Used Car Price Prediction

Team Members:

Venkata Manikanta Monic Kamisetty (A20446683)

Jesleen Sonia Pradeep Kamalesh (A20448891)

Venkata Sai Akshay Kishore Khanderao (A20458999)

Introduction:

The price of a used car is quite hard to predict as it depends on various specifications. Though the sellers come up with a price, they are not sure if the right price is evaluated. Current approaches consider only a few specifications that predict the price which may be a rough calculation. A Used Car Price Prediction will help in predicting the price based on the worthiness of the car. It is useful to different clients based on their requirements. A car dealer would be able to offer quality service if they understand the needs of the buyers. A customer who is looking to buy or sell a car wouldn't be deceived. It will also help online estimators to predict prices with a better model.

Problem Statement:

To check the price of a used car, the worthiness of a car should be considered. The factors like model, condition, manufacturer, year, etc. influence the price prediction. The aim of this project is to predict car prices by implementing better models using machine learning algorithms. The "Used Car Dataset" is used for predicting the price of the used car. The model is developed to produce a reasonable price based on the most relevant features.

Data Description:

The data used for this project (Used Car Dataset) is taken from Kaggle. The data contains vehicle listing from craigslist.org. Craigslist is the world's largest collection of used vehicles for sale.

The Data consists of 26 columns and 458213 rows.

A brief description of each column is as follows:

id – Unique ID given to every ad and is the primary key to the data set.

url – URL form where the data of that row has been taken from.

region – Region where the vehicle is available.

price — Price is given in US dollar and has not been adjusted for inflation.

year — The year in which the car was manufactured

manufacturer — with 43 unique businesses engaged in the manufacture of automobiles.

model — The exact model of the vehicle. Like sierra classic 2500hd.

condition — The condition of the car; excellent, good, fair, like new, salvage, new.

cylinders — The number of cylinders in the car engine ranging from 3 to 12. Also has the ‘other’ category too.

fuel — There were five types of fuel, ‘diesel’, ‘gas’, ‘electric’, ‘hybrid’, and ‘other’.

odometer — This is the distance that the car has traveled after it being bought.

title_status – The cars also had 6 types of statues; ‘clean’, ‘lien’, ‘rebuilt’, ‘salvage’, ‘parts only’, and ‘missing’.

transmission – Transmission type automatic or semi-automatic.

VIN – Vehicle identification number.

drive — There are 3 types of drive transmissions: ‘4WD, ‘FWD’, and ‘RWD’. (Four-wheel drive, forward wheel drive, and rear-wheel drive.)

size – Size of the vehicle.

type — This feature identifies if a vehicle is an SUV or a mini-van. There 13 unique values in this feature.

paint_color – Paint color of the vehicle.

Image_url – URL of the image posted.

description – description of the vehicle posted

state — The state is political territory and is represented in short form in the data set. Like “fl” is used for the state of Florida.

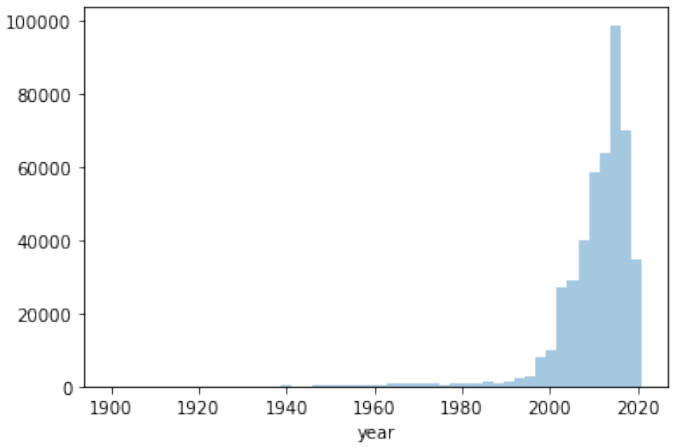
lat, long — When both features are combined, they give the location of where the car is being sold at.

posting_date – Posting date of the vehicle for resale.

Progress

Phase 1: EDA and Data Cleaning

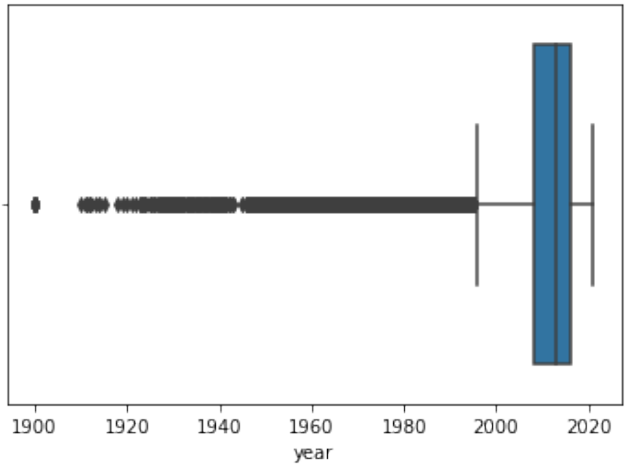
- *Year*



From the plot, we can state that most of the cars in the dataset are manufactured between the years 2000 and 2020.

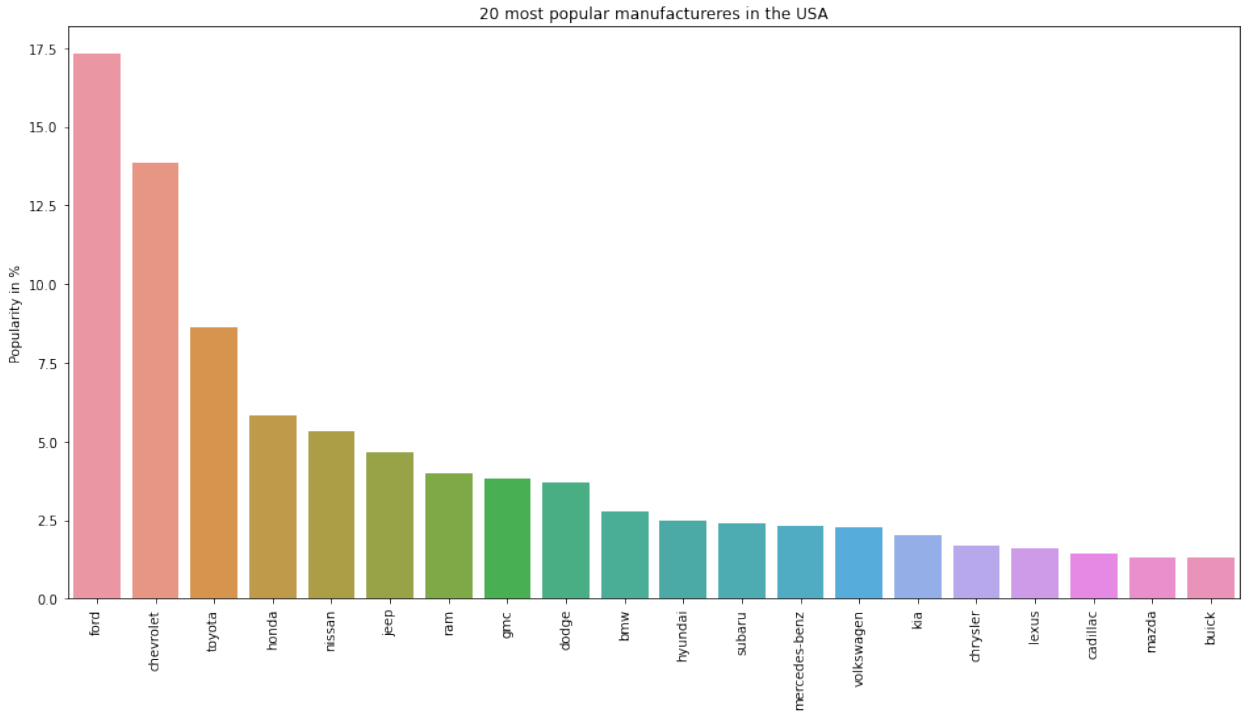
It is better to keep the original values given by the dataset any time than replacing the NaN values with mean or median. For the sake of this, we found out the year with the fewer missing values (1050). So, we dropped those training examples. Now we have a feature with the original data.

Detecting Outliers



From the boxplot, it is clear that most of the outliers are before 1995. So, it is better to remove outliers as we know they will highly affect the model's prediction.

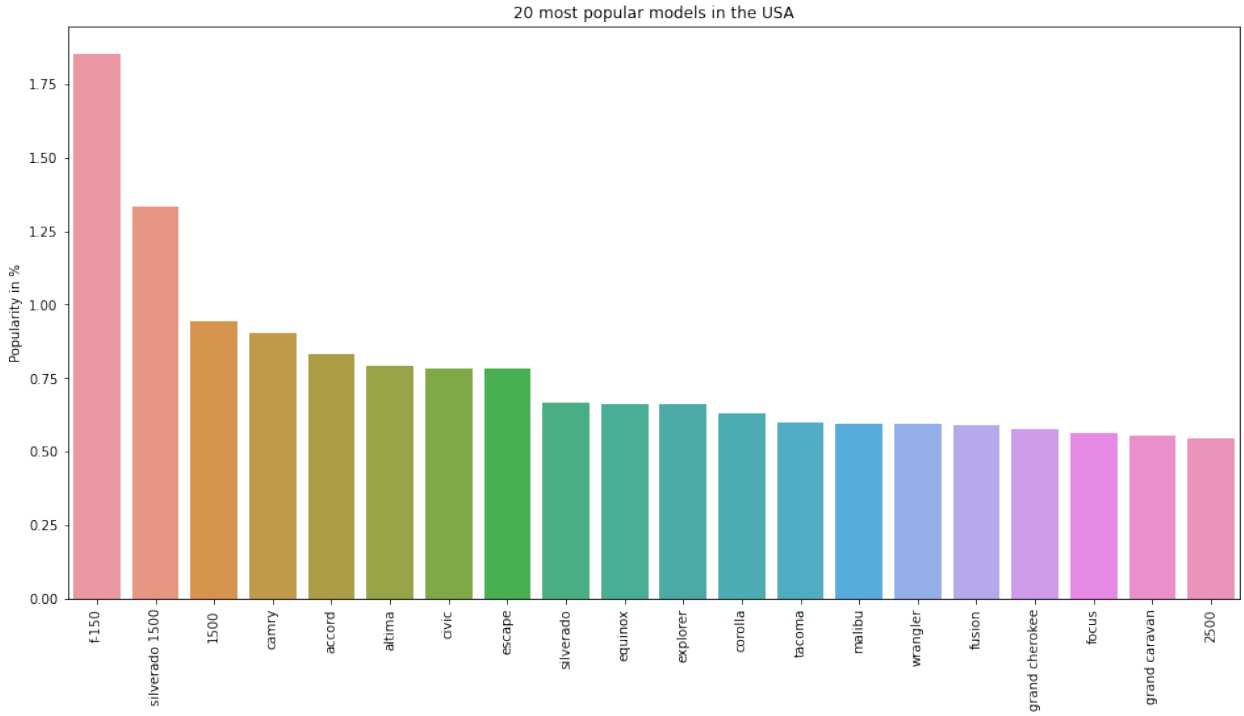
- *Manufacturer*



We have taken the top 20 popular manufacturers in the USA to plot the popularity of the vehicle brands. From the above bar plot, we can conclude that ford and Chevrolet are the most popular brands.

We observed a slight increase in missing values (14655) for the manufacturer compared to the year. As the manufacturer plays important role in predicting the price of a vehicle it's better not to fill the missing values by traditional filling mechanisms (filling with mode, forward fill, etc.) So, we decided to replace the null values with “unknown”. The reason for this is tree-based algorithms (bagging, random forest) are good at detecting null values by this small change. This may increase the prediction accuracy by tree algorithms.

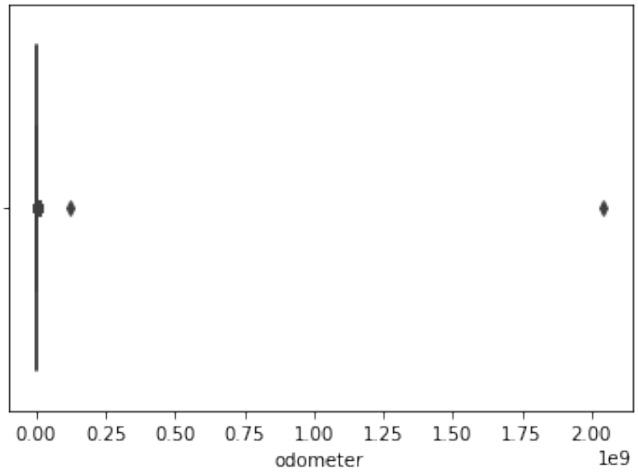
- *Model*



We have taken the top 20 popular models in the USA to plot the popularity of the vehicle models. From the above bar plot, we can conclude that f-150 and Silverado 1500 are the most popular models.

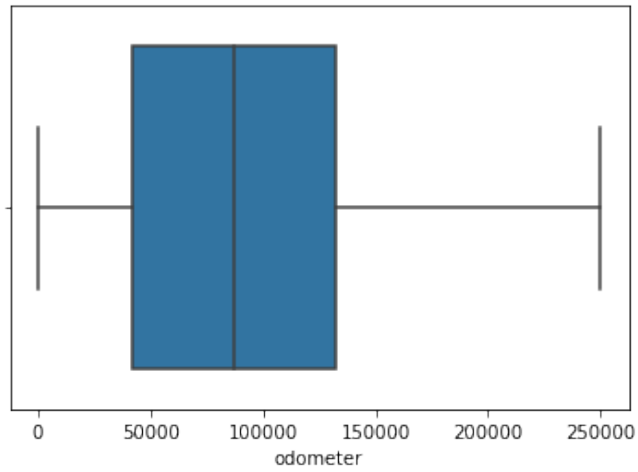
For the same reason as the manufacturer, we filled the null values with “unknown”.

- *Odometer*

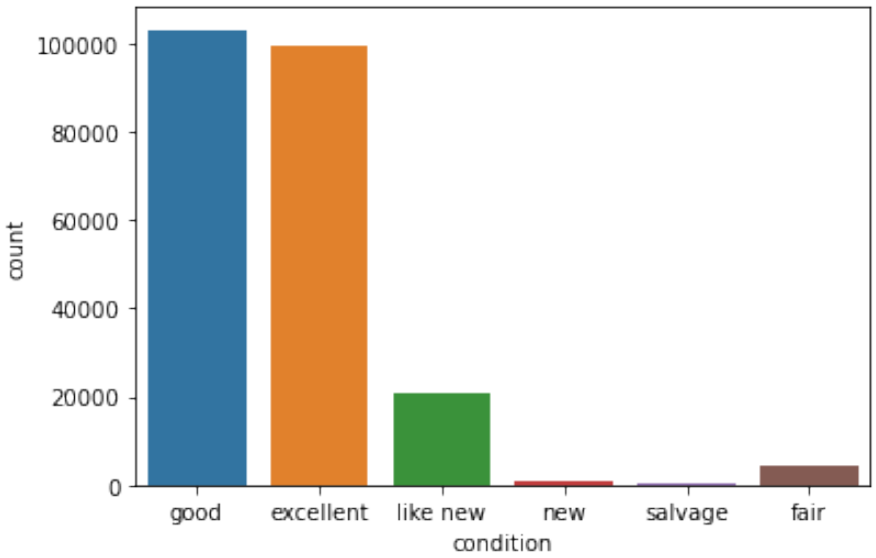


From the boxplot, we can observe two outliers will highly affect the prediction so, it is safe to remove those outliers.

We considered the odometer values that are less than 250000 and greater than 10. Because of this reason all the null values have also been removed.



- *Condition*



Based on the condition plot, a significant amount of the vehicles in the dataset are in good condition followed by vehicles in excellent condition. But the fewer number of vehicles in salvage and new makes the price prediction of this category a little off the chart.

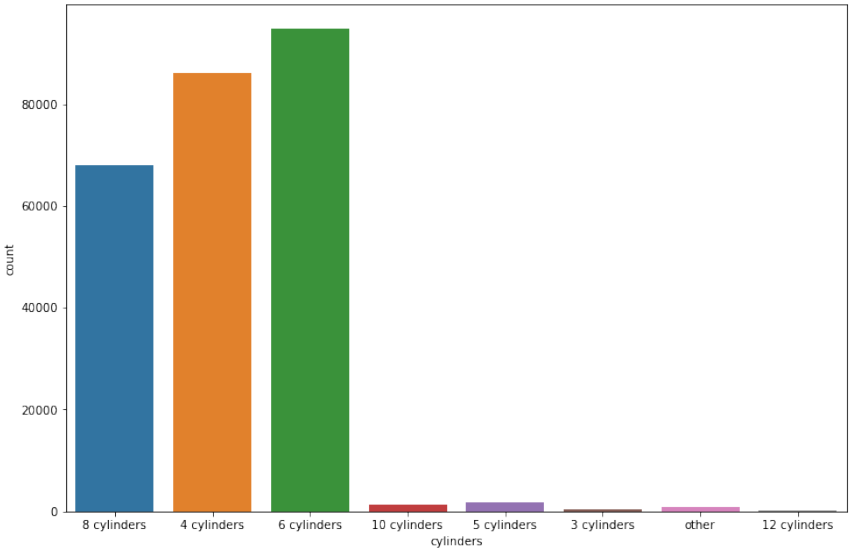
This problem can be avoided by a little common knowledge on people as they don't tend to buy salvage vehicles. To compensate for the fewer number of new vehicles we added the 'new' category to the 'like new' category.

There are a significant amount of null values in condition (192940). There is a high chance of losing a lot of information about the price of a vehicle if the null values are dropped or replaced with Mode (since it is a categorical feature).

So, we did a domain search and found out that the condition of the car depends on the odometer value. Based on the value of the odometer the values of condition are either replaced or null values are filled.

Approach: We calculated the mean values of the odometer for each category of condition. Then, a condition is set for the odometer values based on the mean values of each category of condition. Then they are replaced with the appropriate category of condition.

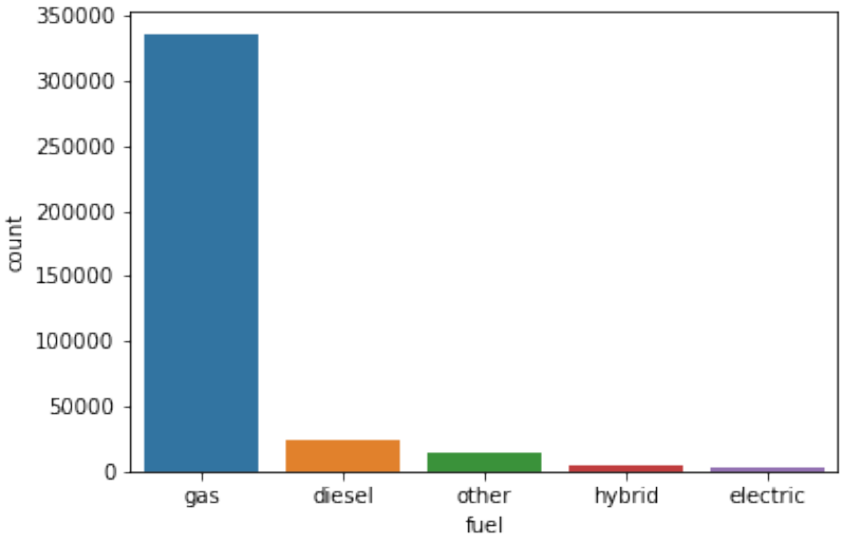
- *Cylinders*



Based on the cylinder plot, most of the vehicles that are coming to resale are the vehicles with 6 cylinders followed by 4 cylinders.

As there are a significant amount of null values (128249) in the cylinders feature, it is not a good approach to fill the null values with mode. So, it is better to replace them with 'unknown' as the tree algorithms will detect null values based on this value.

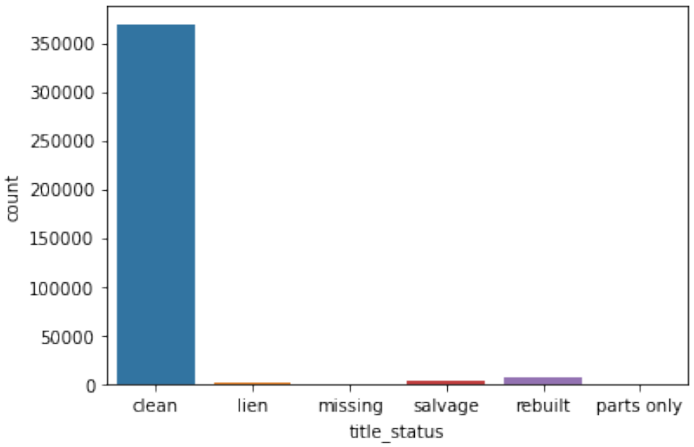
- *Fuel*



Based on the fuel plot, most of the vehicles in the USA run with gas followed by diesel.

As there are very few null values in the fuel feature (2172), replacing those with mode doesn't affect the prediction much.

- *title_status*

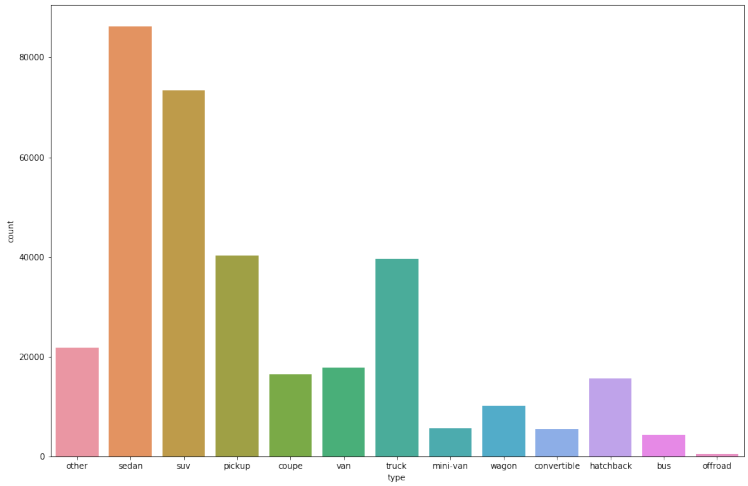


Based on title_status, most of the vehicles in the USA are clean.

As there are very few null values in the title_status feature (1582), we replaced the null values with the mode of the feature.

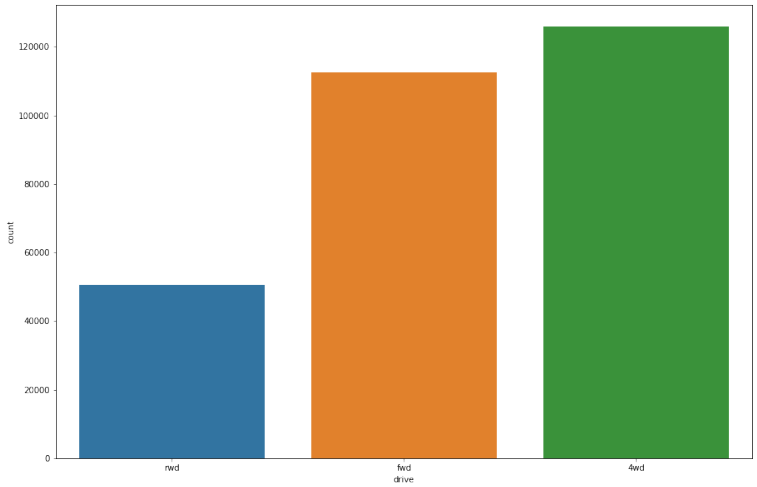
- Type:

This feature gives us information about the type of the car
Ex: sedan, SUV, pickup etc.
From the histogram, we can see that from all the cars the sedans are the most popular cars followed by SUVs and trucks and sedans having the highest average resale price. The type of car feature has a lot of missing values and they are filled by checking the description feature for the car type and the extracted value is assigned to the Type feature. At last, few data points which still had null values were discarded.



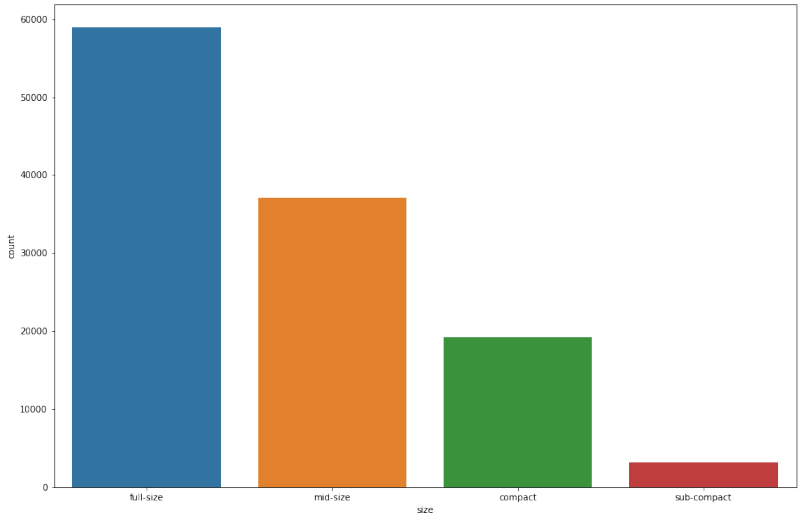
- Drive

  This feature gives us information about the Drive of the car it is classified into 3 types i.e. rear-wheel drive, front-wheel drive & 4-wheel drive. Four-wheel drives are most popular with the highest sales and highest average price followed by front-wheel drive. The rear-wheel-drive Is the least popular having the least average price. The Drive feature has a lot of missing values and they are filled by checking the description feature for the car drive and the extracted value is assigned to the Drive feature. At last, few data points which still had null values were discarded.



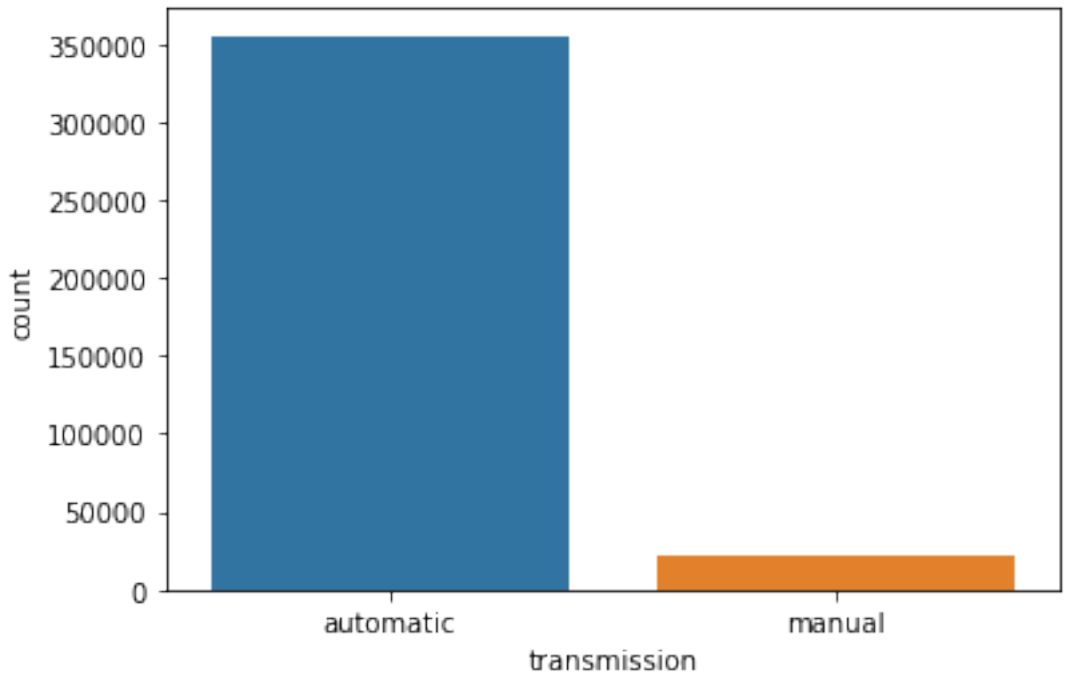
- Size

  The cars are divided into classes based on the sizes they are broadly classified into 4 groups full-size, mid-size, compact, sub-compact. Below is the distribution of the cars. The Size feature has a lot of missing values and they are filled by checking the description feature for the car driver and the extracted value is assigned to the Drive feature. At last, few data points which still had null values were discarded.



With the full-size being the most popular both in terms of several sales and high average resale value closely followed by mid-size in the USA and subcompact being the least preferred segment int the used cars.

- Transmission
  The cars are divided into classes based on the sizes they are broadly classified into 2 groups Automatic and Manual. With the automatic being the most popular type of transmission. The transmission feature had few missing values as the data is skewed towards the automatic value so the few missing values are also filled using the Mode values.



Future:

Phase 2: Modelling and Validation

Phase 3: Deployment