Ramakrishna Danda
DSC530-T301 Data Exploration and Analysis (2205-1)
Week 12: Term Project

Statistical/Hypothetical question:-

The higher the price of the car the higher the deprecation.

Outcome of EDA:

Up on close inspection, Mercedes- Benz make which is higher in price, in fact shown lower depreciation compared to other high-priced luxury cars like BMW , Jaguar and Porsche.

What do you feel was missed during *the analysis?*

Felt the size of the data is not large enough to decipher the permutations

Were there any variables you felt could have helped in the analysis?

Insurance rating would have helped to see the safety of the car & would play a role on deprecation aspect.

Were there any assumptions made you felt were incorrect?

The assumptions of higher priced luxury cars would depreciate lot higher than regular cars was found to be incorrect.

What challenges did you face, what did you not fully understand?

Log CDF was not as flat line as I thought it would be. Have to read it over and over to understand if the approach was correct and has to go back and forth to cross verify the understanding.

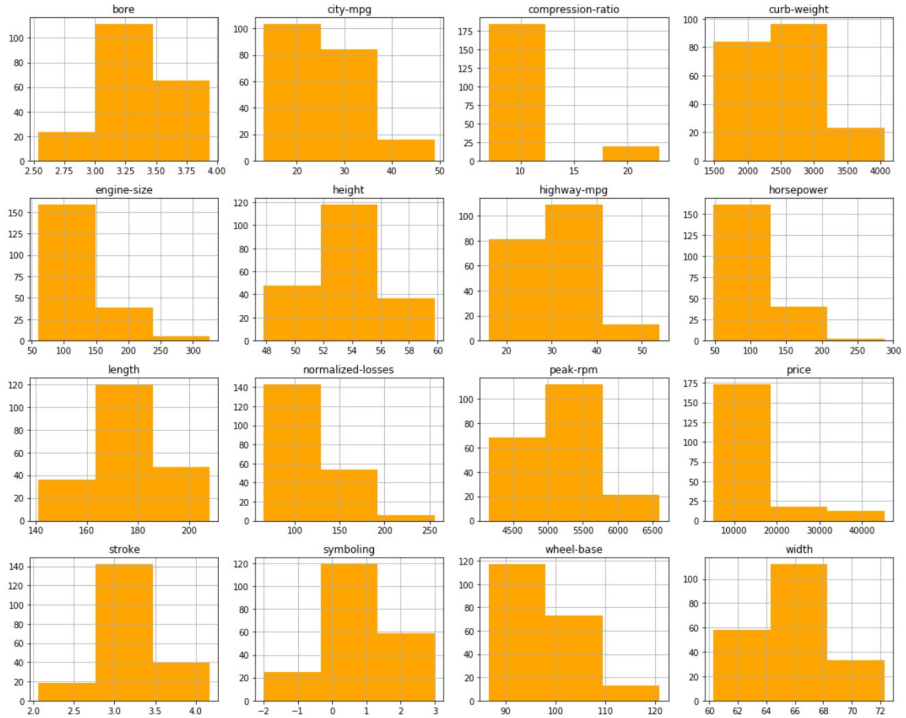
#A minimum of 5 variables in your dataset used during your analysis (for help with selecting, the author made his selection on page 6 of your book). Consider what you think could have an impact on your question – remember this is never perfect, so don't be worried if you miss one (Chapter 1).

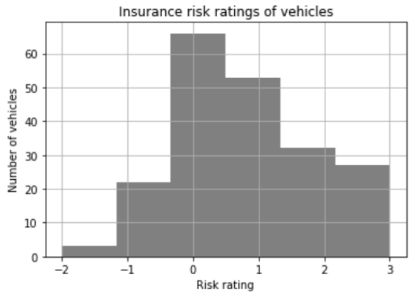
```
amdf.describe() #Show raw auto dataframe descriptions( count of rows, mean , SD and quartiles of data)
```

}]:

	symboling	wheel-base	length	width	height	curb-weight	engine-size	compression-ratio	city-mpg	highway-mpg
count	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000
mean	0.834146	98.756585	174.049268	65.907805	53.724878	2555.565854	126.907317	10.142537	25.219512	30.751220
std	1.245307	6.021776	12.337289	2.145204	2.443522	520.680204	41.642693	3.972040	6.542142	6.886443
min	-2.000000	86.600000	141.100000	60.300000	47.800000	1488.000000	61.000000	7.000000	13.000000	16.000000
25%	0.000000	94.500000	166.300000	64.100000	52.000000	2145.000000	97.000000	8.600000	19.000000	25.000000
50%	1.000000	97.000000	173.200000	65.500000	54.100000	2414.000000	120.000000	9.000000	24.000000	30.000000
75%	2.000000	102.400000	183.100000	66.900000	55.500000	2935.000000	141.000000	9.400000	30.000000	34.000000
max	3.000000	120.900000	208.100000	72.300000	59.800000	4066.000000	326.000000	23.000000	49.000000	54.000000

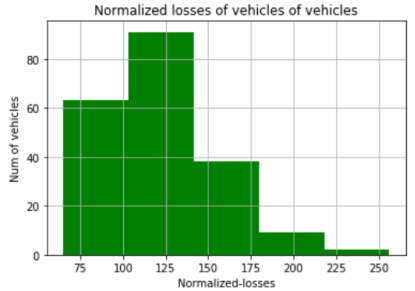
*** #Include a histogram of each of the 5 variables – in your summary and analysis, identify any outliers and explain the reasoning for them being outliers and how you believe they should be handled (Chapter 2).





Observation:-

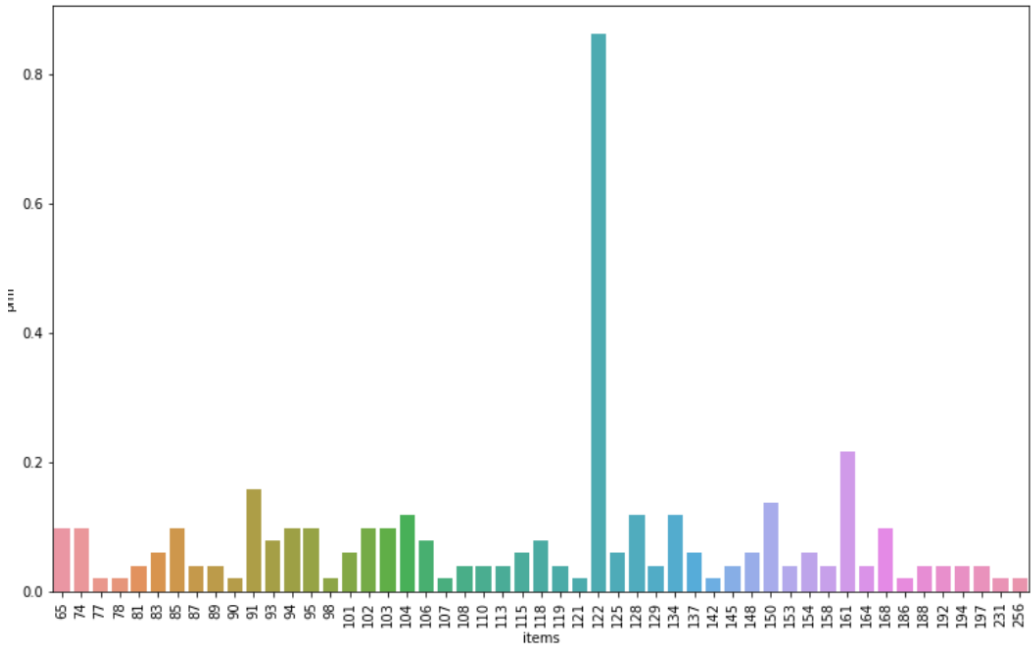
There are more cars in the range of 0 and 1 than other distributions

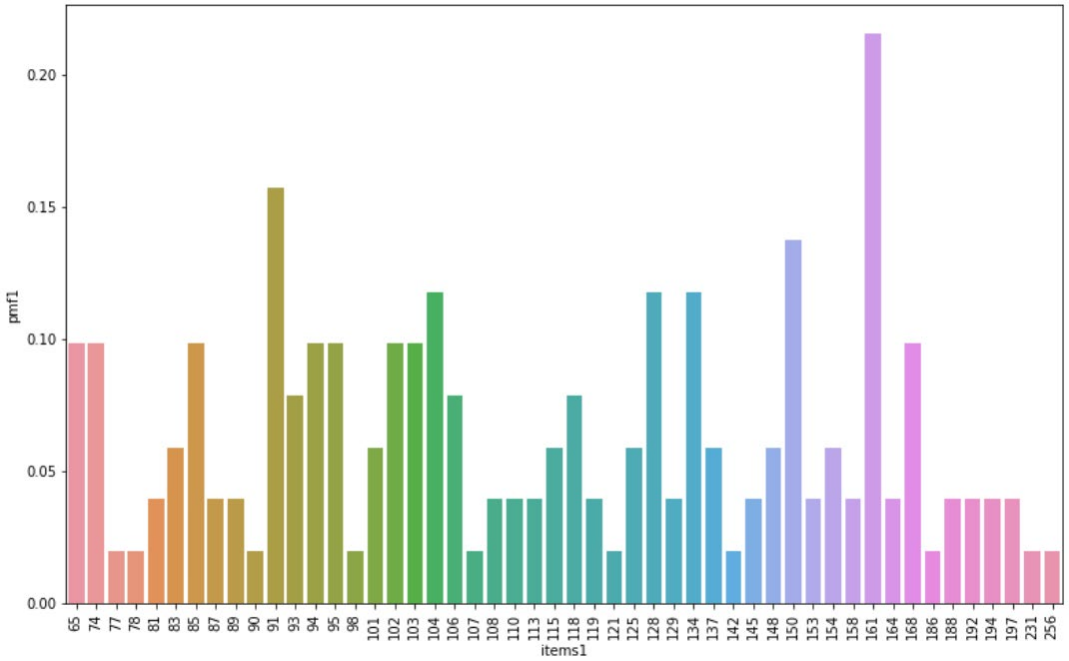


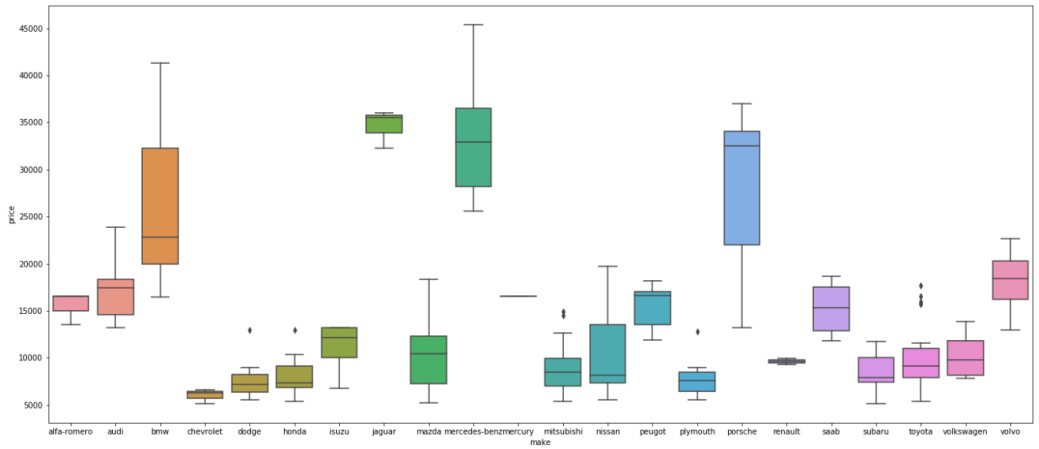
Observation:-

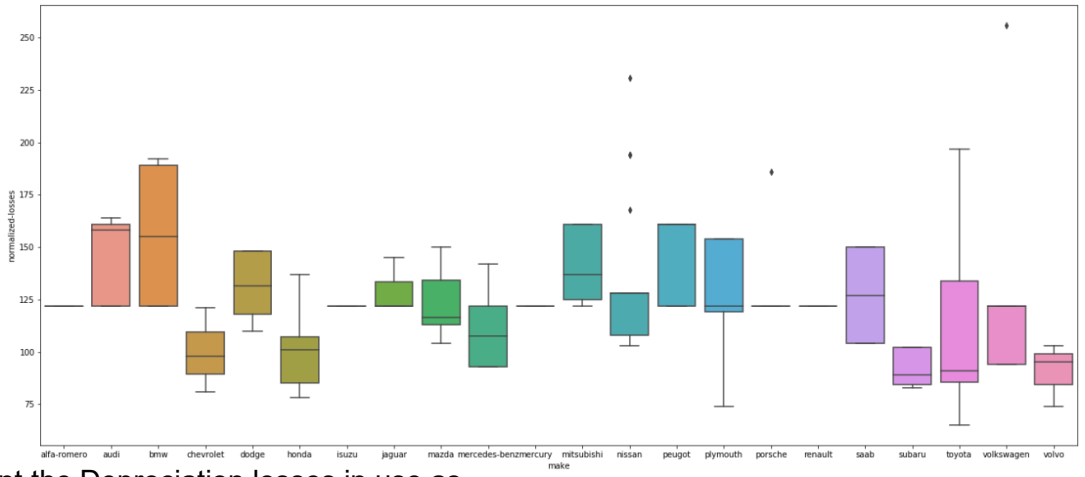
more vehciles in the range of 100 to 150 than any other bucket

*** #Using pg. 29 of your text as an example, compare two scenarios in your data using a PMF. Reminder, this isn't comparing two variables against each other – it is the same variable, but a different scenario. Almost like a filter.

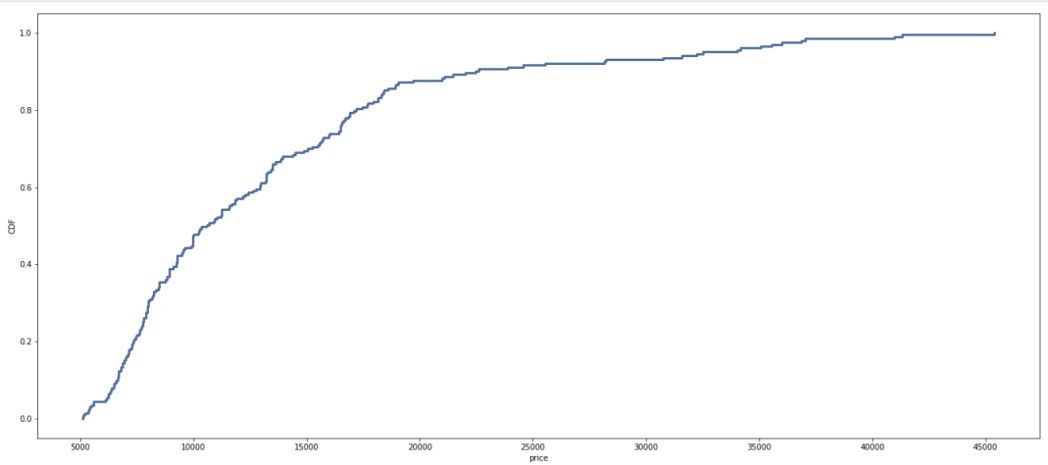








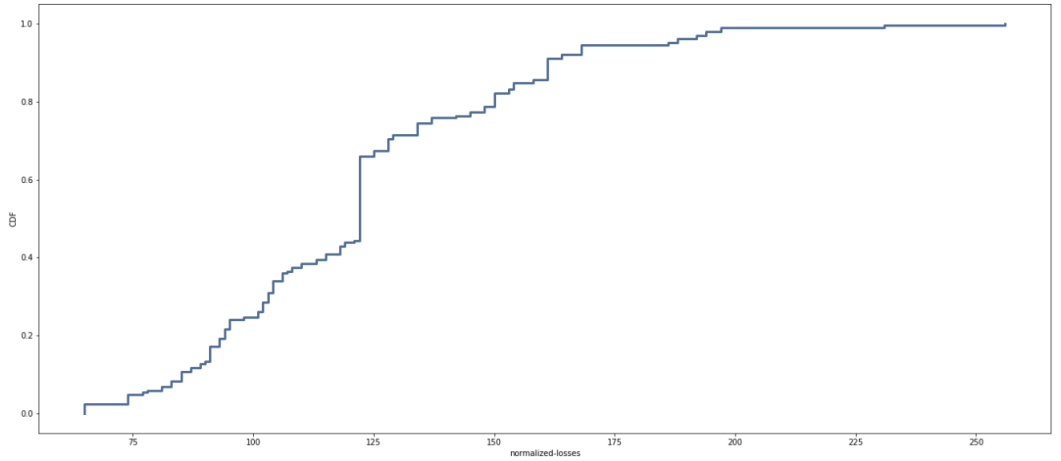
Observation:-From the above two box plots, it is evident the Depreciation losses in use as compared to other cars are high for BMW,AUDI, Mistibushi ,SAAB and Peugot.¶
while Toyota , Honda, Chevrolet and volvo are depreciated quite less in the entire lot of cars.
From price of the car and to depreciation value, Mercedez-Benz wins well in high cost luxary cars



<Figure size 576x432 with 0 Axes>

Based on Price CDF About 90% of car are below 20000, whcih is much higher than the mean and median values (13241.9 & 10595)

and the last 10% is above 40000 prices

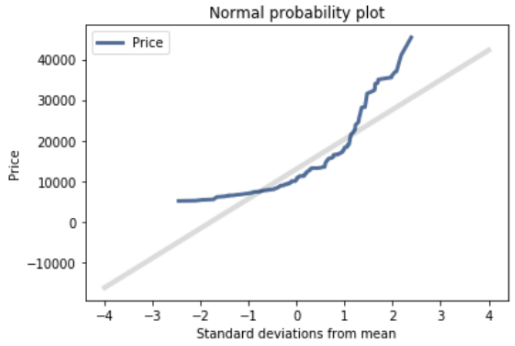


<Figure size 576x432 with 0 Axes>

Based on CDF of normalized loses, it is evident Losses does not deviate much post 122 and remaind there for most part of 50-70% of cars life

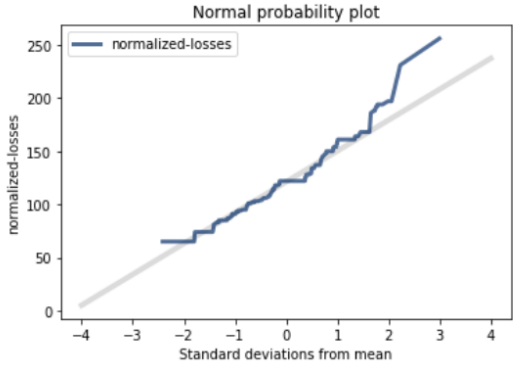
Best bang for the money is buying a used car right about 45% past the life of the car, there on the depreciation is not much

*** #Plot 1 analytical distribution and provide your analysis on how it applies to the dataset you have chosen (Chapter 5).



Observation:-

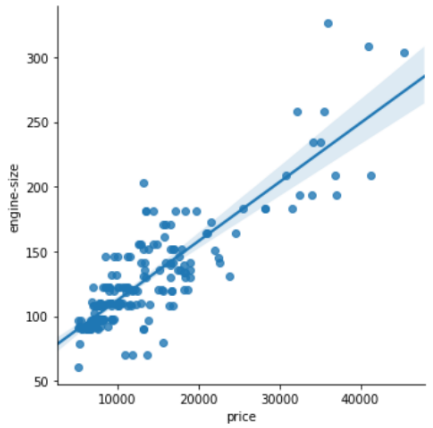
The normal probability plot shows lower end of car prices are cheaper than normal mode and cars are pricier at the higher end



Observation:-

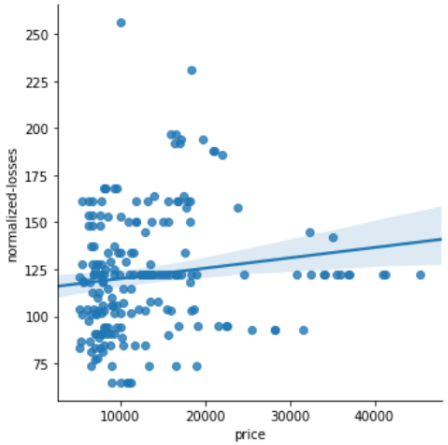
Depreciation is higher as the higher end luxary cars, which deviates from Normal distributions

*** #Create two scatter plots comparing two variables and provide your analysis on correlation and causation.
#Remember, covariance, Pearson's correlation, and Non-Linear Relationships should also be considered during your analysis (Chapter 7).



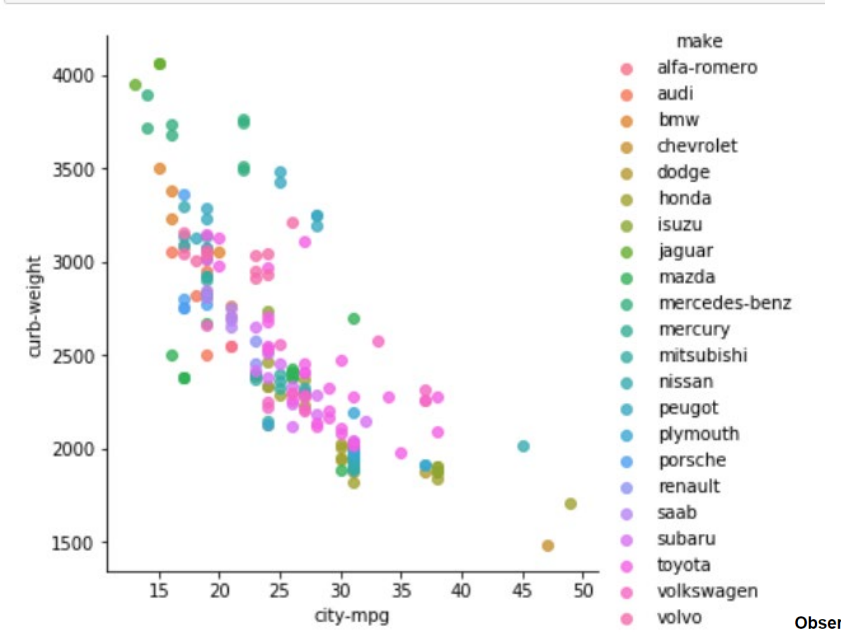
Observation:-

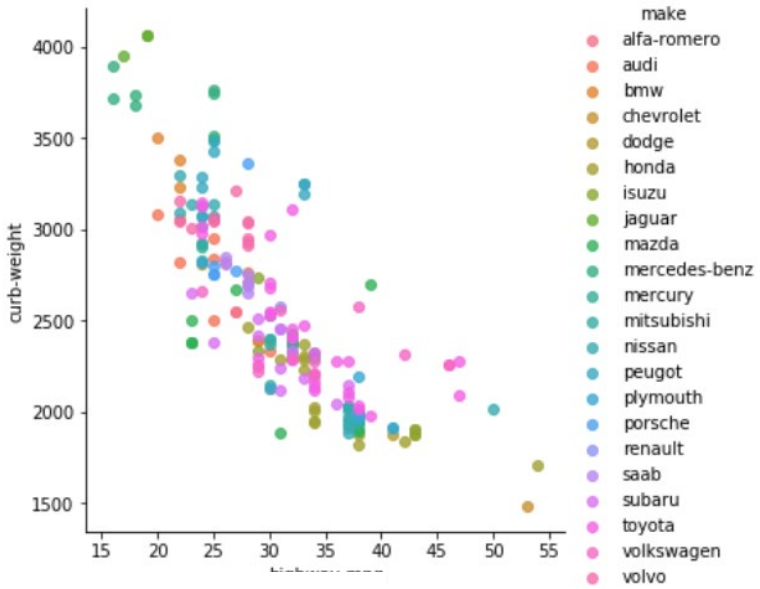
the higher the price, the higher the horsepower (linear)



Observation:-

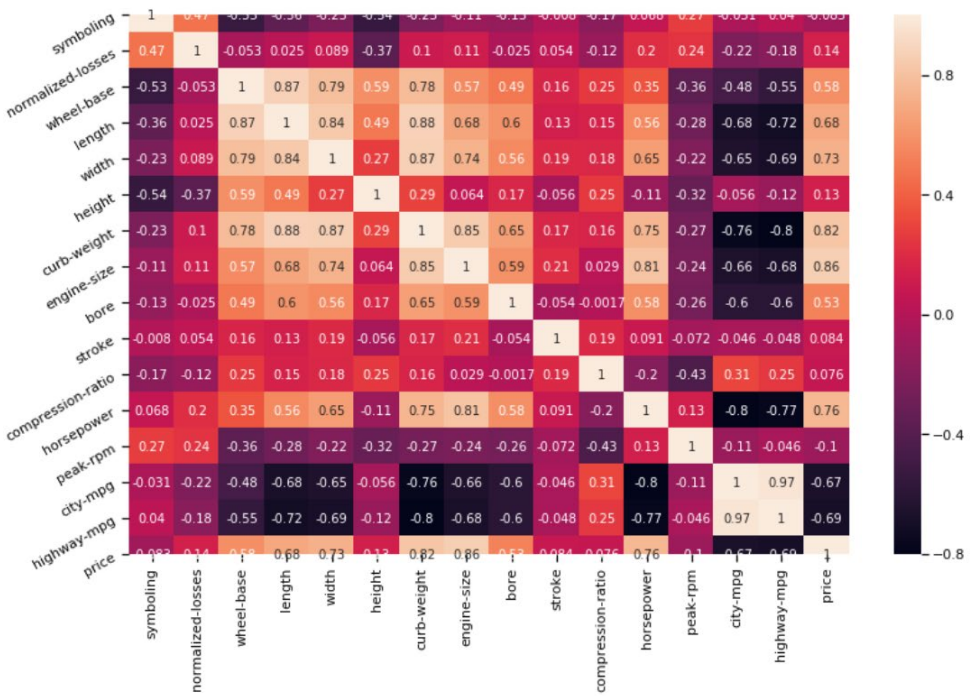
No linear relationship, but it appears deprecations for higher end cars remain flat after 25000 price treshold

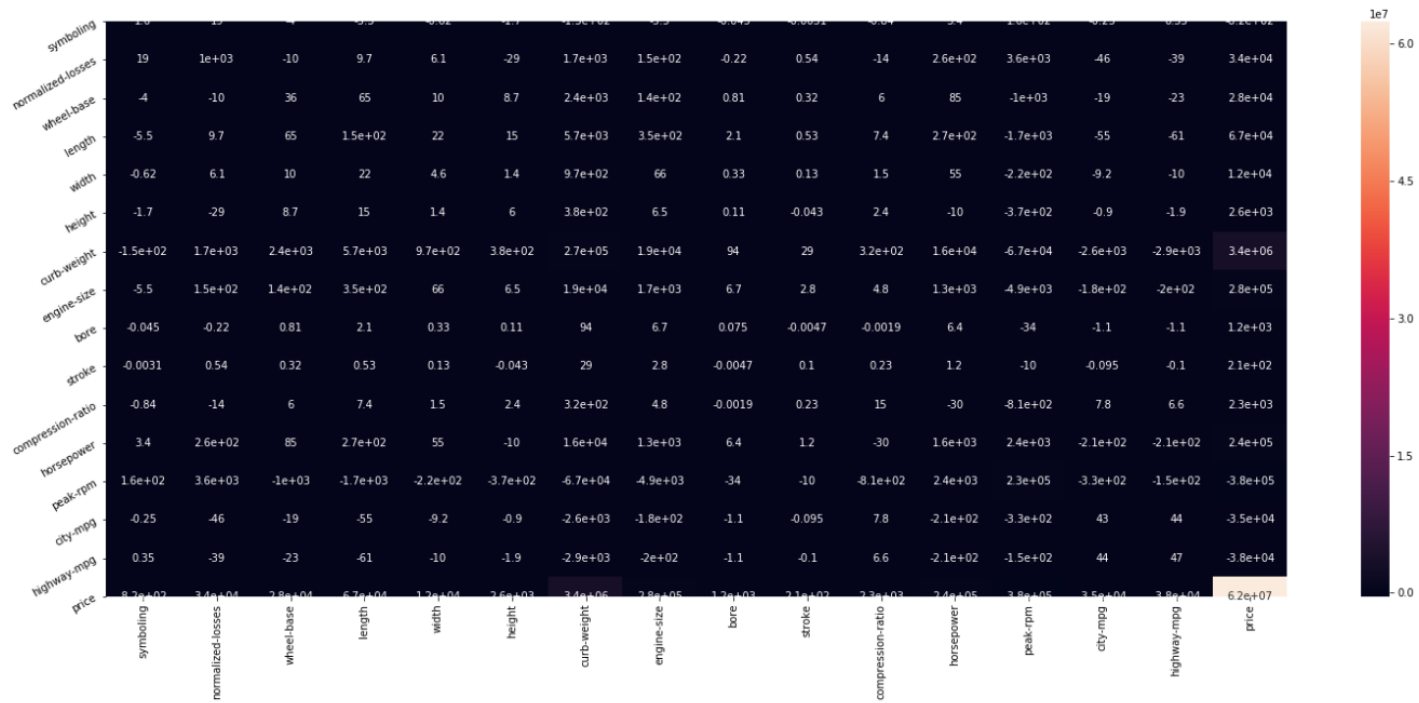




Observation:-

Both city and highway miles are effected by curb weight of the car





Observation:-

Price is more correlated with engine size and curb weight of the car along with price(0.86 and 0.82)

Wheel base is correlated with length and width of the car (0.87 and 0.79)

Millage is inversely corelated to horsepower

*** #Conduct a test on your hypothesis using one of the methods covered in Chapter 9.

Hypothisis is front wheel drive vehicles give higher millage than rare wheel drive and all wheel drive

So, Null Hypothisis is front wheel drive gives less millage than rare wheel and all wheel drive

```
ht = DiffMeansPermute(data) #callong permutation function
pvalue = ht.PValue() #get the p-value
pvalue #display P value
```

```
0.015
```

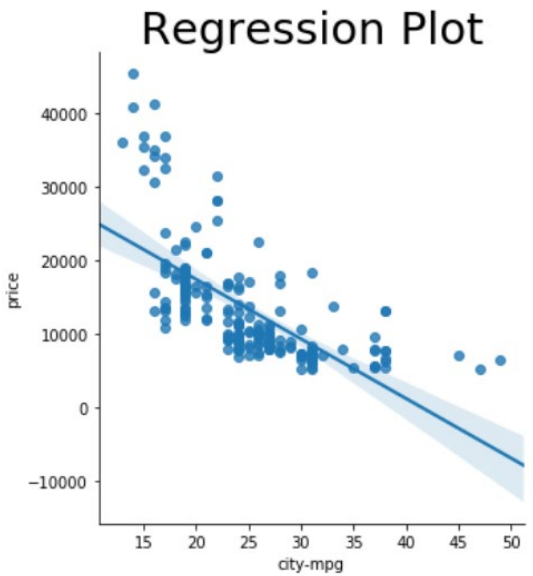
```
group_data=amdf[["drive-wheels","city-mpg"]].groupby(by=["drive-wheels"],as_index=False).mean() #group by data for drive whee
group_data
```

	drive-wheels	city-mpg
0	4wd	23.111111
1	fwd	28.288136
2	rwd	20.578947

Observation:-

from Permutation test above, it is evident that fwd gives less millage is proven right with very low P value. So, it did not happen by sampling issues and also looking at the actual values, front wheel drive has avg of 28.28 compared to 20.57 for rare wheel drive and 23.11 for all wheel drive.

*** #For this project, conduct a regression analysis on either one dependent and one explanatory variable, or multiple explanatory variables (Chapter 10 & 11).




```
results_formula.summary()
```

08]:

OLS Regression Results

Dep. Variable:	Q("price")	**R-squared:**	0.776
Model:	OLS	**Adj. R-squared:**	0.771
Method:	Least Squares	**F-statistic:**	171.2
Date:	Fri, 29 May 2020	**Prob (F-statistic):**	4.13e-63
Time:	23:28:24	**Log-Likelihood:**	-1957.6
No. Observations:	203	**AIC:**	3925.
Df Residuals:	198	**BIC:**	3942.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-9218.0489	4436.913	-2.078	0.039	-1.8e+04	-468.380
Q("horsepower")	18.1823	13.472	1.350	0.179	-8.384	44.748
Q("curb-weight")	3.8011	1.194	3.183	0.002	1.446	6.156
Q("engine-size")	101.0721	14.478	6.981	0.000	72.522	129.623
Q("highway-mpg")	-65.3384	74.212	-0.880	0.380	-211.686	81.010

Omnibus:	25.282	**Durbin-Watson:**	0.979
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	64.960
Skew:	0.512	**Prob(JB):**	7.84e-15
Kurtosis:	5.575	**Cond. No.**	4.38e+04

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.38e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Observation:-

R-squared: is 0.776, which is a strong indication of price to 'horsepower', 'curb-weight', 'engine-size', 'highway-mpg'

with highway mpg negative and intercept is statistically insignificant with a negative value and hight R-squared.

This data set consists of three types of entities:
(a) Technical specs of an auto in terms of various characteristics
(b) Automobile assigned insurance risk rating
(c) Normalized/Depreciation losses in use as compared to other cars
Symbolling rating: It is a degree to which autos are marked based on the risk and price index. It could go up and down. It’s also referred by actuaries/insurance auditors as "symbolling"
Normalized losses: Depreciation losses in use as compared to other cars
Make: Manufacturer name
fuel-type: Fuel type uses by auto. Gas or Diesel
Aspiration: Naturally aspirated engine or turbo charges/super charged
Num-of-doors: Number of doors to the automobile
Body-style: Sedan/Coupe/SUV/Hatchback/Hardtop...etc
Drive-wheels: Front wheel drive, rare wheel drive, all wheel drive
Engine-location: Location of engine
Wheelbase: Length of the car from front wheel to rare wheel
Length: Length of the car from bumper to bumper
Width: Width of the car side to side
Height: Hight of the car from ground to roof
Curb-weight: Empty vehicle weight with fuel
Engine-type: Engine type, dual overhead, single overhead cams, etc
Num-of-cylinders: Number of cylinders in the engine 4,6,8 or 12
Engine-size: Volume of air and fuel that's pushed through the engine by its cylinders fuel-system
Bore: Diameters of each cylinder
Stroke: stroke is the length that it travels when moving from bottom position to the top position.
Compression-ratio: Ratio of the volume of the cylinder and the combustion chamber when the piston is at the bottom, and the volume of the combustion chamber when the piston is at the top.
Horsepower: The power an engine produces
Peak-rpm: Rotation per min of the engine at its peak
City-mpg: Millage per gallon in city driving conditions
Highway-mpg: Millage per gallon in highway driving conditions
Price: Price of the vehicle

Raw data file

Microsoft Excel
ma Separated Valu