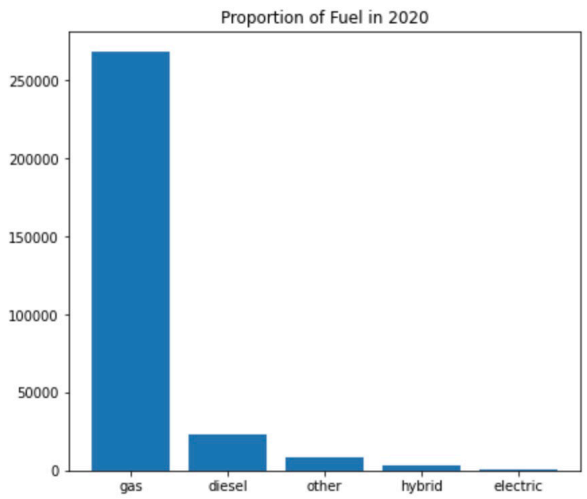


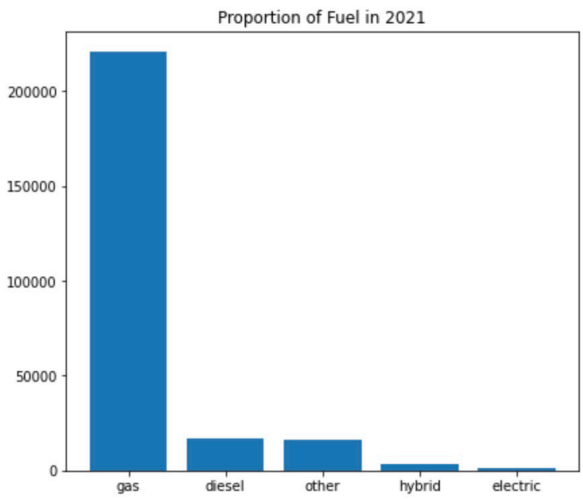
Then we synthesize the data and visualize the price distribution, mean and median, and compare 2020 and 2021. The average price of a car increased from $14018 to $17806 from 2020 to 2021. The median price of a car has increased from $11500 to $13996 from 2020 to 2021. Quite evidently, the price has increased over the 1 year indicating that the sale of expensive and luxurious cars has increased, or the market rate of cars has themselves gone up due to market inflation. The numbers also indicate that sales worth more than $25000 have increased, thus increasing the average and median car price.



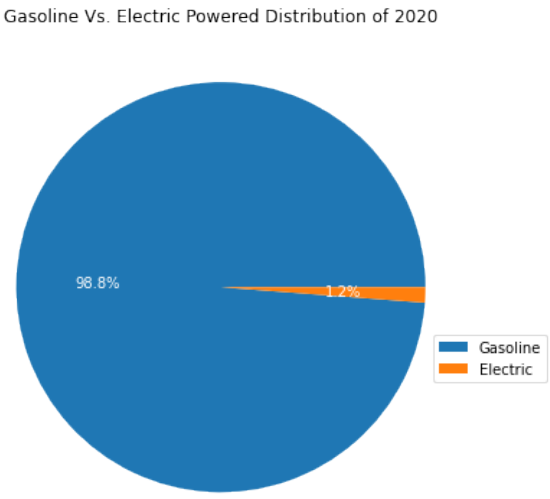


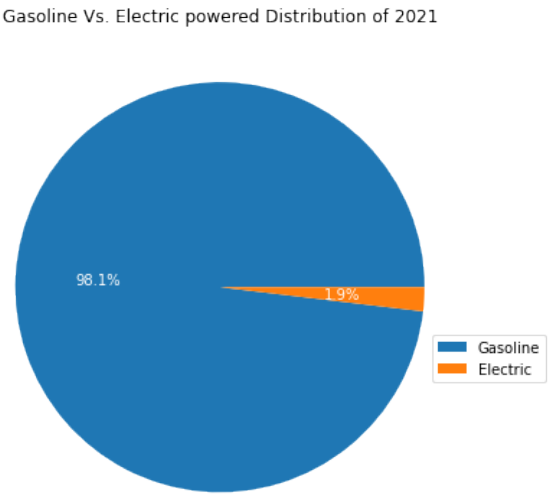
We want to understand how much the type of auto fuel contributes to consumer purchase desire, so we extract the frequency of auto fuel types and visualize the distribution of fuel types using a bar chart.





By extracting data specific to gasoline and electric vehicles and visualizing the distribution of 2020 and 2021 manufacturers, we got Gasoline Vs. Electric fuel proportion (2020): 81.7; Gasoline Vs. Electric fuel proportion (2021): 52. For Gasoline Powered cars we considered both gas and diesel in that category, and for electric cars, we considered both electric and hybrid cars. The Gasoline vs Electric proportion has decreased from 2020 to 2021 by almost 2 times. This is a strong indication that the Sales of Gasoline-powered cars have declined whereas Electric vehicles have taken a steep rise in terms of sales.





4. Data Preprocessing

4.1 Missing Value

There are different degrees of missing variables for each category. The simplest way to deal with missing values is to drop all missing values or remove columns with too many missing values. But these variables may have an impact on used car prices, so in order to ensure that important variables are not lost and keep sufficient data, we filled in missing values of all categorical variables as 'UNKNOWN'.

4.2 Encoding Categorical Feature

There are 11 categorical variables. We need to convert these categorical variables to numerical variables. Most current projects on used car prices usually use LabelEncoder to solve this problem. But LabelEncoder is suitable for the case where the variable is ordinal. For the categorical variables in our dataset, most features are not ordinal. So, we use One-Hot Encoding to solve the problem.

However, we found that there are too many values in the ‘model’ variable. It will generate too many columns which will slow down the running speed. Our solution is grouping model names so that we will have less values and less dummy variables. We use the FuzzyWuzzy package to calculate the similarity ratio between two model names. And group model names based on the similarity ratio.

Finally, we got 70 groups of model names and 179 dummy variables in total.

4.3 Normalization

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. We should use normalization when the features are in different ranges, just like our numerical features, especially ‘odometer’, its value as high as 400,000 miles. Without normalization, the model result will be affected. So, we use MinMaxScaler to normalize the numerical features.

4.4 Train Test Split

In this process, the test size is 0.33, so there are about 170,000 in train data and 85,000 in test data.

5. Model

In this section, different machine learning algorithms are used to predict price/target variables. We applied machine learning models as a framework for data analysis. The dataset is supervised data, which refers to fitting a model of the dependent variable to the independent variable, with the goal of accurately predicting the dependent variable for future observation or understanding the relationship between the variables. The following is the five models we used and their RMSE and R- square results:

1. Linear Regression

   In linear regression, relationships are modeled using linear predictive functions, and the unknown model parameters are estimated from the data. This model is called a linear model.

   ```
   RMSE on train data: 8068.372105103096
   RMSE on test data: 8161.262392068513
   R-square on train data: 0.6738
   R-square on test data: 0.6672
   ```

2. Linear Regression after Log Price

   ```
   RMSE on train data: 8272.253598578518
   RMSE on test data: 8338.427380921941
   R-square on train data: 0.6134
   R-square on test data: 0.6082
   ```

3. Decision Tree Regressor

   ```
   RMSE on train data: 364.9904027260283
   RMSE on test data: 7607.603157641777
   R-square on train data: 0.9993
   R-square on test data: 0.7108
   ```

4. Random Forest Tree Regressor

   A random forest is a classification algorithm consisting of many decision trees. It uses bagging and features randomness in the construction of each tree to try to create a forest of unrelated trees.

   ```
   RMSE on train data: 2145.214320349884
   RMSE on test data: 5762.668948051185
   R-square on train data: 0.9769
   R-square on test data: 0.8341
   ```

5. XGBoost

   XGBoost is an implementation of gradient boosting decision trees designed for speed and performance. This powerful algorithm can be found in its scalability, which drives fast learning through parallel and distributed computing and provides efficient memory usage.

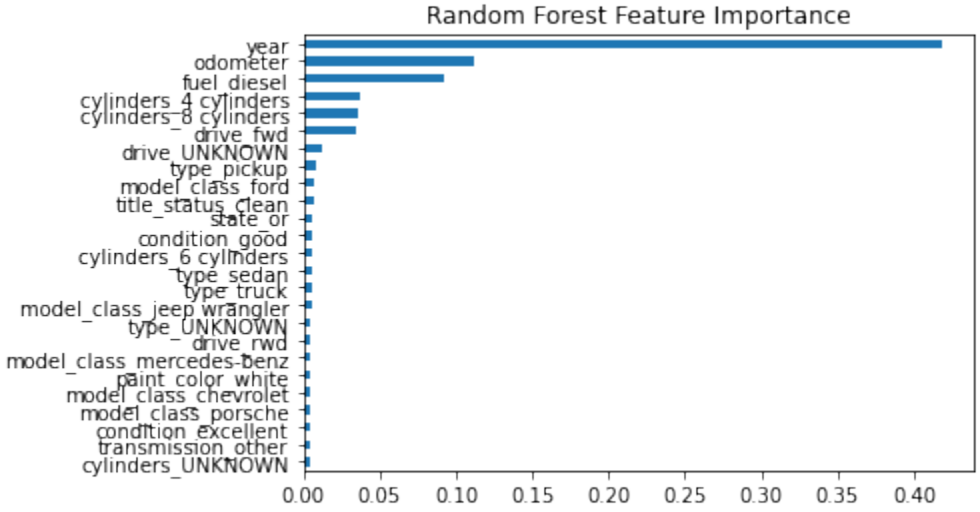
   ```
   RMSE on train data: 7448.273104117994
   RMSE on test data: 7580.64177012461
   ```

```
R-square on train data: 0.722
R-square on test data: 0.7128
```

By comparing the accuracy of each model, we choose to use the Random Forest Regression model to make a price prediction.

	RMSE_train	RMSE_test	R2_train	R2_test
Linear Regression	8272.25	8338.43	0.6134	0.6082
Decision Tree	364.99	7533.07	0.9993	0.7164
Random Forest Regression	2138.58	5756.26	0.9771	0.8344
XGBoost Regression	7448.27	7580.64	0.722	0.7128

The top 25 important features:



6. Final Goal

By comparing the accuracy, we finally choose the Random Forest model for price prediction. Our approach is to determine whether the price of a used car is a good deal or a bad deal by comparing the listing price and the predicted price of this used car. We create functions to check whether the deal is good (the predicted price should be higher than the actual price).

If the listing price is less than the predicted price, the deal is considered a good deal; if the listing price is greater than the predicted price, the deal is considered a bad deal.

Finally, we use a function to select the cheapest car based on the "cars.com" dataset.

7. Conclusion

To sum up, this study uses different models to predict used car prices. We also learned a lot of practical modeling in this class to help us solve future problems. However, more reliable predictions could be produced if more data could be collected. Secondly, there may be more characteristics that could serve as good predictors. For example, here are some features that could also be taken into prediction: the number of doors, gasoline/mile (per gallon), mechanical and exterior refurbishment time/quality, and appraisal trade ratio.

Reference

1.Vinodshiv. "Used Car Price Prediction - 20 Years Data." Kaggle, Kaggle, 16 Aug. 2021, https://www.kaggle.com/vinodshiv/used-car-price-prediction-20-years-data.

2. Anerisavani. "Eda and Price Prediction of Used Vehicles." Kaggle, Kaggle, 3 Nov. 2020, https://www.kaggle.com/anerisavani/eda-and-price-prediction-of-used-vehicles.

3. Suddhu. "Cleaning Data + Eda." Kaggle, Kaggle, 31 Dec. 2018, https://www.kaggle.com/suddhu/cleaning-data-eda.

4. Ismailsefa. "Used Cars Data Analysis and Visualization (EDA)." Kaggle, Kaggle, 27 June 2021, https://www.kaggle.com/ismailsefa/used-cars-data-analysis-and-visualization-eda.

5. Shi, Hua. "Data Visualization: How to Plot a Map with Geopandas in Python?" Medium, Medium, 28 Oct. 2020, https://melaniesoek0120.medium.com/data-visualization-how-to-plot-a-map-with-geopandas-in-python-73b10dcd4b4b.

6. ashmani999. "Preowned Car Price Prediction." Kaggle, Kaggle, 25 Feb. 2021, https://www.kaggle.com/ashmani999/preowned-car-price-prediction/notebook.

7. Anil, Panwar Abhash. "Used Car Price Prediction Using Machine Learning." Medium, Towards Data Science, 26 Apr. 2021, https://towardsdatascience.com/used-car-price-prediction-using-machine-learning-e3be02d977b

8. Gokce, Enes. "Predicting Used Car Prices with Machine Learning Techniques." Medium, Towards Data Science, 10 Jan. 2020, https://towardsdatascience.com/predicting-used-car-prices-with-machine-learning-techniques-8a9d8313952.

9. "Fuzzy String Matching Python: Levenshtein Distance, String Approximate, & Matching Examples." DataCamp Community, https://www.datacamp.com/community/tutorials/fuzzy-string-python.

10. Anerisavani. "Eda and Price Prediction of Used Vehicles." Kaggle, Kaggle, 3 Nov. 2020, https://www.kaggle.com/anerisavani/eda-and-price-prediction-of-used-vehicles.