1. Introduction

Austin is the capital city of the U.S. state of Texas. Incorporated on December 27, 1839, it is the 11th-most populous city in the United States, the fourth-most-populous city in Texas, and the second-most-populous state capital city (after Phoenix, Arizona) [1][2]. Austin is also the one of the most car-dependent large Texas cities. According to the 2011-15 from the U.S. Census Bureau's American Community Survey, only 6.6 percent of the household in Austin didn't have vehicles [4]. The growth of demand for vehicles is inescapable in this city.

Currently, there are about 280 million vehicles in operation in the United States, an increase of about 1.6 percent [3]. The rising demand for used vehicles means that used vehicle inventories are declining. Thus, it is advantageous for a company or person who may want to buy or sell a used car to predict the price of the used car and explore the car dealerships within certain distance.

According to Flex Fleet [7], there are approximately 4.2 million pickup trucks in the Texas, which make Texas become second largest truck market (California is the top). In this study, we will find out if truck is one of the most popular posted car models among the dealers. Additionally, the location data and Craigslist used car dataset might contribute to the analysis of used car price. This study aims to predict the price of a used car, create a map to locate the used car dealer in Austin area, and find the most popular car model.

2. Data Description

Data Source

Craigslist is the world's largest collection of used vehicles for sale, and a dataset which includes every used vehicle entry within the United States on Craigslist was created for a school project. This used car data set is published in Kaggle, and free to download.[5]

This dataset contains 458213 records from Craigslist covers 51 states in USA. As this study is focusing on predict cars price in Austin, we will only use the data in Texas. There are 26 attributes including build year, manufacturer, mode, condition, etc.

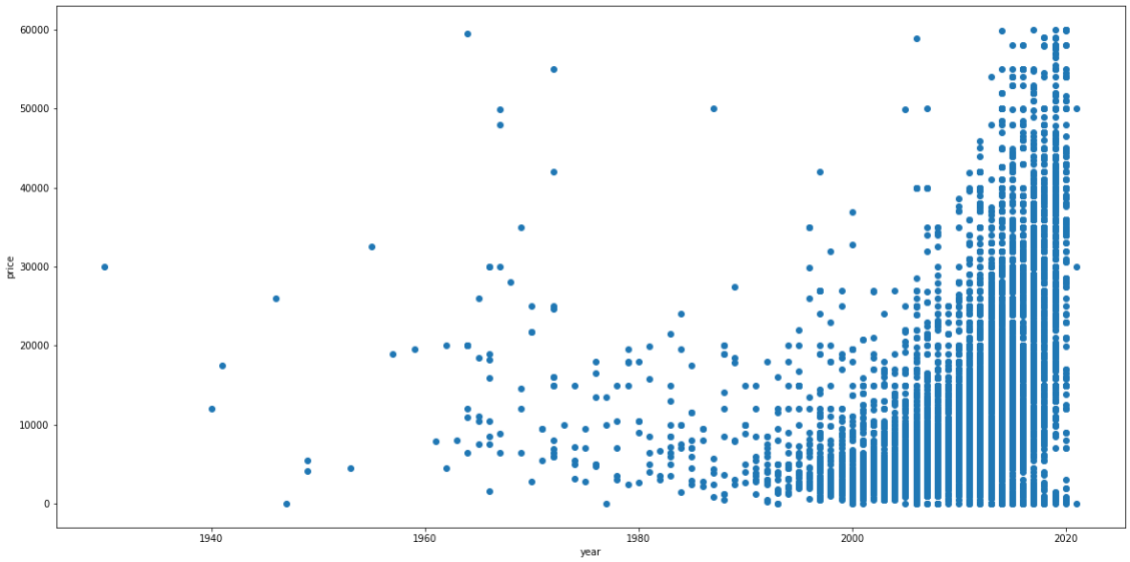
To obtain the car dealers' information within the study area, Forsquare API will be used to identify the dealerships.[6]

Data Cleaning

The Craigslist used car dataset contains some attributes that should not impact the price of the vehicle, such as VIN number, URL, image URL, ect. Those attributes will be removed before creating the predicted model.

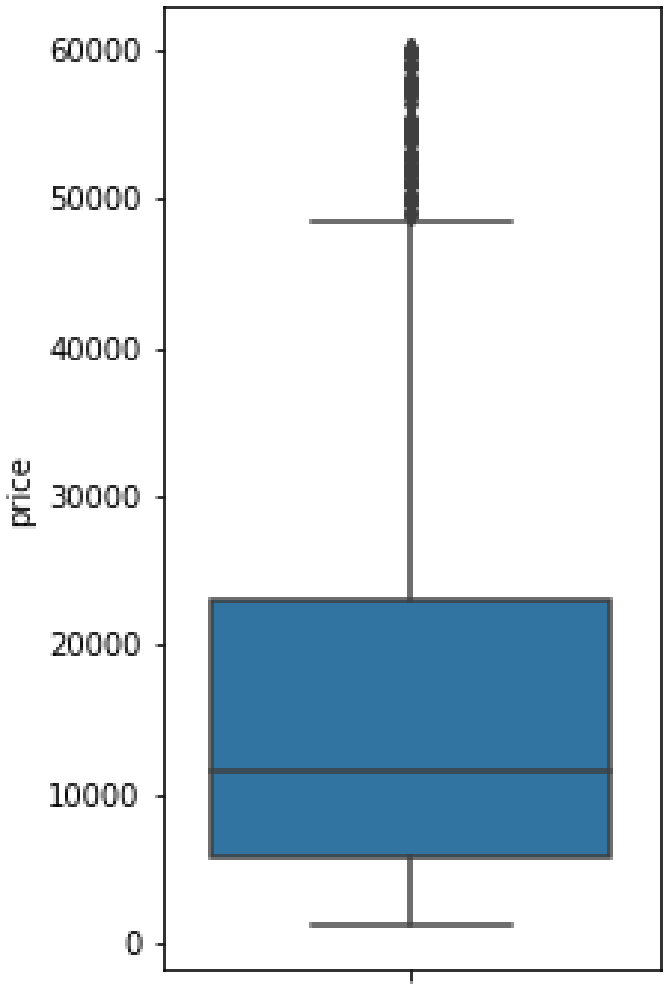
There are some observations in the Craigslist used car dataset includes missing values which will be remove from the dataset.

As shown in the image below, the build year of the vehicles various from 1930 to 2021. However, over 95% of the observations were built after 2000, so we will only include vehicles that build with in past 30 years. Also, there is only 1 observation in 2021, so we will remove it from the dataset.



As shown in the data, the minimum value of used car price is 0, and the maximum value is 3.6152e+09. However, 99% of the used cars have price below 59995. Therefore, we will remove

the observations with price over 60000 or below 1000. The boxplot below shows the price of the used car.

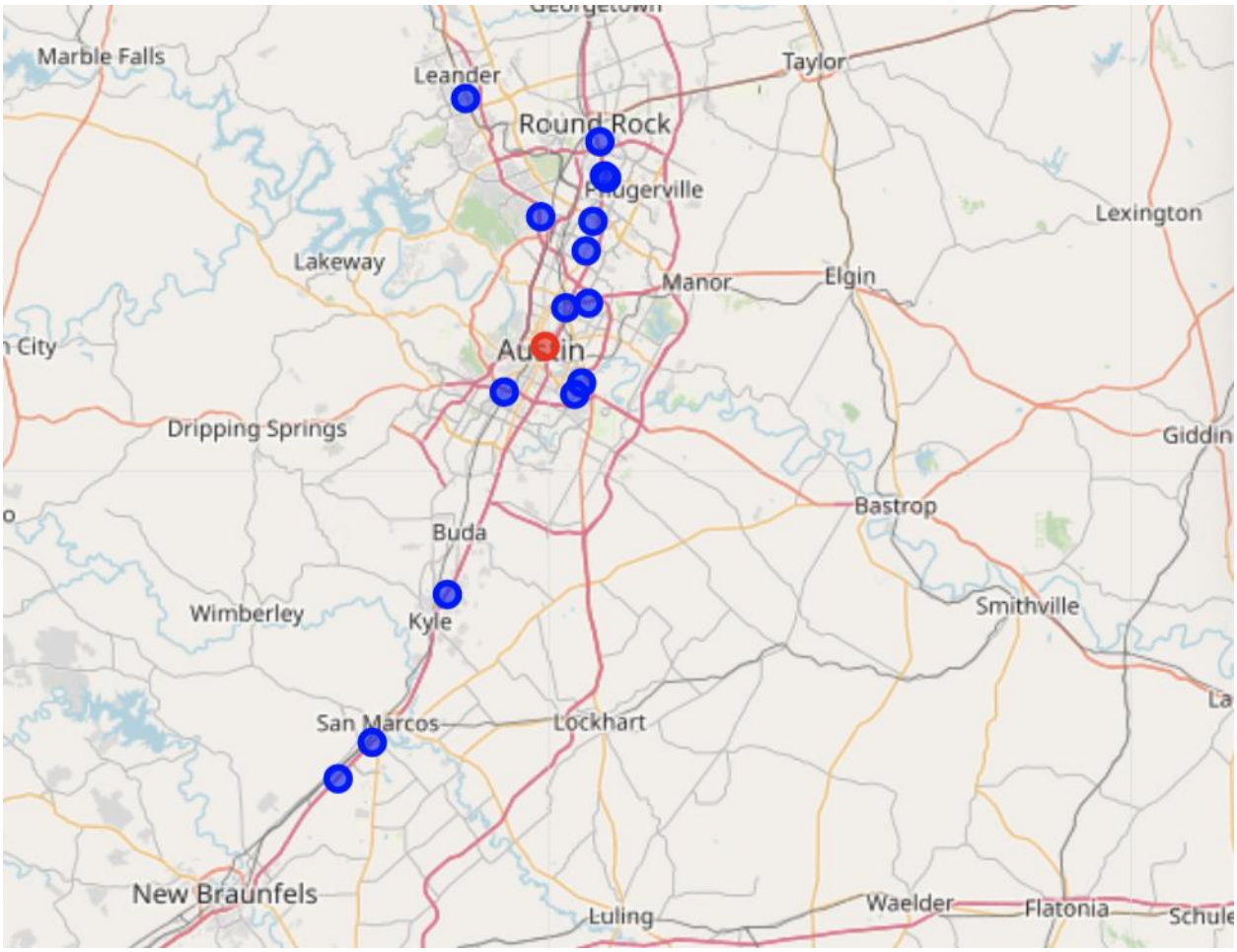


There are some categorical variables in the data, such as condition and fuel. In order to predict the price of the used car, those categorical variables will be transferred to numerical values by applying label encoding.

3. Methodology

Identify the most comment car models

The address of Texas Capitol will be used as the searching location, and the Forsquare API returns all the car dealer information within given radius. There were 15 car dealers returned by the request. To visualize geographic data, a map was created by python folium library based on latitude and longitude values of each record. As shown in the map below, the Texas Capitol at 1100 Congress Ave, Austin, TX is shown in red, and the car dealers are in blue:



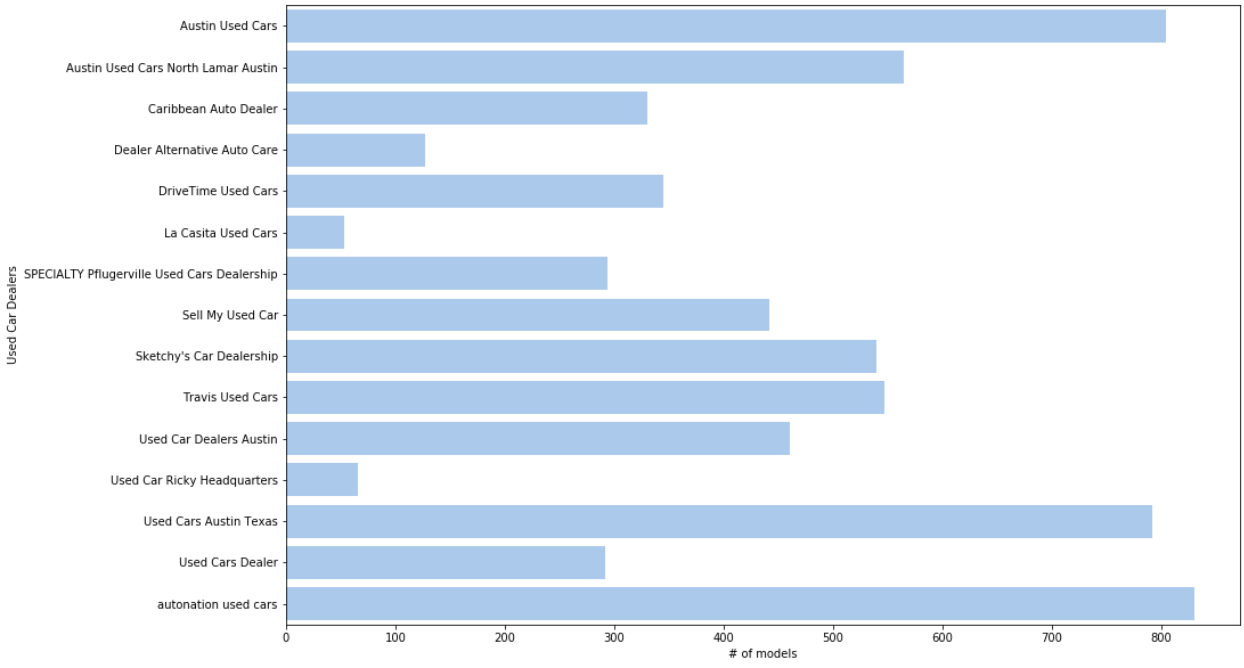
To find the most popular car model near the dealers, the Craigslist used car data were merged into the dealer table based on given latitude and longitude information. If the used car is within 5 miles of a dealer, we will link it to that dealer, and appended it to the merged table.

The dataset contains manufacturer and model of each used car, and we will combine those two attributes to obtain the type of that car. This attribute will be shown as “model” in the merged table. Below is the first 5 rows of the merged table.

name	address	lat	long	Model	latitude	longitude
Used Car Dealers Austin	1100 Congress Ave, Austin, TX	30.320618	-97.688241	SUBARU FORESTER	30.294913	-97.698755
Used Car Dealers Austin	1100 Congress Ave, Austin, TX	30.320618	-97.688241	CHEVROLET TAHOE	30.313479	-97.698669

name	address	lat	long	Model	latitude	longitude
Used Car Dealers Austin	1100 Congress Ave, Austin, TX	30.320618	-97.688241	SUBARU WRX	30.376400	-97.707800
Used Car Dealers Austin	1100 Congress Ave, Austin, TX	30.320618	-97.688241	BUICK SKYLARK	30.331600	-97.700400
Used Car Dealers Austin	1100 Congress Ave, Austin, TX	30.320618	-97.688241	JEEP GRAND	30.343200	-97.738975

The image below shows the number of vehicles appended to each dealer. In summary of this data, there were total 6487 vehicles find near the dealers and 445 unique models were linked to the dealers.



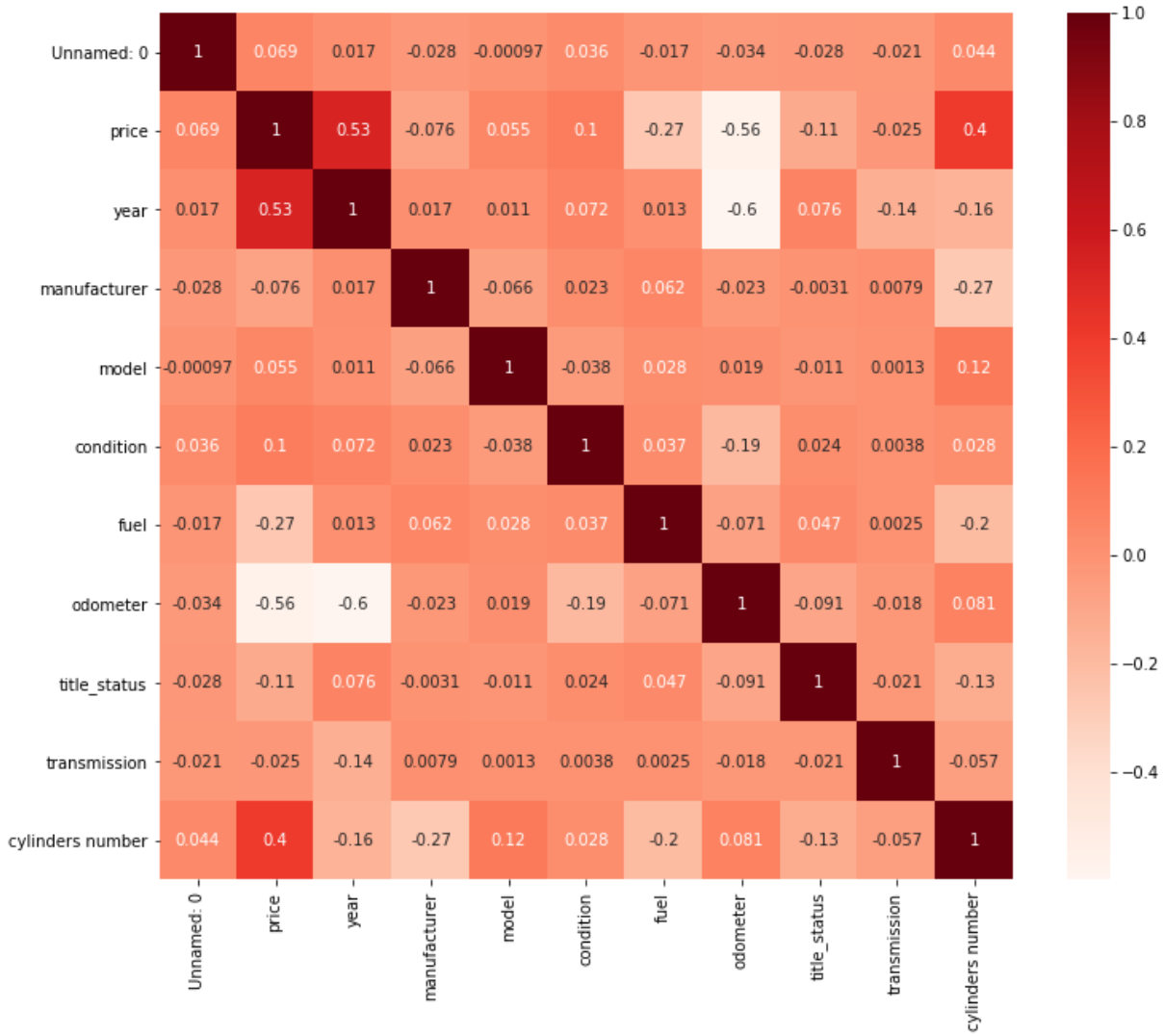
One-Hot Encoder was applied to the data, to help us find the most comment car model posted for each dealer. The table below shows the top 5 car models for each dealer.

Name	1st Most Common Vehicle Model	2nd Most Common Vehicle Model	3rd Most Common Vehicle Model	4th Most Common Vehicle Model	5th Most Common Vehicle Model
Austin Used Cars	TOYOTA CAMRY	HONDA CIVIC	TOYOTA TACOMA	NISSAN ALTIMA	RAM 1500
Austin Used Cars North Lamar Austin	TOYOTA COROLLA	CHEVROLET TAHOE	BMW 3	VOLKSWAGEN TIGUAN	CHEVROLET SILVERADO
Caribbean Auto Dealer	RAM 2500	TOYOTA COROLLA	CHEVROLET SILVERADO	TOYOTA CAMRY	NAN TEXAS
Dealer Alternative Auto Care	CHEVROLET SILVERADO	JEEP WRANGLER	FORD MUSTANG	TOYOTA TACOMA	FORD F-150
DriveTime Used Cars	RAM 2500	CHEVROLET SILVERADO	TOYOTA COROLLA	CHEVROLET TAHOE	FORD F-150
La Casita Used Cars	MERCURY GRAND	RAM 1500	DODGE DURANGO	FORD F250	CHEVROLET TAHOE
SPECIALTY Pflugerville Used Cars Dealership	TOYOTA COROLLA	CHEVROLET SILVERADO	CHEVROLET TAHOE	TOYOTA CAMRY	FORD F-150
Sell My Used Car	TOYOTA COROLLA	TOYOTA CAMRY	CHEVROLET TAHOE	FORD F-150	CHEVROLET SILVERADO
Sketchy's Car Dealership	RAM 1500	HONDA CIVIC	TOYOTA TACOMA	FORD RANGER	HYUNDAI GENESIS
Travis Used Cars	RAM 1500	HONDA CIVIC	FORD RANGER	TOYOTA TACOMA	HYUNDAI SONATA
Used Car Dealers Austin	CHEVROLET SILVERADO	CHEVROLET TAHOE	VOLKSWAGEN TIGUAN	TOYOTA CAMRY	VOLKSWAGEN PASSAT

Name	1st Most Common Vehicle Model	2nd Most Common Vehicle Model	3rd Most Common Vehicle Model	4th Most Common Vehicle Model	5th Most Common Vehicle Model
Used Car Ricky Headquarters	CHEVROLET SILVERADO	FORD F-150	FORD MUSTANG	FORD EXPLORER	NISSAN ALTIMA
Used Cars Austin Texas	CHEVROLET SILVERADO	GMC SIERRA	CHEVROLET TAHOE	RAM 1500	HONDA CIVIC
Used Cars Dealer	FORD F-250	RAM 2500	FORD F-150	FORD F-350	RAM 3500
automation used cars	JEEP WRANGLER	CHEVROLET SILVERADO	ROVER SPORT	MERCEDES-BENZ S-CLASS	AUDI Q5

Predicting used car price

To predict the price, we will build a multiple linear regression model. There are several attributes that can be included in the model, and a filter method will be used for variable selection. The image below shows the correlation matrix with Pearson Correlation.



As shown in the correlation matrix, year, condition, cylinders, fuel, odometer and title_status have relatively high positive and negative relationship with price. However, one of the assumptions of linear regression is that the independent variables need to be uncorrelated with each other, so we will need to check the correlation between each variable.

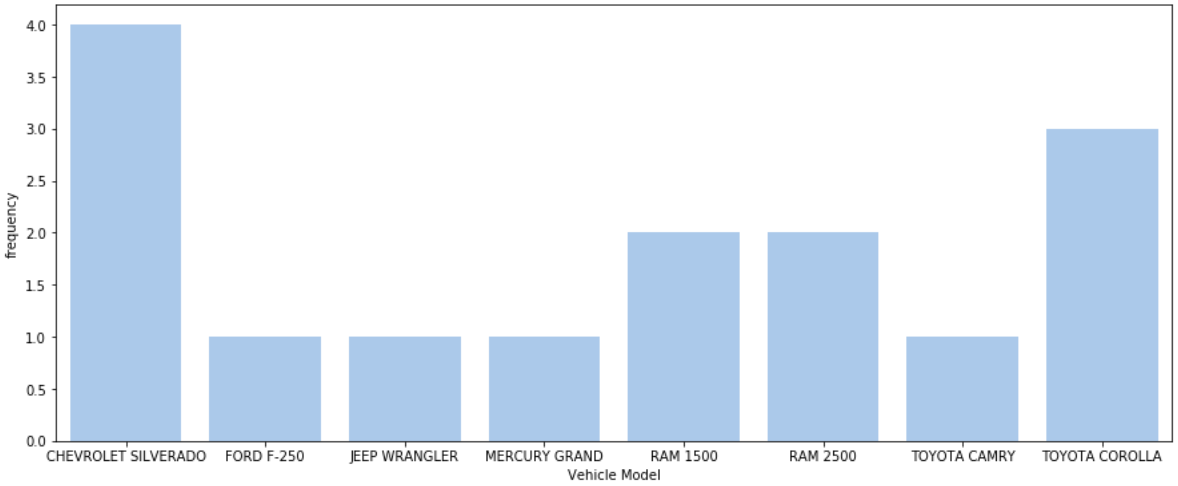
As shown in the table below, the correlation between odometer and year is -0.60, indicating that there is significant correlation between the variables. Thus, we will not include odometer in the model.

	year	condition	cylinders	fuel	title_status	odometer
year	1	0.072026	-0.12454	0.01269	0.075958	-0.59794
condition	0.072026	1	0.042909	0.036515	0.023886	-0.19076
cylinders	-0.12454	0.042909	1	-0.19982	-0.10877	0.092357
fuel	0.01269	0.036515	-0.19982	1	0.046953	-0.07076
title_status	0.075958	0.023886	-0.10877	0.046953	1	-0.09075
odometer	-0.59794	-0.19076	0.092357	-0.07076	-0.09075	1

To test the accuracy of the model, the data were split into train set and test set. We used 20% of the data as test set and trained the model with 80% of the data.

4. Results and Discussion

As the summarize table shown in the previous section, the most comment car model posted is CHEVROLET SILVERADO. The frequency of each model identified as most comment car model is shown in the image below. As we expected, trucks are very popular in Austin, as 60% of the most comment car models are truck.



Let's check our predicted model for used car price. After training the model with training set, the test set was used to test the accuracy of the model. The independent variables including year, condition, transmission, cylinders, title_status. The p-values of variables are shown below. All the p-values are below 0.05, indicating that the null hypothesis can be rejected, and all the independent variables can significantly impact the price.

Independent Variable	P-value
const	0.00E+00
year	0.00E+00
condition	2.99E-11
cylinders	0.00E+00
fuel	3.33E-9
title_status	2.02E-28

The mean absolute error, mean squared error, root mean squared error and R^2 of test set are shown below.

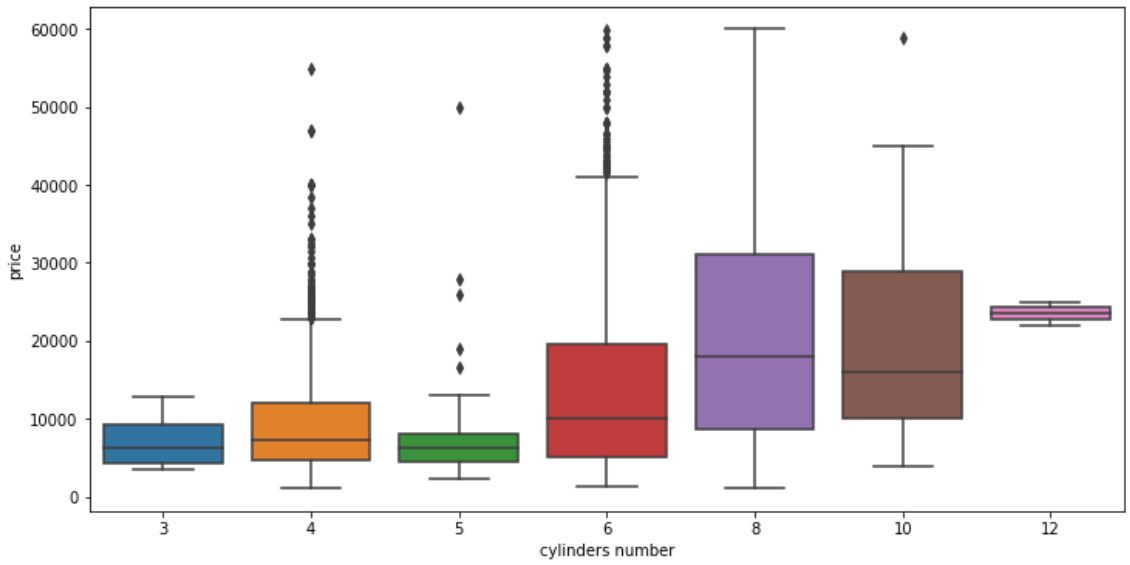
Mean Absolute Error	5682.12
Mean Squared Error	56566562.60
Root Mean Squared Error	7521.07
R2	0.58

After confirming the performance of the model, we used all the data to train the model. The table below summarizes the coefficients of the regression model.

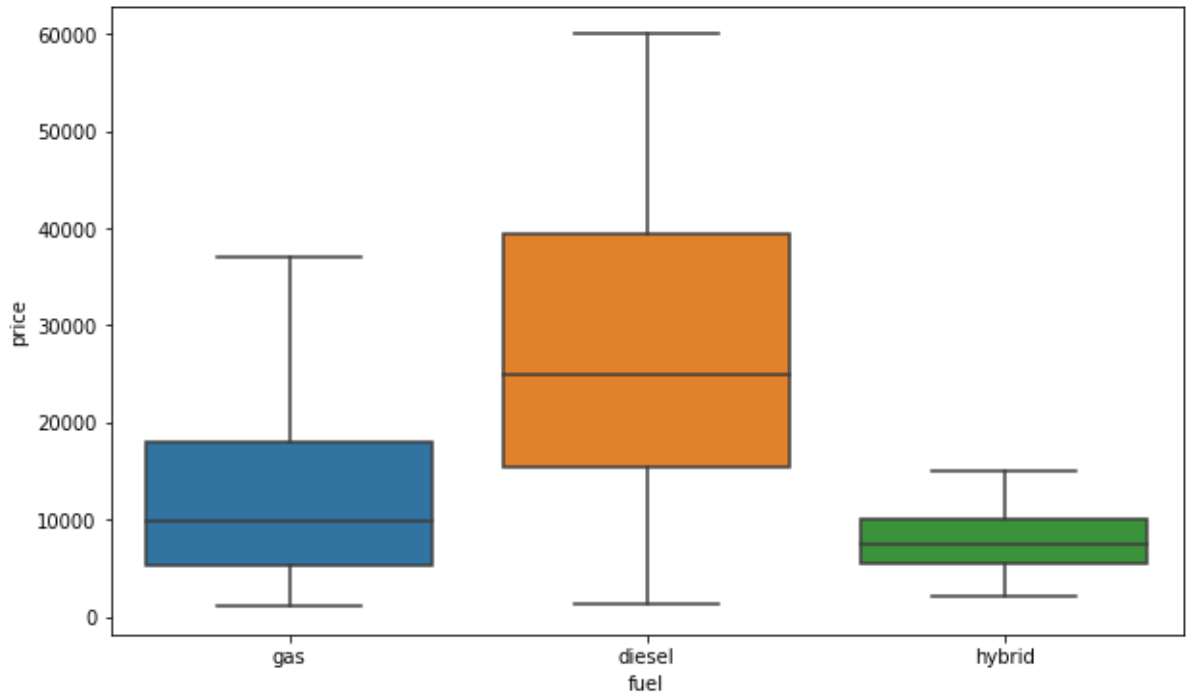
Intercept	-2484668.99
year	1236.82
condition	535.03
cylinders number	3163.44
fuel	-8209.31
title_status	-1254.56

As shown in the image below, vehicles build in later year trends to have higher price because the newer model will cause higher price. In our model, the coefficient of year is 1236.82, indicating that the price of a vehicle will be $ 1236.82 higher than the model from previous year.

The number of cylinders also can impact the price. As shown in the image below, vehicles with more cylinders will have higher price. One unit increase of cylinders will cause $535.03 increase of price.



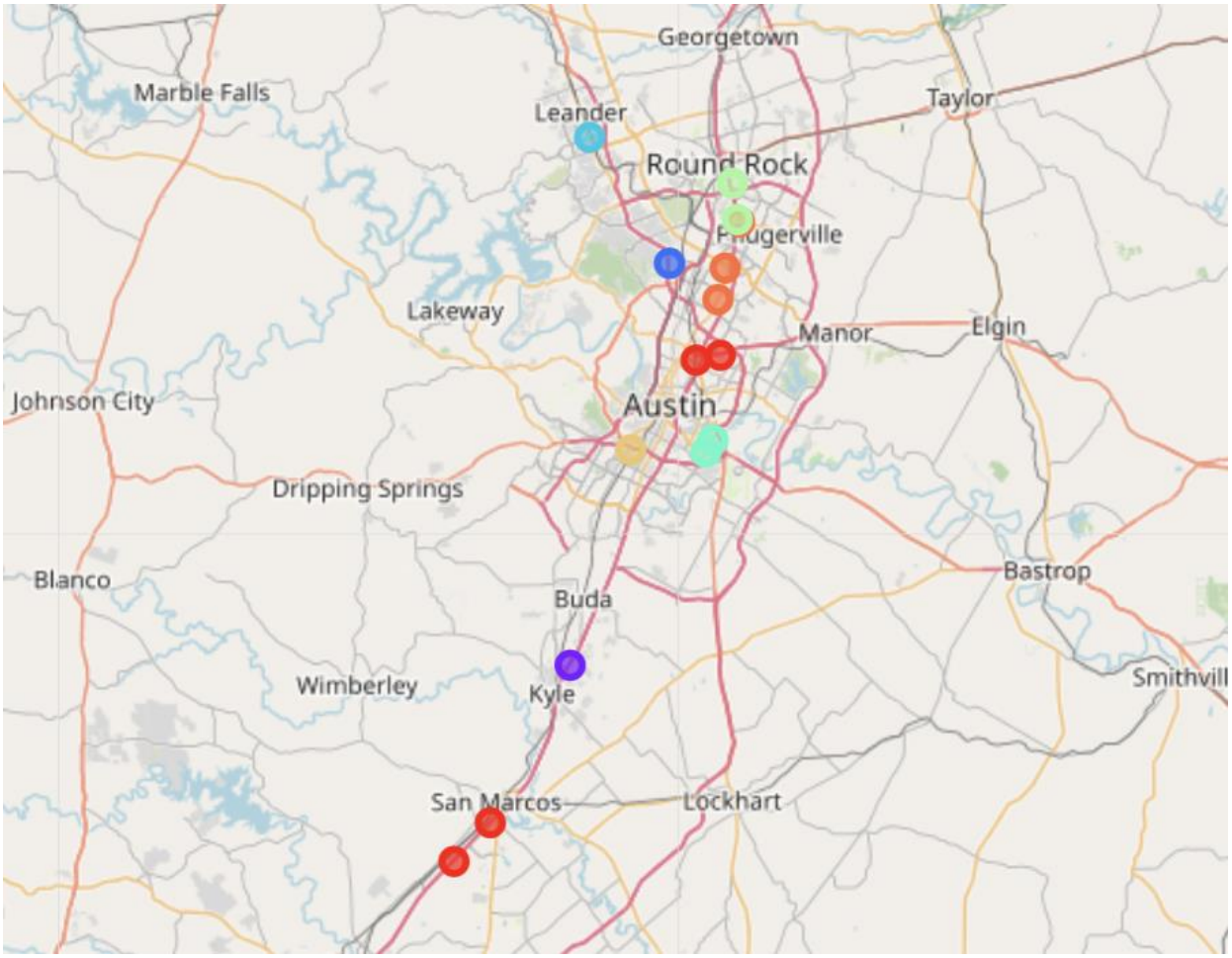
As shown in the model, fuel has negative coefficient. Thus, the vehicles with diesel trend to have higher price, and the usage of gas or hybrid will decrease the price. The price could be 8209.31 lower for a vehicle uses gas and 16418.62 for hybrid compared with a vehicle uses diesel.



5. Conclusion

Per the discussion in the previous sections, below are the conclusions of this study.

There are 15 car dealers returned by the Forsquare API, and 504 unique models were posted near those dealers. The most comment car model posted is CHEVROLET SILVERADO, and 60% of dealers' most comment car model is truck. The relatively lower gas price and large number of farm owners the reasons for this preference. The map below shows the location of each dealer with the most comment car model.



A predicted multiple linear regression model was created in this study to predict used car price. As shown in the model, year, condition, cylinders, fuel and title status can impact the price of a used car with 95% significant level. The factor that will increase the price including newer model, higher cylinder number, usage of diesel, clean title and good condition.

References:

1. "Top 50 Cities in the U.S. by Population and Rank". infoplease.com. Retrieved January 27, 2014.
2. "City of Austin - Austin History Center: When was Austin founded?". www.austinlibrary.com.
3. “U.S. new and used car sales 2010-2019”. https://www.statista.com/statistics/183713/value-of-us-passenger-cas-sales-and-leases-since-1990/
4. “Austin drives to top of list of most car-dependent cities in Texas”. https://austin.culturemap.com/news/citylife/06-20-17-car-free-city-report-us-census-bureau-austin/
5. Used Car Dataset
6. Forsquare API
7. Flex Fleet