ST1009 - Exploratory Data Analysis

Descriptive Analysis of US Cars

Group 24

Group Members:

Madhumali J. T. D. - s15540
Madhuwantha H. H. N. - s15541
Maitipe A. P. - s15469
Malinda R. A. - s15724

Table of contents

page

Introduction

This report includes a descriptive analysis of a data set of cars for sale in USA, scraped from AUCTIONEXPORT.COM.

Following data set includes 11 variables. Those are,

· Price-selling price of the vehicle mentioned in the advertisement

· Brand-Brand of the car

· Model-model of the car

· Years-Registration year of the vehicle

· Title Status- - This feature included binary classification as clean title vehicles and salvage insurance

· Mileage- Miles traveled by vehicle

· Color- Color of the vehicle

· VIN-Vehicle identification number, a collection of 17 characters(numbers & capital letters)

· Lot- an identification number assigned to a particular quantity or lot of material from a single manufacturer. For cars, a lot number is combined with a serial number to form the Vehicle Identification Number.

· State/City-The location which the car is being available for purchase

· Condition-Time left

This report includes three main titles. They are

1. The Dataset
2. Descriptive Analysis
3. Conclusions

The title 'Data set' grabs more details about the data and each variable. Under the title 'Descriptive Analysis' there you can find the Mean, Median, Quartiles, details about the skewness etc. of each variable and corresponding charts are provided.

Finally, the conclusions section will give a clear idea about which brand and models are abundant, what kind of car is suitable for your budget, which are the cars with high condition, which one is worth its price and Finally overall which one is better for you to buy.

The Dataset

This data set includes information about 28 brands of clean and used vehicles for sale in the USA. There are 11 variables in the data set such as price, brand, model, years, lot, state, etc.

Brand, model, title_status, vin, state and condition variables consist of data in the text format.

The variable price shows the sale price of the vehicles. It is a quantitative data which is in numeric form. Brand of the car is given by the variable brand. Brand is a categorical variable that holds categorical data like toyota, bmw and ford. Variable model gives the model of the car and it also a categorical variable. The next variable, year, shows the vehicle registration year and we can identify the new and old vehicles from this. Title status gives data whether a certain vehicle is a clean vehicle or a salvage insurance vehicle. Variable mileage gives the data about how many miles traveled by a car and it is quantitative. We can see the vehicle identification number in the vin variable. It is a collection of 17 characters having both digits and simple letters. Also there is a variable called lot. This is an identification number assigned to a particular quantity. It is a qualitative variable. State gives the location in which the car is being available for selling. It is a qualitative variable. Variable condition implies the remaining time for the selling.

From all these variables anyone can get a clear idea to purchase a vehicle.

Descriptive Analysis

1) Price

Statistics

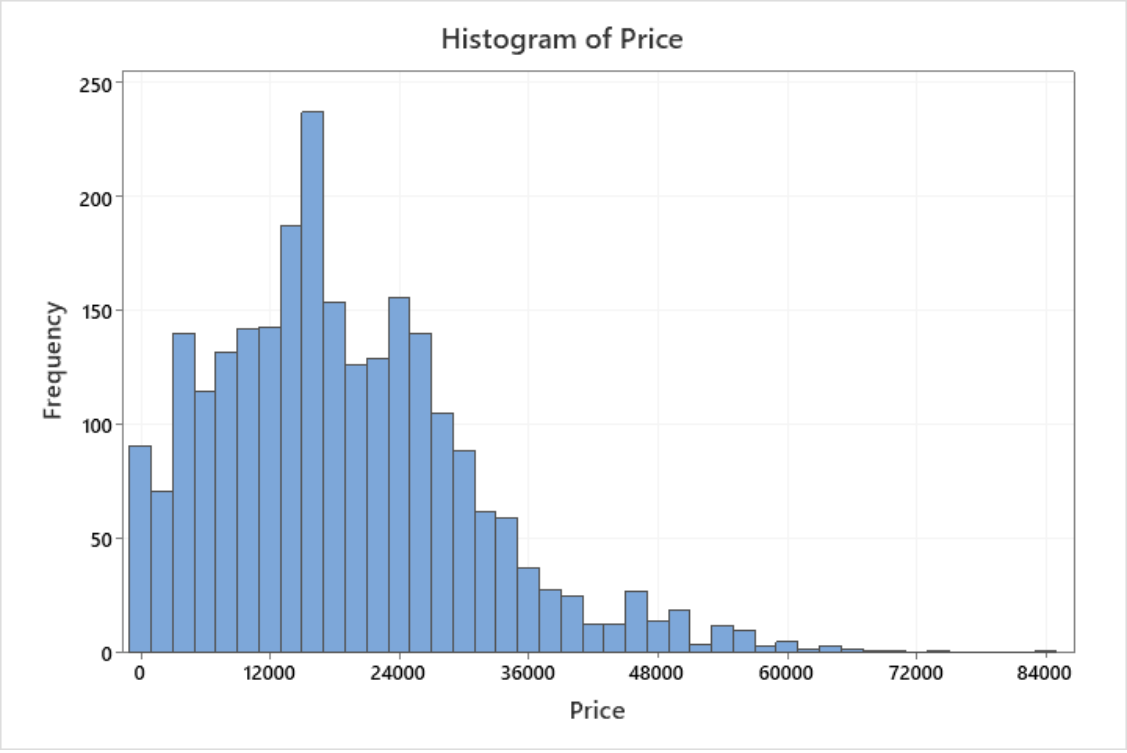
Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum	Skewness
Price	2499	0	18768	242	12116	0	10200	16900	25600	84900	0.92

The mean of the Price is 18768 and the median of the Price is 16900. There were no missing values in this variable. The Price ranges from 0 to 84900. The IQR is 15400.

Lower limit = Q1 – 1.5*IQR = 10200 - 1.5*15400 = -12900

Upper limit = Q3 + 1.5*IQR = 25600 + 1.5*15400 = 48700

As there are prices greater than 48700, there are outliers, and it is clearly visible from the histogram as well.

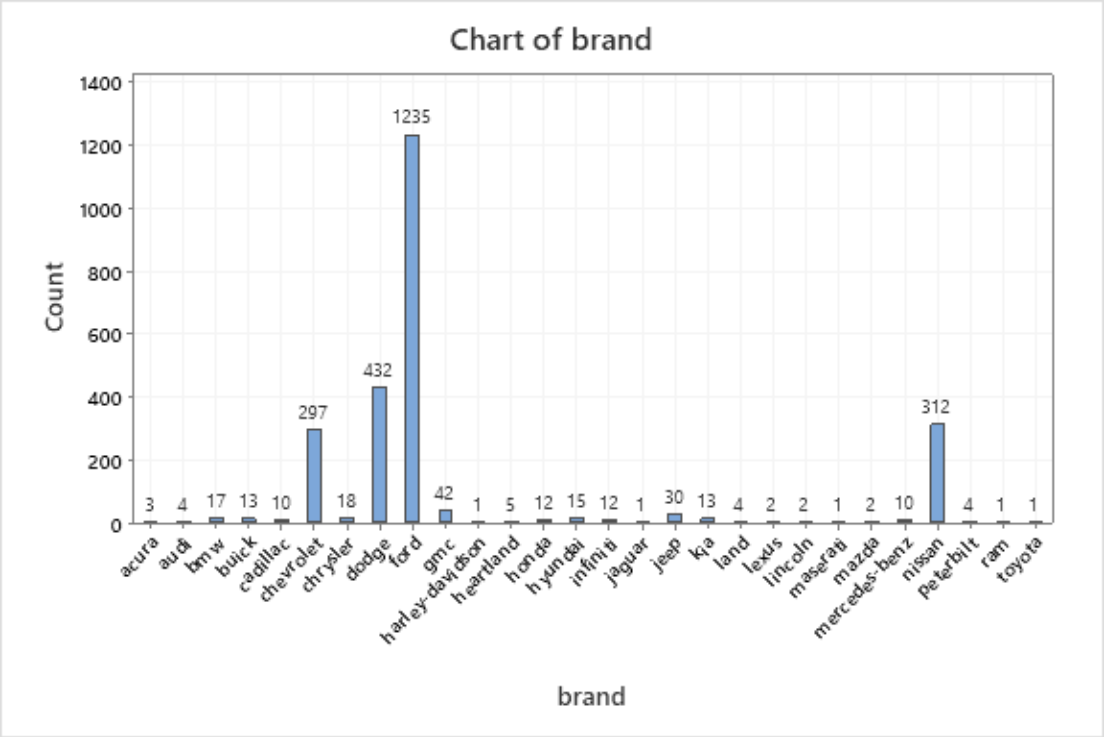


Mean of Price = 18 768

Median of Price = 16 900

Mean > Median, Therefore this is a Positively Skewed distribution.

2) Brands

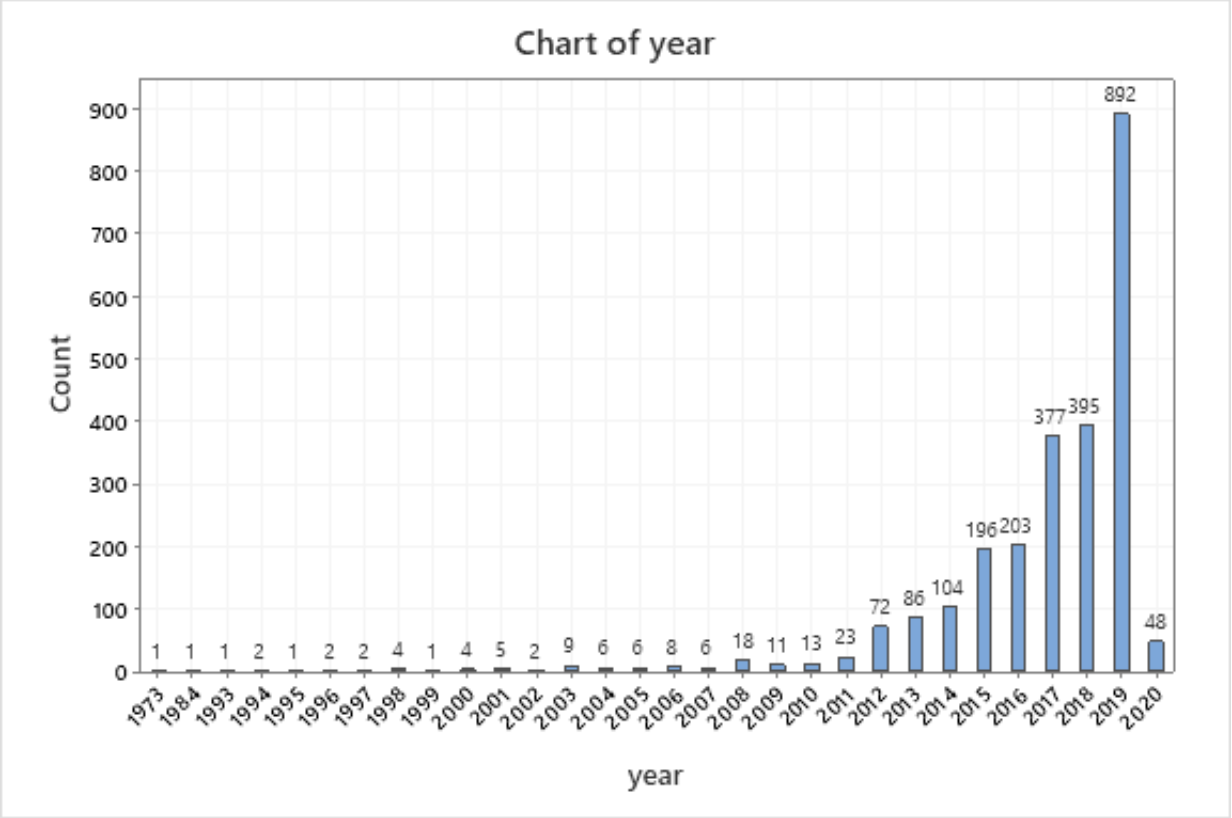


The graph depicts the brand of cars in the USA. Most popular brand is ford. Other than this Chevrolet, Dodge and Nissan brands are also popular. Considerable number of popular brands are Gmc, Jeep, BMW and chrysler. Other all brands are least popular in the USA.

3) Model

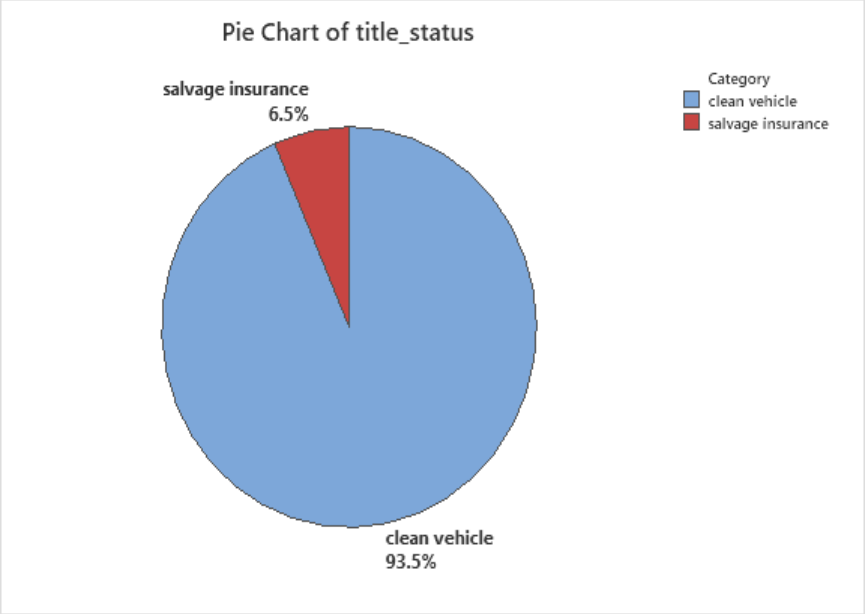
Tally table depicts the model of the cars. Most of the cars are door models. There are 651 vehicles in that model. The next largest number of model is f - 150. Considerable number of models are caravan, doors, durango, fusion and journey. Other brands are not as popular as the above brands. Therefore, door model cars are most popular in the USA.

4) Year



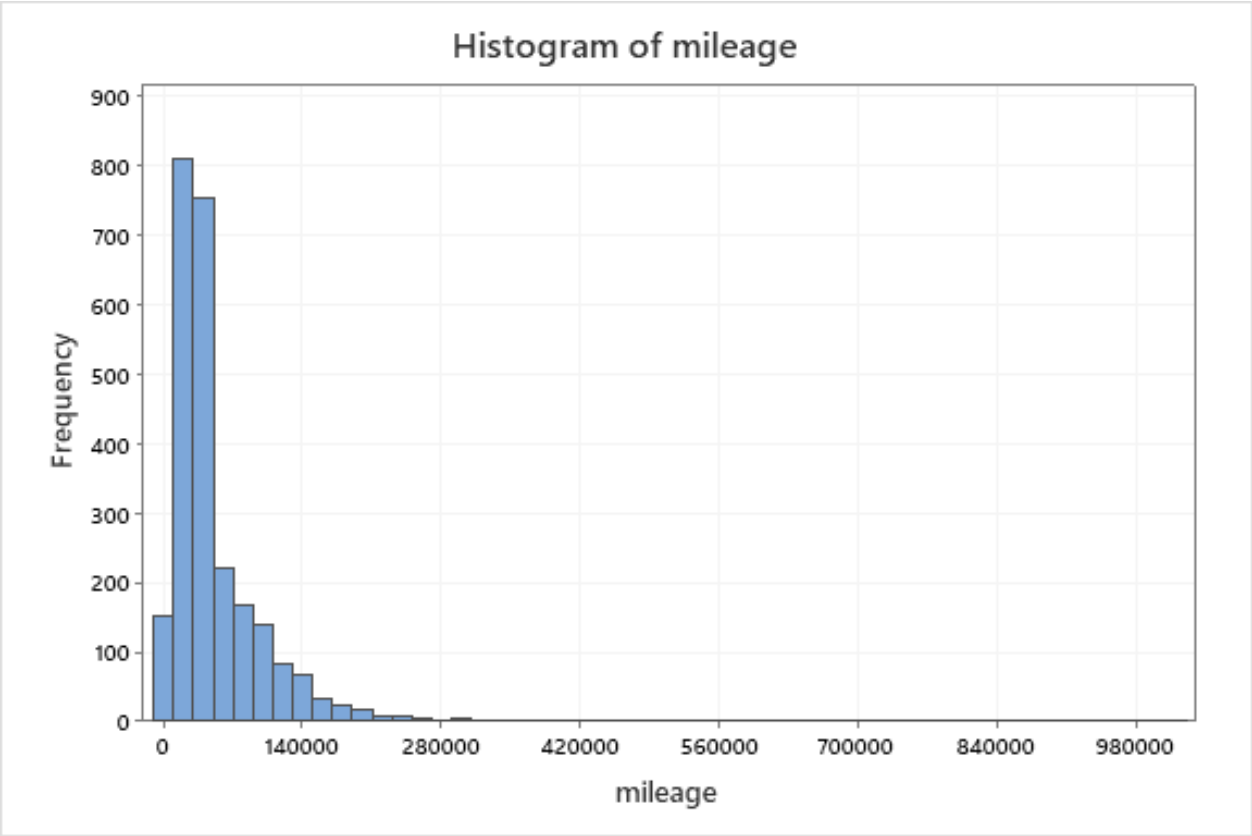
This graph shows vehicle registration year. Most vehicles register in 2019.There are 892 vehicles in that year. Other than this year, there are more vehicles registered in 2017 and 2018. Considerable number of vehicles were registered in 2012, 2013, 2014, 2015, 2016 and 2020. All of the other years have the least number of vehicles registered.

5) Title_Status



This feature included binary classification, which are clean title vehicles and salvage insurance. It seems from the chart that most cars are clean vehicles. This percentage is 93.5%. There are 6.5% salvage insurance vehicles.

6) Mileage



Statistics

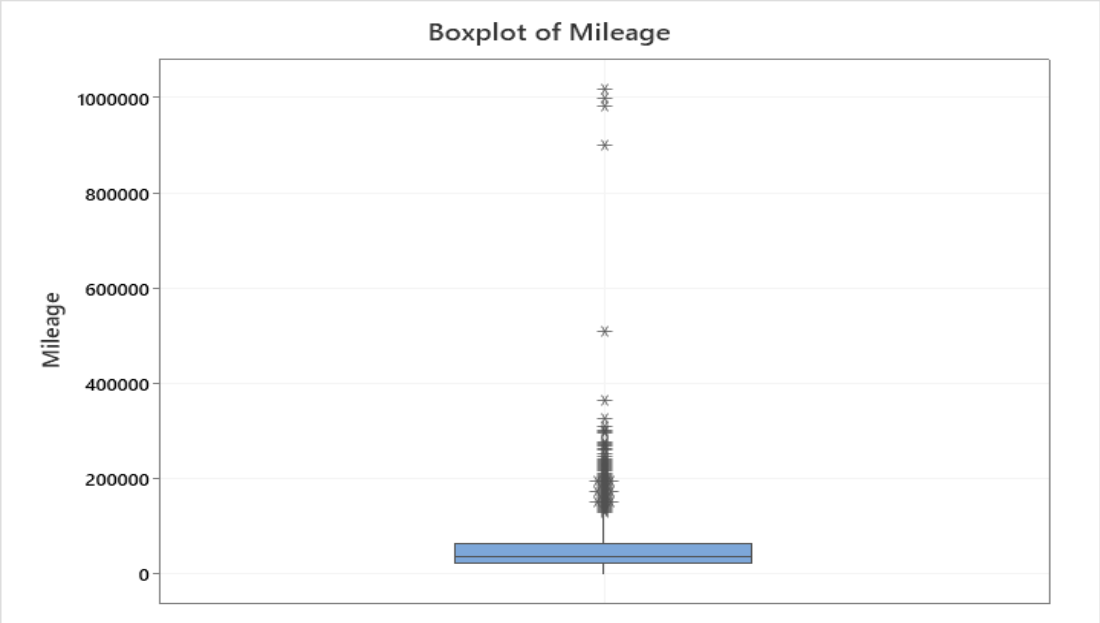
Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum	Skewness
Mileage	2499	0	52299	1194	59706	0	21406	35365	63500	1017936	7.08

The mean of the Mileage is 52299 and the median of the Mileage is 35365. There were no missing values in this variable. The Mileage ranges from 0 to 1017936.The IQR is 42094.

Lower limit = Q1 – 1.5*IQR = 21406 - 1.5*IQR = -41735

Upper limit = Q3 + 1.5*IQR = 63500 + 1.5*IQR = 126641

As there are Mileage greater than 126641, there are outliers, and it is clearly visible from the boxplot as well.



7) Color

Tally

color	Count
beige	5
billet silver metallic clearcoat	3
black	516
black clearcoat	2
blue	151
bright white clearcoat	2
brown	15
burgundy	1
cayenne red	2
charcoal	18
color:	5
competition orange	1
dark blue	1
glacier white	1
gold	19
gray	395
green	24
guard	1
ingot silver	1
ingot silver metallic	4
jazz blue pearlcoat	1
kona blue metallic	1
light blue	1
lightning blue	1
magnetic metallic	6
maroon	1
morningsky blue	1
no_color	61
off-white	2
orange	20
oxford white	4
pearl white	1
phantom black	1
purple	1
red	192
royal crimson metallic tinted clearcoat	1
ruby red	1
ruby red metallic tinted clearcoat	2
shadow black	5
silver	300
super black	3
tan	1
toreador red	1
triple yellow tri-coat	3
turquoise	1
tuxedo black metallic	2
white	707
white platinum tri-coat metallic	2
yellow	9
N=	2499

Tally table depicts the color of the cars. Most of the cars are white color.The next largest number of cars is black. Considerable number of cars are grey,silver,red and blue.Other colors are not as popular as the above colors. Therefore black cars are most popular in the USA.

8) Vin

9) Lot

10) State / City

Tally

state	Count
alabama	17
arizona	33
arkansas	12
california	190
colorado	21
connecticut	25
florida	246
georgia	51
idaho	2
illinois	113
indiana	14
kansas	4
kentucky	9
louisiana	11
maryland	4
massachusetts	27
michigan	169
minnesota	119
mississippi	24
missouri	46
montana	1
nebraska	4
nevada	85
new hampshire	4
new jersey	87
new mexico	4
new york	58
north carolina	146
ohio	31
oklahoma	71
ontario	7
oregon	27
pennsylvania	299
rhode island	2
south carolina	64
tennessee	26
texas	214
utah	10
vermont	2
virginia	90
washington	14
west virginia	21
wisconsin	94
wyoming	1
N=	2499

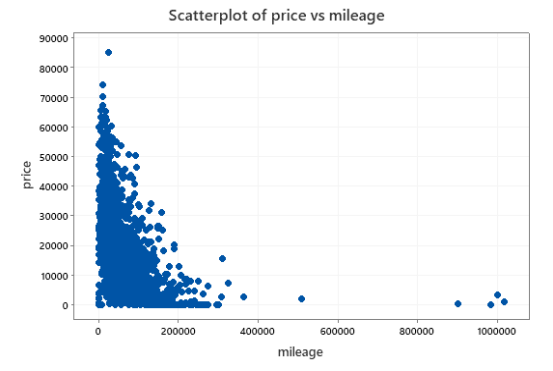
This tally table shows the location in which the car is being available for purchase. Most of the cars available in pennsylvania. The next largest number of cars available in Texas and Florida. Considerable number of cars available in california, illinois, michigan, minnesota and north carolina. All of the other areas have the least number of cars.

11) Condition

Tally

condition	Count
1 days left	91
1 hours left	3
1 minutes	15
10 days left	23
11 days left	42
12 days left	8
12 hours left	1
13 days left	1
14 hours left	108
15 days left	4
15 hours left	8
16 hours left	36
16 minutes	1
17 hours left	76
18 hours left	48
19 hours left	45
2 days left	832
2 hours left	26
20 hours left	67
21 hours left	492
22 hours left	57
23 hours left	16
24 hours left	9
27 minutes	1
28 minutes	1
29 minutes	18
3 days left	137
3 hours left	2
30 minutes	1
32 minutes	1
34 minutes	7
36 minutes	1
4 days left	16
4 hours left	1
47 minutes	2
48 minutes	2
5 days left	6
5 hours left	16
53 minutes	1
6 days left	52
6 hours left	12
7 days left	43
7 hours left	7
8 days left	82
9 days left	58
9 minutes	3
Listing Expired	20
N=	2499

This tally table shows the time of the cars. Most of the cars have 2 days left. The next largest number of cars have 21 hours left. Considerable number of cars have 14 hours, 3 days left, 1 days left, 17 hours left. All of the other cars have the least amount of time.



In this scatter plot, we can see the relationship between mileage and price. It seems that low mileage cars have higher prices than high mileage cars.

Conclusions