Contents

[**Project Goals** 1](#_Toc152359009)

[Data Set 1](#_Toc152359010)

[Exploratory Data Analysis (EDA) 1](#_Toc152359011)

[Data Cleaning Steps 2](#_Toc152359012)

[Machine Learning 7](#_Toc152359013)

[Model Selection 8](#_Toc152359014)

[Predictive Analysis 10](#_Toc152359015)

[**Resulting Conclusion** 10](#_Toc152359016)

**Project Goals**

1. Explore the used car dataset information, shape, features.
2. Look at correlations between price (dependent variable) and various independent variables to determine highest correlation to price.
3. Determine the top 5 -10 variables that are highly correlated to driving the price.
4. Identify trends in associated to price.
5. Determine best model to predict uses car prices based on best features.
6. Determine accuracy to best model

# ****Data Set****

The imported Used Car Data Set consists of 426880 and 18 columns. The dataset contains features such as ['id', 'region', 'price’, manufacturer', 'model', 'condition', 'cylinders', 'fuel', 'odometer', 'title\_status', 'transmission', 'VIN’, 'drive', 'size', 'type', 'paint\_color', 'state'].

The dataset was sourced from Kaggle.com; however, it was downloaded from Module 10 Try-It Exercise.

# Exploratory Data Analysis (EDA)

I’ve listed the data set features below after looking at the shape of the dataset and unique values for each feature. These key features include:

**ID**— Unique ID given to every associated with each car.

**Price**— Price is given in US dollar.

**Year**— The year in which the car was manufactured. The years range from Date Range: 1900-01-01 to 2022-01-01.

**Manufacturer**— 42 various manufactures associated with this dataset.

**Model**— 29649 unique models

**Transmission –** The transmission three types associated with this dataset. These included transmission mapping = {1: 'automatic', 2: 'manual', 3: 'other'}

**Condition**— The assessment of the car’s condition which are indicated in six different descriptions. These include ['excellent' 'fair' 'good' 'like new' 'new' 'salvage']

**Size** – This dataset included only four size types which included: size mapping = {1: 'full-size', 2: 'mid-size', 3: 'compact', 4: 'sub-compact'}

**Type –** The dataset included 13 different types that were used to identify unique features. These included: type\_mapping = {1: 'sedan', 2: 'SUV', 3: 'truck', 4: 'pickup', 5: 'other', 6: 'coupe', 7: 'hatchback', 8: 'wagon', 9: 'van', 10: 'convertible', 11: 'mini-van', 12: 'bus', 13: 'offroad'}

**Fuel**— Various fuel types indicated include ['gas' 'hybrid' 'diesel' 'other' 'electric']

**Odometer**— This current miles recorded while driving.

**Title Status**— The title status included six different types. These were listed as: ['clean' 'rebuilt' 'salvage' 'lien' 'missing' 'parts only']

**Type**— The description on a certain type such as van, truck, convertible, etc. There 13 unique types in this dataset.

**Drive** – The dataset included three different drive features. These included: drive\_mapping = {1: '4wd', 2: 'fwd', 3: 'rwd'}

**Cylinders**

**VIN**

**Paint Color**

**State**

**The dataset included null values along with categorical descriptions and numerical values.**

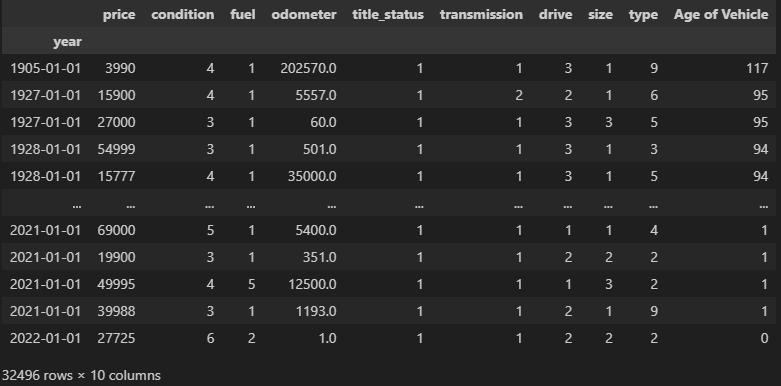
**The vh.describe() function used price, odometer, year, and ID to calculate the statistical values of the std.**

# ****Data Cleaning Steps****

* + 1. I first added column “Age of Vehicle” for quick reference for someone asking, “How old is the car?” The dealership would quickly determine the age versus looking up the year. The age ranged from (0 – 117 years)

## Adding Age of Vehicle column and calculating the age of each vehicle in the data set.

vh\_cleaned['Current Year'] = 2022 vh\_cleaned['Age of Vehicle'] = vh\_cleaned['Current Year'] - vh\_cleaned.index.year

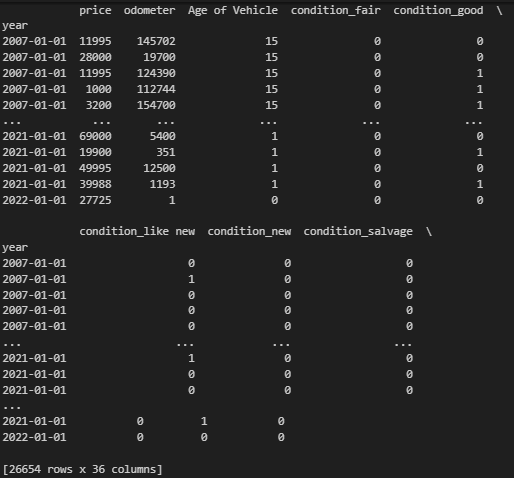


* + 1. I only wanted to look at the overall effect (independent features) that have an impact on the price and therefore I reduced the dataset to only show these features:

## I only want to use valuable columns that have an overall effect of the price.

vh\_new = vh\_cleaned.loc[:, ['price', 'condition', 'fuel', 'odometer', 'title\_status', 'transmission', 'drive', 'size', 'type', 'Age of Vehicle']]

* + 1. Instead of using the “get dummies” option due to the increase in the dataset columns. This would have given me a total of 36 columns versus the final 10. I replaced each feature with a numerical value. I did this for every feature that I chose would best be used in the analysis of predicting the price in used cars.

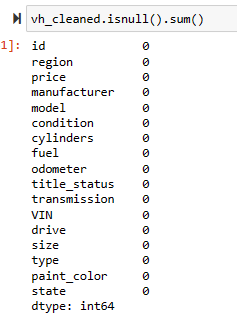


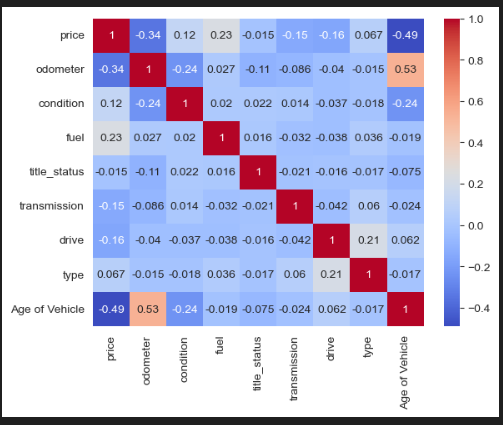
* + 1. I reduced the dataset to only look at cars 15 years and newer.

## Filter the dataframe based on the date range for only 15 years of age

vh\_new = vh\_new.loc['2007':'2022']

* + 1. I also dropped the null values, redundant rows, and drop any row that have price of $0 dollars. I checked to confirm.



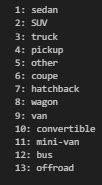
1. I also set the index to year. I did not change the column name to date as I did not see this would make a significance difference.
2. I created a final dataframe called vh\_final to use throughout the remainder of the Exploratory Data Analysis and Prediction Modeling.
3. I wanted to look at the highest correlations to price to give me an idea which features are driving the price.

Looking at the correlation matrix, the odometer ranked first, then fuel, and type. I then looked at the odometer to see how this trended over the Age of the Vehicle. Age does have an impact on price.

A graph showing the age of vehicles and odometer

Description automatically generated

I also looked at the top 10 cars based on odometer and age of vehicle.

A graph of different colored bars

Description automatically generated with medium confidence

The analysis shows for Odometer “truck” ranked highest and for Age of Vehicle “hatchback” ranked highest.

I wanted to see the trend in fuel types. Gas is mostly; however, I see a decrease in diseal. Hybrid is still new; however it’s not below the cutline (0) which indicates it could have potential to increase. Also, the age of the cars and inventory on hand also plays a role.

A graph of a graph showing a trend of fuel types

Description automatically generated

# Machine Learning

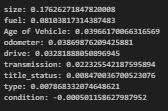
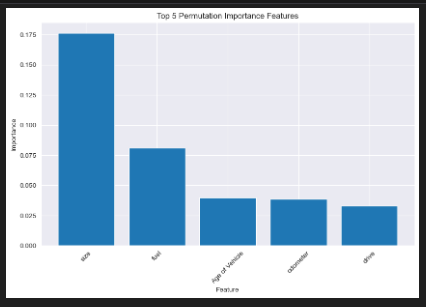
My goal was to determine the best model based on the features of the dataset to provide an estimated prediction of what a car dealership should be “pricing” their cars. Steps I took during this phase include:

* Conducted the service decomposition on the time series data.
* Conducted the AdFuller Test. Results concluded that the data was stationary.
* Conducted the ACF and PACF against the STD of the dataset.
* Conducted train/test split on the dataset.

# Model Selection

**I am evaluating TWO models.**

* + - 1. Linear Regression **using Permutation Importance. The results are below. I plotted the top 5.**

* + - 1. **Random Forest** using Permutation Importance. The results are below. I plotted the top 5.

**A screenshot of a computer program

Description automatically generated** **A graph of blue rectangular bars

Description automatically generated**

The results varied between the two models so I will look at which model is best to use for predictive analysis.

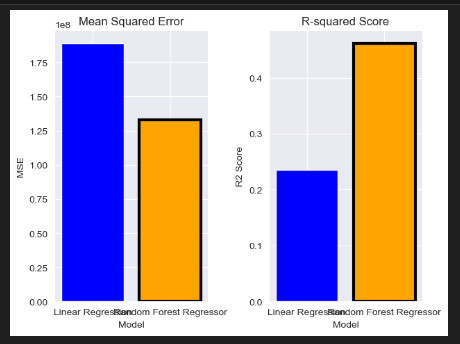
The Random Forest was the best model under evaluation with results of:



Random Forest

Linear Regression

Plots of both models are below. The best model is outlined in black.



# Predictive Analysis

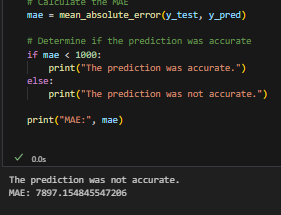
Based on the model selection using the Linear Regression, it was determined the model was inaccurate. Below is the predictive chart showing predictive amounts much lower than the asking price.

A graph with blue and orange lines

Description automatically generated

**Resulting Conclusion**

After looking/assessing the prediction and conducted the Mean Absolute Error calculation to determine if this prediction was accurate. It was not accurate based on the results below:



The market value of a car can vary based on several factors; however, I have learned from past experiences if Kelly Blue Book says a certain price, the dealership is not required to sell a car at that price due to variable like supply and demand, capitalism, free market, and areas with higher income.

For the purpose of this exercise, I chose the best model based on cross validation between Linear Regression and Random Forest. I applied the parameters to the model for an estimated prediction. I then validated its accuracy to determine if the model I used was sufficient. The results concluded I should try other types of models.