Machine Learning Application in Predicting Used Cars Price

DSCI 3415 Fundamentals of Machine Learning

Habeeba Hossam
Data Science
The American University in Cairo
Cairo, Egypt
habibahossam@aucegypt.edu

Youssif Abuzied
Computer Engineering
The American University in Cairo
Cairo, Egypt
youssif.abuzied@aucegypt.edu

Data Overview

US Used cars dataset.

Link:
https://www.kaggle.com/datasets/ananaymital/us-used-cars-dataset

This data set provides data for used cars in the United States. The observations in this dataset were collected through crawling the website of Cargurus inventory. This dataset contains more than 3 million samples of cars' features and their prices. This dataset contains about 66 features which is a great number. Some of these features include the mileage, fuel type, miles traveled per gallon in the city and the highway, interior and exterior color of the car, etc. It included all features that might have any relation with the car's price. Also, almost all the features are described precisely which will help facilitate data cleaning and help us decide what features to include.

Features Description

As aforementioned, the dataset consists of 66 features which will be listed with their detailed description below.

(vin [string] ; Vehicle Identification Number is a unique encoded string for every vehicle. Read more at https://www.autocheck.com/vehiclehistory/vin-basics. ***back_legroom*** [string]; legroom in the rear seat. ***bed*** [string]; indicates the category of bed size (open cargo area) in a pickup truck (e.g., short, medium or long). Null usually means the vehicle isn't a pickup truck. ***bed_height*** [string]; bed height in inches. ***bed_length*** [string] bed length in inches. ***body_type*** [string]; body type of the vehicle (e.g., SUV / Crossover, Sedan, Pickup Truck, Hatchback, Minivan, Coupe, Van, Wagon and Convertible) ***cabin*** [string]; category of cabin size(open cargo area) in pickup truck. Eg: Crew Cab, Extended Cab, etc. ***city*** [string]; city where the car is listed. ***City_economy*** [float]; fuel economy in city traffic in km per litre. ***combine_fuel_economy*** [float]; combined fuel economy is a weighted average of City and Highway fuel economy in km per liter. ***daysonmarket [integer]***; days since the vehicle was first listed on the website. The days that the cars spent listed on the website until it was sold. ***dealer_zip*** [integer]; zip Code of the dealer. ***description*** [string]; vehicle description on the vehicle's listing page. ***engine_cylinders*** [string]; the engine configuration. ***engine_displacement*** [float]; the measure of the cylinder volume swept by all of the pistons of a piston engine, excluding the combustion chambers. ***engine_type*** [string]; indicates the engine configuration. ***exterior_color*** [string]; exterior color of the vehicle, usually a fancy one same as the brochure. ***fleet*** [boolean]; indicates whether the vehicle was previously part of a fleet or not. **frame_damaged** [boolean]; specifies whether the vehicle has a damaged frame. ***franchise_dealer*** [boolean]; demonstrates whether the dealer has franchises. ***franchise_make*** [string]; the company that owns the franchise (e.g., Honda, Chevrolet, Mazda, GMC, Nissan, etc.). ***front_legroom*** [string]; the legroom in inches for the passenger seat. ***fuel_tank_volume*** [String]; fuel tank's filling capacity in gallons. The last two mentioned features are quantitative; however, it is indicated as string by the main owner of the dataset as string because a suffix of "in" and "gal" respectively is added to all values. ***fuel_type*** [string]; dominant type of fuel ingested by the vehicle. ***has_accidents*** [boolean]; demonstrates whether the vin has any accidents registered. ***height*** [string]; height of the vehicle in inches. ***highway_fuel_economy*** [float]; fuel economy in highway traffic in km per litre. ***horsepower*** [float]; horsepower is the power produced by an engine. ***interior_color*** [string]; Interior color of the vehicle, usually a fancy one same as the brochure. ***isCab*** [boolean] Whether the vehicle was previously taxi/cab. ***is_certified*** [boolean]; specifies whether the vehicle is certified. To better explain, certified cars are covered through warranty period. ***is_cpo***

[boolean]; pre-owned cars certified by the dealer. Certified vehicles come with a manufacturer warranty for free repairs for a certain time period. Read more at https://www.cartrade.com/blog/2015/auto-guides/pros-and-cons-of-buying-a-certified-pre-owned-car-1235.html. ***is_new*** [Boolean]; if True means the vehicle was launched less than 2 years ago. ***is_oemcpo*** [Boolean]; pre-owned cars certified by the manufacturer. Read more at https://www.cargurus.com/Cars/articles/know_the_difference_dealership_cpo_vs_manufacturer_cpo. ***latitude*** [float]; latitude from the geolocation of the dealership. ***length*** [string]; length of the vehicle in inches. ***listed_date*** [date M/DD/YYYY]; the date the vehicle was listed on the website. ***listing_color*** [string]; dominant color group from the exterior color. **listing_id** [integer]; listing id from the website. ***longitude*** [float]; Longitude from the geolocation of the dealership. ***main_picture_url*** [string]; link to picture of the vehicle. ***Major_options*** [string]; the major options in the vehicle (e.g.,Leather Seats, Sunroof/Moonroof, Bluetooth, Backup Camera, Heated Seats, etc.) ***Make_name*** [string]; the manufacturer of the vehicle. ***Maxuim_seating*** [string];the maximum number of seating that can be accommodated in the car. ***Mileage*** [integer]; the number of miles traveled or covered is measured in miles per gallon (mpg).***Model_name*** [string]; the model of the vehicle (e.g., Corolla, Grand Cherokee, Sportage, E-Class, ect.). ***Owner_count*** [integer]; the number of previous owners.***(48) Power*** [string]; power produced by an engine (hp @ RPM) .***Price*** [integer]; price in USD. ***(50) Salvage*** [boolean]; specifies whether a vehicle has an official indication that it has been damaged and is considered a total loss by an insurance company that paid out on a damaged vehicle claim. ***saving _amount*** [integer]; the amount of money saved by pucrhasing the used vehicle in compassion to its average market price. ***seller_rating*** [float]; the seller's rating, takes any value between 1 and 5. ***Sp_id*** [integer]; service parts dealer id. ***Sp_name*** [string]; service parts dealer name. ***Theft_title*** [boolean]; indicate whether the vehicle was stolen and repaired or not. ***(56) Torque*** [string]; force that can cause an object to rotate about an axis. ***Transmission*** [string]; transmission of vehicle (e.g., CVT, A, M). ***Transmission_display*** [string]; type of transmission (e.g., Automatic, Manual). ***trimID*** [integer]; Trim levels are used by manufacturers to identify a vehicle's level of equipment or special features. Those levels are given id by the vehicle manufacturer. ***Trim_name*** [string]; trim levels are given names by the vehicle manufacturer. ***Vehicle_damage_category*** [string]; indicates the damage category of the vehicle. ***Wheel_system*** [string]; the wheel system of the vehicle (e.g., FWD, AWD, 4WD). ***Wheel_system_display*** [string]; specifies the wheel system display (e.g., Front-Wheel Drive, All-Wheel Drive, Four-Wheel Drive). ***Wheelbase*** [string]; the wheelbase is the horizontal distance between the centers of the front and rear wheels in inches. ***Width*** [string]; the width of the vehicle. ***year*** [integer]; model year of the vehicle.

As noticed, several quantitative features are viewed as string due to the fact that they have a suffix like “in” or “gal”. This issue will be solved during the parsing of each feature.

Data Cleaning and Preprocessing

The data cleaning and preprocessing is conducted on several steps that build on each other. It is salient to note that due to logistical matter, the local machines of the researchers were not equipped to handle such a massive csv file of size 9.98 GB. The website Split csv allowed the researchers to download the data into multiple lighter csv files.

As previously mentioned, the main challenge encountered is the enormous size of the data. Upon close inspection of the features, the researchers hypothesized that the feature ***description*** was redundant. To test this hypothesis, a random sample of 1000 observations was thoroughly read and analyzed. The principle takeaway is that almost all of the information mentioned in ***description*** were already available in other features. Thus, the researchers removed the feature ***description*** which resulted in a decrease of file size by 85% and this step allowed easier exploratory data analysis of the whole dataset.

Upon joining the whole dataset and conducting through data analysis, the researchers made the decision to remove vehicles that serve in transportation of produce goods; such as pickup trucks and vans (which made up almost 17.39% of the total observations). In addition, any observations that had a null value in the feature ***body_type*** were removed to ensure that all the left vehicles were of either one of the following types; SUV / Crossover, Sedan, Hatchback, Minivan, Coupe, Wagon and Convertible. Furthermore, it must be noted that all features related to the bed that is usually associated with vehicle transporting goods were removed from the dataset. To be more specific, the features ***bed***, ***bed_height*** and ***bed_length.***

Since the main goal of this research is to build a robust model to predict the price of used cars, some features that are either irrelevant or with a majority of null values will be

dropped. To elaborate, ***main_picture_url*** will be removed as no picture or image processing will be conducted. As for ***combine_fuel_economy***, ***is_certified***, ***cabin***, ***is_oemcpo***, ***Vehicle_damage_category*** and ***is_cpo*** , they were removed due to the fact that the percentage of missing values in these features is over 90%.

A random sample of 200,000 observations was selected to carry out any further analysis or parsing.

Some of the very significant statistics and figures about the sample are on the ***body_type***. The seven categories include the following numbers of observations; SUV / Crossover (114892), Sedan (60398), Hatchback (7122), Minivan (6449), Coupe (5676), Wagon (3351), Convertible (2112) and no missing values as insure in earlier steps. The researchers had a particular interest in comparing the mean price of different ***body_type*** as illustrated in figure (1).

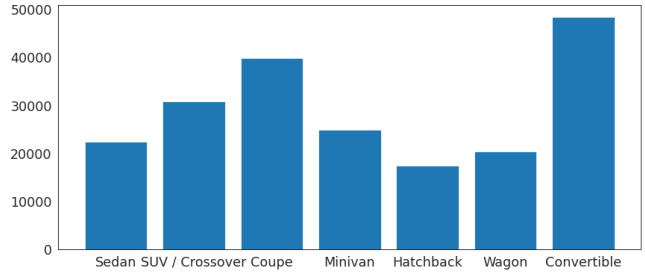


Figure (1)

As evident from the above plot, Convertible cars have the highest mean followed by Coupe then SUV / Crossover. Wagon and Sedan appear to almost have the same price mean. Hatchback cars have the lowest price mean and minivan have the second least price mean.

The researchers were also curious about the maximum price value for different body Types as shown in figure (2).

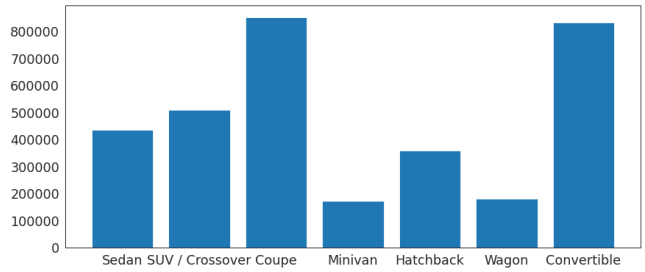


Figure (2)

As demonstrated in the graph above, Coupe and Convertible cars have the highest price points followed by SUV/Crossover, sedan and hatchback. The minivan and wagon have the lowest maxim price values.

The very opposite plot which is the minimum price values for different Body Types is very exciting as demonstrated in figure (3).

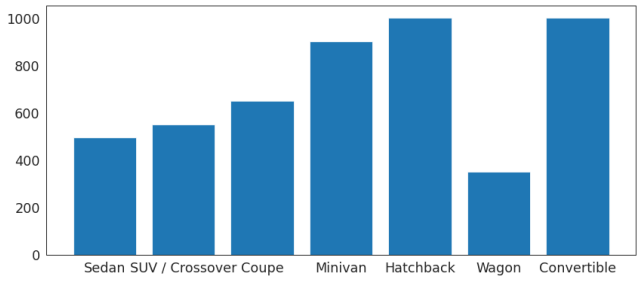


Figure (3)

The cheapest car in the sample is of type wagon followed by Sedan, SUV/ Crossover and Coupe. The most expensive vehicles are of type Hatchback, Convertible and Minivan.

The standard deviation of prices according to the body type is an important graph to investigate, figure (4).

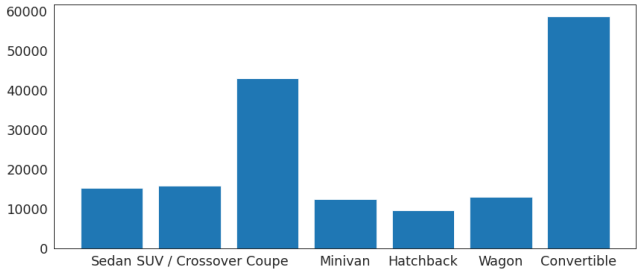


Figure (4)

The biggest standard deviation can be noticed in the price of Convertibles followed by Coupe cars. The rest of car types have almost the same deviation.

As evident from the above graphs the body type influences the price in different ways, hence the feature will not be removed but rather properly encoded to optimize the available features. As ***body_type*** is a categorical feature, hot encoding was used and then dropped as all data was captured and coded as 0 and 1.

A number of features required to be parsed before any analysis could be done. The suffix "in" which stands for Inch was removed from the features ***back_legroom***, ***front_legroom***, ***height***, ***length***, ***wheelbase*** and ***width***. While "gal" stands for Gallon was removed from the feature ***fuel_tank_volume***. Furthermore, "seats" was

dropped from the feature ***maximum_seating***. The feature ***power*** was parsed in a special way to preserve its two main components (hp and RPM). To elaborate, two columns were created to capture all of the available data; the first one is named ***power_in_horses*** and the second is named ***RPM***. Similarly, the feature ***torque*** was parsed to preserve its two crucial components (lb-ft and RPM). Two features were created to store all the data; the first one is named ***torque_ib_ft*** and the second feature is named ***torque_RPM***. As for the parsing ***listed_date***, only the month and the year were kept and the day was removed to decrease the huge number of unique values. The feature ***interior_color*** was parsed in a very unique way. The number of unique values for this feature is caused by the very specific description of the color. Thus, the original or the core values were kept which reduced the number of unique values. For instance "Fancy red" is simply "red". It is crucial to note that each feature's parsing was completed independently in a separate section of the source code; it was mentioned collectively for the benefit of consistency and easy reading of this article.

In cases when outliers are present, substituting the missing value with the feature's median value is a frequent method for impugning missing data for numerous features. The expectation is made for categorical data to which the mode is used.

The second feature that needed some work to be in its proper format is ***back_legroom***. The box plot of the feature is shown in figure (5). The missing values (7570) will be replaced by the median (38.0). The correlation coefficient between ***back_legroom*** and ***price*** is 0.17.

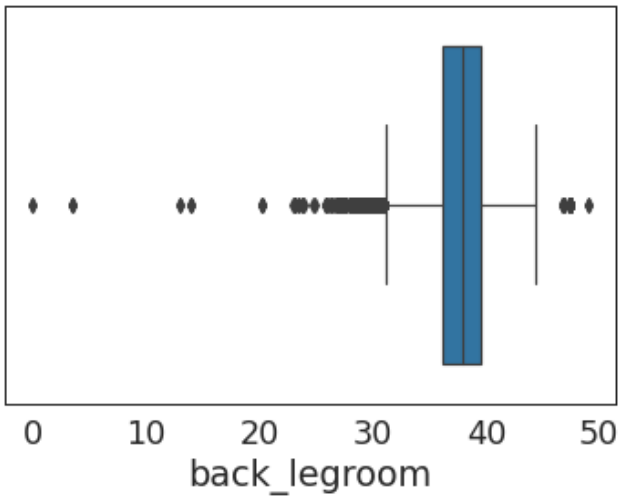


Figure (5)

The researchers were persistent on learning the statistics and figures after the standardization process; the mean (3.6), The standard deviation (0.99 almost 1), The minimum value (-15.1), the maximum value (4.5), The first quartile (-0.6), The median (0.146), the third quartile (0.7). The correlation coefficient between ***back_legroom*** and ***price*** is 0.175. As for the distribution histogram and box plot of ***back_legroom*** it can be observed in figure (6).

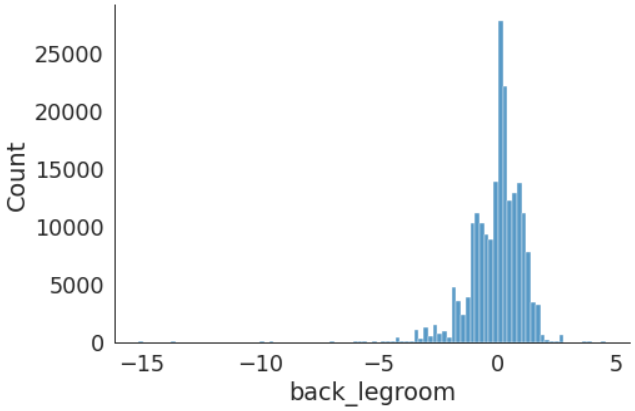


Figure (6)

Interestingly, the observations in ***back_legroom*** can be best described as following an almost beta or normal distribution, as demonstrated in figure (7).

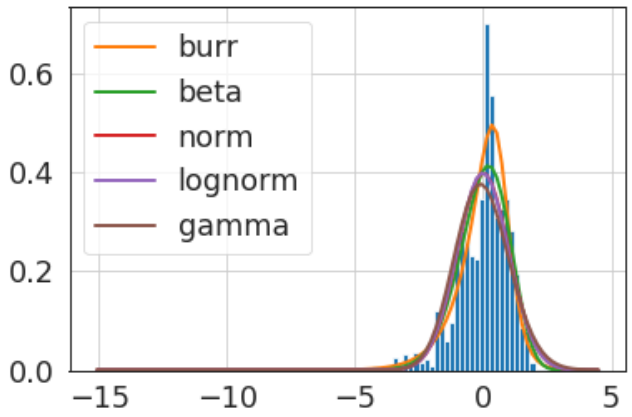


figure (7)

Regarding the feature ***front_legroom,*** the box plot clearly shows the outliers of the data, figure (8). The missing values (7521) will be replaced by the median (41.8). The correlation coefficient between ***front_legroom*** and ***price*** is 0.003.

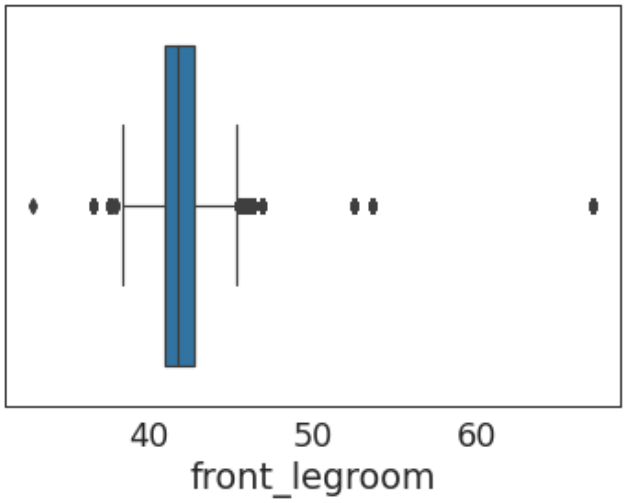


Figure (8)

After data imputation standardizing the data, the following are the important statistics; the mean (-1.5), The standard deviation (1), the minimum value (-6.4), the maximum value (17.6), the first quartile (-0.6), The median (-0.179), the third quartile (0.7). As for the distribution histogram of ***front_legroom*** it can be observed in figure (9).

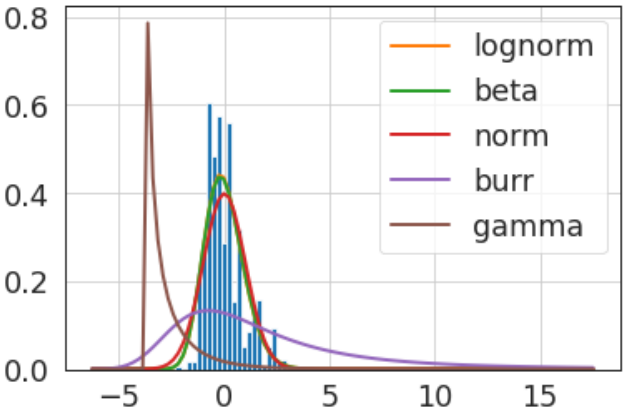


Figure (9)

As shown in figure (10), the data in ***front_legroom*** best fit the description of the following distributions; log-normal, Beta, Normal.

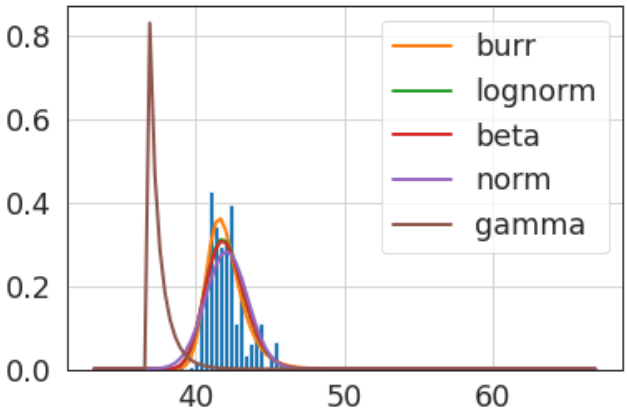
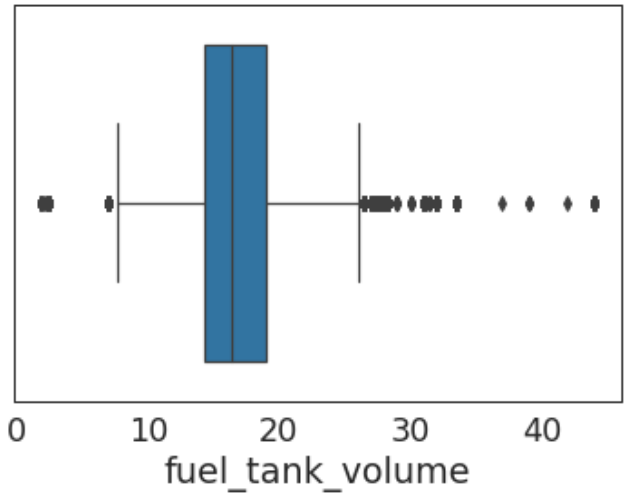


Figure (10)

Concerning the feature fuel_tank_volume, the box plot clearly shows the outliers of the data, figure (11). The missing values (7464) will be replaced by the median (16.4). The correlation coefficient between ***fuel_tank_volume*** and ***price*** is (0.3).



After data imputation standardizing the data, the following are the important statistics; the mean (2.7e-16), The standard deviation (1), the minimum value (-4.1), the maximum value (7.31), the first quartile (-0.6), The median (-0.17), the third quartile (0.5). As for the distribution histogram of front_legroom it can be observed in figure (12).

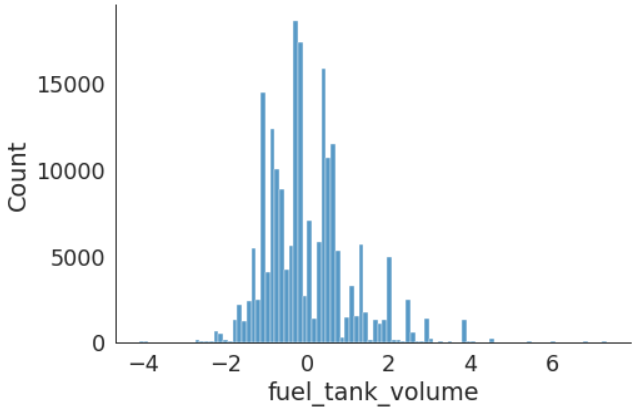


Figure (12)

As shown in figure (13), the data in front_legroom best fit the description of the following distributions; Burr, log-normal,Gamma ,Beta, Normal.

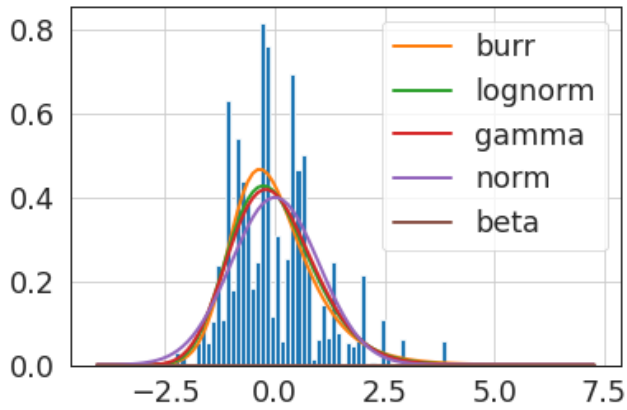


Figure (13)

Concerning the feature height, the box plot clearly shows the outliers of the data, figure (14). The missing values (7462) will be replaced by the median (16.4). The correlation coefficient between ***height*** and ***price*** is (0.2).

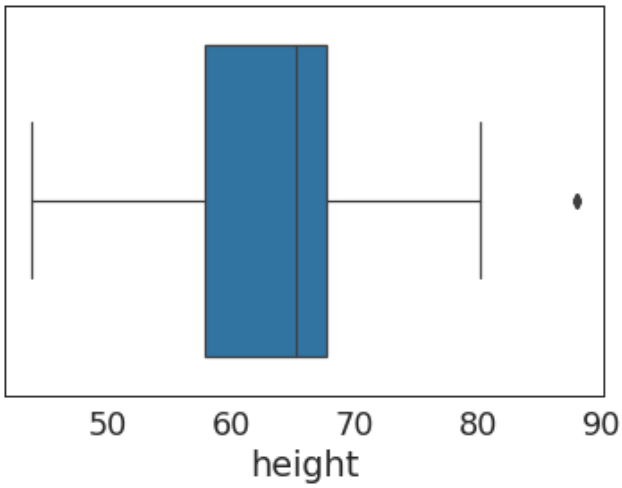


Figure (14)

After data imputation standardizing the data, the following are the important statistics; the mean (1.4e-15), The standard deviation (1), the minimum value (-3.4), the maximum value (4.2), the first quartile (-1.02), The median (0.2), the third quartile (0.7). As for the distribution histogram of front_legroom it can be observed in figure (15).

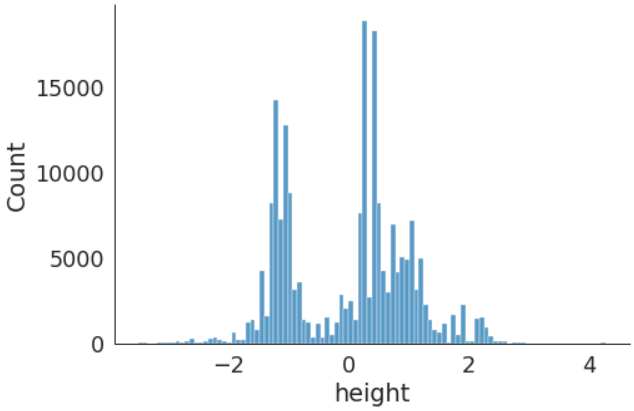


Figure (15)

Concerning the feature length, the box plot clearly shows the outliers of the data, figure (16). The missing values (7462) will be replaced by the median (16.4). The correlation coefficient between length and price is (0.3).

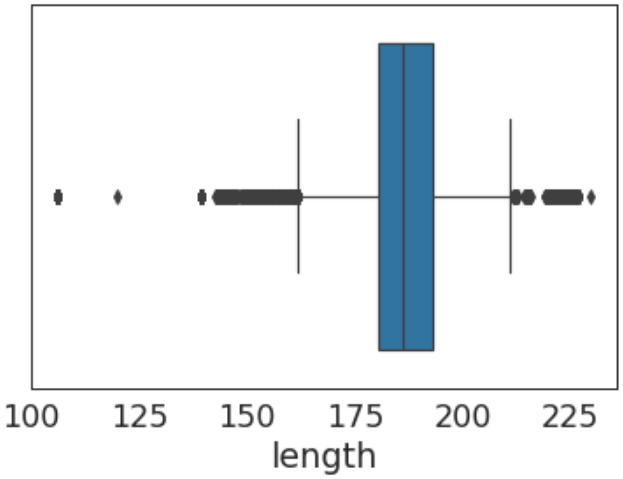


Figure (16)

After data imputation standardizing the data, the following are the important statistics; the mean (1.1e-15), The standard deviation (1), the minimum value (-6.9), the maximum value (3.7), the first quartile (-0.5), the median (-0.019), the third quartile (0.5). As for the distribution histogram of front_legroom it can be observed in figure (16).

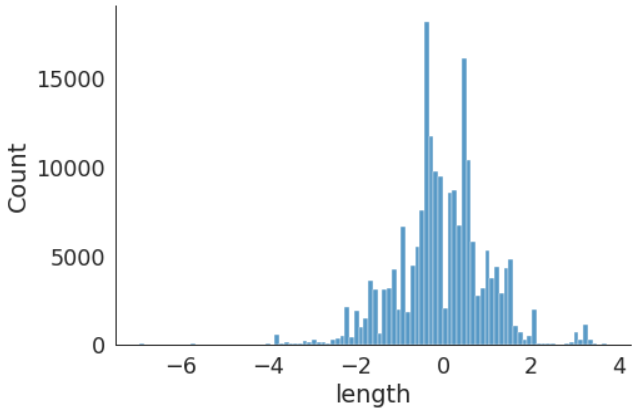


Figure (17)

As shown in figure (18), the data in front_legroom best fit the description of the following distributions; Log-normal,Gamma ,Beta, Normal.

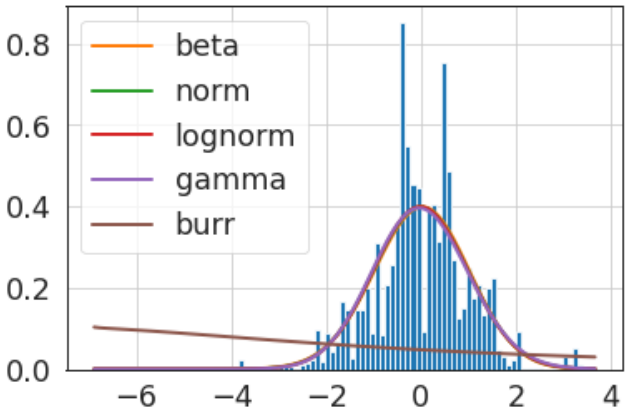


Figure (18)

Concerning the feature ***maxuim_seating***, the box plot clearly shows the outliers of the data, figure (19), the following are the important statistics; the mean (5.4), The standard deviation (1.08), the minimum value (2), the maximum value (9), the first quartile (5), the median (5), the third quartile (5). As for the distribution histogram of front_legroom it can be observed in figure (19). The final step in the process is conducting hot encoding.

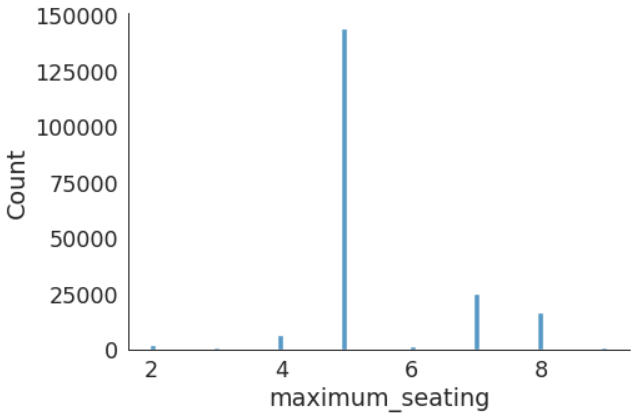


Figure (19)

As for the feature ***Wheelbase***, its box plot indicates the presence of multiple outliers as shown in figure (30). So the missing values were replaced by the median equal to 109.3. The correlation coefficient between ***Wheelbase*** and ***price*** (0.35).

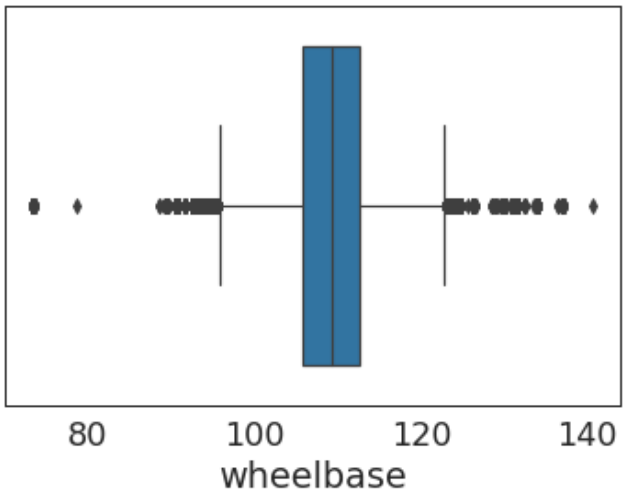


Figure (20)

After standardization of the feature, some of the important statistics and graphs include; mean (-1.9e-15), standard deviation (1), minimum value (-5.9), maximum value (5.05), first quartile (-0.6), third quartile (0.4) and the histogram of the distribution is shown in figure (21).

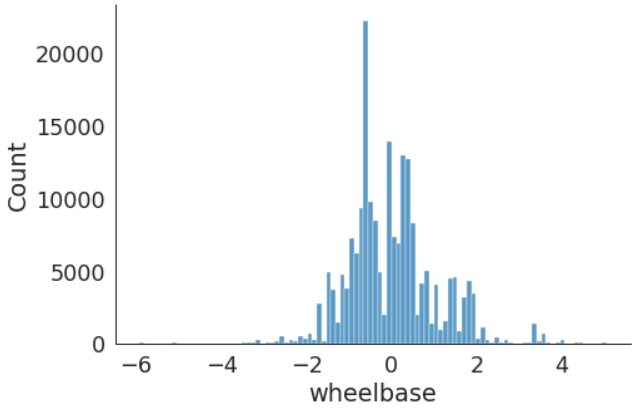


Figure (21)

The best well commonly known distributions that fairly describe the feature are Burr, the Lognorm, and Normal distribution as demonstrated in figure (22).

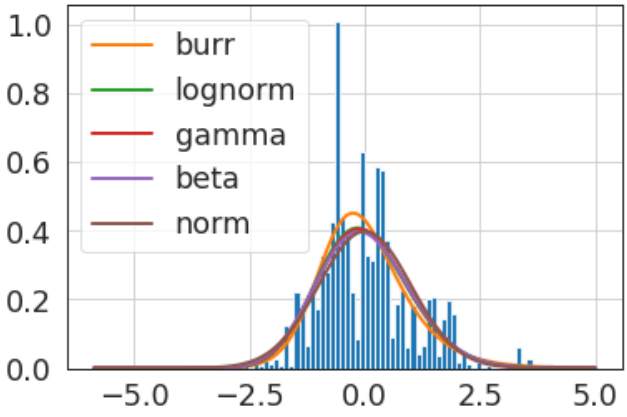


Figure (22)

Although the ***width*** feature shows only one outlier in its boxplot, figure (23), the median will be used to replace the missing value for its robustness and for the sake of consistency with the other features.

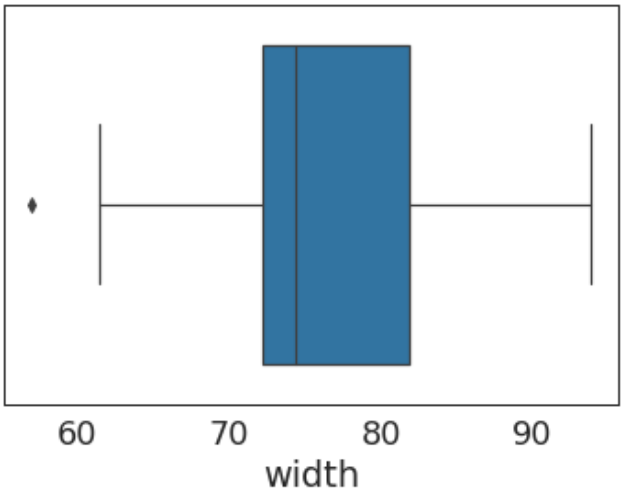


Figure (23)

After data imputation and standardization, the feature ***width*** can be best described using the following figures; mean (-2.53e-15), standard deviation (1.0), minimum value (-3.1), maximum value (2.7), first quartile (-0.6), third quartile (0.7), correlation coefficient between ***Width*** and ***price*** (0.3), and the histogram of the distribution is shown in figure (24).

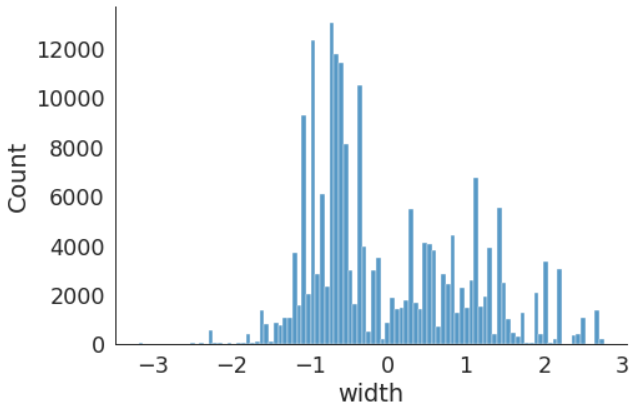


Figure (24)

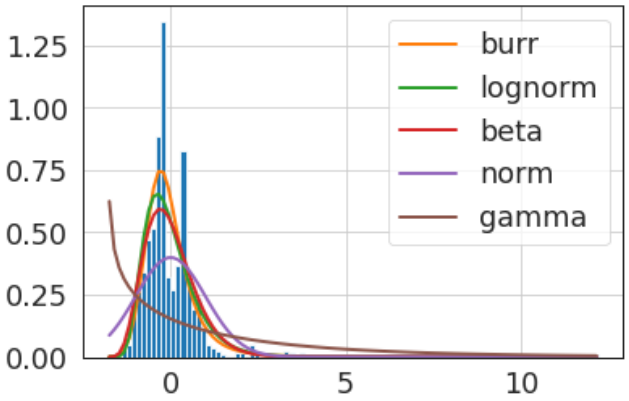


Figure (26)

After data imputation using the median and the standardization, these figures are obtained for ***city_fuel_economy***; mean (7.560174708487466e-17), standard deviation (1.0), minimum value (-1.8), maximum value (12.2), first quartile (-0.5), third quartile (0.3), and the histogram of the distribution is shown in figure (25). Among the well known continuous distribution, Burr, Lognorm, Gamma, Beta and Normal can roughly describe the feature, figure (26).

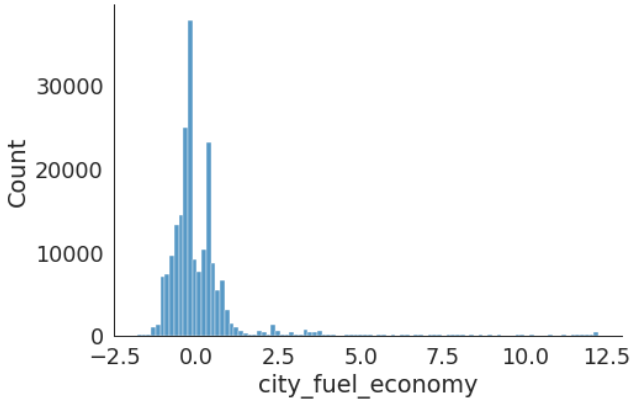


Figure (25)

Interestingly, the feature ***daysonmarket*** does not contain any missing values. The correlation coefficient between ***daysonmarket*** and ***price*** (0.08). The mean (76.51) is significantly higher than the median (36.0) which signals that there are outliers in the data. This is confirmed by the box plot of the feature, figure (27). After standardizing the feature, these statistics are obtained; mean (1.9930723738070812e-17), standard deviation (1.0), minimum value (-0.7), maximum value (23.0), first quartile (-0.5), third quartile (0.06) , and the histogram of the distribution is shown in figure (28).

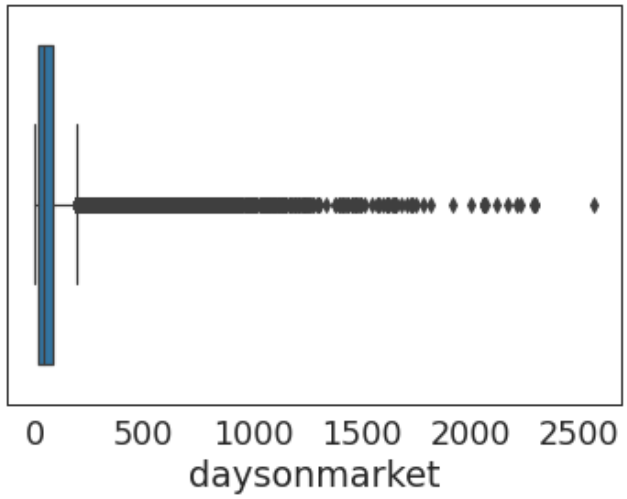


Figure (37)

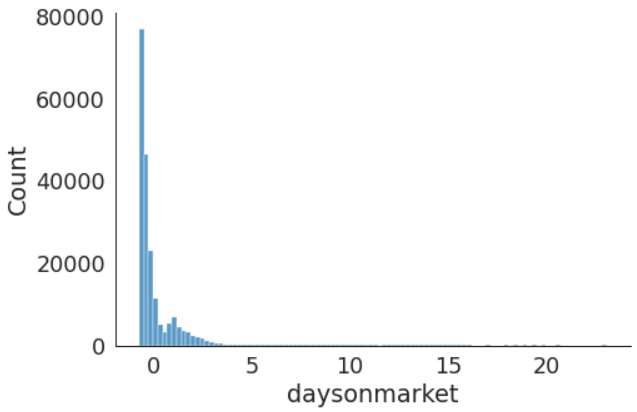


Figure (38)

As for features about the car's engine, ***engine_displacement*** has 8643 missing values which will be replaced by the median as the data contain multiple outliers as shown in the horizontal box plot, figure (39). The correlation coefficient between ***engine_displacement*** and ***price*** (0.31),

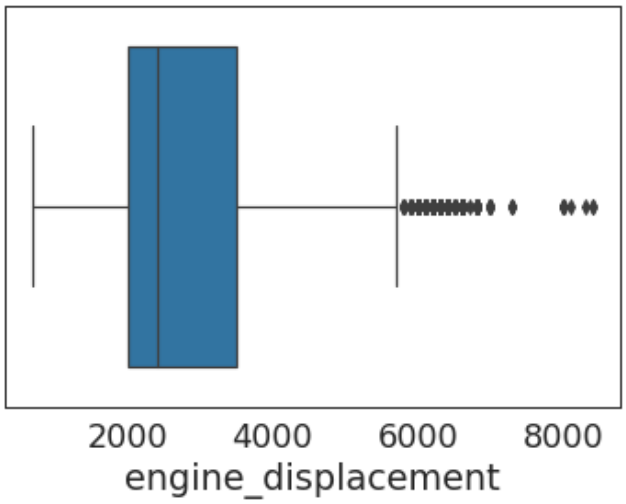


Figure (39)

After standardizing the feature, these statistics are obtained; mean (-1.9 e-16), standard deviation (1.0), minimum value (-1.9), maximum value (5.80), first quartile (-0.59), median (-0.19),third quartile (0.90),and the histogram of the distribution is shown in figure (40).

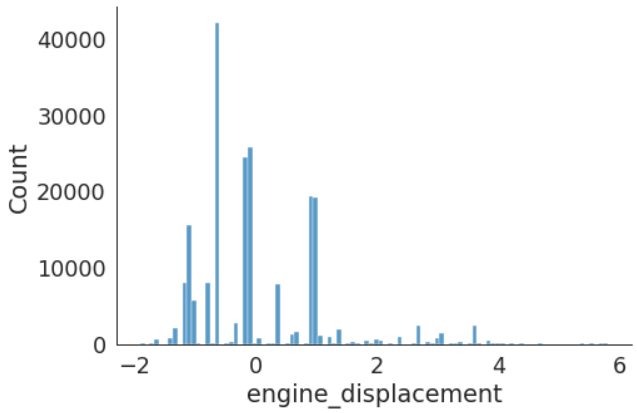


Figure (40)

The feature ***highway_fuel_economy*** has 23620 missing values which are to be replaced by the median (30.0), as it has several outliers as shown in figure (41). The correlation coefficient between ***highway_fuel_economy*** and ***price*** demonstrates a negative relationship (-0.19).

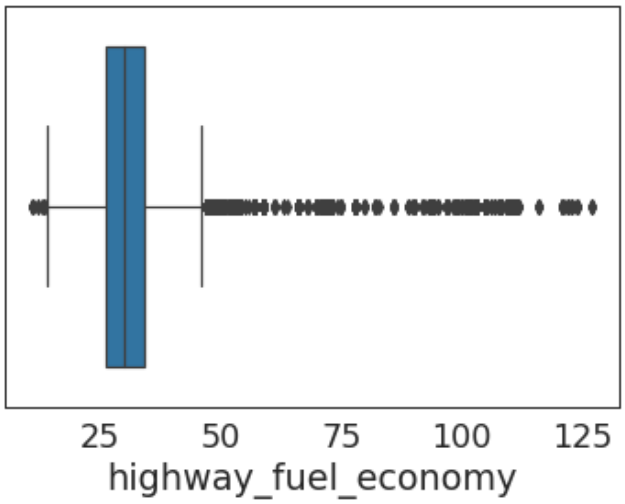


Figure (41)

After standardizing the feature, the important statistics are as follows; mean (-2.3 e-16), standard deviation (1.0), minimum value (-2.7), maximum value (13.37), first quartile (-0.48), median (-0.06), third quartile (0.34),and the histogram of the distribution is shown in figure (42).

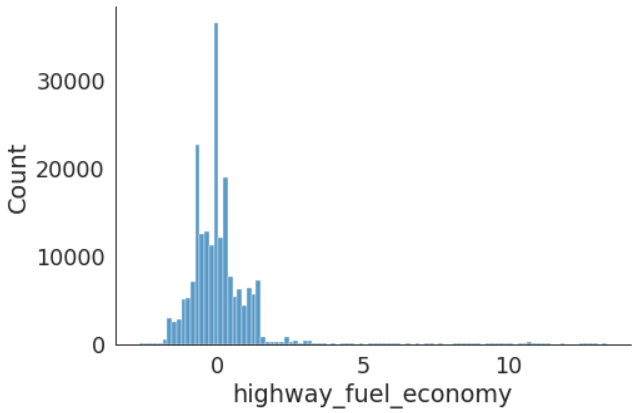


Figure (42)

One could argue that the data might be roughly described using the Lognorm, Beta, Gamma and Normal distribution as evident in figure (43).

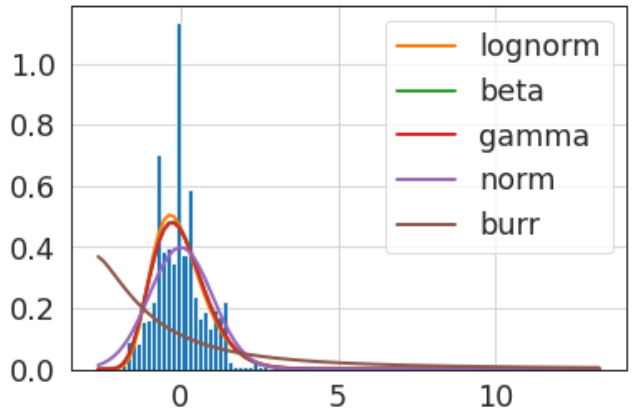


Figure (43)

The feature ***horsepower*** has 8643 missing values which are to be replaced by the median (201.0), as it has several outliers as shown in figure (44). The correlation coefficient between ***horsepower*** and price demonstrates a positive relationship (0.58).

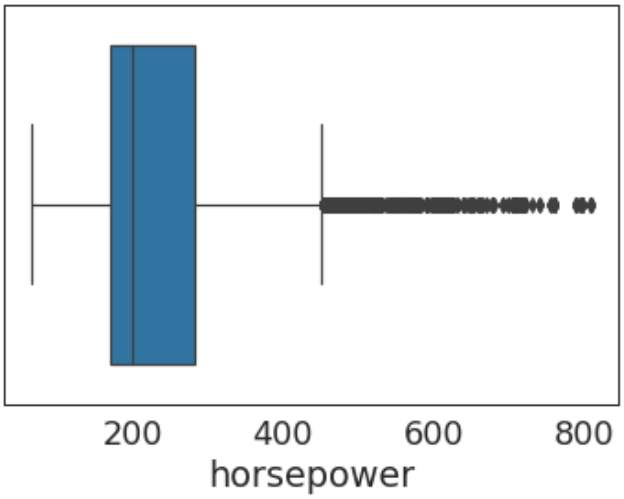


Figure (44)

After standardizing the feature, the important statistics are as follows; mean (-3.3e-17), standard deviation (1.0), minimum value (-2.02), maximum value (7.38), first quartile (-0.7), median (-0.32), third quartile (0.69),and the histogram of the distribution is shown in figure (45).

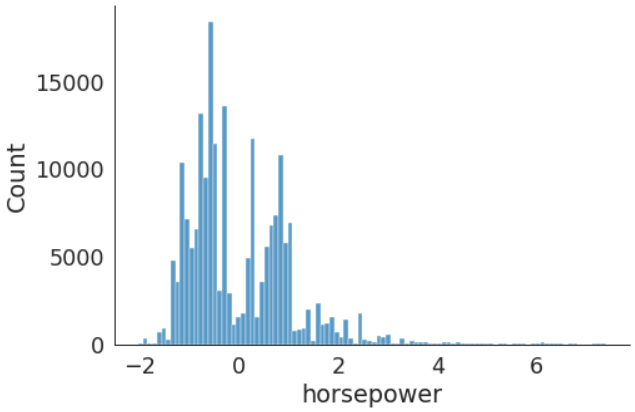


Figure (45)

One could argue that the data might be roughly described using the Lognorm, Burr Beta, Gamma and Normal distribution as evident in figure (46).

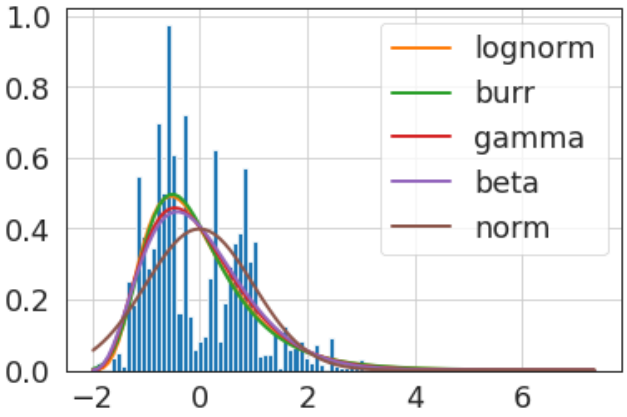


Figure (46)

The feature mileage has 8643 missing values which are to be replaced by the median (12146.0), as it has several outliers as shown in figure (47). The correlation coefficient between highway_fuel_economy and prince demonstrates a positive relationship (-0.45).

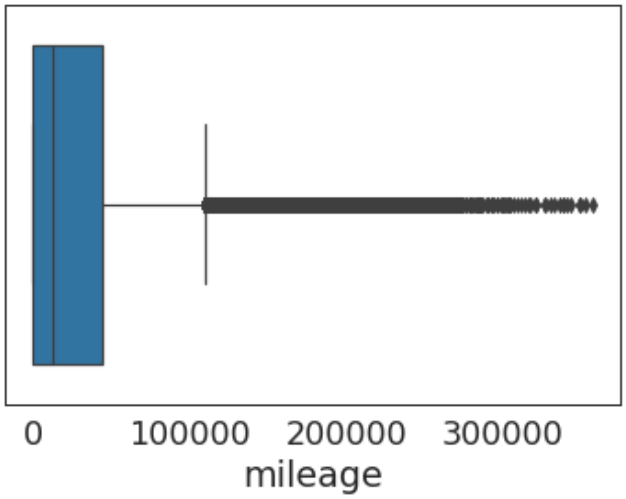


Figure (47)

After standardizing the feature, the important statistics are as follows; mean (1.95e-17), standard deviation (0.99), minimum value (-0.7), maximum value (7.6), first quartile (-0.7), median (-0.4), third quartile (0.27),and the histogram of the distribution is shown in figure (48).

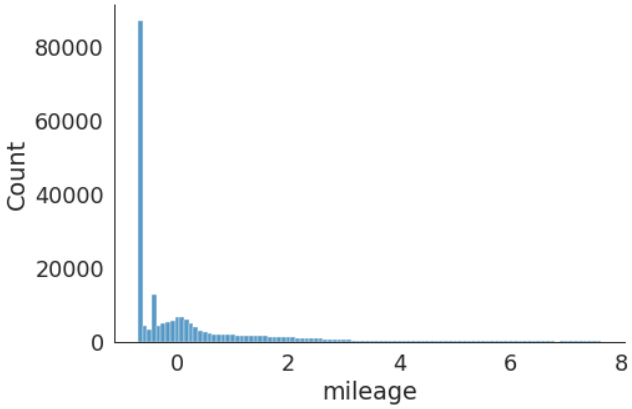


Figure (48)

As mentioned earlier, the feature ***power*** was captured in two columns ***power_in_horses*** and ***RPM***. Although each will be analyzed separately, it is natural that they both have the same number of missing values which is 29634.

The missing values for the feature ***power_in_horses*** are to be replaced by the median (208.0), as it has several outliers as shown in figure (49). The correlation coefficient between ***power_in_horses*** and ***price*** demonstrates a positive relationship (0.53).

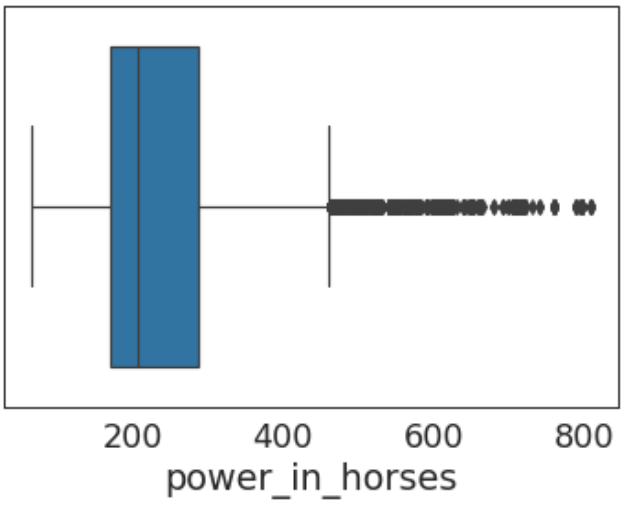


Figure (49)

After standardizing the feature, the important statistics are as follows; mean (-5.5e-17), standard deviation (1.0), minimum value (-2.12), maximum value (7.64), first quartile (-0.7), median (-0.26), third quartile (0.68),and the histogram of the distribution is shown in figure (50).

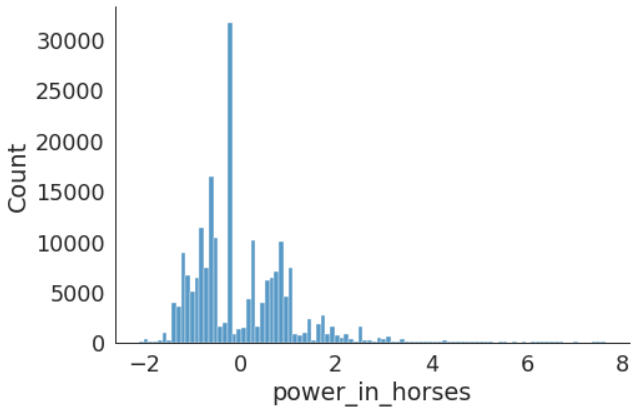


Figure (50)

Regarding the missing values for the feature ***RPM***, they are to be replaced by the median (208.0), as it has several outliers as shown in figure (51). The correlation coefficient between ***RPM*** and ***price*** demonstrates a negative relationship (-0.07).

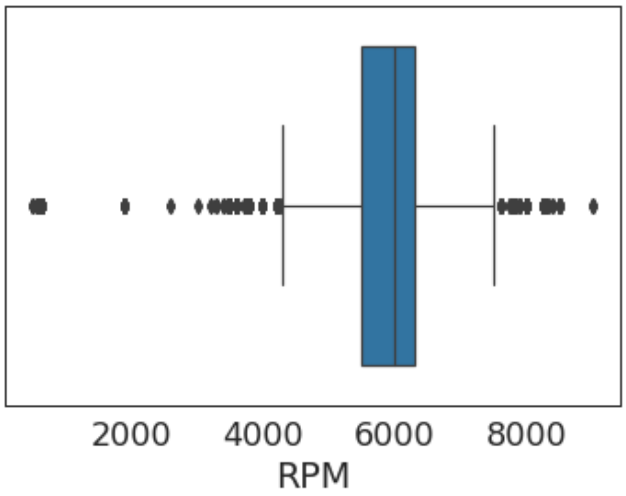


Figure (51)

After standardizing the feature, the important statistics are as follows; mean (-4.05e-16), standard deviation (1.0), minimum value (-9.23), maximum value (5.34), first quartile (-0.49), median (0.18), third quartile (0.53),and the histogram of the distribution is shown in figure (52).

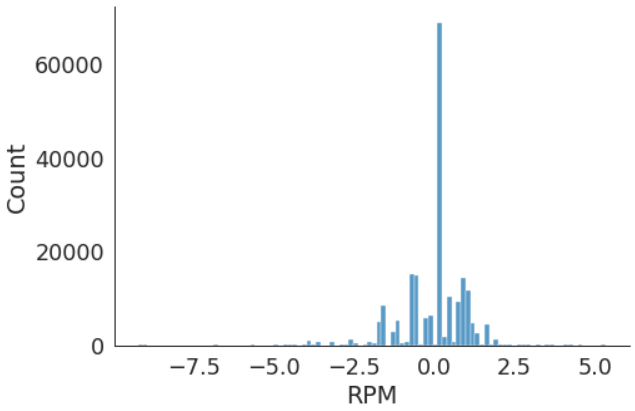


Figure (52)

It is salient to note that the feature power was dropped as all its data was properly captured in two features (***power_in_horses*** and ***RPM***).

Concerning the feature ***torque, it*** was captured in two columns: ***torque_ib_ft*** and ***torque_RPM***. Although each will be analyzed separately, it is natural that they both have the same number of missing values which is 32563.

The missing values for the feature ***torque_ib_ft*** are to be replaced by the median (236.0), as it has several outliers as shown in figure (52). The correlation coefficient between ***torque_ib_ft*** and ***price*** demonstrates a positive relationship (0.54).

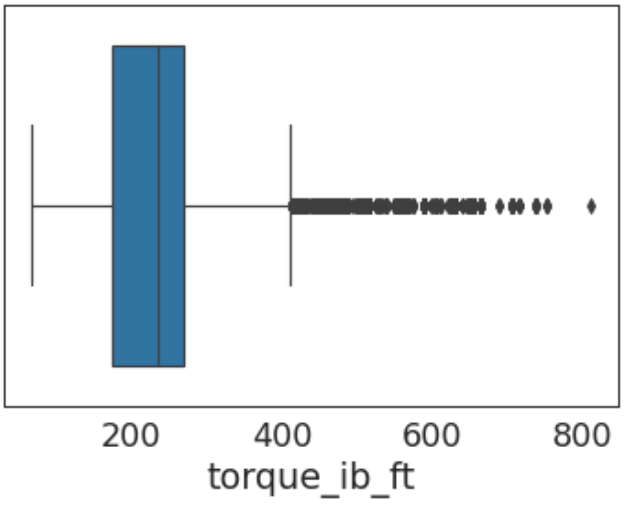


Figure (52)

After standardizing the feature, the important statistics are as follows; mean (1.6e-16), standard deviation (0.99), minimum value (-2.09), maximum value (7.29), first quartile (-0.7), median (0.027), third quartile (0.36),and the histogram of the distribution is shown in figure (53).

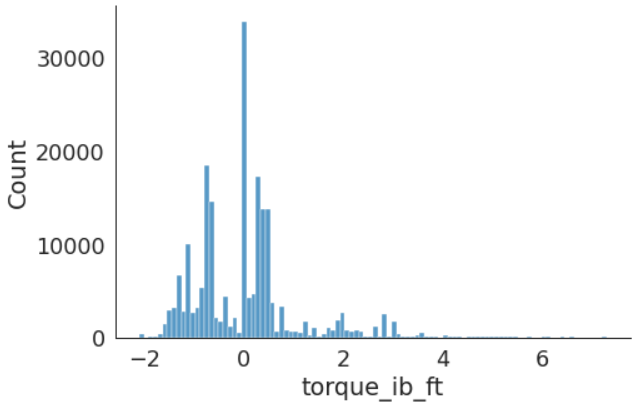


Figure (53)

The missing values for the feature ***torque_RPM*** are to be replaced by the median (4000.0). It is important to note that ***torque_RPM*** does not contain any outliers, as shown in figure (54); however, to stay consistent and follow the same pattern of the data imputation of other features, the median will be the measure of center. The correlation coefficient between ***torque_RPM*** and ***price*** demonstrates a negative relationship (-0.14).

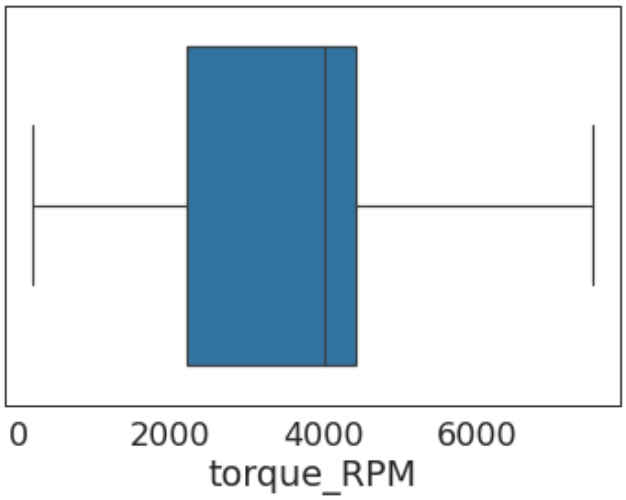


Figure (54)

It is intriguing to observe that the imputing missing data with median creates some outliers in the data, Figure (55). This is probably because of the bias created at the center of the data.

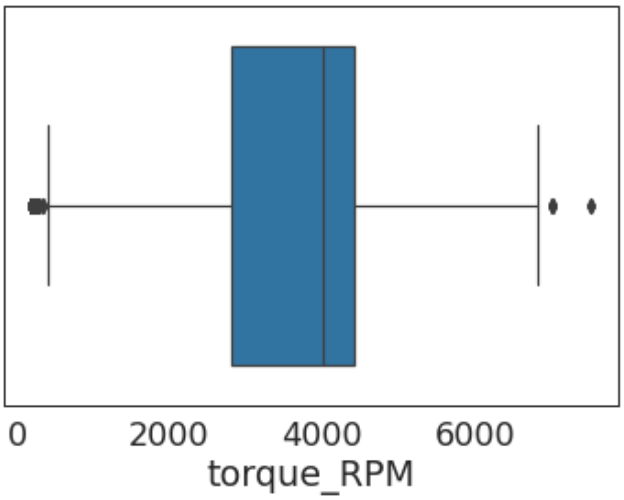


Figure (55)

After standardizing the feature, the important statistics are as follows; mean (1.4e-16), standard deviation (0.99), minimum value (-2.79), maximum value (3.27), first quartile (-0.6), median (0.36), third quartile (0.69),and the histogram of the distribution is shown in figure (55).

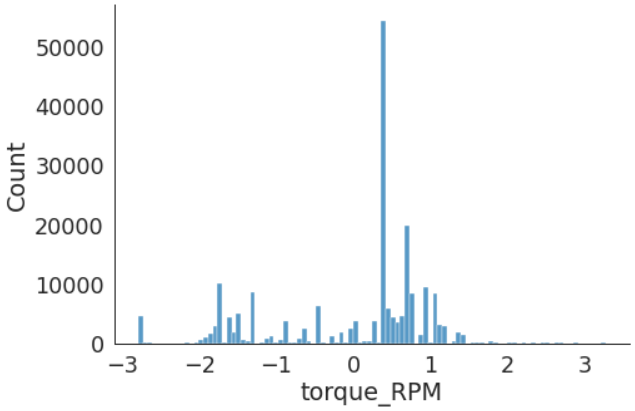


Figure (55)

As for the feature ***seller_rating***, it has 2706 missing which will be replaced by the median value (4.3). Using the median does not solve existing outlier problems or create any new ones. The correlation coefficient between ***seller_rating*** and ***price*** demonstrates a positive relationship 0.07. After standardization, some of the important statistics are; mean (1.56e-15), standard deviation (1.0), minimum value (-6.35), maximum value (1.42), first quartile (-0.52), median (0.14), third quartile (0.64), and the histogram of the distribution is shown in figure (56).

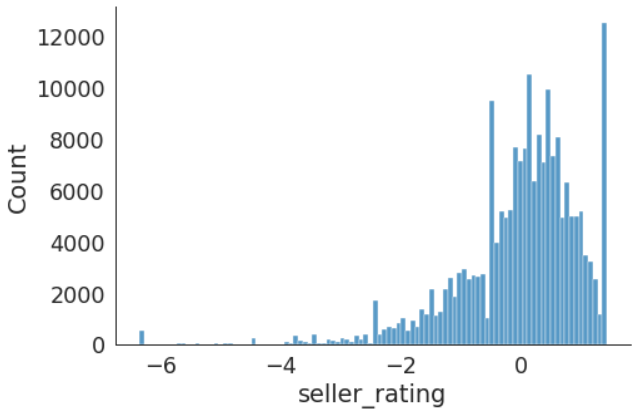


Figure (56)

One could argue that the data might be roughly described using the Beta, Burr, Normal, Lognorm, Gamma (skewed to the left), figure (57).

As for the feature ***city***, it carries a significant semantic importance as cities differ in standards of living and pricing. The feature ***city*** has 4188 unique values and no missing values for the sample at hand. As the number of unique values is extremely big, the top 100 mentioned or frequent cities will be left with their names while the other cities will be cited as "others". In this fashion, the needed one hot encoding will be done in a time and memory efficient manner.

Regarding the boolean feature ***fleet***, the number of cars that were not part of a fleet (83535) is bigger than the number of cars that were (24797). The number of missing values is extremely very high (91668) making 45% of the data. The researchers were unsure how to properly handle the aforementioned problem. The final decision reached is to conduct one hot encoding having; ***fleet nan***, ***fleet False***, ***fleet True***. This is an attempt and acknowledgment that not all observations have complete data and that the model built to predict prices has to capture this uncertainty.

As for the features ***engine_type*** and ***engine_cylinders***, the two features have the same exact values for each observation. Thus, one of them had to be removed (engine_cylinders). The number of unique values that ***engine_type*** can take are distributed in the following manner; I4 (111391),V6 (48606), V8 (9600), I4 Hybrid (5853), H4 (5307), I3 (4561),V6 Flex Fuel Vehicle (3080), I6 (2327), V8 Flex Fuel Vehicle (1308), I4 Flex Fuel Vehicle (780), I4 Diesel (542), H6 (368), I5 (346), V6 Hybrid (236), V6 Diesel (226), V12 (110), V10 (76), I2 (63), W12 (47), I6 Diesel (39), V8 Hybrid (10), H4 Hybrid (10), V8 Diesel (9), W12 Flex Fuel Vehicle (9), R2 (6), I4 Compressed Natural Gas (4), I6 Hybrid (3), I5 Diesel (2), W8 (1). Then, one hot encoding is conducted and the feature ***engine_type*** is dropped as all data is captured.

As for the feature ***exterior_color***, it has an extremely high number of unique values (7774). Hot encoding each and every unique value will result in a dataset with huge dimensionality. One way to solve this matter is to use the feature ***listing_color*** which assigns each car to a color that matches the dominant exterior color. The unique values of ***listing_color*** are 15 and distributed in the following manner; WHITE (39838), BLACK (39566), UNKNOWN (27667), SILVER (26971), GRAY (26865), BLUE (17094), RED (16197), GREEN (1758), BROWN (1519), ORANGE (867), GOLD (823), TEAL (427), YELLOW (293), PURPLE (99), PINK (16). The following step is to conduct one hot encoding and to drop both ***exterior_color*** and ***listing_color***.

The feature ***franchise_dealer*** is a boolean one that did not require any pre-processing as there were no missing values. The number of franchise dealers (160719) is significantly higher than the number of non franchise dealers (39281).

The features ***make_name*** and **model_name** have no missing values as well as 68 and 907 unique values, respectively. Hot encoding was conducted and the columns were dropped.

The feature ***franchise_make*** has 39,512 missing values and 47 unique values. Upon conducting one hot encoding the column was dropped.

The features ***transmission*** and ***transmission_display*** have the same number of missing values (4198) but different numbers of unique values. To elaborate, the number of unique values for ***transmission*** is 5 which is disturbed as follows; A (154348), CVT (36925), M (3634), Dual Clutch (895) and missing values (4198). As for the number of unique values for ***transmission_display*** it is 37. Hot encoding is conducted on both features and then they are dropped.

The features ***wheel_system*** and ***wheel_system_display*** have the same number of missing values and the same number of unique values. The unique values for ***wheel_system*** are distributed in the following manner; FWD (101039), AWD (56082), 4WD (20260), RWD (13093), 4X2 (2666) and missing values (6860). The unique values for ***wheel_system_display*** distributed in the following manner; Front-Wheel Drive (101039), All-Wheel Drive (56082),

Four-Wheel Drive (20260), Rear-Wheel Drive (130933), 4X2 (2666) and missing values (6860). As it is evident, the two features capture the same exact data but ***wheel_system*** uses abbreviation and ***wheel_system_display*** uses the full word. After thorough analysis, the two features are the same exact so to ease any future analysis the one with the abbreviation will be kept and ***wheel_system_display*** is dropped. Hot encoding will be applied to ***wheel_system***.

The feature ***year*** has no missing values and a positive correlation coefficient with ***price*** (0.36). The box plot of the data shows extreme outliers (figure 56) and the histogram (figure 57) of the standardized data shows that it could be best described by the following distribution; Normal and Burr (with skewness to the left).

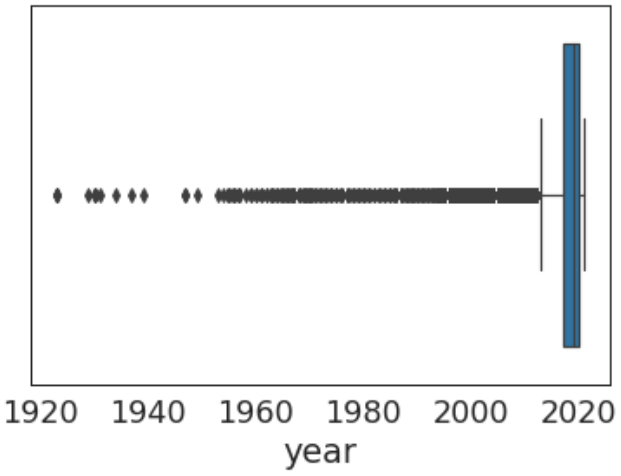


Figure (56)

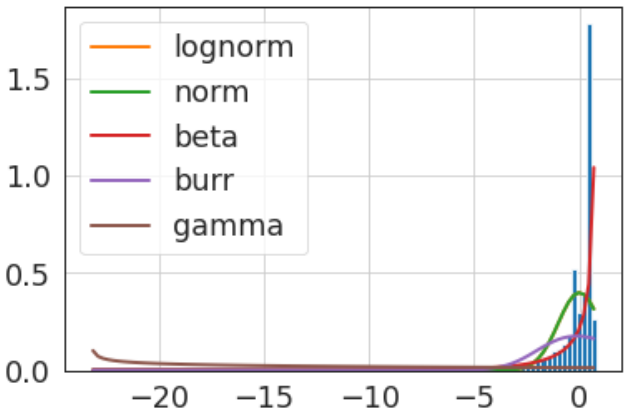


Figure (57)

On the other hand, the feature ***savings_amount*** shows skewness to the right and can be roughly described as a normal distribution, figure (58).

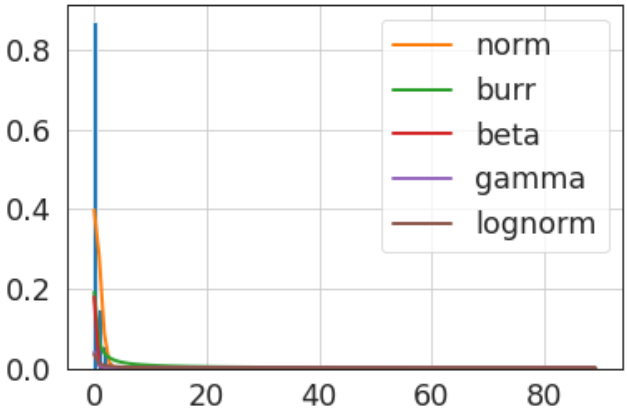


Figure (58)

Some of the important statistics are the standardized data of the feature ***savings_amount***; mean(-5.04), minimum value (-0.516), maximum value (89.6), first quartile (-0.51), median (-0.5) and third quartile (0.24).

As for ***listed_date,*** only the month and year were kept to reduce dimensionality. We treated the listed_date feature as a categorical feature. We used one hot encoding to treat the 74 unique values of this feature.

To extract the most out of the features ***longitude*** and ***latitude***, the distance from the selling location of the car to the center of the city they belong to. Using the Google API, the city's longitude and latitude will be obtained. Using the computed distance some statics and figures were computed; [ADD Figures and statistics].

The following features were removed for being redundant or irrelevant; ***vin***, ***listing_id***, ***dealer_zip.*** The first two are ids that do not contribute to the price, while enough information is given on the dealer and there is no need to keep ***dealer_zip.***

The feature ***is_new,*** is boolean and does not require any pre-processing as there are no missing values.

As for the features, ***frame_damaged***, ***isCab***, ***has_accidents***, ***fuel_type***, ***salvage*** , ***theft_title*** and ***interior_color*** the major pre-processing conducted is mainly hot encoding with accounts for missing values if they exist.

The feature ***major_options*** has 134 unique extra features, each extra feature will have a column and 1 indicates that the vehicle possesses this property.

Four-Wheel Drive (20260), Rear-Wheel Drive (130933), 4X2 (2666) and missing values (6860). As it is evident, the two features capture the same exact data but ***wheel_system*** uses abbreviation and ***wheel_system_display*** uses the full word. After thorough analysis, the two features are the same exact so to ease any future analysis the one with the abbreviation will be kept and ***wheel_system_display*** is dropped. Hot encoding will be applied to ***wheel_system***.

Owner_count feature: This feature had about 96737 missing values and about 12 unique values. We chose to assign the missing values to the mod of this data. Since this feature follows discrete distribution of data,we chose to treat this feature as a categorical feature instead of numeric and we used one hot encoding to deal with it

Salvage Feature: This feature is a boolean feature and it has 91668 missing values. Since the missing values are huge in this feature, we decided to use one hot encoding with this feature. The categories are Salvage True, Salvage False, Salvage Nan.

TrimID Feature: This feature is a categorical feature and it has about 17084 unique values and 4069 missing values. Because of the huge amount of unique values, we decided to include only the 100 values with the highest frequencies and the other values will be under others category. We used one hot encoding with this feature.

Trim Name Feature: This feature is a categorical feature and it has about 3892 unique values and 4632 missing values. Because of the huge amount of unique values, we decided to include only the 100 values with the highest frequencies and the other values will be under others category. We used one hot encoding with this feature.

SP_ID Feature: This feature is a categorical feature and it has about 22429 unique values and 7 missing values. Because of the huge amount of unique values, we decided to include only the 100 values with the highest frequencies and the other values will be under others category. We used one hot encoding with this feature.

SP_Name Feature: This feature is a categorical feature and it has about 21796 unique values and 0 missing values. Because of the huge amount of unique values, we decided to include only the 100 values with the highest frequencies and the other values will be under others category. We used one hot encoding with this feature.

Final List of chosen Features

(1) City: We decided to keep the city feature since cars prices differ from a city to another and also the demand for different cars differ. (2) Back legroom: We chose the back legroom as we ended up with a correlation between the legroom feature and cars of different legrooms have different prices. (3) Front legroom: We will keep this feature as it might have an effect on the price. (4) Fuel tank volume: Fuel tank volume can greatly impact the price of a car. So, we decided to keep this feature. (5) Height: Since height can affect the price of the car as measured by the calculated correlation coefficient, we decided to keep it. (6) Length: We kept length because it can affect the prices of the cars. (7) Width: We decided to keep width as it can affect the prices as seen by the correlation coefficient. (8) City Fuel economy : Since the fuel consumption can affect prices, we decided to keep this feature. (9) Highway Fuel Economy: It is the same as feature 8 but it is the fuel consumption on highways. (10) Horsepower: We kept this feature as it can affect the prices. (11) Maximum Seating: We kept this feature as cars of different numbers of seats can affect the prices. (12) Wheel base feature: This feature measures the diameter of the wheel base. So, we decided to keep it as it can affect the prices and it showed a considerable correlation with the price. (13) Days on market: As this feature showed a correlation with the price, we kept it. (14) Engine Displacement: We will keep this feature as it showed a correlation with the price. (15) Horsepower: This feature showed the greatest correlation with the price. As a result, we will include it. (16) Mileage: This feature indicates the distance traveled by this car. So, it can affect the price of the cars so we kept it. (17) Power: This feature is composed of two parts which are the RPMs and the horses. We divided it into those two parts and kept it. (18) Torque: It was also composed of two parts which was the torque in Ib.Ft and the RPMs and both can affect the price so we divided this feature into those two parts. (19) Seller ratings: We kept this feature as the rating of a particular seller can affect the price of the cars. (20) Body Types: Since cars of different bodies can have different ranges of prices as shown by the previous graphs, we decided to keep this feature and preprocess it with one hot encoding. (21) Fleet: Since the fleet feature can be decisive in the cars price, we decided to keep it and to use one hot encoding to deal with the missing values by creating three categories which are: Fleet True, Fleet False, Fleet nan. (22) Engine Type: We kept this feature as the different engines have different prices (23) Listing color: Since the color might have an effect on the price, we decided to keep

this feature. (24) Interior color: Inner color of cars cause fluctuations in the prices. Thus, we kept this feature. (25) Franchise Dealer: Since whether the dealer is franchised or not can affect prices, we agreed to keep this feature. (26) Make Name: Since cars of different makes (BMW, Nissan, etc) have different prices, we decided to keep this feature. (27) Model name: Since some car models are more expensive than others, we decided to keep this feature. (28) Franchise make: We decided to keep this feature as it might have an effect on the prices. (29) Transmission: Cars have different transmissions such as Manual and Automatic. And transmission can be a decisive factor in the prices of the cars. Thus, we will keep this feature. (30) Transmission Display: We decided to keep this feature as it might have an effect on the prices. (31) Wheel system: This feature is kept as it can affect prices. (32) Wheel system Display: We chose to keep this feature since it might affect the prices. (33) Model year: It is usually the case that cars of recent model years have higher prices. Also, the calculated correlation value agrees with this. So, we will keep this feature. (34) Saving amount: We kept this feature as it can affect prices. (35) Frame damaged: This feature can affect the prices, so we kept it. (36) Has accidents: This feature can affect the prices, so we kept it. (37) IsCab: This feature can affect the prices, so we keep it in the data set. (38) Fuel type: We agreed to keep this feature as different fuel types can have different prices. (39) Owner count: We chose to keep this feature since it might affect the prices. (40) Salvage: We chose to keep this feature since it might affect the prices. (41) Theft Title: We kept this feature since it can affect prices (42) Trim name: this feature had a lot of unique values but we decided to keep it and to include the 100 values with the highest frequencies and the others we put them into a category named others. (43) Trim ID: this feature had a lot of unique values but we decided to keep it and to include the 100 values with the highest frequencies and the others we put them into a category named others. (44) Sp name: this feature had a lot of unique values but we decided to keep it and to include the 100 values with the highest frequencies and the others we put them into a category named others. (45) SP ID: this feature had a lot of unique values but we decided to keep it and to include the 100 values with the highest frequencies and the others we put them into a category named others. for the features 42-45, we kept them as they have an effect on the prices (46) Major options: We decided to include this feature as prices of cars differ with different numbers of major options. (47) Listing Date: We decided to include this feature as it may affect the prices.

Added Features:

(1)Distance from city center: We added this feature to make use of the latitude and longitude features as we used them to calculate the distance between the selling location of the cars and the center of city they belong to.

Removed Features:

(1)Latitude and (2) Longitude: Since we used those two features to add the distance from the center of city feature, we decided to remove them after getting the distance. (3) Exterior color: Since the exterior color feature had great dimensionality and the listing color feature included only the color group each value of the exterior color feature belonged to based on the description of the data set, we decided to remove it. (4) VIN: Since this feature was a unique value for each vehicle, we decided to remove it as it has no impact on the price. (5) Listing_ID: Since this feature was a unique value for each vehicle, we decided to remove as it has no impact on the price.