Used Car Price Prediction

“Used Car Price Prediction” is using the “Used Car Dataset” to predict the longevity and worthiness of a used car based on several attributes. The commercial importance of predicting the resale value is high.

**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.

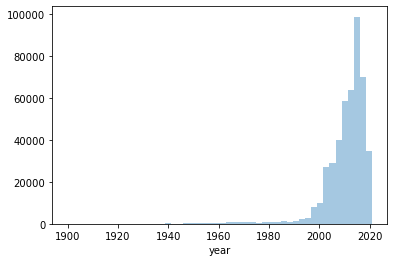
lat, long — When both features are combined, they give the location of where the car is being sold at.

posting\_date – Posting date of the vehicle for resale.

**Progress**

*Phase 1: EDA and Data Cleaning*

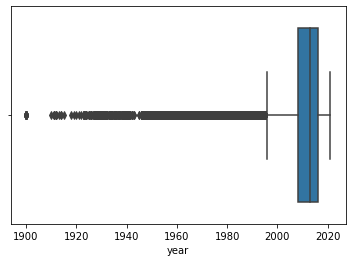
* *Year*



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

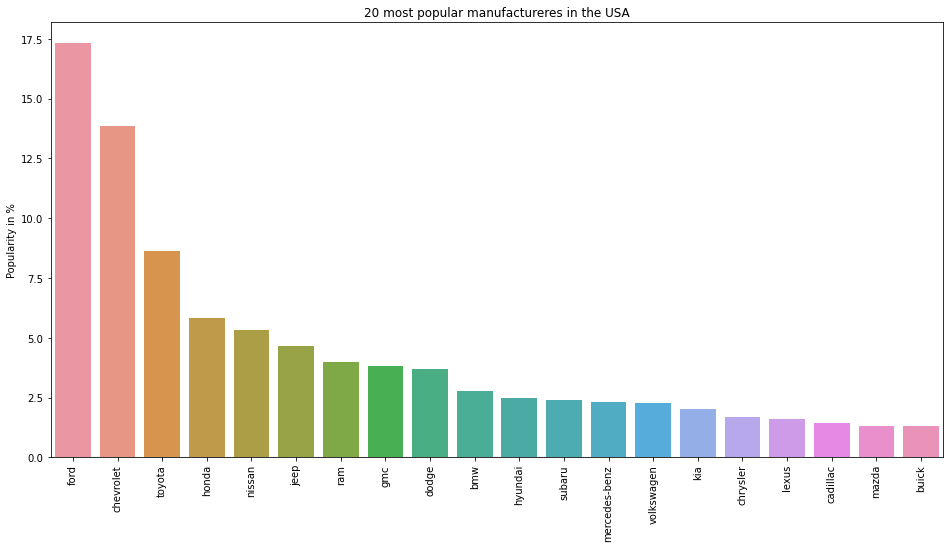
Its better to keep the original values given by the dataset any time than replacing the NaN values with mean or median. For the sake of this we found out year with the less missing values (xxx). So, we dropped those training examples. Now we have a feature with the original data.

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 models prediction.

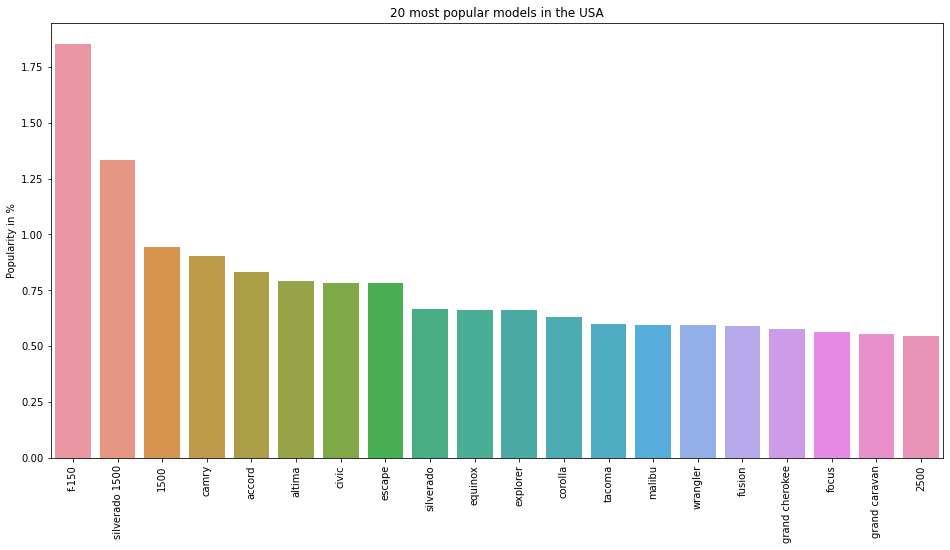
* *Manufacturer*



We have taken 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 manufacturer compared to year. As, manufacturer plays important role in predicting the price of a vehicle its 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 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 manufacturer, we filled the null values with “unknown”.

* *Odometer*



From the boxplot, we can observe there are two outliers which 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.

