Group 66
Chloe Liu and Zeyu Deng

An Exploratory Study of CA Used Car Market

1. Introduction

Buying an automobile is always a time-consuming process because you need to do great amounts of research, compare a variety of vehicle choices, and go through all kinds of preparation work to get a good deal. However, buying a used vehicle is even trickier. Every used car has its unique history that largely decides its conditions, functions, and thus value, which you must pay extra attention to avoid getting ripened off. The largest concern risen here is: Is this offer a really good deal?

2. Objective

How can you tell if a price is really a good deal? When you do a used car research, the websites would allow you to filter the results based on keywords, or car features such as brand, engine, location, year, mileage, body style, etc. For this project, we want to answer the questions and investigate which of these features are more likely to decide the car's premium power, and if we filter the searching results based on specific preferences on these used car features, would we be able to obtain an overall accurate prediction of the car's premium power (that allows us to compare and determine if the dealer's offer is really a good deal)? (Additionally, based on results, we hope to provide some recommendations for ...)

3. Data Mining

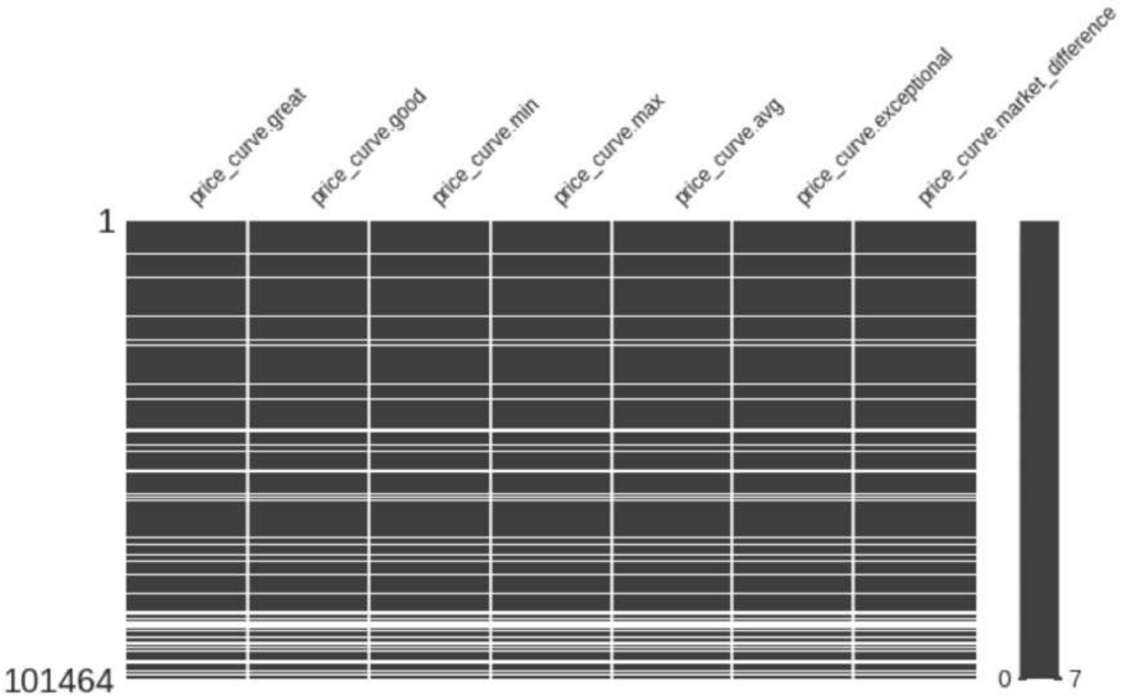
To answer the questions above, we used data mining and picked one of the major online car buying websites, Truecar.com, and grab data by making API requests utilizing the requests package. We found public but hidden APIs of Truecar.com through the Web developer console, and by systematically generating URLs and making requests, we obtained JSON responses containing details for all listings within certain miles from a certain postcode, i.e. we chose 5000 miles from 95616. After that, we collected a nested collection. We transformed the collection into data, and differed the columns into five main dimensions: 'vehicle', 'dealership', 'pricing_flag', 'pricing', 'price_curve'. We parsed the data and extracted relevant listings only in California. This dataset was saved as 'only_ca.csv' as our original reference.

The only_ca dataset contains 109774 listings with 81 columns, and during the data cleaning, we basically made the following adjustments:

- Filtered data only in the recent 10 years (2009 until now)
- Converted some data types, e.g. changed the vehicle.year to integers
- Removed duplicate or irrelevant variables
- Dealt with missing values

For variables with missingness, we imputed missing values for some of the variables based on their properties. For example, we replaced missing values for the continuous variables of the vehicle's mpg with their means and set some missing values to constants so that they do not hinder other computations. In addition, we discovered some patterned missingness in the dataset. There are seven variables of price curves having 13308 NAs in common, and by visualizing them in a matrix, we discovered that their corresponding listings are in the same group. Another group of 361 listings also showed noticeable NA patterns in several variables, revealing the fact that they are from the same dealer.

Figure 3.1 - Missingness Matrix of Price_curve variables: The matrix below clearly demonstrates that there is a group of vehicles lacking price curve information simultaneously.

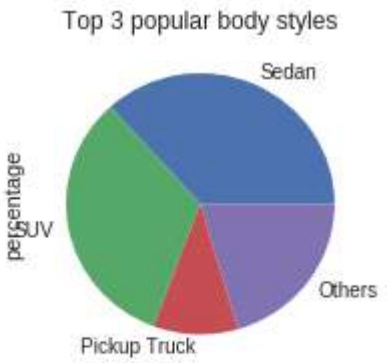


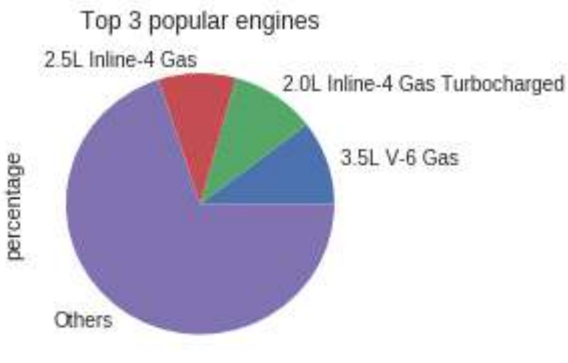
After dealing with missing values, we saved a copy of the cleaned data in cleandata.csv, and the cleaned data has 101464 listings with 47 columns containing the vehicle, pricing, and dealership information for each car, and we also explored the data with summary statistics, frequency counts, and visualization to discover relationships between used car features.

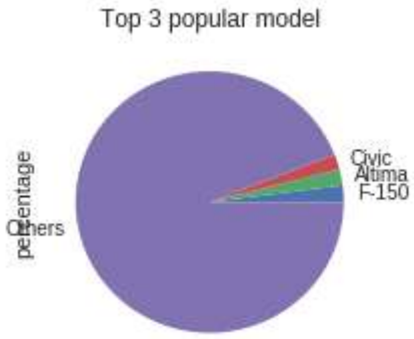
4. Data Exploration

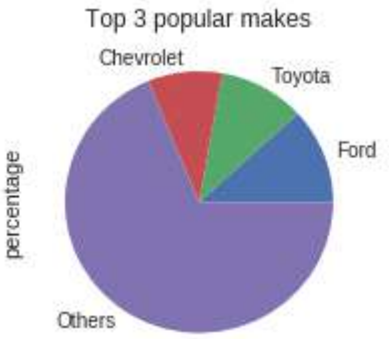
For data exploration, we first examined the relationships between the number of listings and several categorical car features. Utilizing groupby() function, we sorted out three keys with the most listings for each feature, and these keys can provide a sense of some hot trends in the CA used car market.

Figure 4.1 ~ 4.4 - Pie charts: Based on the number of listings, vehicles with specific features seem to be more popular in the used car market in California. Revealed by the pie charts below, it seems that Sedan and SUV are dominant body styles that each has one-third of the market share









Also, we visualized another variable of interest, the average market price, along with some important features like drive train, mileage, and body style

Figure 4.5 - Average market price vs. drive train: The scaled price distribution below indicates that the FWD vehicles generally have lower market prices than those with RWD, and RWD vehicles take larger proportions of premium cars. The gap in the plot represents the rapid drop in density of listings when the price goes over $100,000.

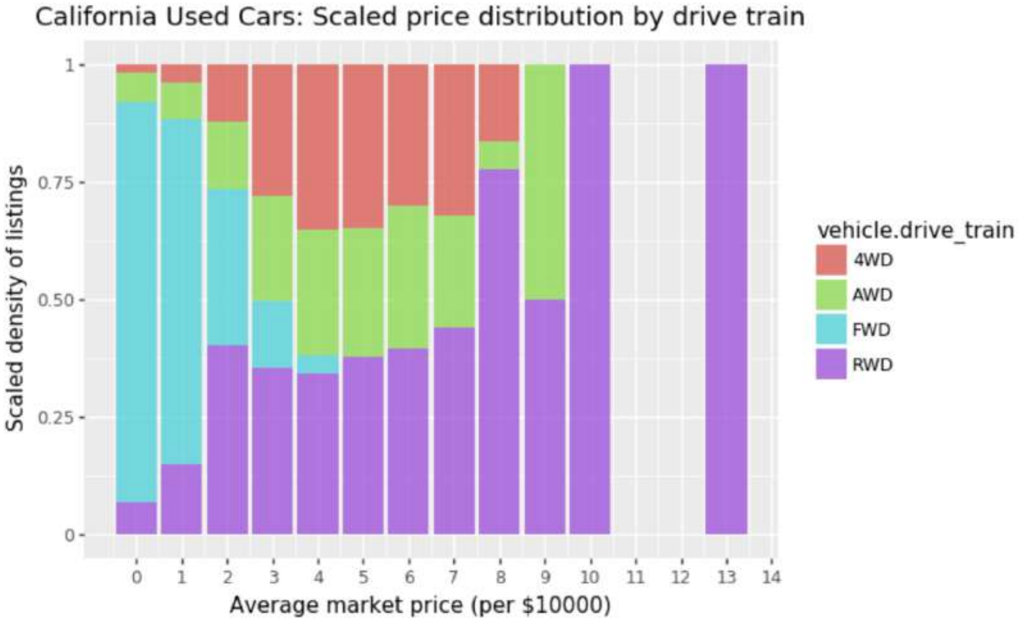


Figure 4.6 - Average market price vs. mileage: For the plot below, we changed the variable mileage into intervals of 30000 mileages and plotted it against the average market price as

shown below. The plot demonstrates a rapid drop in vehicle price as mileage grows larger, and this relationship implies that mileage might be a good predictor of used car prices.

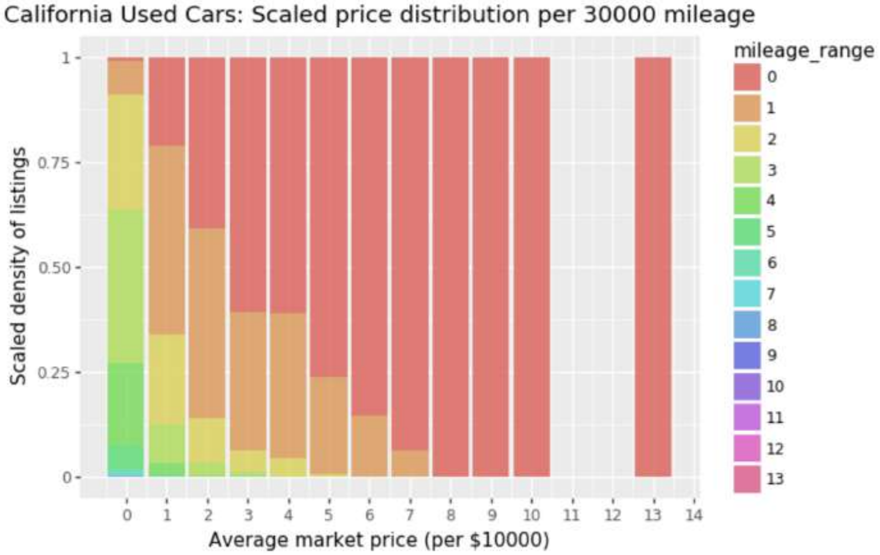
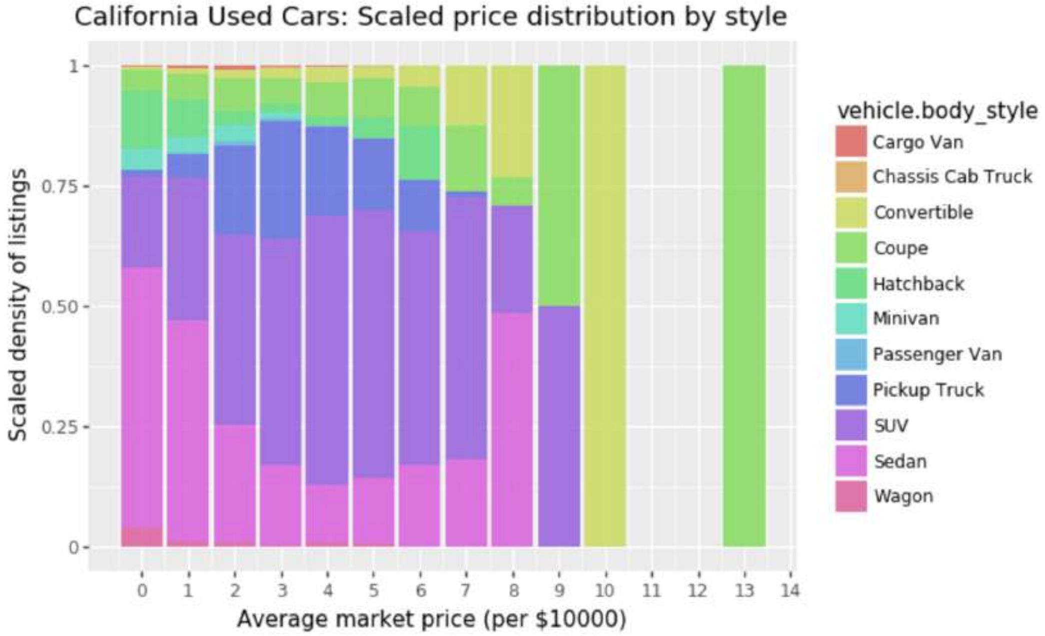
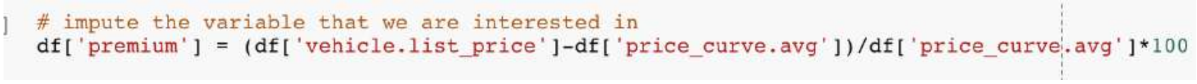


Figure 4.7 - Average market price vs. body style: For the plot below, we can tell that vans are generally priced quite low in the used car market, while SUV, Sedan, and pickup trucks take major proportions of middle-priced vehicles, and Coupes and Convertibles are more like the premium cars with higher market prices.



5. Exploratory Data Analysis with Machine Learning

To answer the research question, we generated a new numeric variable called 'premium' by calculating the percentage that the listed price for a vehicle is above or below the average market price.



```
]  # impute the variable that we are interested in
   df['premium'] = (df['vehicle.list_price']-df['price_curve.avg'])/df['price_curve.avg']*100
```

Then, we defined a categorical variable called 'premium_power', which represents the variable that we are primarily interested in, by separating the data according to whether the 'premium' is above or below one standard deviation σ from its mean μ.

Finally, we can separate the cars into three categories:
(1) those of normal premium power, with $\mu - \sigma$ < 'premium' < $\mu + \sigma$
(2) those of strong premium power, with 'premium' > $\mu + \sigma$
(3) those of weak premium power, with 'premium' < $\mu - \sigma$

To predict our variable of interest, we chose the decision tree modeling, a classical machine learning usd to predict the category of an instance, and we transformed all the categorical variables into domain variables to do categorical regression on the premium power by building a decision tree model. With the help of the sklearn package, we let the model automatically select the most effective variable among all the variables to make each classification decision for each group of the instance. The closer an variable is near the root node, it matters more for classification. One good standard to measure if a decision tree model is reliable can be the difference between training prediction accuracy rate and testing prediction accuracy rate. The smaller the better. Therefore, if we are able to build a reliable model, we can know which variables are indeed important for distinguishing different cars.

During the process, we dropped irrelevant variables such as those about the dealership and other prices, since we are only interested in what and how the vehicle features influence the premium power. After the data was selected, we took out the ‘premium_power’ from the data as the label, transformed both the data and label into proper types, and made sure there was no NAs in each of them.

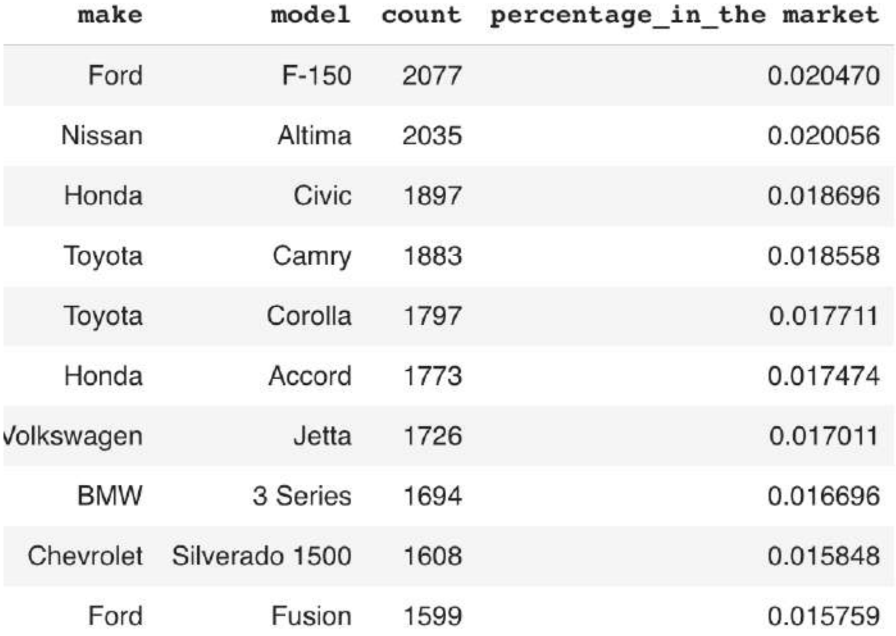
When designing the decision tree model, we randomly split the data into the training set and testing set, which took 80% and 20% of the data respectively. We limited the maximum depth to 3 to prevent the tree from becoming too big because big tree model can lead to overfitting problems. To check to prediction outcomes, we counted the number of cases that they have a difference large than one percent between accuracy rate on the training set and testing set. Besides, we also calculated the average accuracy rate on the testing set.

To make the result more obvious, we used 100 different seeds to randomly split the data set. then, we counted the frequency of all the factors appeared in the first three layers of the decision tree to see which factors are the most important ones when predicting a car’s premium power.

6. Key Findings

After data exploration and decision tree modeling, we had some interesting discoveries about the used car market, just as shown below:

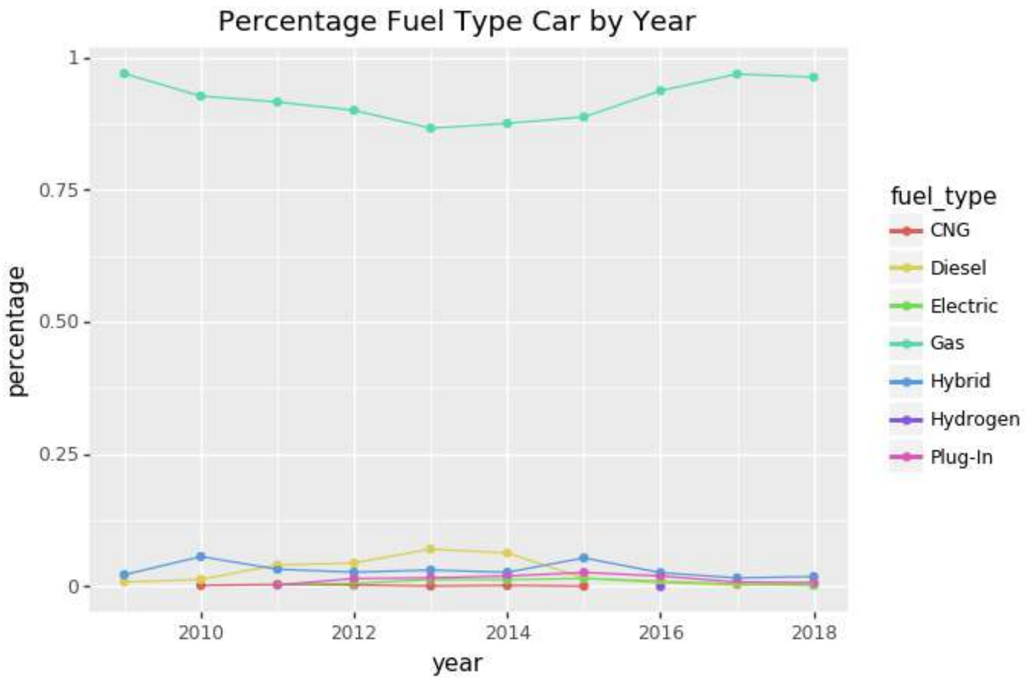
Figure 6.1 - Top 10 popular used cars for sale in CA:



make	model	count	percentage_in_the market
Ford	F-150	2077	0.020470
Nissan	Altima	2035	0.020056
Honda	Civic	1897	0.018696
Toyota	Camry	1883	0.018558
Toyota	Corolla	1797	0.017711
Honda	Accord	1773	0.017474
Volkswagen	Jetta	1726	0.017011
BMW	3 Series	1694	0.016696
Chevrolet	Silverado 1500	1608	0.015848
Ford	Fusion	1599	0.015759

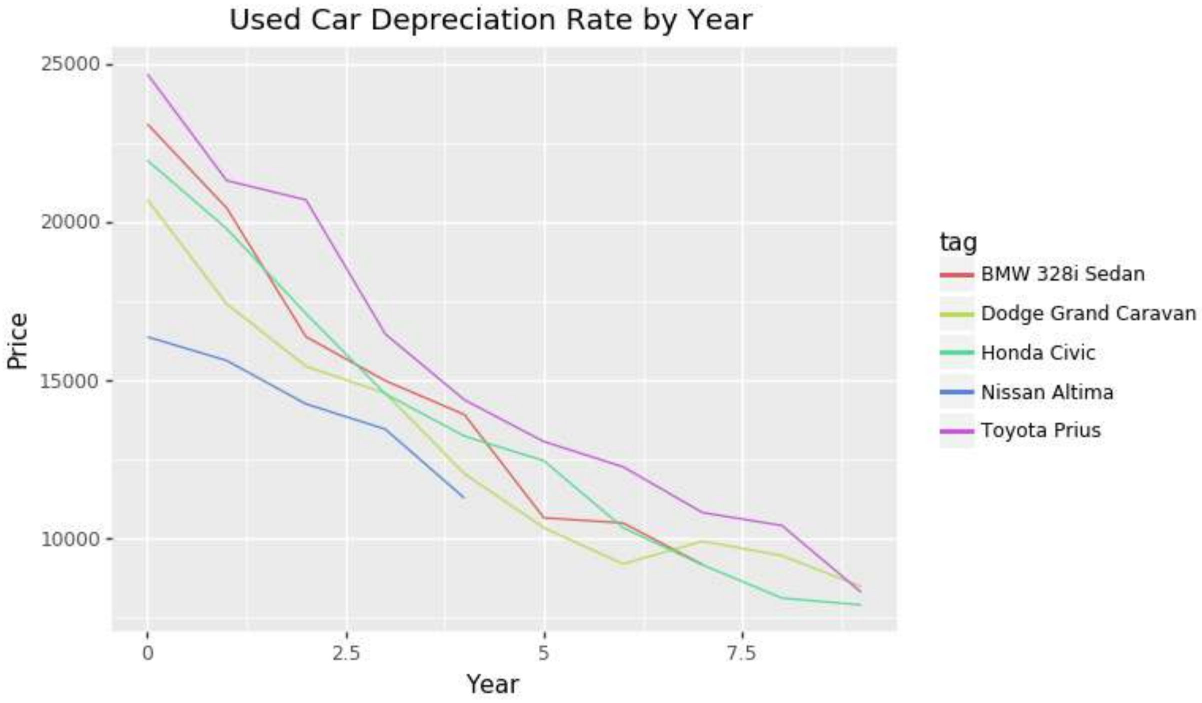
After we got the top 10 popular cars, we noticed that 50% of them are Japanese cars, 30% of them are American cars, and the remaining are European cars. The Japanese cars are Economy type sedan. The most popular vehicle is Ford F150 pickup, which is well used in work and daily commute. It’s not a surprise to see this result as this car is easily seen on the street. Besides, BMW 3 Series is one of the most popular models in the market, and as an entry-level luxury car, its moderate price, elegant appearance, luxurious interior, as well as good handling are very popular among young people.

Figure 6.2 - Percentage of fuel type by Year:



From this plot above, we can see that between 2009 and 2014, the percentage of gas-using cars has a significant drop, while diesel cars increased since the gas price experienced an increase from that period. After 2014, the gas price dropped again which led gas cars to increase.

Figure 6.3 - Used car depreciation rate by year:



From the graph, it seems that buying a new car might not be a good investment as you lose your money every year, and your car would only worth half of the original price in 5 years.

Comparing several popular car models, Toyota Prius is most valuable car in the first 8 years, and most cars' price would become very low after 8 years as the condition of the car gets bad.

Figure 6.4 - Bay Area used car dealer heatmap: Bay area is an area full of mountains, so the location of car dealers can also reflect the relationship between population and economy. The area with more people is also likely to have stronger economy. And because silicon valley is the heart of technology development, so the salary level will be higher than other areas. We can see there are two places which have the most car dealers, one in San Francisco and the other in Cupertino. Because many hi-tech companies are in San Francisco, there are more people in SF who can afford a car. Whereas Apple has set their company at Cupertino, which makes Cupertino also have many car dealers.

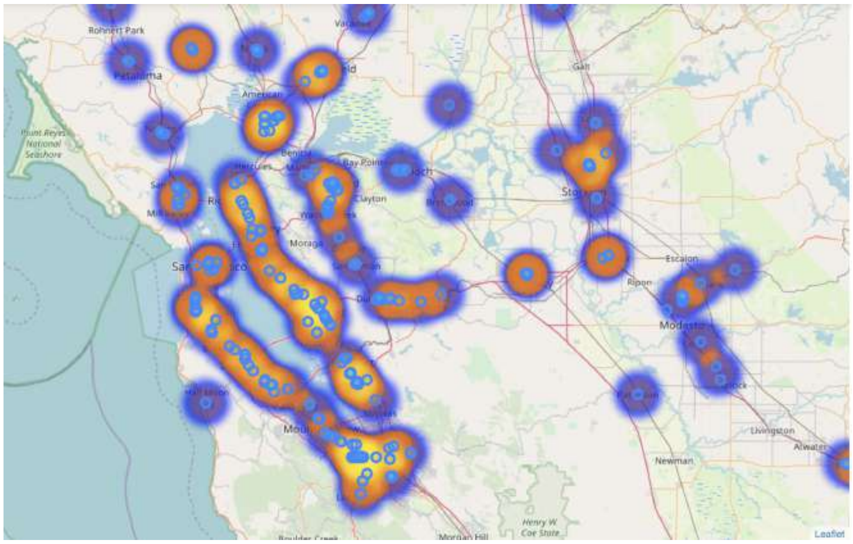


Figure 6.5 - Sacramento area used car dealer heatmap: Sacramento is the capital of CA but with the lower level of economic development than the silicon valley, so the heat map is looser compared to silicon valley and most car dealers prefer places that have better economic conditions. For Davis, many students and university professor are living here so the heat map is sort of concentrated here.

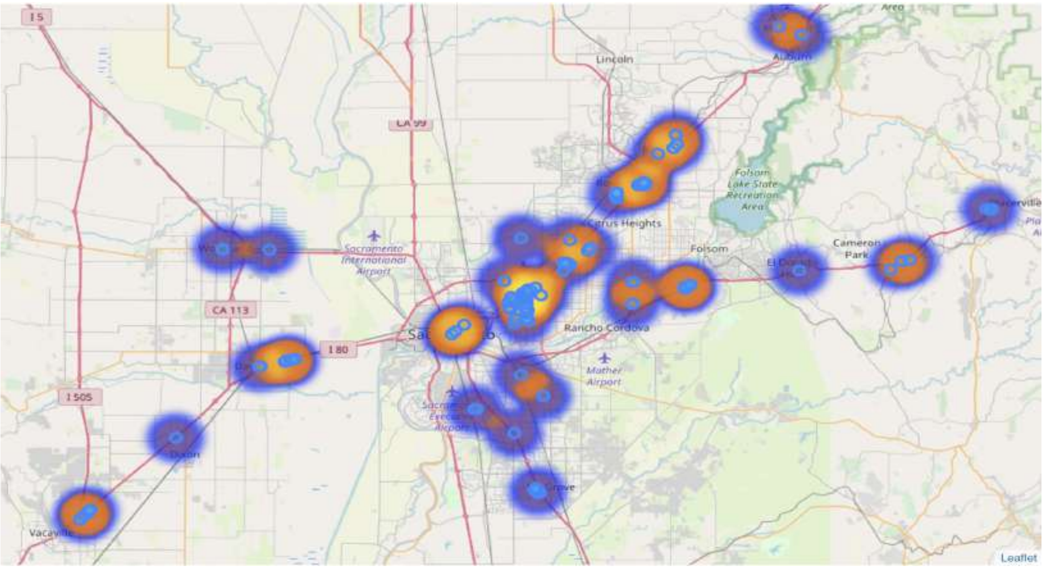


Figure 6.6 - South California used car dealer heatmap: Because south California is mostly plain, the car dealers are concentrated here as well. The middle of Los Angeles has a better level of economic development, so car dealers are likely to settle here.

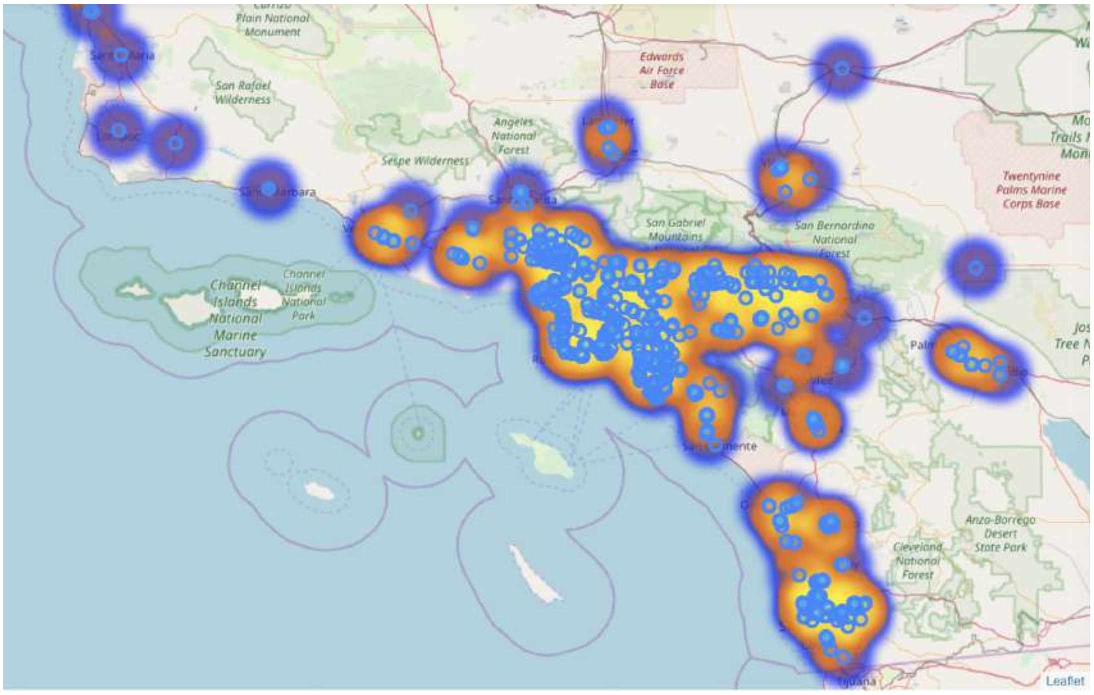
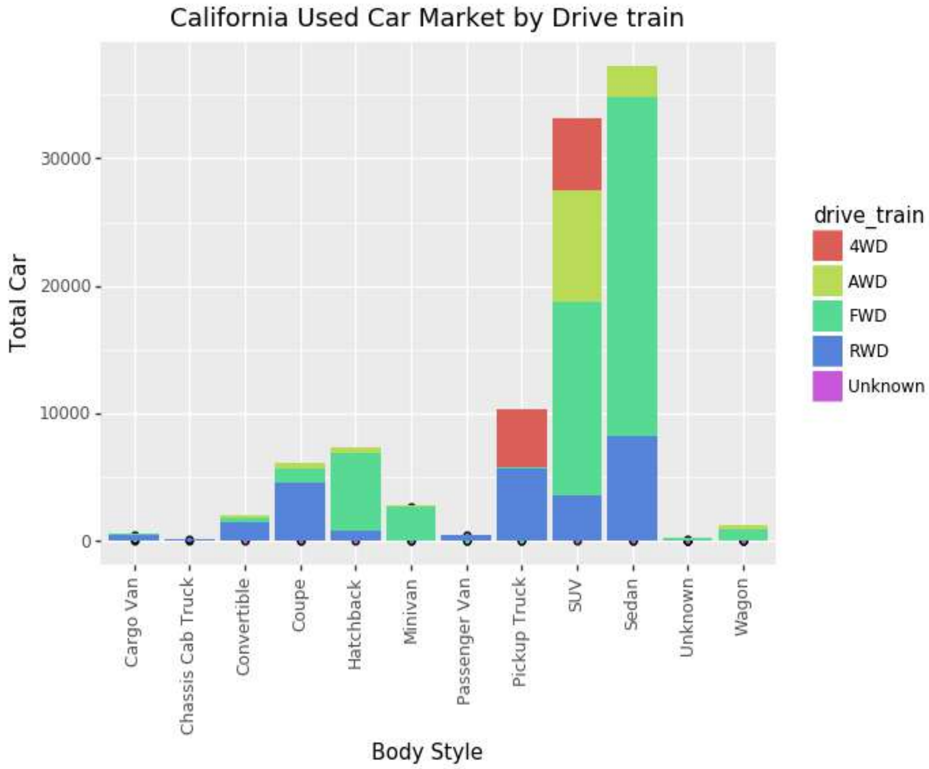
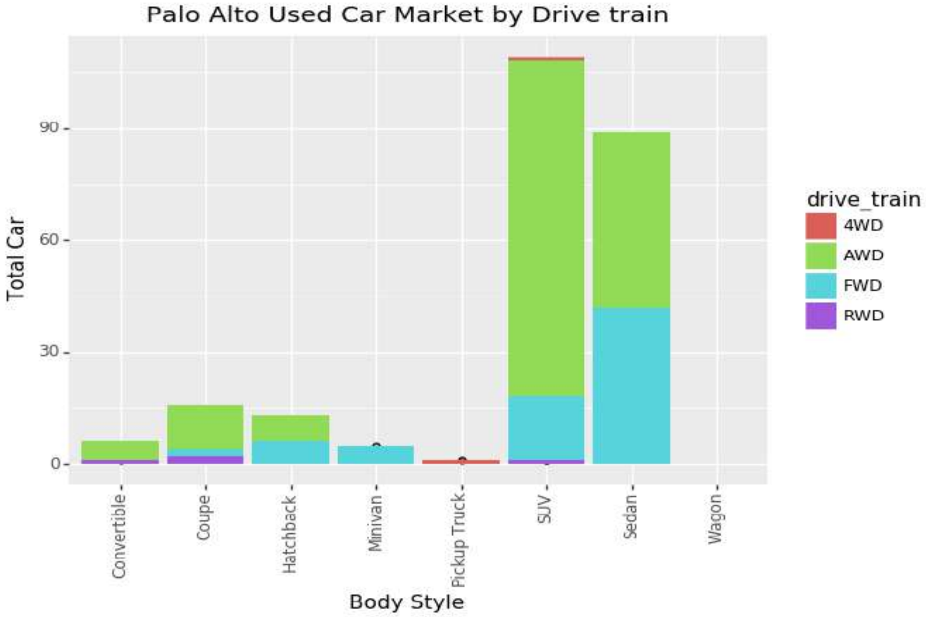

Plain National Monument
San Rafael Wilderness
Sespe Wilderness
Angeles National Forest
Edwards Air Force Base
San Gabriel Mountains
San Bernardino National Forest
Twentynine Palms Marine Corps Base
Channel Islands National Marine Sanctuary
Channel Islands National Park
Anza-Borrego Desert State Park
Cleveland National Forest
Tijuana
Leaflet

Figure 6.7 - California used car market by drive trains:



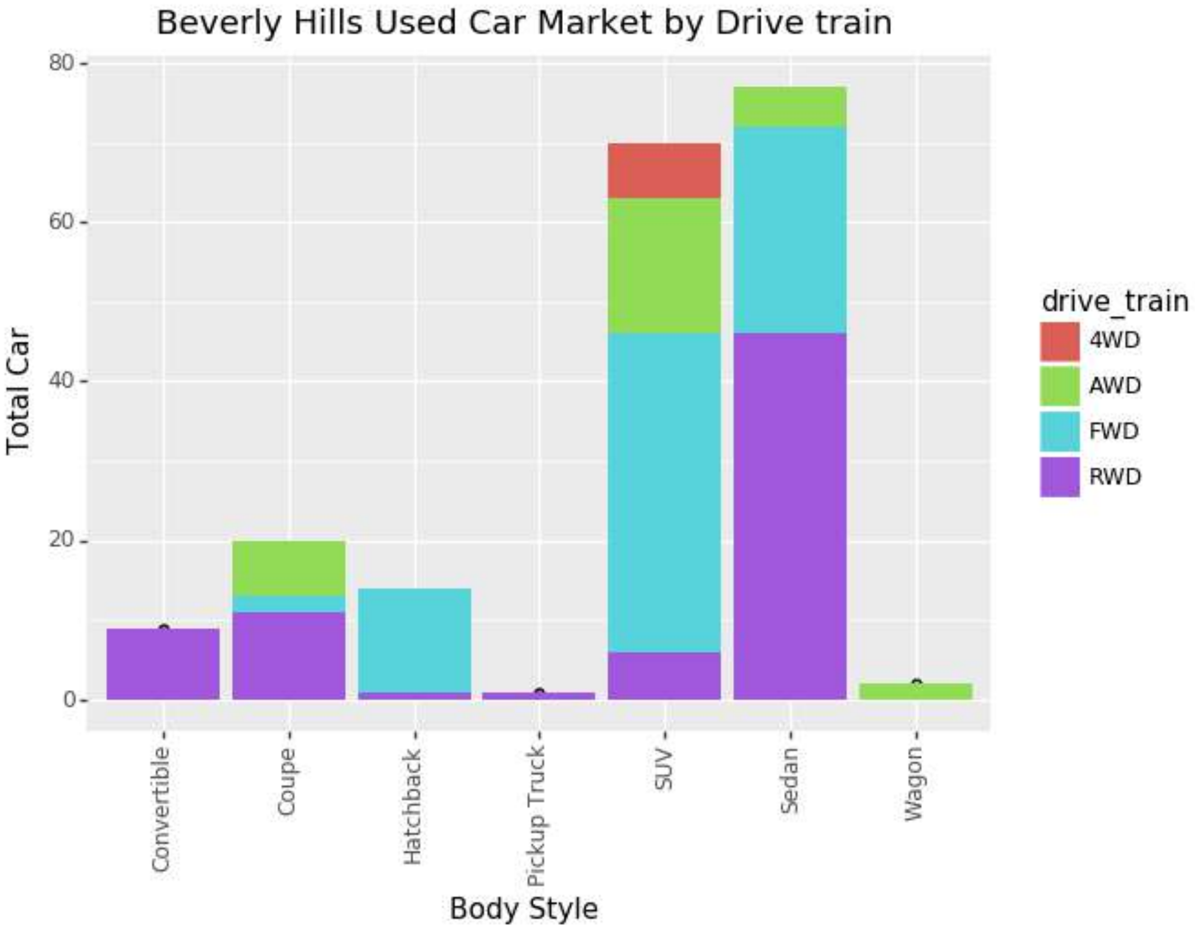
In California, FWD is the most popular car, and the next most popular one is the FWD SUV. At SUV, the 4WD and AWD are relatively popular but not in sedan, because most sedans are designed to run in cities but SUV has a wider range of usage. People like to drive SUV at the iced or slippy road at winter or climb the mountain so people who purchase SUV will focus more on its performance on handling different kinds of driving conditions.

Figure 6.8 Palo Alto used car market by drivetrain:



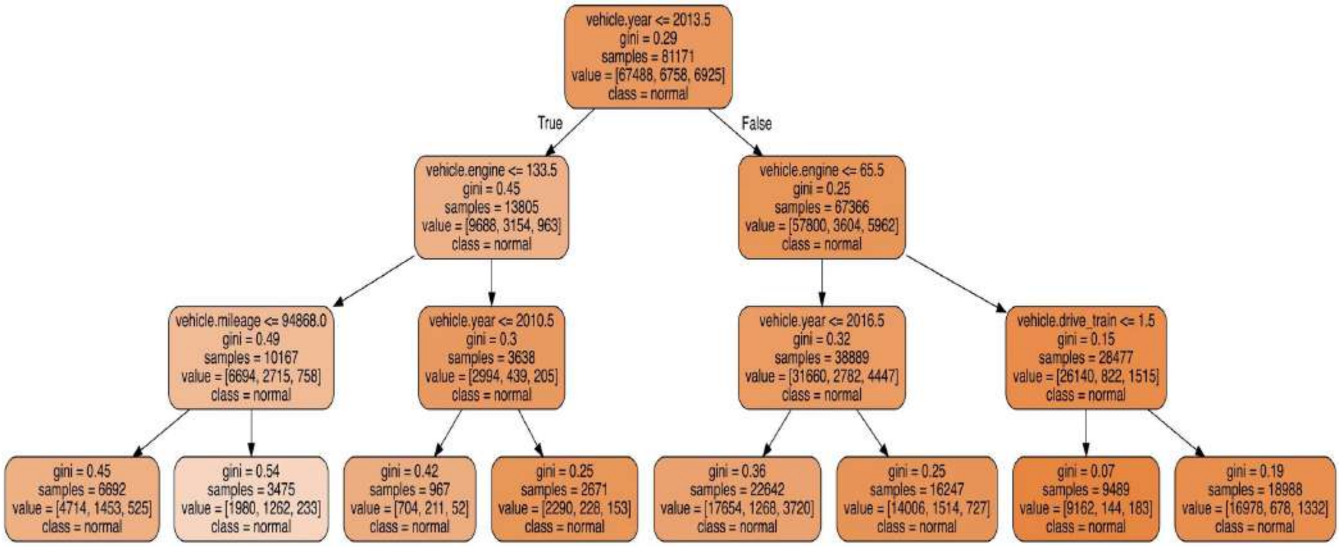
Palo Alto is a relatively rich city in northern California. People more likely to purchase SUV rather than sedan cars. Most SUVs are AWD. The percentage of AWD and FWD at sedan cars are almost the same, which means people are focus more on quality rather than the price of cars. For Coupe and Convertible, most cars are AWD. In conclusion, people at Palo Alto likes to buy AWD cars and it is totally different than the whole of California.

Figure 6.9 Beverly Hills used car market by drivetrain:



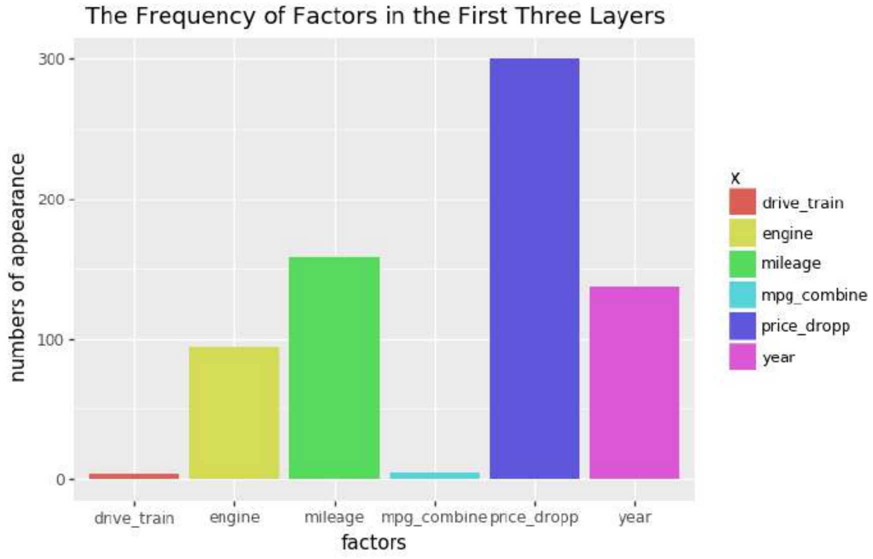
Beverly Hills is a relatively rich city in southern California. The people who lived here prefer RWD cars. And the FWD SUV is taking majority percentage because Beverly Hills is mostly a plain area and it will not snow here, people do not have a high interest on buying 4WD and AWD which Palo Alto does.

Figure 6.10 Decision Tree Model



According to the decision tree plot, we can see that the vehicle's year, engine, mileage and driving training are crucial when predicting the premium power. This result is coincide with what we find in the data exploration part, where we found that these variables have relations with used cars price. Besides, based on common sense, we know most used car website put factors like year, mileage and engine into the filter buttons of the users' webpage. Therefore, our results are also matched with this fact. However, since the training and testing data is randomly selected, this single graph is not convincible to tell whether these factors are indeed the most important ones. Therefore, we counted the frequency for factors shown in the first three layers of the decision model, which is shown in the next graph.

Figure 6.11 - Important Factors from Decision Tree results:



After building 100 different decision trees with 100 different seeds, we catch no case where the difference between predicting accuracies training and testing is larger than 1%. Besides, the mean of predicting accuracies on testing set is 83%, which is relatively acceptable since we only use the information about the vehicle itself, instead of including all other information. Therefore, we think the decision tree modeling is reliable and decided to analyze the variables based on it.

According to the graph and compared with the result in the single decision tree model, we can see that factors like the engine, mileage, and year are indeed crucial when determining used cars' premium power. However, driving training is not as important as those factors. Besides, we can see that price dropping appeared in the first three layers very frequent. It is a variable indicating whether there is a drop in the used cars' price. Its high frequency makes sense, since a used car is obviously losing its premium power when its price is dropping.

7.Conclusion and Final Thoughts

We can tell if a price is really a good deal for a used car based on factors of a vehicle itself, such as the year, engine, mileage and price_dropped. We can conclude that a vehicle's year, mileage, engine and price dropping are the most important factors to determine a used car's premium power. It is a good time for customers to avoid an unfair price when the price is dropping given the same condition on other factors. It is also necessary for customers to highly consider the used car' year, mileage and engine type when judging a price.

However, since we do not have the business data about the actual sales price, we cannot analyze how these factors influence the used cars prices or how the trends of used car market in a certain area. If we have the actual sales price in the future, we want to explore more about how to predict a used car's price and find out the reason why some good deals take place!