Predicting United States Used Car List Prices from Craigslist Data

Zachary Kekoa

Table of Contents

Executive Summary

In the world of e-commerce, predictive analytics have become a hot topic for optimizing search algorithms and for being able to give users the best deals. Craigslist has been one of the top competitors in e-commerce for the past 20 years. However, one area where they are currently lacking is being able to let their users know whether the price that is presented to them is fair. My focus on this project will be on used car list prices from Craigslist.

Many variables, both categorical and numerical, are available through Craigslist's website and can influence the list price of a vehicle. These factors include year, manufacturer, model, condition, cylinders, fuel, odometer, title status, transmission, drive, size, type, paint color, and state. This analysis will look at used car data on Craigslist from April 4th, 2021 – May 4th, 2021 to determine which of these variables influences the list price of a used vehicle the most.

Using regression techniques, I was able to create 3 machine learning models for 5 different regions of the United States (west, south, Midwest, northeast, and combined US) for a total of 15 models. I also used polynomial regression and basic data analysis to determine that there is no statistical evidence of a difference in list price between manual transmission vehicles and automatic transmission vehicles. The following report will go into the details of this process.

April 30, 2023

Profile of the organization and background of the opportunity

Primary Company Details:

Founded – 1995
Founders – Craig Newmark
Headquarters – San Francisco, CA
IPO – none, private company
Categories – E-Commerce, Classifieds, Forums

Address:

989 Market Street
#200
San Francisco, CA 94103
United States

Company Communication:

Phone Number: (415) 566-6394
Fax Number: (415) 504-6394
Website: http://www.craigslist.org/

Business Description:

Craigslist is a classified advertisements website with various topics relevant to housing, services, jobs, résumés, items for sale, and discussion forums (Wikimedia Foundation, 2015). Craigslist is completely free for the buyer and seller to use and almost anything can be listed on the website. Another advantage to using Craigslist is that there are a wider variety of items listed than there would be on a typical classified page. There are also many listings that are free, as long as the buyer is willing to pick the item up. Craigslist also acts as a vehicle for connecting users. There are discussion forums and community pages, like Reddit, where users can come together to talk about and seek help on almost any topic imaginable. Additionally, Craigslist allows users to list job openings as well as hire freelancers for gigs, similarly to Fiverr.

Financials:

Latest Financial Data – December 2022
Revenue – $660 million
Assets – N/A
Liabilities – N/A
Return on Equity – N/A
Earnings – N/A
No. of Employees – 50

Key Executives:

Jim Buckmaster, 60, Chief Executive Officer
Craig Newmark, 70, Founder

Major Competitors:

eBay
Facebook Marketplace
Amazon
Carvana
Indeed.com
Apartments.com

Project Plan

Business/Analysis Opportunity:

Using a dataset web scraped from Craigslist over a one-month period from April 4th to May 4th, 2021 that includes 426,881 used car listings, I will be analyzing a couple of research questions and create predictive analytic regression models. The models produced will provide an insight into how fair user listings are compared with expected listings to enable the company to engineer a system for buyers to identify fair used car offers.

Research Questions

Given the current state of the car market, more people are looking for better tools to find the best deal for their budget. Craigslist is still a highly used website for buyers to find their desired used vehicles. This project will address the following research questions.

RQ1: Which attributes influence the list price of a used car, and can the price be predicted?

One of the biggest features of other used car competitors is that they include some kind of price comparison as to whether the listed price is fair or not. Consumers are always wanting to know that they are getting the best deal possible for their needs and their budget. Currently Craigslist doesn't have any kind of feature like this, not just on used cars, but across their entire website. When consumers have the ability to compare the prices of products, it makes them feel confident they are getting the best deal and more likely to buy. If they discover their purchase is not a smart move, they will purchase elsewhere. The goal of this research question is to find the variables that aid in predicting the list price as well as create an effective model to be able to engineer a price comparison tool.

RQ2: Are cars with manual transmissions worth more than those with automatic transmissions?

Many buyers of used cars are enthusiasts who value finding vintage cars with specific attributes such as the number of cylinders, the type of transmission, what kind of fuel the car runs on, etc. Sometimes the ideas of these people can become skewed with false information and ideas because of rumors. One such idea is that manual transmission cars are worth more than cars with automatic transmissions. The goal of this research question is to provide mathematical and statistical evidence as to whether there is any truth to this statement.

Hypotheses

H1: Various factors such as manufacturer, condition, fuel type, transmission, vehicle type, state, etc. will influence the list price.

There are a wide variety of quantitative and qualitative features in this dataset that will be useful in predicting the list price. Machine learning techniques such as linear regression or random forests will be good tools to test this hypothesis.

H2: Automatic cars will have a higher list price than manual cars.

Typically, cars with automatic transmissions cost more to repair than manual cars which makes them more expensive overall. However, cars with manual transmissions tend to have better resale values which may an impact on this project as the dataset is dealing specifically with used cars.

Data

The data is from Kaggle and was originally web scraped from Craigslist's website using used car data from April 4th, 2021 – May 4th, 2021. The dataset includes 426,880 entries of used cars and has both qualitative and quantitative fields. The fields that will be used in this project are price, year, manufacturer, model, condition, cylinders, fuel, odometer, title status, transmission, drive, size, type, paint color, and state.

Price

The *price* variable is the price that the car is listed as being valued. This variable will be the label for the machine learning model analysis.

Year

The *year* variable is the year that the car is manufactured.

Manufacturer

The *manufacturer* variable is the name of the manufacturer of the car. For example, Ford, Honda, etc.

Model

The *model* variable is the name of the model of the car.

Condition

The *condition* variable describes the quality that the vehicle is in. The values for this variable are fair, good, salvage, excellent, like new, new.

Cylinders

The *cylinders* variable describes the number of cylinders that the car has. The values for this variable are 3, 4, 5, 6, 8, 10, 12, and other.

Fuel

The *fuel* variable is the type of fuel that the car runs on. The values for this variable are gas, diesel, hybrid, electric, and other.

Odometer

The *odometer* variable is the amount of miles that are on the odometer. This variable is a specific numeric value. However, for this analysis I will transform this variable into different levels. For example, 0-50k, 50k-100k, etc.

Title Status

The *title status* variable is the accident history of the vehicle. The values of this variable are clean, rebuilt, salvage, lien, missing, and parts only.

Transmission

The *transmission* variable describes what kind of transmission the car has. The values of this variable are manual, automatic, and other.

Drive

The *drive* variable is the type of drive train the vehicle has and what wheels the power is transferred to. The values of this variable are 4 wheel drive, front wheel drive, and rear wheel drive.

Size

The *size* variable describes how big the vehicle is. The values of this variable are full-size, mid-size, compact, and sub-compact.

Type

The *type* variable is a combination of the body style along with the size of the car. The values of this variable are sedan, SUV, pickup, truck, coupe, hatchback, wagon, van, convertible, mini-van, offroad, bus, and other.

Paint Color

The *paint color* variable is self-explanatory as it is the color of the vehicle. The values for this variable are white, black, silver, blue, red, grey, green, brown, yellow, orange, purple, and custom.

State

The *state* variable is the state from which the listed item originates from. The states included in this variable are all 50 US states and the District of Columbia.

Methodology

For research question 1, I am trying to predict what variables influence the list price of the vehicle. To achieve this goal, it will be beneficial to segment the data and create linear regression and random forest algorithms.

For the process of segmentation, there are a variety of options for segmentation as the dataset includes multiple categorical variables. For example, I could segment the data by manufacturer, size, odometer, condition, fuel type, etc. This analysis will be using segmentation by either state or manufacturer. Many of the other segmentation options listed previously have a

intuitive result. For example, segmenting by condition, I would expect the better condition cars to have a higher list price. A similar situation presents itself by segmenting by year, odometer, and title status. However, this does not mean that these variables should not be analyzed with respect to price in this or future projects.

For research question 2, I am trying to determine whether manual transmission cars have a higher list price on average compared to cars with an automatic transmission. There are 2 methods that I will use to accomplish this. I will first compare prices in my EDA and then I will create a segmented linear regression model by transmission to determine if there is a statistically significant difference between the prices.

Computational Methods and Outputs

As mentioned above, I will use both a linear regression model and a random forest model for research question 1 and compare which model is better at predicting the list price. These models will be segmented by either state or manufacturer. For the linear regression model, I will use a K-fold cross-validation technique to prevent overfitting. Since I am using a random forest as the second model, there is no need for cross-validation as that is done internally. For research question 2, the Beta value for *transmission* in my linear regression model will determine which transmission type has a higher value.

The outputs for research question 1 are prediction models for the list price of used cars on Craigslist as well as the accuracy of the models. The output for research question 2 is the value of how much more or less valuable a manual transmission car is than an automatic transmission car.

Output Summaries

RQ1: Which attributes influence the list price of a used car, and can the price be predicted?

The goal of this analysis is to identify variables that affect the list price of a used car on Craigslist. This will include a linear regression model as well as a random forest model along with the prediction accuracy of each model. With the linear regression analysis, a table will be included with the independent variables, the beta values, and the level of significance for each variable as it relates to list price of each vehicle. Each of the quantitative variables will be included with scatterplots directly comparing them to list price. An example of the linear regression output would be the following:

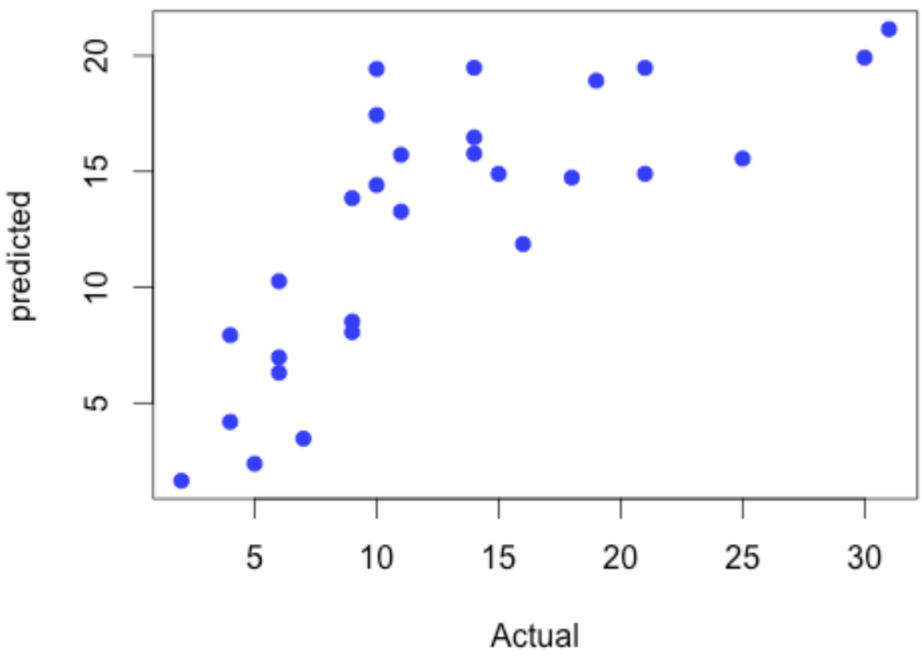
OLS Regression Results

Dep. Variable:	Monthly Nominal (USD)	R-squared (uncentered):	0.925
Model:	OLS	Adj. R-squared (uncentered):	0.917
Method:	Least Squares	F-statistic:	120.0
Date:	Tue, 09 Mar 2021	Prob (F-statistic):	5.01e-50
Time:	17:00:35	Log-Likelihood:	-724.27
No. Observations:	107	AIC:	1469.
Df Residuals:	97	BIC:	1495.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Workweek (hours)	-5.7263	4.335	-1.321	0.190	-14.329	2.877
GDP per capita	0.0148	0.002	5.912	0.000	0.010	0.020
Cost of Living Index	8.0372	2.331	3.448	0.001	3.410	12.664
Stability	-3.1099	2.132	-1.459	0.148	-7.341	1.122
Rights	5.6809	2.161	2.629	0.010	1.392	9.970
Health	1.1002	1.485	0.741	0.460	-1.846	4.047
Safety	-0.0602	1.486	-0.041	0.968	-3.009	2.888
Climate	-1.2375	1.328	-0.932	0.354	-3.874	1.399
Costs	-1.9052	2.078	-0.917	0.361	-6.029	2.218
Popularity	4.3583	1.733	2.515	0.014	0.919	7.798

Omnibus:	14.866	Durbin-Watson:	2.162
Prob(Omnibus):	0.001	Jarque-Bera (JB):	40.070
Skew:	0.372	Prob(JB):	1.99e-09
Kurtosis:	5.904	Cond. No.	5.03e+03

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 5.03e+03. This might indicate that there are
strong multicollinearity or other numerical problems.



RQ2: Are cars with manual transmissions worth more than those with automatic transmissions?

The analysis will identify whether there is a significant difference in price between automatic and manual transmission cars. This will include the beta value for the transmission variable. I will also include statistical summaries of list price for both automatic and manual transmission vehicles. The results will be a section from the first picture as shown above under RQ1.

Campaign Implementation

E-commerce has grown at an exponential rate since the start of the 21st century. Even though the majority of sales of stores across the world occur on-site, the percentage of online growth continues to grow every year (Zippia, n.d). With high inflation prices, more Americans are currently buying secondhand goods (Yahoo, n.d). While Craigslist is a major competitor in used goods, they are falling behind in a few areas to their competitors.

Creating predictive models for research question 1: “Which attributes influence the list price of a used car, and can the price be predicted?” allows Craigslist to engineer a classification system for the value of the car. For example, when a user goes on Carvana to buy a car, every listing has a meter that shows whether the price for the car is a poor, fair, or good value given other attributes. Craigslist could use this feature to enhance all of their listings, not just used car sales.

Addressing research question 2: “Are cars with manual transmissions worth more than those with automatic transmissions?” allows Craigslist to give greater confidence to consumers with greater knowledge of the item they are buying. A feature could be added to compare the same vehicle with models with different transmissions to give customers an expected value for each type of vehicle.

Review of the Literature

Machine learning and predictive analytics continue to grow as a useful tool for companies, E-commerce is an industry which rely heavily on these techniques for their success. This project will use these techniques in the form of regression and random forest algorithms. In this part of the project, I will be researching the relationship between these techniques and used car sales and Craigslist.

Used car sales date as far back as 1898 and is currently the largest retail sector in America. However, C.J Moore, a writer for Automotive News, stated that "Used-vehicle sales in 2022 tumbled to their lowest numbers in nearly a decade, and volume is poised to fall further this year if volatile economic conditions and consumers' resulting affordability concerns continue to hamper the market". Moore goes on to state that used car sales have not been this low since 2013. This presents a problem as it seems as though looking into predictive analytics for used cars might not be a productive venture right now.

With this being said, used car sales will never become obsolete as people will always need cars, at least in the foreseeable future. In fact, as used car sales decline, it becomes more important to use predictive analytics to try to determine what will happen to used car sales in the coming months and year. In 2020, C.K Puteri and L.N Safitiri, students at the Institut Teknologi Sepuluh Nopember Surabaya, created a linear regression model for used car sales in Indonesia using year, mileage, color, transmission, and city. Their model was able to predict used car sales with an accuracy of 74.6%. This analysis shows that there is evidence that predictive analytics can be used effectively to predict used car sales.

Craigslist has been a leader in E-commerce since 2000 and is still widely used for people to buy and sell items of all kinds. Predictive analytics has been used to study Craigslist before this project. Geoff Boeing, a writer for Economy and Space, conducted a study of whether housing and rental listings were segregated by social class. He found that Craigslist concentrated a majority of their listings and that white, wealthier, and better-educated communities overrepresented the listings. This study took more of an ethical approach to their project whereas my project will focus on predicting the price of cars. However, in an extension of my project, I can do a similar analysis as my dataset includes longitude and latitude values for each listing.

When it comes to predictive analytics with cars on Craigslist, I am not the first person to try to predict the list price. In 2021 Ronald Wahone, a data scientist at Asurion, conducted a data analysis of used car data on Craigslist to try to find the best deal for himself. While he did do an extensive amount of EDA, he did not do any predictive analytics to try to predict price. Instead, he used bar charts analyzing paint color, histograms showing year distribution by city, and a correlation matrix of all the variables to try to gain insight from the data. Mina Omobonike, a machine learning and AI engineer for Medium, went to the next level and created 8 machine learning models for predicting the price of a used car on Craigslist. All of her models had accuracies of over 95%. My goal with this project is for my models to be just as effective at predicting the price of used car sales on Craigslist.

Exploratory Data Analysis

The data I am using for this project was from kaggle.com and was originally web scraped from craigslist.com. This dataset consists of 426,880 observations and 26 variables of structured used car data from April 4th, 2021 – May 4th, 2021 in the United States. My first step was choosing the necessary variables, which are as follows:

region: the name of the county from which the listing originates from

price: the listed price of the vehicle which ranges in value from 0 – 3,736,928,711 (this is unusually high and will be addressed later on in the EDA)

year: the year the car was manufactured which ranges from 1990 - 2022

manufacturer: the name of the vehicle manufacturer, which there are 43 unique values

model: the name of the model of the vehicle, for which there are 29,668 unique values

condition: the quality that the listed car is in; either 'good', 'excellent', 'like new', 'fair', 'new', or 'salvage'

cylinders: the number of cylinders (chamber where fuel is combusted and creates power) that a vehicle has; either 3, 4, 5, 6, 8, 10, 12, or other

fuel: the type of fuel that the vehicle runs on; either 'gas', 'diesel', 'hybrid', 'electric', or 'other'

odometer: the number of miles that are currently on the vehicle; ranging from 0 – 10,000,000 (this unusually high value will also be addressed later in the EDA)

title_status: the title that the vehicle is registered as; either 'clean', 'rebuilt', 'salvage', 'lien', 'missing', and 'parts only'

transmission: the type of transmission that the vehicle has; either ‘automatic’, ‘manual’ or ‘other’

drive: what part of the car is the power sent to; either ‘4wd’ (4 wheel drive), ‘fwd’ (front wheel drive), or ‘rwd’ (rear wheel drive)

size: the size that the vehicle is classified as; either ‘full-size’, ‘mid-size’, ‘compact’, or ‘sub-compact’

type: the type of car that the vehicle is classified as; either ‘sedan’, ‘SUV’, ‘pickup’, ‘truck’, ‘coupe’, ‘hatchback’, ‘wagon’, ‘van’, ‘convertible’, ‘mini-van’, ‘offroad’, ‘bus’, or ‘other’

paint_color: the color that the car is painted; either ‘white’, ‘black’, ‘silver’, ‘blue’, ‘red’, ‘grey’, ‘green’, ‘custom’, ‘brown’, ‘yellow’, ‘orange’, and ‘purple’

state: the state where the listing originates; this includes values from all 50 United States

posting_date: the timestamp when the listing went live

My first step was to handle missing or N/A values. I looked to see how many N/A values were present for each variable and I found that the size, condition, cylinders, drive, and paint_color variables had over 100,000 N/A values which would heavily impact my analysis. Therefore, I decided to drop these 5 variables from my dataset, resulting in 12 variables left to work with. From there, I dropped all observations which had N/A values in at least 1 column. This dropped my working dataset down to 306,976 observations.

My next step was to transform and engineer the variables that I desired. First, I changed the year and odometer variables into an integer value as they were listed as floats in my dataset.

I then, created a variable called ‘age’ which took the current year and subtracted the year column to get an easier understanding of how old each vehicle is.

My next step was to start the data exploration process. I began by looking at the distribution of the ages of the cars as shown in Appendix A. I noticed that the distribution was extremely skewed to the right with outliers having an age of 50 – 120 years old. I filtered my data to only show cars that were manufactured before 1990. This revealed 5,371 cars fitting this description. These values would negatively impact my models as many cars manufactured before 1990 are considered “classic” cars and are therefore worth more than they would have originally sold for. Since this was a small percentage of my working dataset, I chose to drop these values which brought the number of observations in my working dataset down to 301,605. This resulted in a more appropriate distribution as shown in Appendix B.

A similar situation arose when looking at the distribution of the odometer variable. The initial distribution resulted in Appendix C. Again, there appeared to be extreme outliers in the distribution, so I looked at all the observations where the odometer was over 300,000 miles. This resulted in 1,032 cars which were either a mistake or very unlikely to have reached such a large mileage. Either way, this would negatively affect my models later and I decided to drop these values which resulted in a more appropriate distribution as shown in Appendix D. I was then left with 300,573 observations in my working dataset.

My next step was to look at the correlation between a vehicle’s odometer and the price, since the list price is the variable I’m interested in. An initial scatterplot, Appendix E, revealed that there were extreme list prices that needed a closer analysis. I decided to filter my data to view all the cars that had a list price greater than $100,000, which resulted in 247 cars. I then dropped these values as they would negatively impact my models as well. This resulted in

300,326 values in my working dataset. I created a new scatterplot, Appendix F, for the correlation of vehicle odometer vs the list price and there appears to be a very weak negative correlation, which is to be expected as cars that have been used more are typically worth less. This will be numerically verified later with a correlation coefficient in the analysis portion of this project.

I also noticed many observations where the price variable was 0, so I decided to look at the distribution of price as shown in Appendix G. This revealed a large number of values that had a value of 0 so I filtered my dataset to look at cars whose list price was less than $1,000. This resulted in 29,904 vehicles that fit this description and I chose to drop these values so it would not negatively impact my models. I created an updated distribution of list price as shown in Appendix H and I now have 270,442 values in my working dataset.

I then looked at the median list price by manufacturer as shown by Appendix I. As expected, the manufacturers with the top 5 median list prices were Ferrari ($94,850), Aston-Martin ($42,747.5), Tesla ($37,990), Ram ($31,990), and Porsche ($31,990) which are all high-end car manufacturers with the exception of Ram. The error bars on the graph reveal that Ferraris and Aston-Martins have an extreme standard deviation. After further research, this isn't surprising as most luxury cars tend to depreciate at a much higher rate than common vehicle manufacturers.

Next, I looked at the list price by fuel type as shown by Appendix J. Vehicles with diesel fuel types had the highest median list price at $34,990, followed by electric, gas, and hybrid vehicles. This isn't surprising as most vehicles that run on diesel fuel are large trucks and commercial vehicles which tend to cost more than regular cars due to their sheer size. The median price for electric vehicles was only $6,806 less than diesel cars. This is expected as

electric vehicles are still an emerging industry and those type of cars are more expensive than gas or hybrid vehicles.

My next step was to look at the list price by title status as shown by Appendix K. Cars with a lien title (someone borrowed money from a lender to make car payments) had a higher median list price at $18,000 compared to cars with a clean title at $17,990. One possibility for this could be that more people tend to take out loans for more expensive cars than for affordable cars. As expected, median list price of lien and clean titles were followed by rebuilt ($11,000), salvage ($8,000), missing ($2,850), and parts only ($2,800).

Next, I looked at the list prices by transmission type as shown in Appendix L. The distributions for manual vs automatic transmission cars look almost identical so I need to look closer at the median values. The median value for manual transmission cars is $10,995 compared to $13,999 for automatic cars. Just based on that information alone, it appears that automatic transmission cars are worth more than manual transmission cars, but this will be investigated further when I make my models.

The next graph was the median list price by vehicle type as shown in Appendix M. The 2 types of vehicles that had the highest median list price were pickups ($28,990) and trucks ($26,995) which is expected as these are larger vehicles.

The final graph was the median list price by state as shown in Appendix N. The 4 states that had the highest median list price were West Virginia ($27,990), Washington ($25,999), Montana ($23,732), and Alaska ($22,999). I had hypothesized that this was because more people had pickups and trucks as these states have a lot of rural land where those kinds of vehicles would be used. However, when I looked at the number of vehicle types by state, SUV

was the most popular type of vehicle in Washington, Alaska, and Montana. However, trucks and pickups were still in the top 4 most popular vehicles in each of these states so that seemed to make sense. I also wanted to look at the most popular cars in the states with the lowest median list prices. I looked at New Jersey and Iowa and the trend seems to be that the more sedans that are sold in a state, the lower the median list price is.

Based on this information, my first conclusion that can be drawn is that automatic transmission cars seem to be more valuable than manual cars. I was surprised by the initial exploration of the median list price by state as I had not predicted any of the top 4 states to be West Virginia, Washington, Alaska, or Montana. As I have found this to be the most interesting result in my EDA, I will segment my predictive models by state to see if I can provide more insight into the discrepancies between states.

Methodology

RQ1 – Which attributes influence the list price of a used car, and can the price be predicted?

Regression

My main goal with this research question is to use attributes of used car data from Craigslist, such as odometer, manufacturer, and state, to predict the list price. When trying to predict a numerical value, a regression algorithm is the best approach to accomplishing this task. Regression models take selected attributes from a dataset, splits the data into a test set and a training set, trains the model from the training data, and then uses that data to try to predict the list price for the test dataset.

Some regression techniques are susceptible to outliers, which is why I have already cleaned my numerical variables in my EDA. For this project, I will be doing an 80-20 train/test split of my data. The majority of my independent variables are categorical, so they will be transformed into dummy variables. For example, with transmission there are 3 values; 'automatic', 'manual', and 'other'. I will have $k - 1$ dummy variables with 1 variable being the base, for instance automatic = base, manual = 1, and other = 2. This process will be done for all categorical variables. I will also segment my data based on regions of the United States (Pacific Northwest, Midwest, New England, Southeast, etc.).

To prevent overfitting of my models, I will implement the k-fold cross-validation technique. This means that my dataset will first be randomly sliced into 'k' number of equal parts. Next, my model will run 'k' times and takes one of the 'k' sections as the test data and all other parts as the train data. Then, my model will be fit on the training set and be evaluated on

the test data. The last step is to return the evaluation score. For each model, I will also produce the root mean squared error as well as the accuracy of the model on the test data.

Modeling Techniques:

- Linear/Polynomial Regression
    - For linear regression, the relationship between features and the label is assumed to be linear and the goal is to minimize the sum of the squared residuals between the prediction and the actual value. Polynomial regression is similar, but the features are transformed to better represent the relationship to the label. For example, a quadratic or exponential relationship.
- Random Forests
    - Random forests combine several individual decision trees, which allow for learning non-linear relationships by splitting the dataset into different branches. As the number of branches grow, the better the predictability as it allows for better random selection and a better average of the predictions is produced.
- K-Nearest Neighbors
    - The KNN algorithm groups the closest k data points together, averages them, and then plots the averages as the predictive model. Similar to linear and polynomial regression, the quality of the model is based on the distances between the predicted values and the actual values. However, I might not use this algorithm for my project as it is computationally expensive and tends to perform poorly on large datasets with multiple independent variables.

RQ2 – Are cars with manual transmissions worth more than those with automatic transmissions?

From my EDA, I have done an initial exploration and so far, this theory seems to be wrong. I will continue exploring this theory by segmenting my data by transmission type and see what my models look like. I will be looking for the beta coefficient for the automatic transmission model to be greater than the coefficient for the manual transmission model. I will also look at the model as a whole and see what the expected price per transmission type is assuming all other independent variables are held constant.

Data Visualizations

Multiple Regression

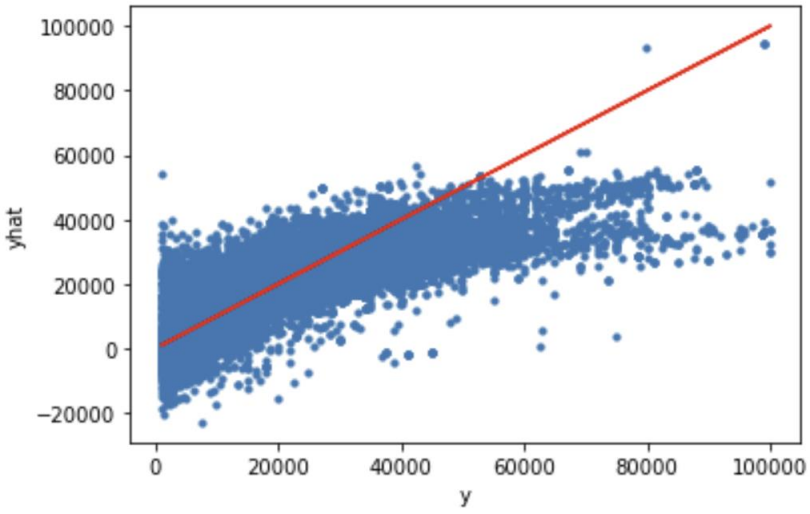
I will first assess the predictability of list price in the west region. The following are the results of the model: *not all variables are shown because of the large quantity of independent variables*

OLS Regression Results

Dep. Variable:	price	R-squared:	0.719
Model:	OLS	Adj. R-squared:	0.719
Method:	Least Squares	F-statistic:	3321.
Date:	Fri, 14 Apr 2023	Prob (F-statistic):	0.00
Time:	15:39:42	Log-Likelihood:	-8.6560e+05
No. Observations:	83051	AIC:	1.731e+06
Df Residuals:	82986	BIC:	1.732e+06
Df Model:	64		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
odometer	-0.0838	0.001	-116.855	0.000	-0.085	-0.082
age	-1000.5945	7.308	-136.916	0.000	-1014.918	-986.271
manufacturer_acura	853.2849	329.915	2.586	0.010	206.653	1499.917
manufacturer_alfa-romeo	5129.9155	643.437	7.973	0.000	3868.783	6391.048
manufacturer_aston-martin	-6592.0295	3251.377	-2.027	0.043	-1.3e+04	-219.355
manufacturer_audi	2864.1844	302.656	9.464	0.000	2270.981	3457.387
manufacturer_bmw	1409.0844	272.828	5.165	0.000	874.345	1943.824
manufacturer_buick	-3090.9006	361.494	-8.550	0.000	-3799.427	-2382.375
manufacturer_cadillac	5648.3053	328.364	17.201	0.000	5004.714	6291.897
manufacturer_chevrolet	-585.9793	251.453	-2.330	0.020	-1078.826	-93.133
manufacturer_chrysler	-3228.2459	348.755	-9.256	0.000	-3911.804	-2544.688
manufacturer_dodge	-1877.7548	287.368	-6.534	0.000	-2440.994	-1314.516
manufacturer_ferrari	3.573e+04	3013.347	11.858	0.000	2.98e+04	4.16e+04
manufacturer_fiat	-9462.7249	579.910	-16.318	0.000	-1.06e+04	-8326.105
manufacturer_ford	-855.1494	248.493	-3.441	0.001	-1342.194	-368.105
manufacturer_gmc	3444.1474	278.140	12.383	0.000	2898.995	3989.300
manufacturer_harley-davidson	-5582.9504	1520.010	-3.673	0.000	-8562.159	-2603.741
manufacturer_honda	-1547.5153	263.805	-5.866	0.000	-2064.571	-1030.460
manufacturer_hyundai	-5875.8085	289.536	-20.294	0.000	-6443.297	-5308.320
manufacturer_infiniti	1412.3954	353.265	3.998	0.000	719.999	2104.792
manufacturer_jaguar	3954.0682	494.598	7.995	0.000	2984.660	4923.476
manufacturer_jeep	1453.4288	270.829	5.367	0.000	922.607	1984.251
manufacturer_kia	-6185.0125	300.280	-20.597	0.000	-6773.560	-5596.465

Based on the adjusted R^2 value, this is a good model with an accuracy above 70%. The F-test value is high and the p-values for the independent variables are below 0.05 which is a good sign. I also made a scatterplot of the actual vs predicted values of the model:



Based on the graph, it seems realistic to say that the model has an accuracy of 71.9%. It also seems that for higher priced cars, the model underestimates the value of them.

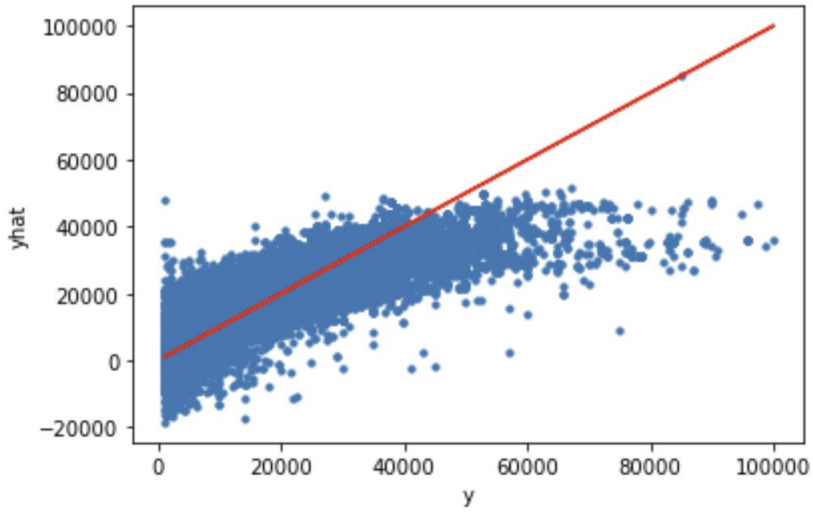
The next region I analyzed was the Midwest region. The following are the results of the model: *not all variables are shown*

OLS Regression Results

Dep. Variable:	price	R-squared:	0.717
Model:	OLS	Adj. R-squared:	0.717
Method:	Least Squares	F-statistic:	2267.
Date:	Fri, 14 Apr 2023	Prob (F-statistic):	0.00
Time:	15:53:02	Log-Likelihood:	-5.7876e+05
No. Observations:	56344	AIC:	1.158e+06
Df Residuals:	56280	BIC:	1.158e+06
Df Model:	63		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
odometer	-0.0619	0.001	-89.388	0.000	-0.063	-0.061
age	-1065.4064	7.344	-145.065	0.000	-1079.801	-1051.011
manufacturer_acura	959.2940	279.242	3.435	0.001	411.977	1506.611
manufacturer_alfa-romeo	3686.6353	570.442	6.463	0.000	2568.566	4804.705
manufacturer_audi	4636.7545	268.839	17.247	0.000	4109.828	5163.681
manufacturer_bmw	2645.9832	237.020	11.164	0.000	2181.423	3110.543
manufacturer_buick	-2691.6341	264.600	-10.172	0.000	-3210.251	-2173.017
manufacturer_cadillac	2067.4635	276.332	7.482	0.000	1525.852	2609.075
manufacturer_chevrolet	-1696.1293	191.899	-8.839	0.000	-2072.253	-1320.006
manufacturer_chrysler	-4738.4575	272.277	-17.403	0.000	-5272.122	-4204.793
manufacturer_dodge	-3814.4505	234.660	-16.255	0.000	-4274.385	-3354.516
manufacturer_ferrari	7.049e+04	4830.545	14.593	0.000	6.1e+04	8e+04
manufacturer_fiat	-9938.8681	861.131	-11.542	0.000	-1.16e+04	-8251.046
manufacturer_ford	-2815.2836	192.049	-14.659	0.000	-3191.701	-2438.866
manufacturer_gmc	-665.1731	225.905	-2.944	0.003	-1107.948	-222.398
manufacturer_harley-davidson	-6140.2133	1902.907	-3.227	0.001	-9869.922	-2410.504
manufacturer_honda	-2996.5748	220.573	-13.585	0.000	-3428.899	-2564.250
manufacturer_hyundai	-6278.3072	264.635	-23.724	0.000	-6796.993	-5759.621
manufacturer_infiniti	1026.1924	303.703	3.379	0.001	430.933	1621.452
manufacturer_jaguar	3700.2558	403.670	9.167	0.000	2909.060	4491.452

Based on the adjusted R^2 value, this is a good model with an accuracy above 70%. The F-test value is high and the p-values for the independent variables are below 0.05 which is a good sign. I also made a scatterplot of the actual vs predicted values of the model:



Once again, it seems realistic to say that the model has an accuracy of 71.7%. It also seems that for higher priced cars, the model underestimates the value of them.

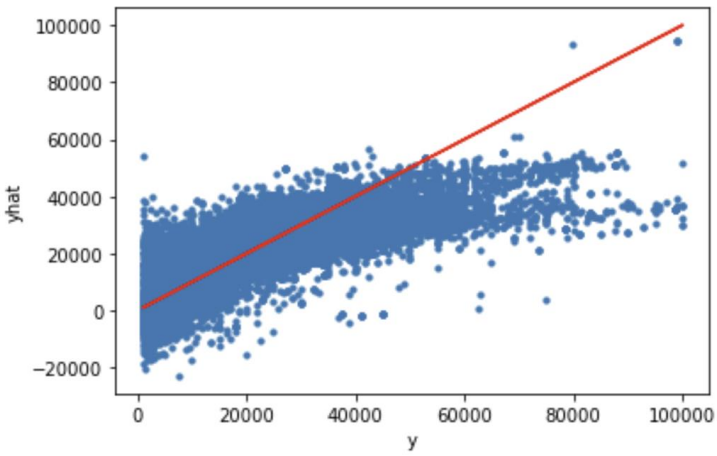
The next region I analyzed was the south region. The following are the results of the model: *not all variables are shown*

OLS Regression Results

Dep. Variable:	price	R-squared:	0.688
Model:	OLS	Adj. R-squared:	0.688
Method:	Least Squares	F-statistic:	3008.
Date:	Fri, 14 Apr 2023	Prob (F-statistic):	0.00
Time:	15:53:35	Log-Likelihood:	-8.8781e+05
No. Observations:	85843	AIC:	1.776e+06
Df Residuals:	85779	BIC:	1.776e+06
Df Model:	63		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
odometer	-0.0669	0.001	-105.955	0.000	-0.068	-0.066
age	-1012.6873	6.782	-149.320	0.000	-1025.980	-999.395
manufacturer_acura	-1076.1550	214.374	-5.020	0.000	-1496.327	-655.983
manufacturer_alfa-romeo	2548.7889	429.250	5.938	0.000	1707.462	3390.116
manufacturer_aston-martin	3.607e+04	2770.426	13.020	0.000	3.06e+04	4.15e+04
manufacturer_audi	3229.7239	210.265	15.360	0.000	2817.606	3641.842
manufacturer_bmw	-277.8733	176.250	-1.577	0.115	-623.321	67.574
manufacturer_buick	-4964.6103	254.599	-19.500	0.000	-5463.623	-4465.597
manufacturer_cadillac	1311.4975	218.378	6.006	0.000	883.479	1739.516
manufacturer_chevrolet	-1709.2497	149.054	-11.467	0.000	-2001.393	-1417.106
manufacturer_chrysler	-5293.2150	254.302	-20.815	0.000	-5791.644	-4794.786
manufacturer_dodge	-2654.5913	197.508	-13.440	0.000	-3041.705	-2267.478
manufacturer_ferrari	7.889e+04	2995.967	26.332	0.000	7.3e+04	8.48e+04
manufacturer_fiat	-1.119e+04	574.839	-19.458	0.000	-1.23e+04	-1.01e+04
manufacturer_ford	-2529.3318	146.703	-17.241	0.000	-2816.869	-2241.794
manufacturer_gmc	-199.4687	184.421	-1.082	0.279	-560.933	161.996
manufacturer_harley-davidson	-1.183e+04	2212.904	-5.344	0.000	-1.62e+04	-7488.561
manufacturer_honda	-3852.6952	169.858	-22.682	0.000	-4185.615	-3519.775
manufacturer_hyundai	-7268.7502	202.088	-35.968	0.000	-7664.840	-6872.660

Based on the adjusted R^2 value, this is an ok model with an accuracy above 60%. The F-test value is high and the p-values for the independent variables are below 0.05 which is a good sign. Some of the variables have p-values above 0.05, but when I removed them from the model, it did not affect the adjusted R^2 value so removing them is unnecessary. I also made a scatterplot of the actual vs predicted values of the model:



Once again, it seems realistic to say that the model has an accuracy of 68.8%. It also agrees with the trend that for higher priced cars, the model underestimates the value of them.

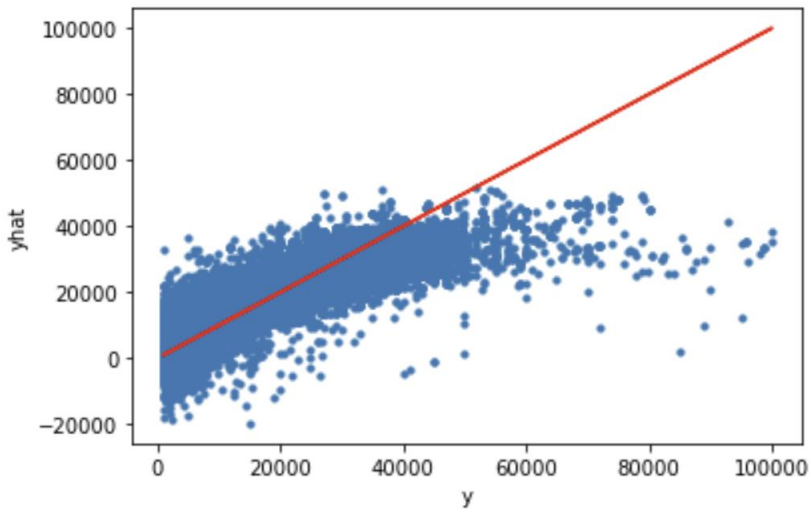
The next region I analyzed was the northeast region. The following are the results of the model: *not all variables are shown*

OLS Regression Results

Dep. Variable:	price	R-squared:	0.733
Model:	OLS	Adj. R-squared:	0.733
Method:	Least Squares	F-statistic:	1967.
Date:	Fri, 14 Apr 2023	Prob (F-statistic):	0.00
Time:	15:54:07	Log-Likelihood:	-4.6147e+05
No. Observations:	45184	AIC:	9.231e+05
Df Residuals:	45120	BIC:	9.236e+05
Df Model:	63		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
odometer	-0.0712	0.001	-88.079	0.000	-0.073	-0.070
age	-955.1973	8.235	-115.997	0.000	-971.337	-939.057
manufacturer_acura	1074.1868	302.186	3.555	0.000	481.897	1666.477
manufacturer_alfa-romeo	3498.5408	655.847	5.334	0.000	2213.069	4784.012
manufacturer_aston-martin	3.686e+04	6438.359	5.725	0.000	2.42e+04	4.95e+04
manufacturer_audi	4219.9528	271.186	15.561	0.000	3688.424	4751.481
manufacturer_bmw	3166.5903	244.240	12.965	0.000	2687.877	3645.304
manufacturer_buick	-1601.0094	333.327	-4.803	0.000	-2254.335	-947.684
manufacturer_cadillac	2949.7391	304.191	9.697	0.000	2353.519	3545.959
manufacturer_chevrolet	127.2169	219.542	0.579	0.562	-303.090	557.524
manufacturer_chrysler	-3427.5782	336.879	-10.174	0.000	-4087.868	-2767.289
manufacturer_dodge	-848.4507	275.300	-3.082	0.002	-1388.043	-308.858
manufacturer_fiat	-9223.2609	637.838	-14.460	0.000	-1.05e+04	-7973.087
manufacturer_ford	-847.0157	216.744	-3.908	0.000	-1271.838	-422.194
manufacturer_gmc	993.9705	260.864	3.810	0.000	482.673	1505.268
manufacturer_harley-davidson	-5546.0976	1872.184	-2.962	0.003	-9215.609	-1876.587
manufacturer_honda	-1719.5810	231.275	-7.435	0.000	-2172.883	-1266.279
manufacturer_hyundai	-5117.2088	263.794	-19.399	0.000	-5634.249	-4600.169
manufacturer_infiniti	1185.0192	322.298	3.677	0.000	553.309	1816.729
manufacturer_jaguar	4795.6361	445.617	10.762	0.000	3922.220	5669.052

Based on the adjusted R^2 value, this is a good model with an accuracy above 70%, the best of all the multiple regression models. The F-test value is high and the p-values for the independent variables are below 0.05 which is a good sign. Some of the variables have p-values above 0.05, but when I removed them from the model, it did not affect the adjusted R^2 value so removing them is unnecessary. I also made a scatterplot of the actual vs predicted values of the model:



Once again, it seems realistic to say that the model has an accuracy of 73.3%. It also agrees with the trend that for higher priced cars, the model underestimates the value of them.

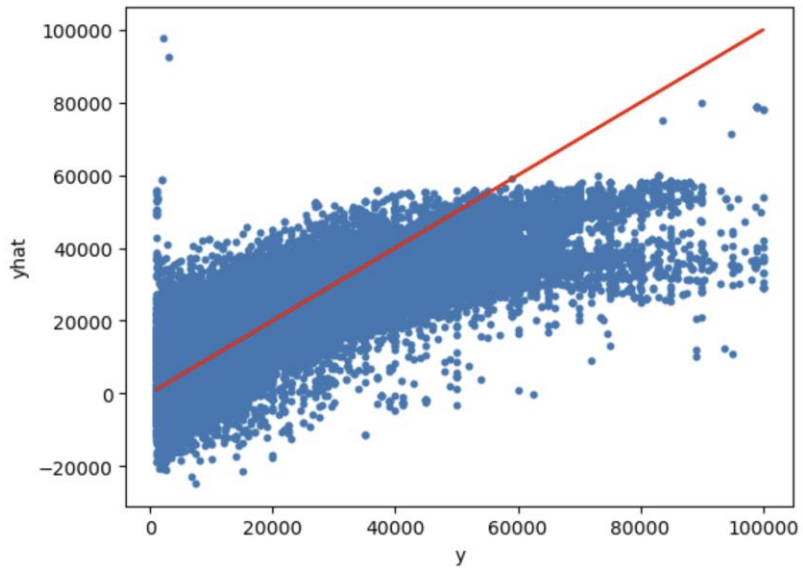
The last region I analyzed was the combined dataset of the US. The following are the results of the model: *not all variables are shown*

OLS Regression Results

Dep. Variable:	price	R-squared:	0.705
Model:	OLS	Adj. R-squared:	0.705
Method:	Least Squares	F-statistic:	9661.
Date:	Mon, 17 Apr 2023	Prob (F-statistic):	0.00
Time:	13:00:54	Log-Likelihood:	-2.7998e+06
No. Observations:	270422	AIC:	5.600e+06
Df Residuals:	270354	BIC:	5.600e+06
Df Model:	67		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
odometer	-0.0733	0.000	-203.711	0.000	-0.074	-0.073
age	-1021.3702	3.760	-271.659	0.000	-1028.739	-1014.001
manufacturer_acura	-107.7093	147.416	-0.731	0.465	-396.641	181.222
manufacturer_alfa-romeo	3189.4488	286.657	11.126	0.000	2627.609	3751.288
manufacturer_aston-martin	1.707e+04	1981.702	8.613	0.000	1.32e+04	2.1e+04
manufacturer_audi	3069.3318	139.267	22.039	0.000	2796.373	3342.291
manufacturer_bmw	830.2995	123.767	6.709	0.000	587.719	1072.880
manufacturer_buick	-3619.2439	158.170	-22.882	0.000	-3929.252	-3309.235
manufacturer_cadillac	2455.9845	147.928	16.603	0.000	2166.049	2745.920
manufacturer_chevrolet	-1544.6568	109.960	-14.047	0.000	-1760.176	-1329.137
manufacturer_chrysler	-4662.7848	158.205	-29.473	0.000	-4972.862	-4352.708
manufacturer_dodge	-2917.4182	131.616	-22.166	0.000	-3175.381	-2659.455
manufacturer_ferrari	5.627e+04	1915.905	29.370	0.000	5.25e+04	6e+04
manufacturer_fiat	-1.088e+04	325.181	-33.458	0.000	-1.15e+04	-1.02e+04
manufacturer_ford	-2333.1336	109.031	-21.399	0.000	-2546.831	-2119.436
manufacturer_gmc	480.1388	126.125	3.807	0.000	232.938	727.340
manufacturer_harley-davidson	-7197.3618	931.852	-7.724	0.000	-9023.767	-5370.956
manufacturer_honda	-3073.0557	118.734	-25.882	0.000	-3305.771	-2840.341
manufacturer_hyundai	-6817.4263	134.686	-50.617	0.000	-7081.407	-6553.446
manufacturer_infiniti	354.8728	157.289	2.256	0.024	46.591	663.155

Based on the adjusted R^2 value, this is a good model with an accuracy above 70%. The F-test value is high and the p-values for the independent variables are below 0.05 which is a good sign. Some of the variables have p-values above 0.05, but when I removed them from the model, it did not affect the adjusted R^2 value so removing them is unnecessary. I also made a scatterplot of the actual vs predicted values of the model:



Random Forest

I will first assess the predictability of list price in the west region. The advantage of using a random forest algorithm is that it not only predicts the dependent variable, but it also finds how important each feature is to the predictability of the label. The following shows the importance of each variable: *not all variables are shown because of the large quantity of independent variables*

```
Variable: age                  Importance: 0.41
Variable: odometer             Importance: 0.16
Variable: fuel_diesel          Importance: 0.15
Variable: type_sedan           Importance: 0.05
Variable: type_pickup          Importance: 0.03
Variable: type_truck           Importance: 0.03
Variable: type_hatchback       Importance: 0.02
Variable: manufacturer_chevrolet Importance: 0.01
Variable: manufacturer_ford    Importance: 0.01
Variable: manufacturer_mercedes-benz Importance: 0.01
Variable: manufacturer_porsche Importance: 0.01
Variable: manufacturer_toyota  Importance: 0.01
Variable: manufacturer_volkswagen Importance: 0.01
Variable: fuel_gas             Importance: 0.01
Variable: fuel_other           Importance: 0.01
Variable: transmission_automatic Importance: 0.01
Variable: transmission_other   Importance: 0.01
Variable: type_SUV             Importance: 0.01
```

The three variables that are the most important are age, odometer, and diesel fuel, which makes sense as age and odometer tend to be negatively correlated with price and diesel fuel vehicles tend to be more expensive as they are usually larger vehicles. The accuracy and RMSE for the model are 81.51% and 2324.81 respectively, which results in a very good model.

Doing the same process for the Midwest region:

```
Variable: age                  Importance: 0.48
Variable: odometer             Importance: 0.19
Variable: fuel_diesel          Importance: 0.03
Variable: transmission_other   Importance: 0.03
Variable: type_pickup          Importance: 0.03
Variable: type_truck           Importance: 0.03
Variable: manufacturer_chevrolet Importance: 0.02
Variable: type_sedan           Importance: 0.02
Variable: manufacturer_audi    Importance: 0.01
Variable: manufacturer_bmw     Importance: 0.01
Variable: manufacturer_ford    Importance: 0.01
Variable: manufacturer_kia     Importance: 0.01
Variable: manufacturer_mercedes-benz Importance: 0.01
Variable: title_status_clean   Importance: 0.01
Variable: transmission_automatic Importance: 0.01
Variable: type_SUV             Importance: 0.01
Variable: type_convertible     Importance: 0.01
Variable: type_hatchback       Importance: 0.01
Variable: type_other           Importance: 0.01
```

Again, the three variables that are the most important are age, odometer, and diesel fuel. The accuracy and RMSE for the model are 79.95% and 1945.66 respectively, which results in a very good model.

Doing the same process for the south region:

```
Variable: age                  Importance: 0.45
Variable: odometer             Importance: 0.18
Variable: fuel_diesel          Importance: 0.06
Variable: type_pickup          Importance: 0.03
Variable: type_sedan           Importance: 0.03
Variable: type_truck           Importance: 0.02
Variable: manufacturer_audi    Importance: 0.01
Variable: manufacturer_bmw     Importance: 0.01
Variable: manufacturer_chevrolet Importance: 0.01
Variable: manufacturer_dodge   Importance: 0.01
Variable: manufacturer_ford    Importance: 0.01
Variable: manufacturer_kia     Importance: 0.01
Variable: manufacturer_lexus   Importance: 0.01
Variable: manufacturer_mercedes-benz Importance: 0.01
Variable: fuel_gas             Importance: 0.01
Variable: fuel_other           Importance: 0.01
Variable: transmission_automatic Importance: 0.01
Variable: transmission_other   Importance: 0.01
Variable: type_SUV             Importance: 0.01
Variable: type_coupe           Importance: 0.01
Variable: type_hatchback       Importance: 0.01
Variable: type_other           Importance: 0.01
Variable: manufacturer_acura   Importance: 0.0
```

Again, the three variables that are the most important are age, odometer, and diesel fuel but diesel does not have nearly as much of an effect on price as it does in the west. The accuracy and RMSE for the model are 80.29% and 2036.69 respectively, which results in a very good model.

Doing the same process for the northeast region:

```
Variable: age                  Importance: 0.5
Variable: odometer             Importance: 0.17
Variable: fuel_diesel          Importance: 0.05
Variable: type_pickup          Importance: 0.04
Variable: transmission_other   Importance: 0.02
Variable: type_sedan           Importance: 0.02
Variable: type_truck           Importance: 0.02
Variable: manufacturer_audi    Importance: 0.01
Variable: manufacturer_bmw     Importance: 0.01
Variable: manufacturer_chevrolet Importance: 0.01
Variable: manufacturer_ford    Importance: 0.01
Variable: manufacturer_hyundai Importance: 0.01
Variable: manufacturer_lexus   Importance: 0.01
Variable: manufacturer_mercedes-benz Importance: 0.01
Variable: manufacturer_nissan  Importance: 0.01
Variable: fuel_gas             Importance: 0.01
Variable: title_status_clean   Importance: 0.01
Variable: transmission_automatic Importance: 0.01
Variable: type_SUV             Importance: 0.01
Variable: type_coupe           Importance: 0.01
Variable: type_hatchback       Importance: 0.01
Variable: type_other           Importance: 0.01
```

Again, the three variables that are the most important are age, odometer, and diesel fuel but diesel does not have nearly as much of an effect on price as it does in the west. The accuracy and RMSE for the model are 82.44% and 1796.56 respectively, which results in a very good model.

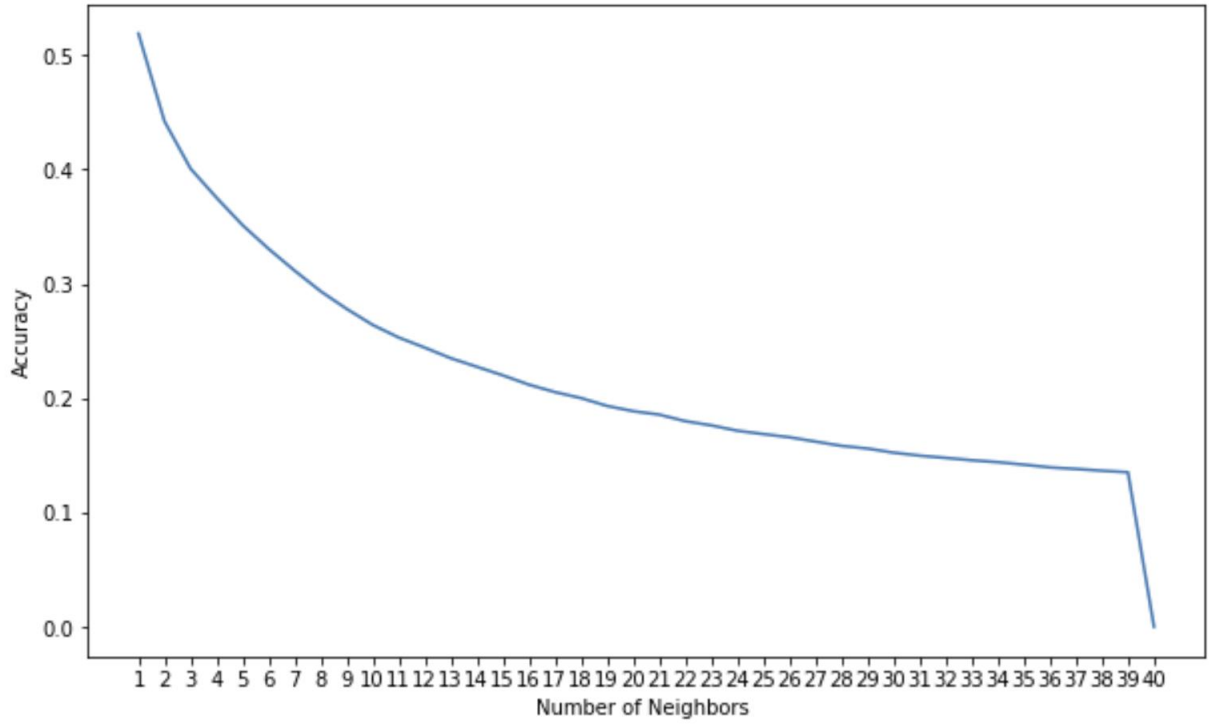
Doing the same process for the combined US:

```
Variable: age                  Importance: 0.46
Variable: odometer             Importance: 0.16
Variable: fuel_diesel          Importance: 0.08
Variable: type_sedan           Importance: 0.04
Variable: type_pickup          Importance: 0.03
Variable: type_hatchback       Importance: 0.02
Variable: type_truck           Importance: 0.02
Variable: manufacturer_audi    Importance: 0.01
Variable: manufacturer_bmw     Importance: 0.01
Variable: manufacturer_chevrolet Importance: 0.01
Variable: manufacturer_ford    Importance: 0.01
Variable: manufacturer_lexus   Importance: 0.01
Variable: manufacturer_mercedes-benz Importance: 0.01
Variable: transmission_automatic Importance: 0.01
Variable: transmission_other   Importance: 0.01
Variable: type_SUV             Importance: 0.01
Variable: region_West          Importance: 0.01
```

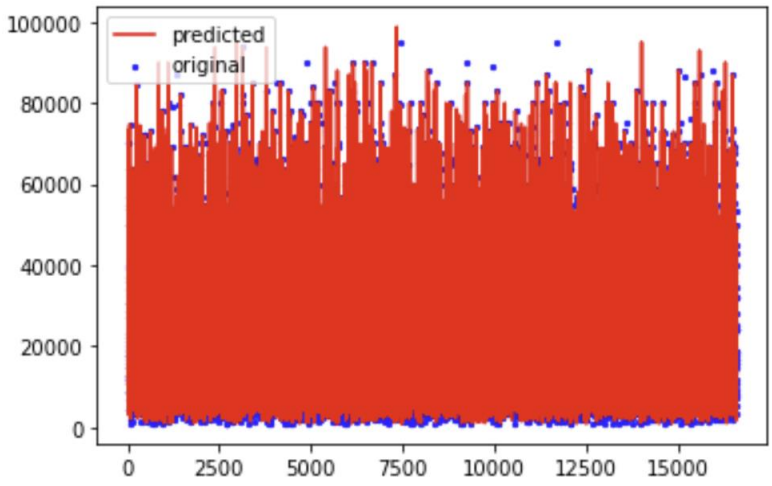
Again, the three variables that are the most important are age, odometer, and diesel fuel but diesel does not have nearly as much of an effect on price as it does in the west. The accuracy and RMSE for the model are 81.47% and 1902.81 respectively, which results in a very good model.

k-Nearest Neighbor

As the KNN algorithm requires hyperparameter tuning, I ran a KNN algorithm for values of k from 1 to 40 and then plotted the accuracies of the model with each k value for the west region:

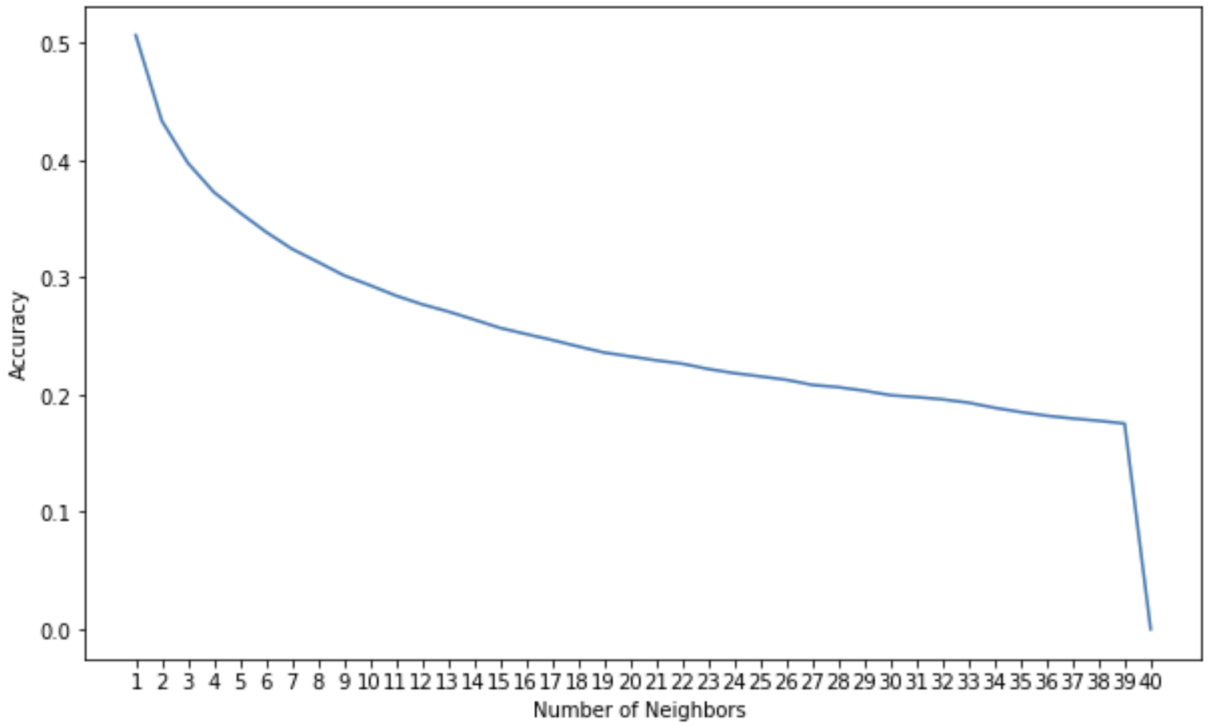


Based on this plot, the best model is when $k = 1$. I then ran the model with $k = 1$ and the resulting accuracy turned out to be 51.91%, which matches the above graph. I then plotted the model against the actual values:

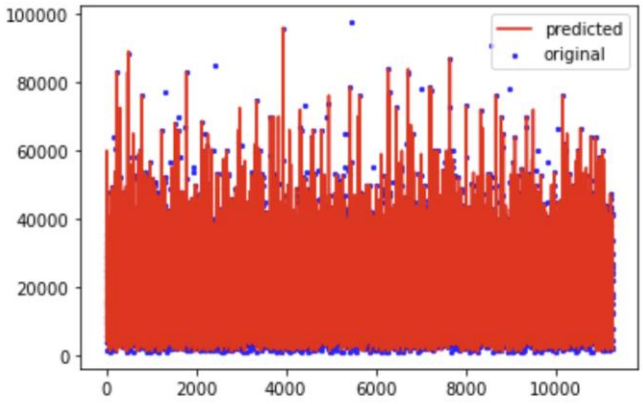


This is a very messy model as k is the smallest value that it can possibly be. Clearly, this model is overfitting the data.

Doing the same procedure for the Midwest region:

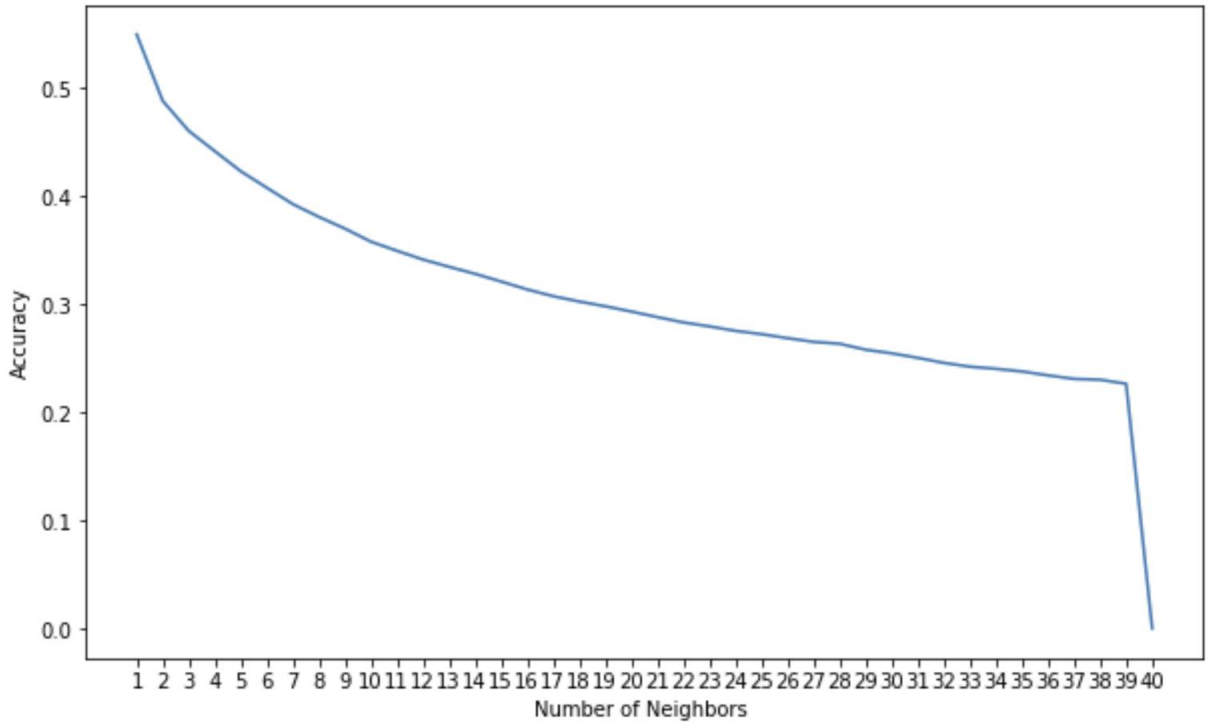


Based on this plot, the best model is when $k = 1$. I then ran the model with $k = 1$ and the resulting accuracy turned out to be 50.69%, which matches the above graph. I then plotted the model against the actual values:

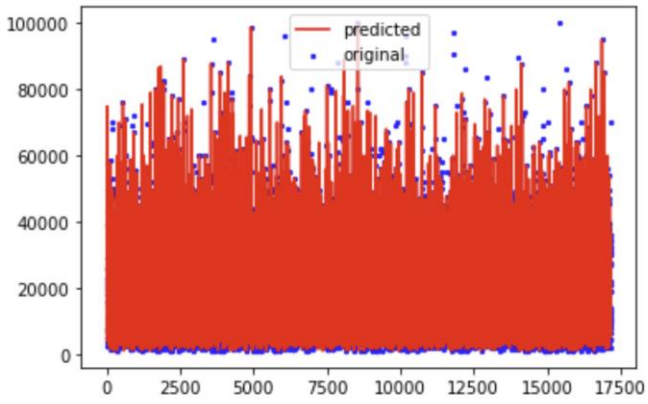


Again, this is a very messy model as k is the smallest value that it can possibly be. This model is also overfitting the data.

Doing the same procedure for the south region:

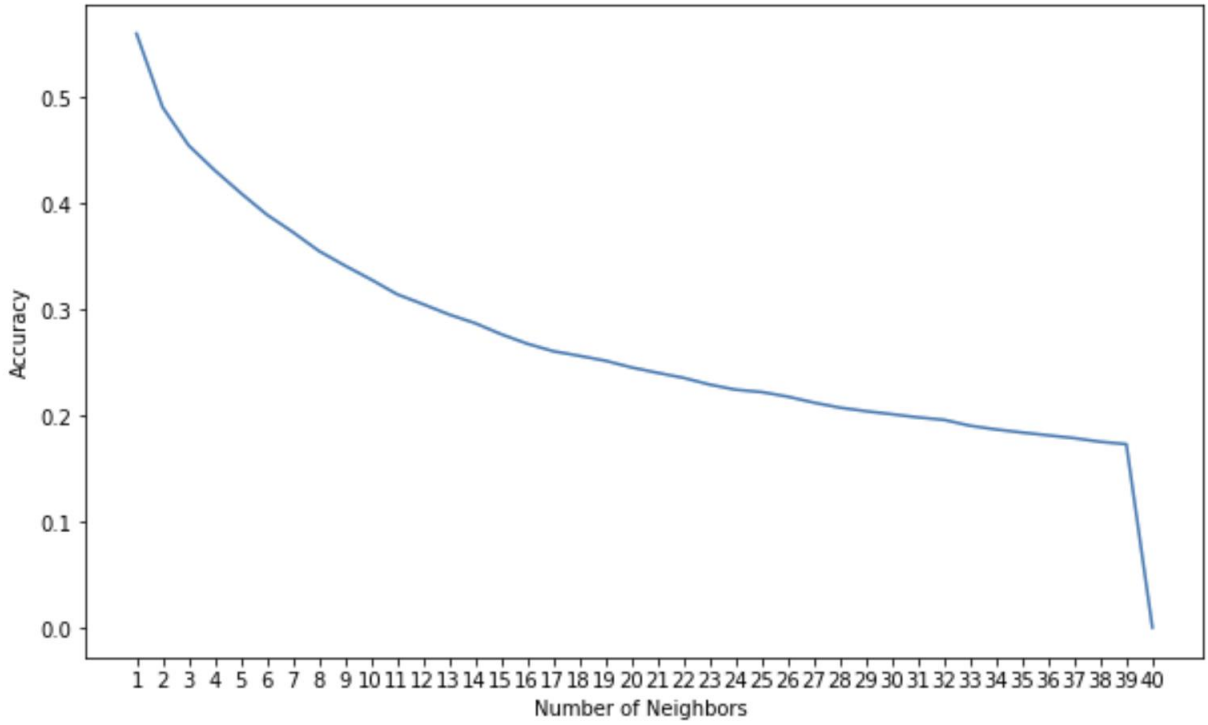


Based on this plot, the best model is when $k = 1$. I then ran the model with $k = 1$ and the resulting accuracy turned out to be 54.98%, which matches the above graph. I then plotted the model against the actual values:

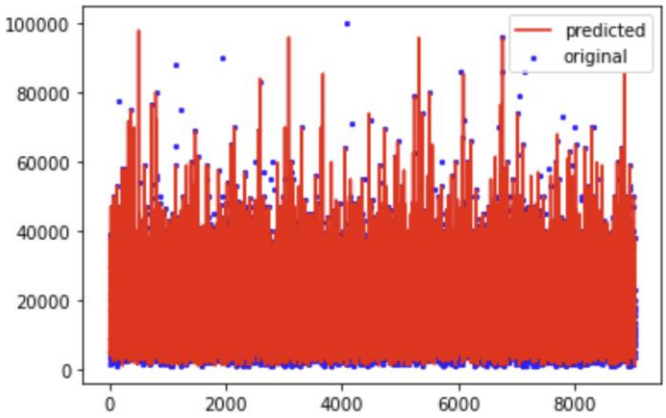


Again, this is a very messy model as k is the smallest value that it can possibly be. This model is also overfitting the data.

Doing the same procedure for the northeast region:

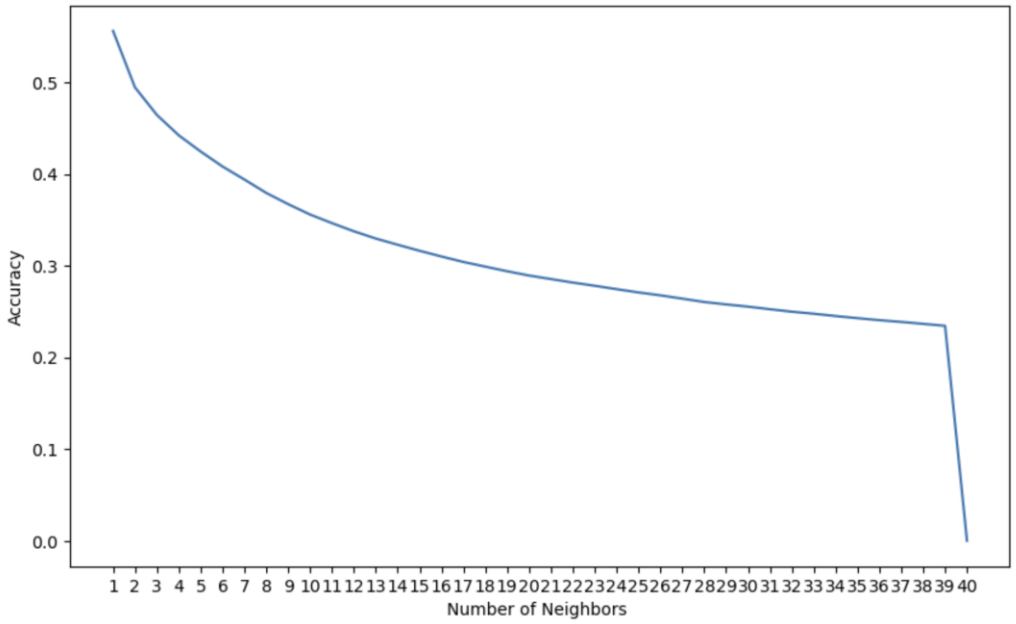


Based on this plot, the best model is when $k = 1$. I then ran the model with $k = 1$ and the resulting accuracy turned out to be 56.04%, which matches the above graph. I then plotted the model against the actual values:

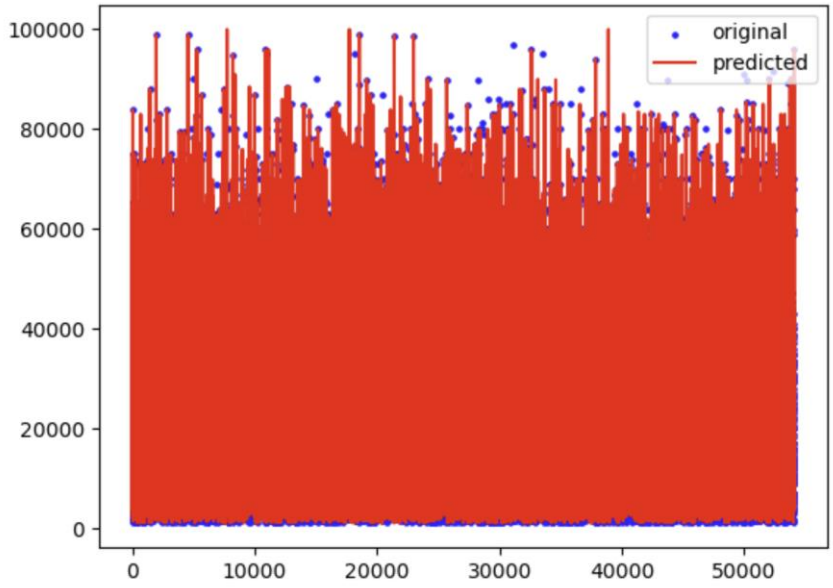


Again, this is a very messy model as k is the smallest value that it can possibly be. This model is also overfitting the data.

Doing the same procedure for the combined US:



Based on this plot, the best model is when $k = 1$. I then ran the model with $k = 1$ and the resulting accuracy turned out to be 55.6%, which matches the above graph. I then plotted the model against the actual values:



Again, this is a very messy model as k is the smallest value that it can possibly be. This model is also overfitting the data.

Analysis

Model Description

Before the modeling process can begin, I first need to prep the dataset. The first step was to remove variables that were either unnecessary or would overcomplicate the model. I removed the 'year' variable because I already feature engineered an 'age' variable which gives the same information, but in a format that is easier to understand and work with. I also removed the 'posting_date' because this project does not include any time series analysis. The last two variables removed were 'model' and 'region' because these categorical variables had more than 100 different categories, which would make the final model messy because of the large amount of encoding of the categorical values.

The next step was to segment my data for my analysis by US region. I created a new variable 'region' which represents 4 US regions; west, midwest, south, and northeast. The states included in the western region are Washington, Alaska, Oregon, California, Hawaii, Montana, Idaho, Wyoming, Colorado, Utah, Nevada, Arizona, and New Mexico. The states included in the midwest region are North Dakota, Minnesota, South Dakota, Iowa, Kansas, Missouri, Nebraska, Wisconsin, Michigan, Illinois, Indiana, and Ohio. The states included in the south region are Oklahoma, Arkansas, Texas, Louisiana, Kentucky, Tennessee, Mississippi, Alabama, West Virginia, Washington D.C., Maryland, Delaware, Virginia, North Carolina, South Carolina, Georgia, and Florida. The states included in the northeast region are New York, Pennsylvania, New Jersey, Connecticut, Rhode Island, Massachusetts, Vermont, New Hampshire, and Maine.

I produced 3 different types of models for each region (a total of 12 models) plus 2 models for automatic and manual transmissions. To prevent overfitting of the model, I used

cross-validation by splitting my dataset into training and testing datasets with a ratio of 80:20, respectively. The model will be created on the training data and the accuracy of the model will be assessed by predicting using the test data.

The independent variables used in the models are as follows:

Variable	Definition
manufacturer	The name of the manufacturer of the vehicle
fuel	The type of fuel that the vehicle runs on
odometer	The number of miles that the vehicle has on it
title_status	The title that the vehicle is registered as
transmission	The type of transmission that the vehicle has
type	The type of car that the vehicle is classified as
age	How old the vehicle is

Originally I had planned to include a neural network, Gaussian process regression, support vector regression, and RANSAC regression techniques as well, but I ran into complications with each model. My issues were that I received an error in Jupyter Notebook telling me I did not have enough memory to create the model, or my computer would crash when trying to run other models, likely due to size of the dataset.

Model Results

RQ1 - Which attributes influence the list price of a used car, and can the price be predicted?

The following table displays the results of each model:

Model Type	Region	Root Mean Squared Error (RMSE)	Accuracy
Random Forest	Northeast	1796.56	82.44%
Random Forest	West	2324.81	81.51%
Random Forest	Total US	1902.81	81.47%
Random Forest	South	2036.69	80.29%
Random Forest	Midwest	1945.66	79.95%
Multiple Regression	Northeast	6673.48	73.30%
Multiple Regression	West	8118	71.90%
Multiple Regression	Midwest	6912.21	71.70%
Multiple Regression	Total US	7611.98	70.50%
Multiple Regression	South	7527.29	68.80%
KNN	Northeast	7120.71	56.04%
KNN	Total US	7863.76	55.60%
KNN	South	8256.96	54.98%
KNN	West	9721.31	51.91%
KNN	Midwest	7904.98	50.69%

Based on the table above, the random forest algorithm performed the best with all models around 80% or above, followed by the multiple regression model and the k-nearest neighbor model. The northeast region consistently had the highest predictability rating across all the models. I decided to include the combined dataset to see if it would perform better and I ended up with mixed performances across all the models. Overall, I would say that in practice, I would use either the random forest or multiple regression models as any model that has a test accuracy of over 70% is a relatively strong model.

RQ2 - Are cars with manual transmissions worth more than those with automatic transmissions?

Some analysis was already done in the EDA which showed that the median list price for automatic transmission cars was higher than those with manual transmission cars. My final step was to look at the beta coefficient in the multiple regression analysis. The coefficient for manual transmission cars in the combined US model was 0.0001355. This means that, assuming all other independent variables are held constant, the mean difference in list price between manual and automatic transmission cars is $0.0001355. Since this number is so small, it cannot be said that there is a significant difference between the list price between manual and automatic transmission cars.

Ethical Recommendations

Machine learning and predictive analytics have become highly sought after skills used by practically every business today. Many executives base their business decision-making on the results of these models, but machine learning is far from a perfect science. After studying the predictability of used car list prices on Craigslist, there are a few ethical issues that must be addressed.

First, this dataset was collected using web scraping. Since web scraping can grab any piece of information from a website, users' personal information may be compromised. For instance, if a web scraper somehow pulled the email or names of the individuals who listed the cars, that would be a breach of privacy and should not be used in any kind of analysis (Collard, 2022). Luckily for this project, none of that information was included in this dataset.

However, something that was concerning about this dataset was that it included latitude and longitude of the listing. This is problematic because it could be potentially used to reveal the location of a user, which would be a breach of privacy. Another problem that can arise from using this information is that the model might develop a bias and predict that cars in lower-income communities are cheaper than in a middle-class neighborhood. For these reasons, I chose not to use this information in my analysis.

The goal of this project is to demonstrate the ability to effectively predict used car listing prices to create a feature on Craigslist that can let users know if the listing is a fair price or not, like other tools used by Carvana or Kelly Blue Book. If this research was not to be used, it could result in customers purchasing products at a bad value. This in turn, would result in a loss of customer confidence in Craigslist as a reliable website for buying used goods and therefore a loss

in business for Craigslist. Customers who buy used goods trust that they are getting a good deal. However, there is always some level of risk involved in these situations as most of the time the two parties do not know each other. Having a "value tool" eliminates this concern for customers not just for used cars, but for any listing on Craigslist's website.

Challenges

Throughout this project, I was faced with many challenges. One issue that came about was figuring out how to deal with the many missing or incorrect values of data. Since this dataset was web scraped, there are many opportunities for incorrect data to occur as all the information on Craigslist's website is entered by a user. Since some columns had more missing values than others, it was relatively easy to simply drop those columns as I was working with a large amount of potential independent variables.

Another issue that I faced was in the modeling section of my project. Many of my models failed to run either due to the size of my data and/or because I did not have enough memory on my laptop. This issue most likely wouldn't happen in the workplace because industry computers usually have a much higher computing power than a conventional laptop. Unfortunately, I had to skip those models for this project.

A third challenge that I had was that when I tried to update my Anaconda Navigator, my Jupyter notebook with my project failed to run models that had run fine previously. I ended up uninstalling and reinstalling Anaconda Navigator and that fixed the issue. Technical errors like these will constantly come up in the workplace, so it is good for me to have experience handling these situations.

Future Work and Recommendations

If I were to come back to this project, something that I would like to do is to research more regression techniques. Since I am still relatively new to the field of data science, there are only so many algorithms and regression techniques that I am familiar with. As my experience in the field continues to grow, my knowledge of machine learning algorithms will also continue to grow and I can approach this problem with potentially better and more optimized algorithms than the ones I chose for this project.

I would also like to include more data from other time periods. Even though my dataset only included a month of used car sales in the US, it still included a lot of observations. However, I would like to use more data, for instance using used car sales in the US throughout an entire year and even over multiple years to do a time-series analysis.

Additionally, I would like to compare Craigslist data to similar companies like Carvana, Kelly Blue Book, or Carfax to see if there is a difference in predictability across companies and try to determine what the cause of the difference would be.

References

20+ Fascinating Online Shopping Statistics [2023]: Online Shopping Vs. In-Store Shopping – Zippia. (n.d.). https://www.zippia.com/advice/online-shopping-statistics/#:~:text=As%20of%202021%2C%20In%2Dstore

Boeing. (2020). Online rental housing market representation and the digital reproduction of urban inequality. *Environment and Planning. A*, *52*(2), 449–468. https://doi.org/10.1177/0308518X19869678

Collard, M. (2022, June 8). *Price prediction for used cars - miun.diva-portal.org*. Price Prediction for Used Cars. Retrieved April 16, 2023, from https://miun.diva-portal.org/smash/get/diva2:1674070/FULLTEXT01.pdf

craigslist: phoenix jobs, apartments, for sale, services, community, and events. (n.d.). Craigslist. Retrieved March 31, 2023, from https://phoenix.craigslist.org/

Moore. (2023). Used-car sales hit near-decade low in 2022; Analysts expect numbers to fall further in 2023. *Automotive News*, *97*(7073), 3–.

Omobonike, M. (2021, May 22). *EXPLORATORY DATA ANALYSIS AND MODEL BUILDING OF CRAIGSLIST USED CAR DATA SET*. Medium. https://minaomobonike.medium.com/data-analysis-and-science-of-craigslist-used-car-dataset-cfe8b0147a51

Puteri, & Safitri, L. N. (2020). Analysis of linear regression on used car sales in Indonesia. *Journal of Physics. Conference Series*, *1469*(1), 12143–. https://doi.org/10.1088/1742-6596/1469/1/012143

RocketReach - Find email, phone, social media for 450M+ professionals. (n.d.). RocketReach. Retrieved March 31, 2023, from https://rocketreach.co/craigslist-management_b5c74243f42e0d19#:~:text=Craigslist%20employs%20432%20employees.

Top 10 Best Alternatives to Craigslist. (n.d.). Investopedia. Retrieved March 31, 2023, from https://www.investopedia.com/articles/personal-finance/091515/4-best-alternatives-craigslist.asp

Using the web to get stuff done: What is Craigslist? GCFGlobal.org. (n.d.). Retrieved March 25, 2023, from https://edu.gcfglobal.org/en/using-the-web-to-get-stuff-done/what-is-craigslist/1/

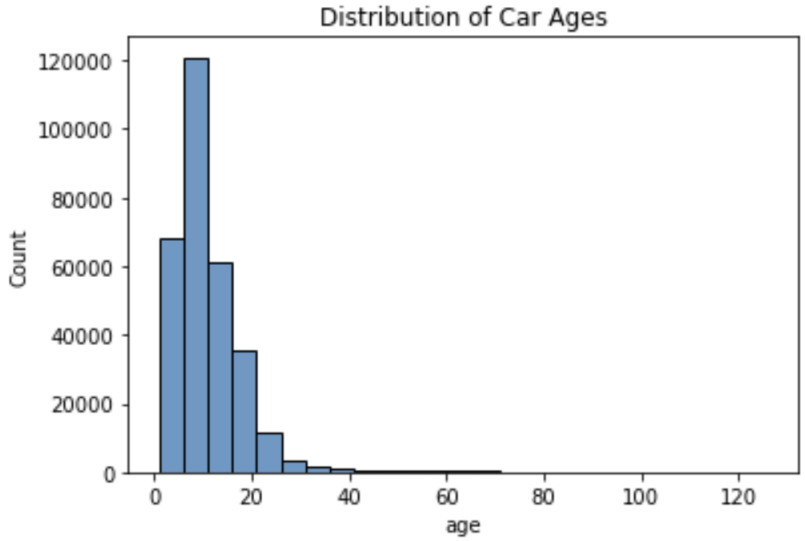
Wahome, R. (2021, October 31). *A Step By Step How To Do A Data Science Project With Craigslist Data: Part 1 — Some Background*. Medium. https://medium.com/@wahomeron/a-step-by-step-how-to-do-a-data-science-project-with-craigslist-data-part-1-some-background-215b4d1efd5f

Wikimedia Foundation. (2023, March 15). *Craigslist*. Wikipedia. Retrieved March 23, 2023, from https://en.wikipedia.org/wiki/Craigslist

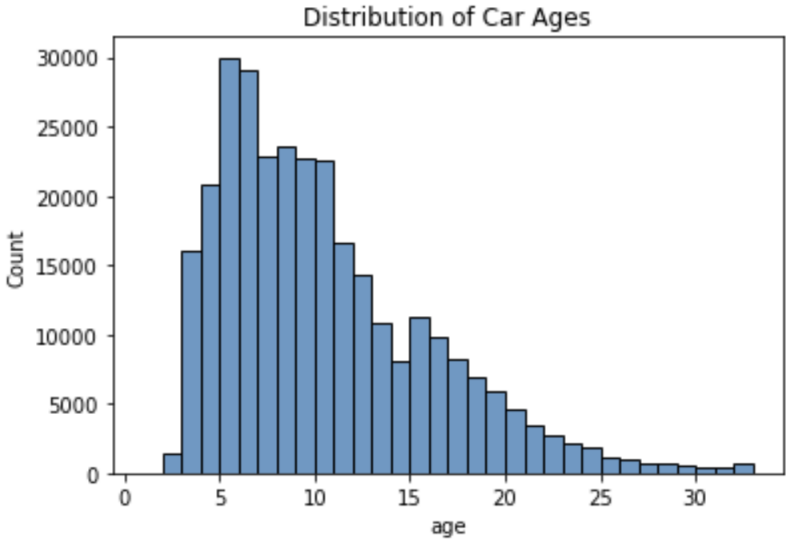
Yahoo is part of the Yahoo family of brands. (n.d.). https://www.yahoo.com/video/getting-thrifty-93-americans-now-135522261.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAEwcMJ5WYGs5Pdabb9V7tjuUv5mXqSbbNrcD6-xRNncQskHsevEkVkIUFBtuXwhj-ofYDuoAQzdQI2DGw1EPneUBr9zCxtV_9Zkc5UZS5WDQvJ_C6SgU8yJXFfy1OyHqJh--B-XHkJgObZ5t0VwtEU5gqoXI5S0DtE5G7JP7vN8i#:~:text=More%20than%20half%20(58%25),of%20the%20growth%20involved%20apparel.

Appendix

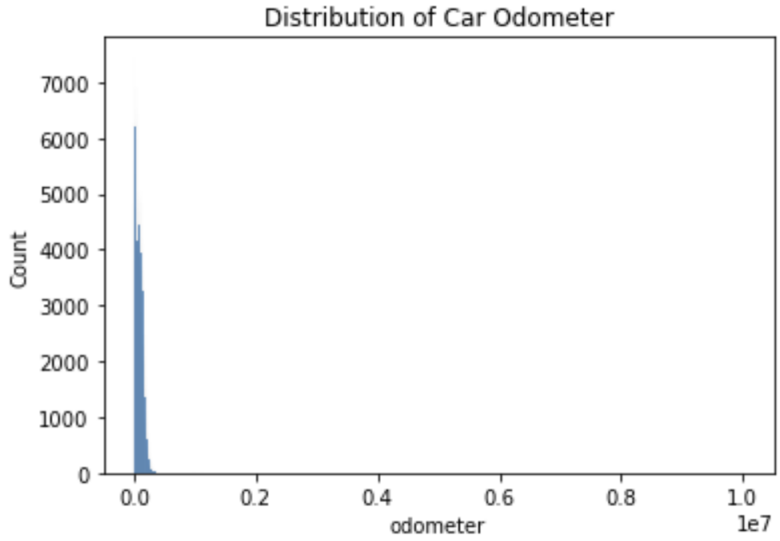
A. “Distribution of Car Ages” Mark I



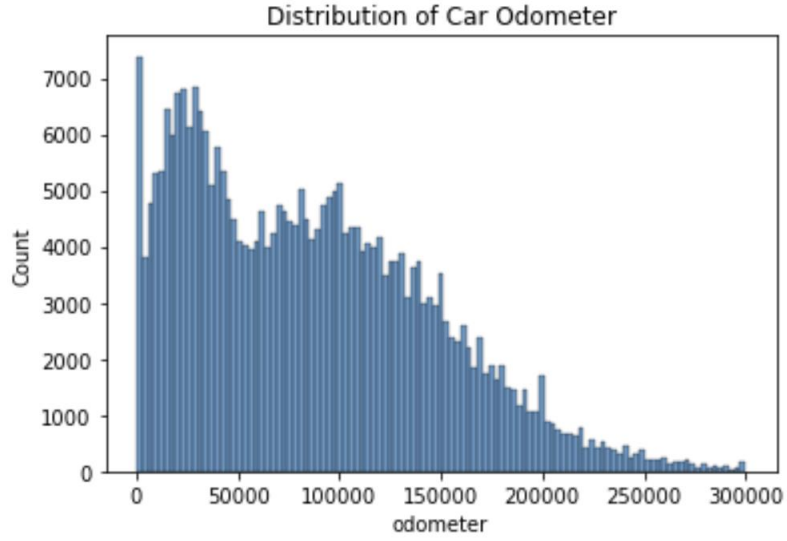
B. “Distribution of Car Ages” Mark II



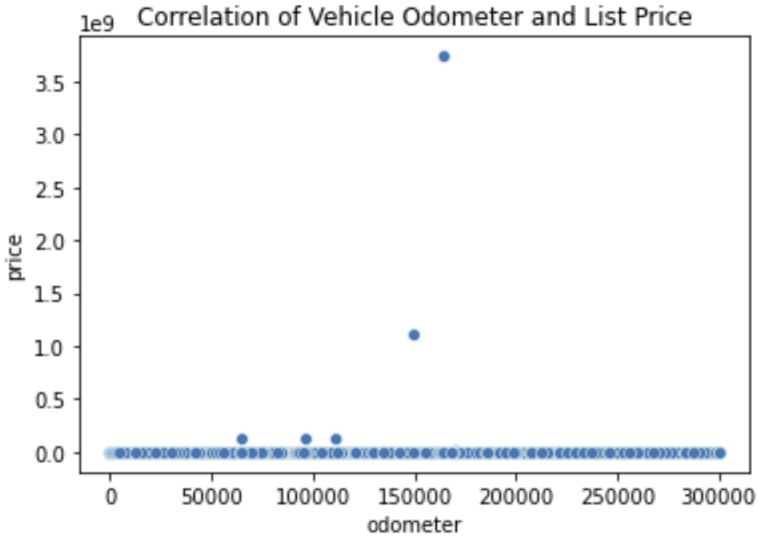
C. “Distribution of Car Odometer” Mark I



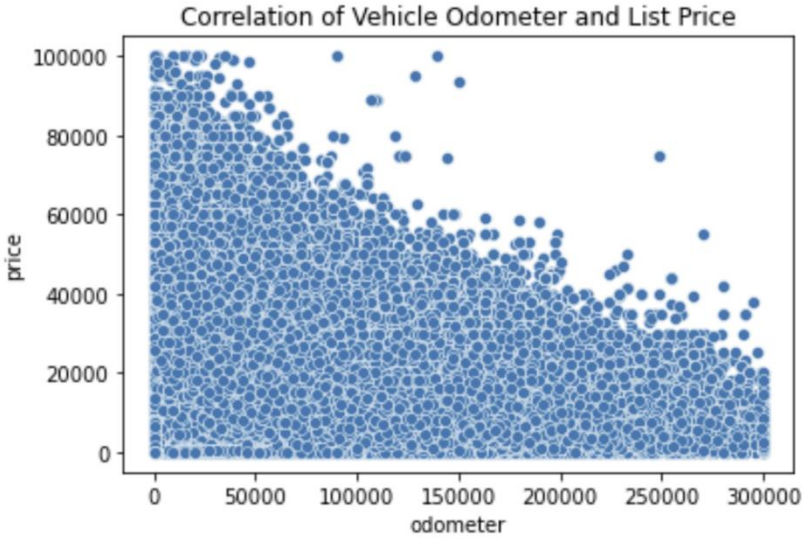
D. “Distribution of Car Odometer” Mark II



E. “Correlation of Vehicle Odometer and List Price” Mark I

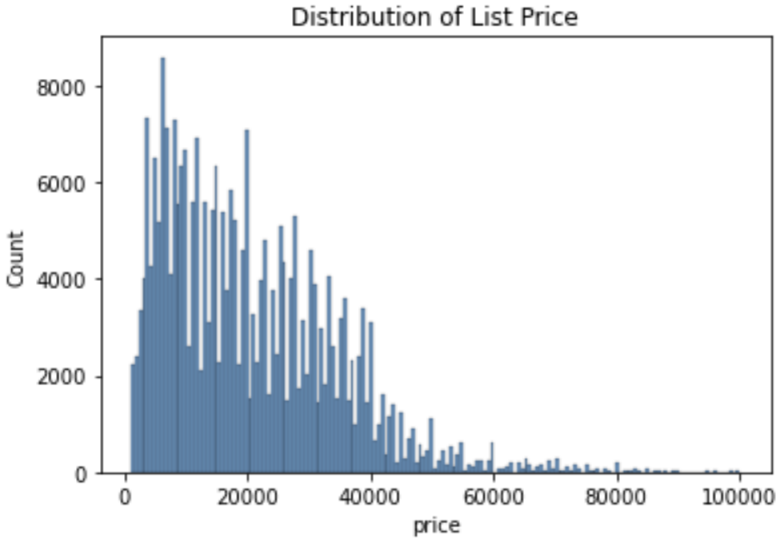


F. “Correlation of Vehicle Odometer and List Price” Mark II

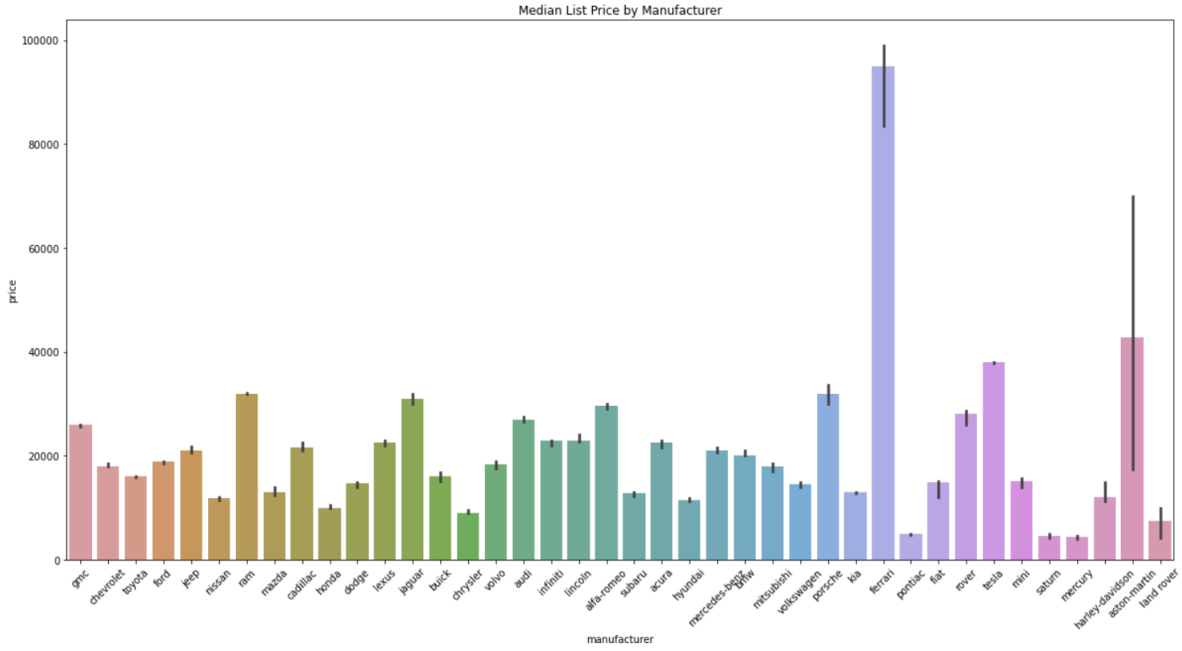


G. “Distribution of List Price” Mark I

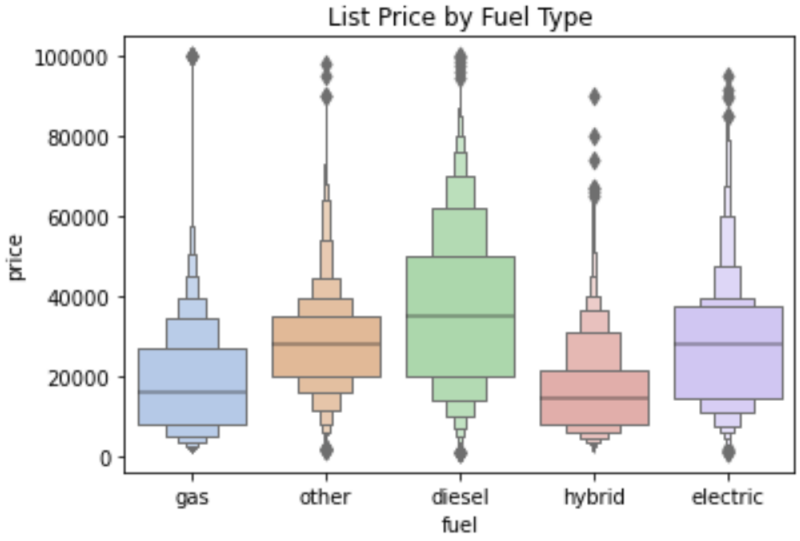
H. “Distribution of List Price" Mark II



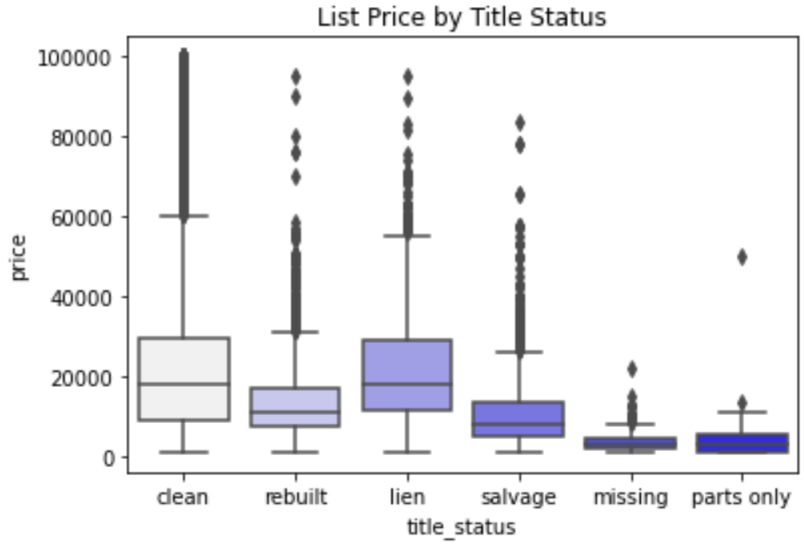
I. “Median List Price by Manufacturer”



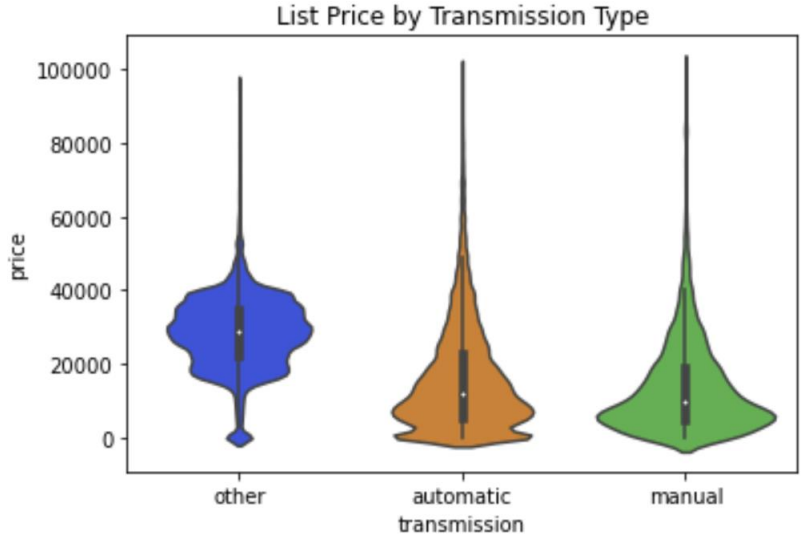
J. “List Price by Fuel Type”



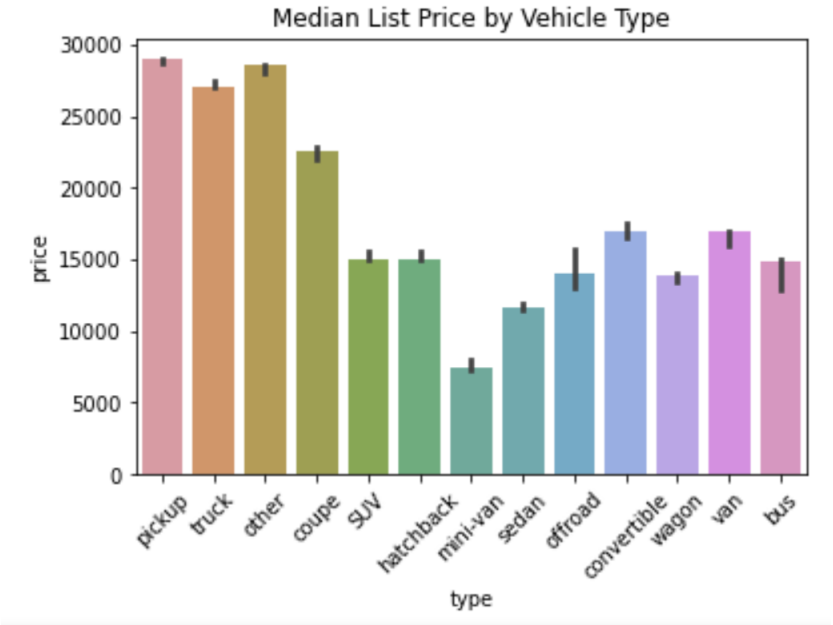
K. “List Price by Title Status”



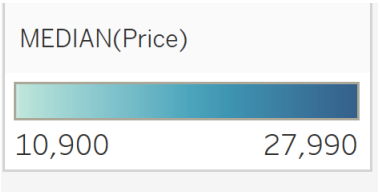
L. “List Price by Transmission Type”

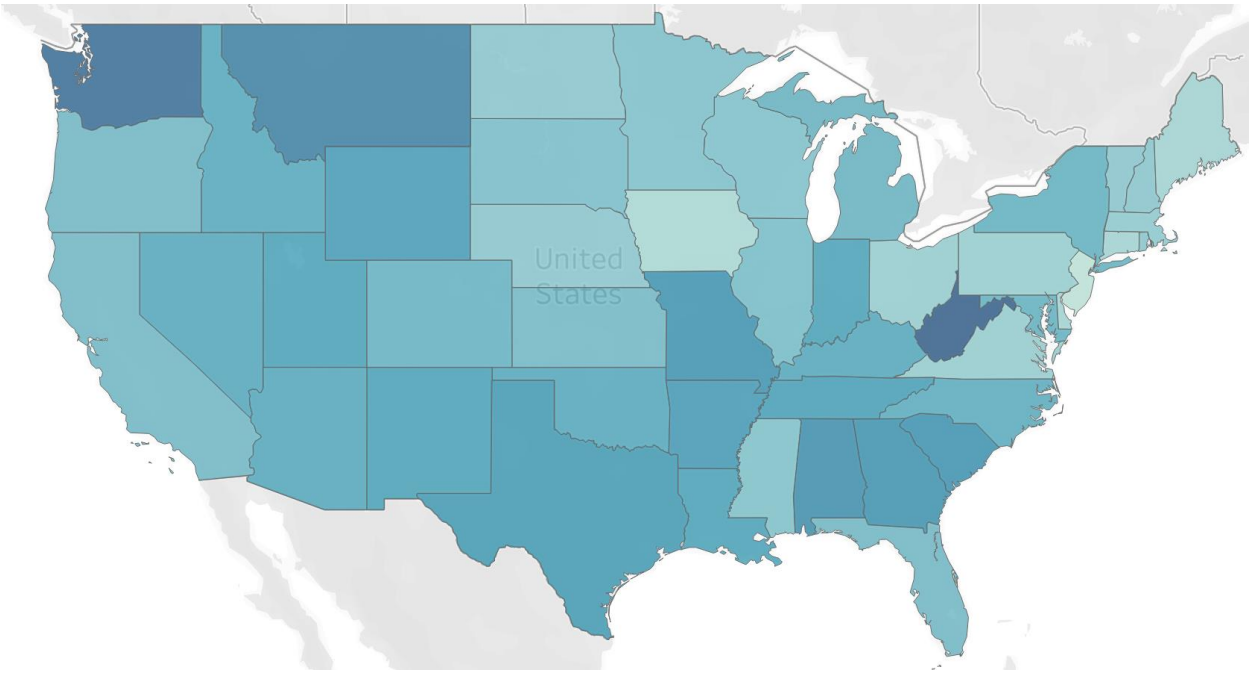


M. “Median List Price by Vehicle Type”



N. “Median List Price by State”






Hawaii