Visualize data in the used car vehicles dataset by using the R studio and Tableau

DynamisHub

Table of Contents

1. Abstract

Big data analysis is the sector which analyze large datasets that are complex and difficult to take valuable information from these data. Therefore, there are some platforms that are available for big data analysis. The right choice of these platforms is useful so as to exclude worthy insights which is the main purpose of big data analysis. The approach of this paper is to provide some valuable information about the R programming language and Tableau and if these platforms are useful for big data analysis. Hence, the dataset we choose is about used car vehicles and visualizing the valuable components so as to show which one is better for big data analysis.

2. Introduction

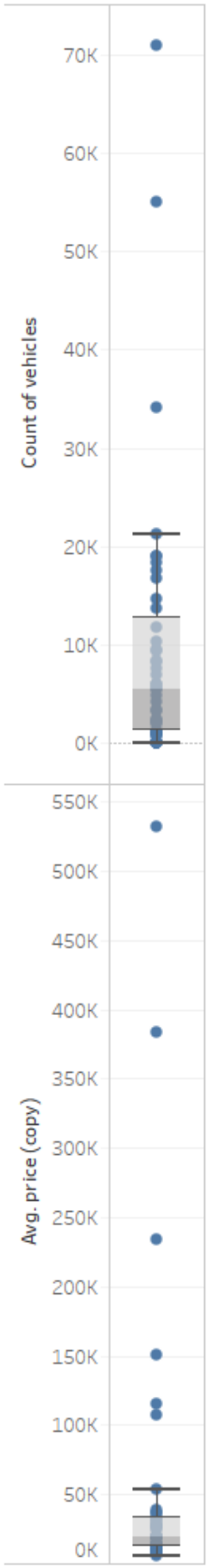
This case is important in order to prove that data analysis is not as simple as many people think, the main purpose is to show that data analysts need to have knowledge about the tools they use. Therefore, we use the R tool and Tableau, two powerful software that gives us the opportunity to visualize our dataset, exclude some valuable insights and analyze the results via statistical methods that provide us these two tools. R software is a programming language that provides statistical analysis and visualizations and Tableau is a powerful visualization tool which is made to make every visualization we need so that analyzing the data faster than other tools. To focus on this dataset we don't use some components like image_url, region_url and description of the vehicles because in the data cleaning process, the approach is that these components are redundant.

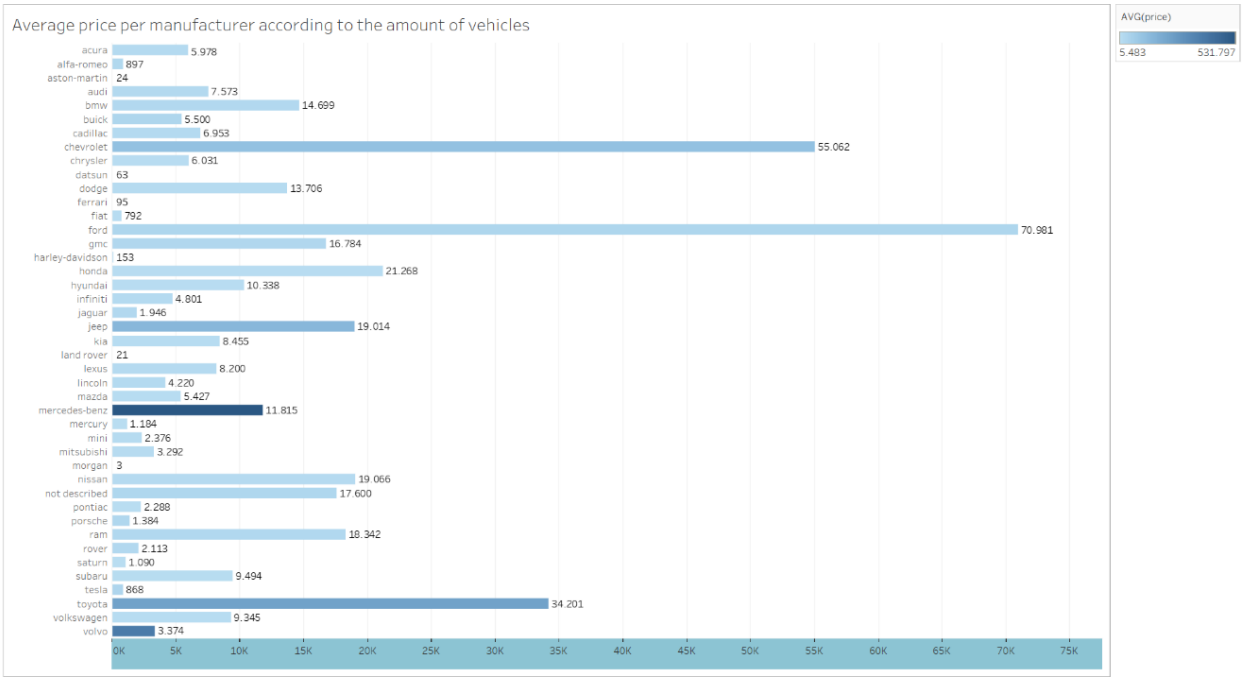
3. What is Tableau

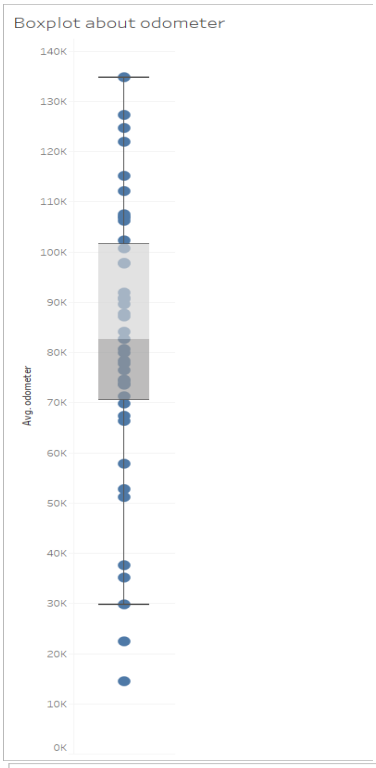
Tableau is a user-friendly platform that is aimed to assist the analysis through visualization. It is easy to use by dragging and dropping the components that are needed to visualize. Tableau is a platform that helps us to make data preparation for statistical analysis by making visualizations and as a result, excluding some valuable insights for the business development.

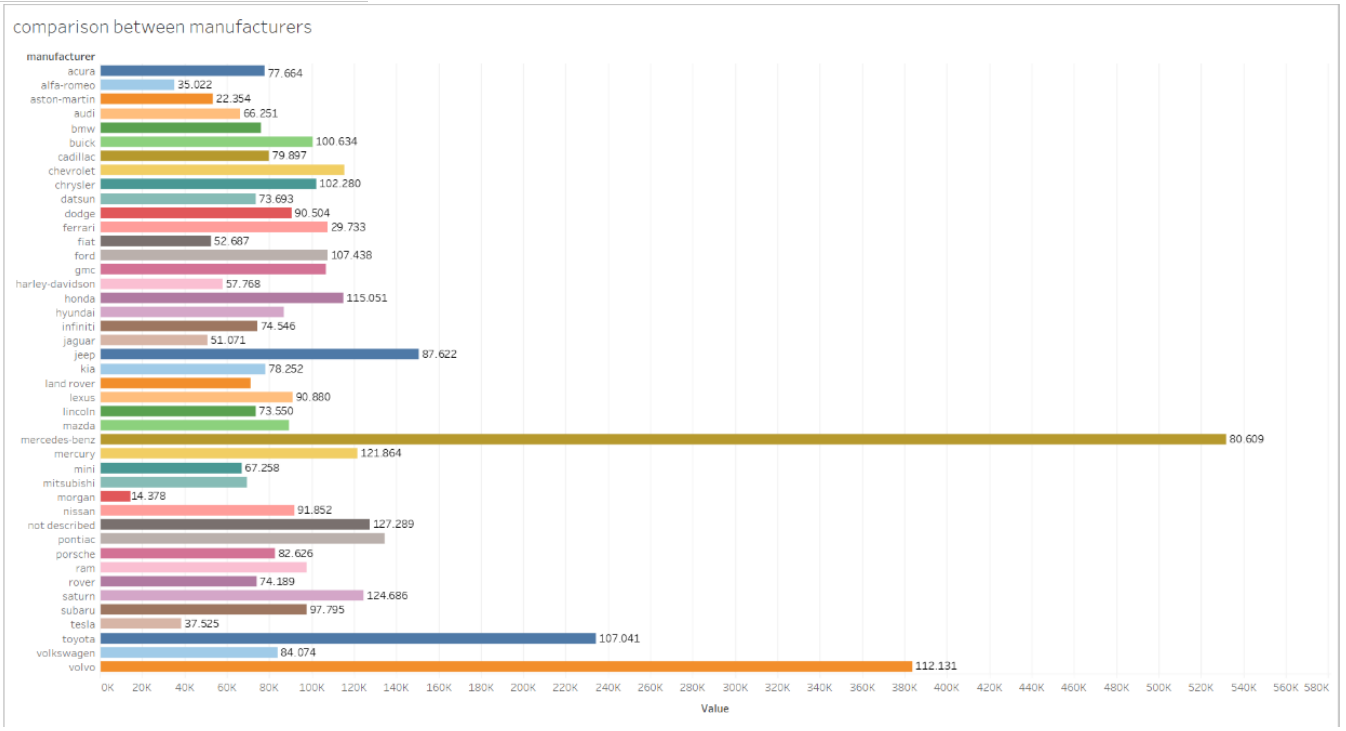
3.1 Visualizations in used cars dataset via Tableau

boxplot about manufacturer average prices and count of vehicles

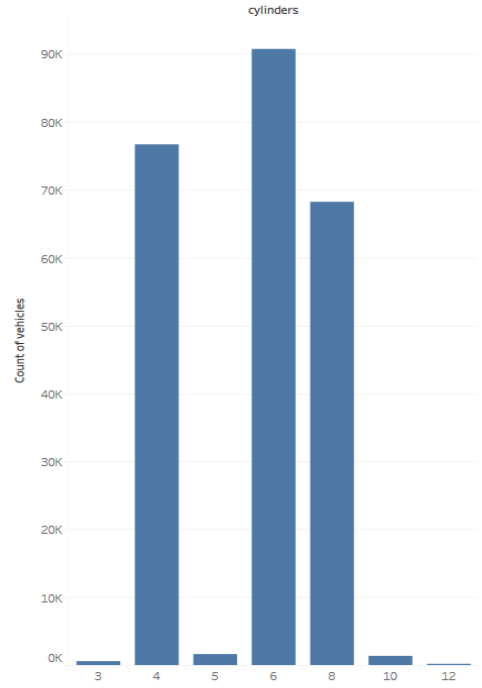



Average price per manufacturer according to the amount of vehicles
AVG(price)
5.483
531.797
acura 5.978
alfa-romeo 897
aston-martin 24
audi 7.573
bmw 14.699
buick 5.500
cadillac 6.953
chevrolet 55.062
chrysler 6.031
datsun 63
dodge 13.706
ferrari 95
fiat 792
ford 70.981
gmc 16.784
harley-davidson 153
honda 21.268
hyundai 10.338
infiniti 4.801
jaguar 1.946
jeep 19.014
kia 8.455
land rover 21
lexus 8.200
lincoln 4.220
mazda 5.427
mercedes-benz 11.815
mercury 1.184
mini 2.376
mitsubishi 3.292
morgan 3
nissan 19.066
not described 17.600
pontiac 2.288
porsche 1.384
ram 18.342
rover 2.113
saturn 1.090
subaru 9.494
tesla 868
toyota 34.201
volkswagen 9.345
volvo 3.374
0K
5K
10K
15K
20K
25K
30K
35K
40K
45K
50K
55K
60K
65K
70K
75K


Boxplot about odometer
Avg. odometer
140K
130K
120K
110K
100K
90K
80K
70K
60K
50K
40K
30K
20K
10K
0K

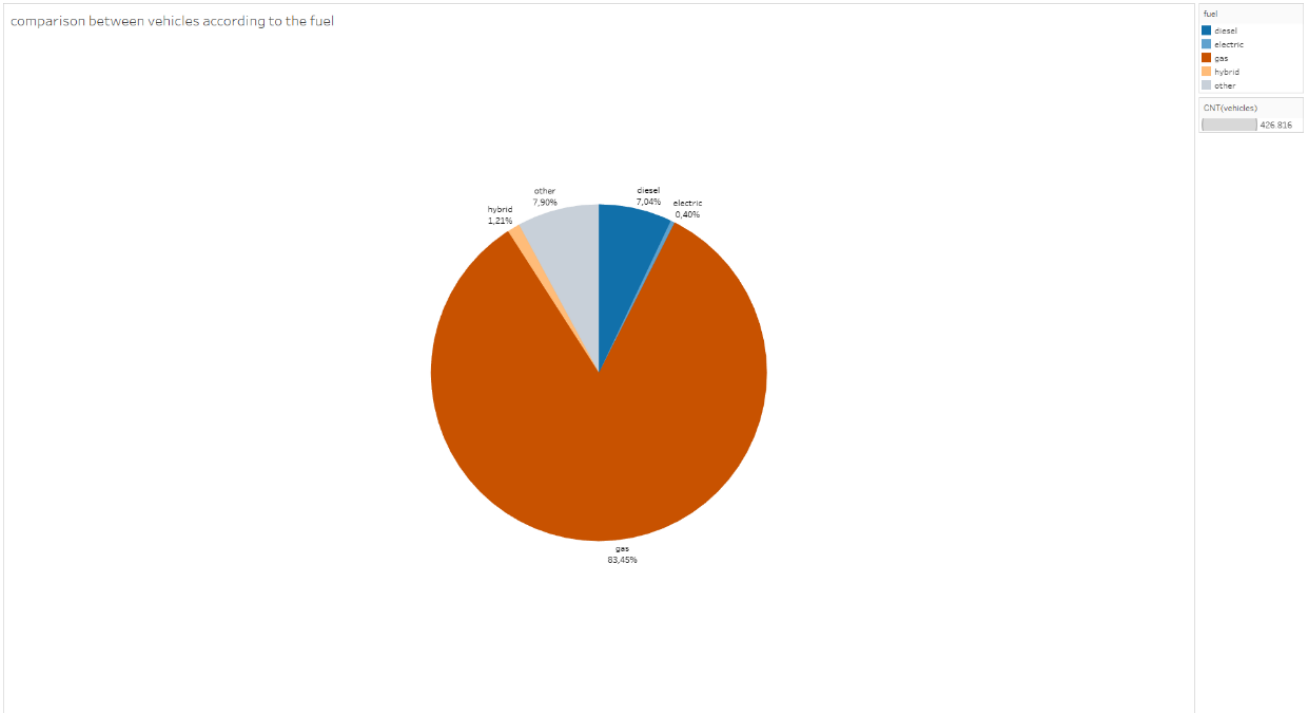

comparison between manufacturers
manufacturer
acura 77.664
alfa-romeo 35.022
aston-martin 22.354
audi 66.251
bmw
buick 100.634
cadillac 79.897
chevrolet
chrysler 102.280
datsun 73.693
dodge 90.504
ferrari 29.733
fiat 52.687
ford 107.438
gmc
harley-davidson 57.768
honda 115.051
hyundai
infiniti 74.546
jaguar 51.071
jeep 87.622
kia 78.252
land rover
lexus 90.880
lincoln 73.550
mazda
mercedes-benz 80.609
mercury 121.864
mini 67.258
mitsubishi
morgan 14.378
nissan 91.852
not described 127.289
pontiac
porsche 82.626
ram
rover 74.189
saturn 124.686
subaru 97.795
tesla 37.525
toyota 107.041
volkswagen 84.074
volvo 112.131
0K 20K 40K 60K 80K 100K 120K 140K 160K 180K 200K 220K 240K 260K 280K 300K 320K 340K 360K 380K 400K 420K 440K 460K 480K 500K 520K 540K 560K 580K
Value

count of vehicles per cylinders



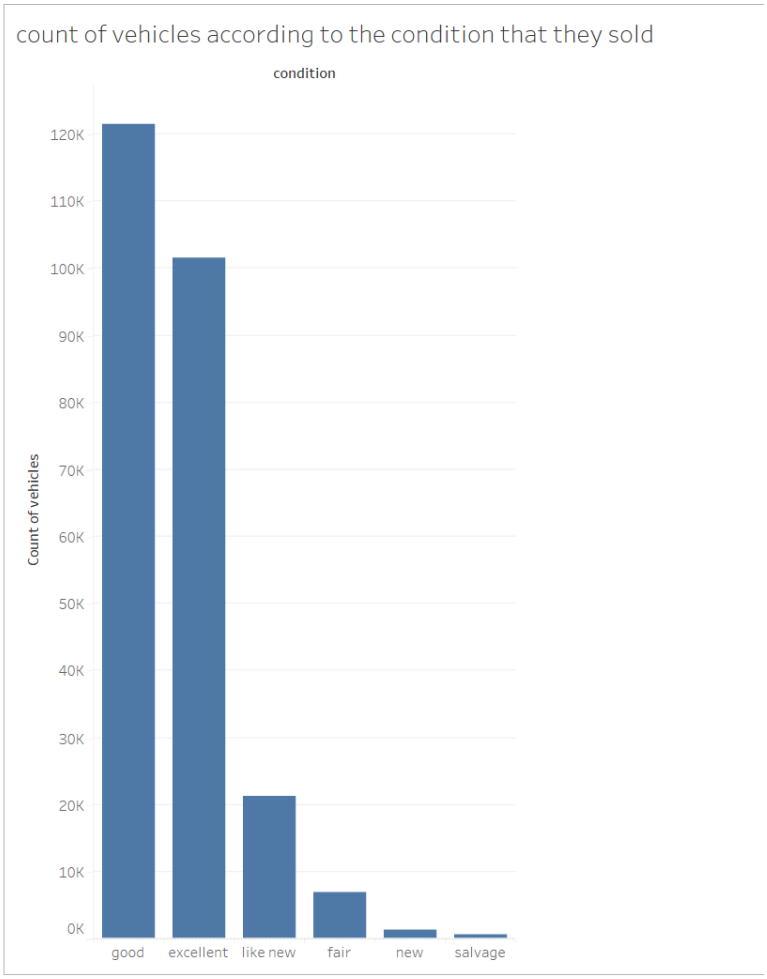
Count of vehicles for each cylinders. The data is filtered on fuel and Exclusions (cylinders,fuel). The fuel filter excludes other. The Exclusions (cylinders,fuel) filter keeps 33 members. The view is filtered on cylinders, which excludes other.

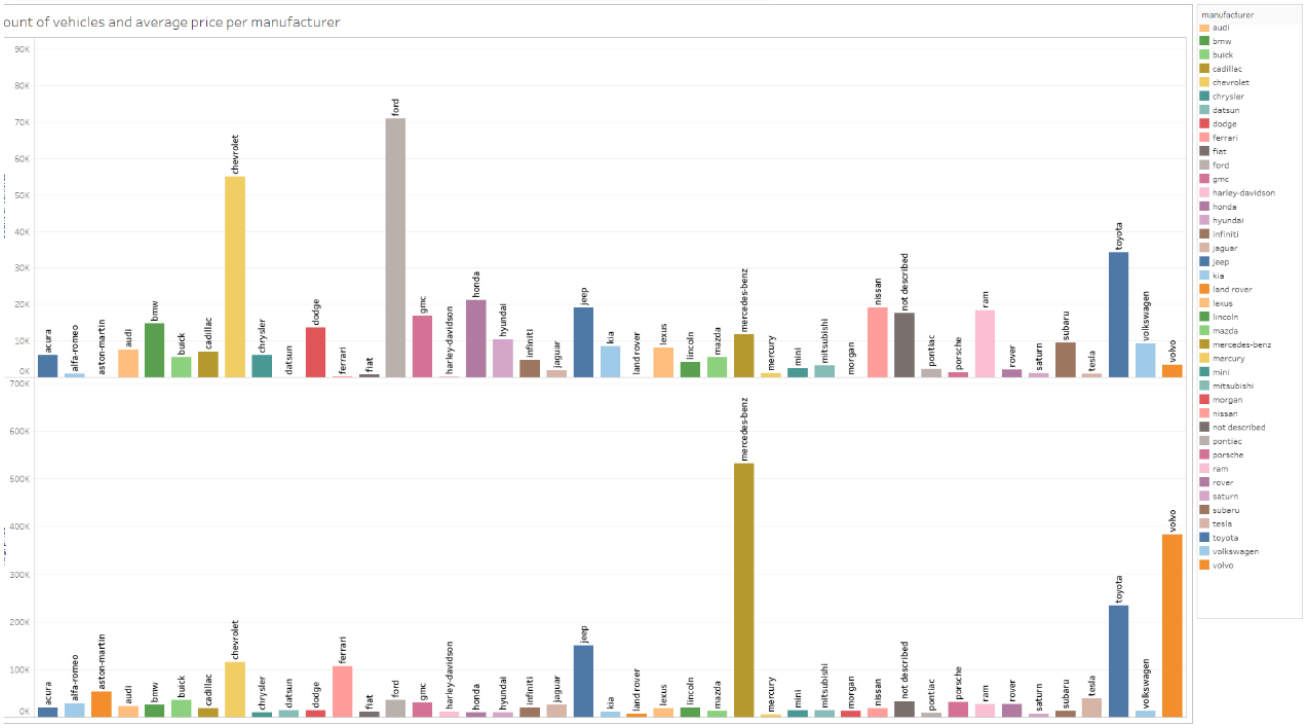
comparison between vehicles according to the fuel



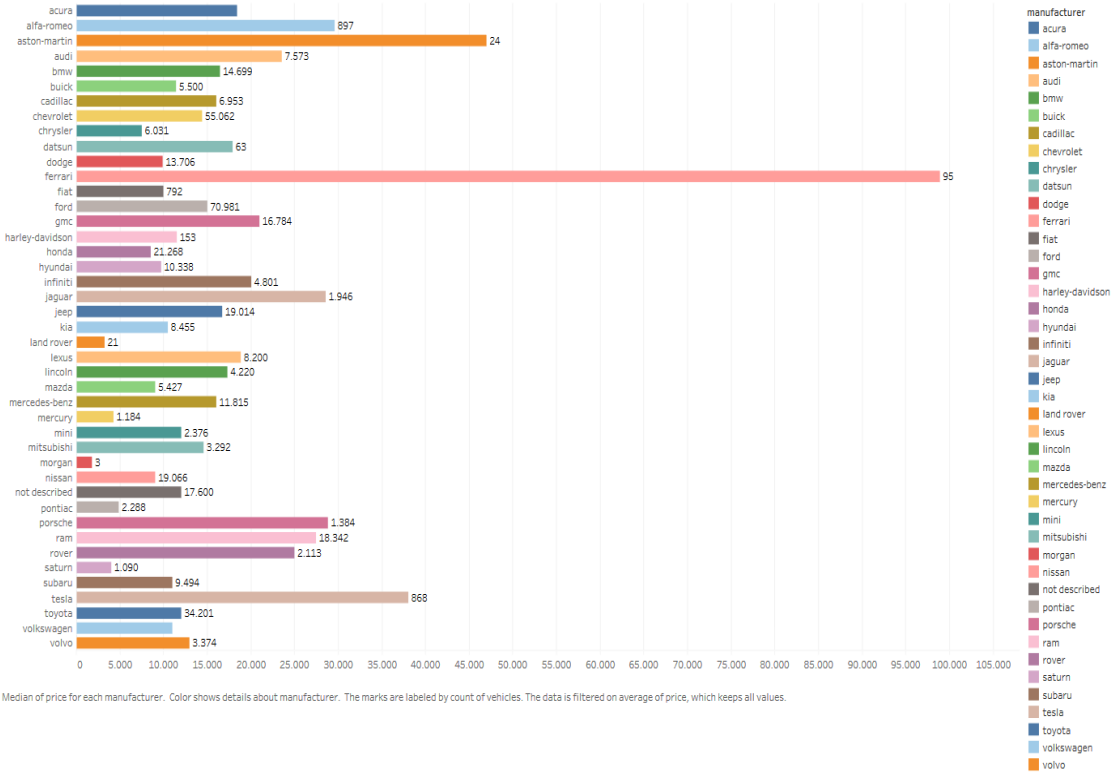
Condition per manufacturer

	condition					
manufacturer	excellent	fair	good	like new	new	salvage
acura	970	73	3.259	175	16	3
alfa-romeo	41	3	694	13		
aston-martin	3		2	5	2	
audi	1.456	48	3.431	291	15	5
bmw	3.482	139	4.843	699	29	7
buick	1.223	100	2.047	275	14	6
cadillac	1.597	87	2.479	355	6	9
chevrolet	13.001	1.072	15.150	3.111	168	65
chrysler	1.661	104	1.567	326	7	9
datsun	22	8	7	3		1
dodge	3.191	260	3.345	746	71	34
ferrari	21	1	14	5		
fiat	175	6	327	56	2	
ford	16.905	1.243	16.033	3.383	147	108
gmc	3.491	273	4.987	692	44	17
harley-davidson	53	2	13	14	2	
honda	6.023	398	5.195	1.163	73	39
hyundai	2.621	78	2.793	514	23	3
infiniti	894	20	2.425	208	10	4
jaguar	265	29	1.210	61	2	1
jeep	4.595	266	4.670	944	52	17
kia	2.041	41	2.343	498	21	5
land rover	4		5	1		
lexus	1.908	32	2.992	355	25	6
lincoln	854	39	1.975	196	3	4
mazda	1.332	70	1.959	228	22	8
mercedes-benz	3.141	100	2.842	612	32	12
mercury	389	68	311	68	1	2
mini	592	9	848	145	4	1
mitsubishi	633	60	1.418	126	85	8
morgan	2					
nissan	5.106	236	4.494	1.131	70	46
pontiac	615	130	590	139	5	10
porsche	371	9	264	91	5	1
ram	3.414	208	3.892	648	39	12
rover	523	25	523	107	9	3
saturn	273	45	299	82	1	
subaru	2.509	160	2.326	497	44	25
tesla	107	1	516	40	3	1
toyota	9.114	485	9.198	1.661	91	56
volkswagen	2.097	155	3.572	425	23	12
volvo	910	87	1.353	122	6	11


count of vehicles according to the condition that they sold
condition
Count of vehicles
120K
110K
100K
90K
80K
70K
60K
50K
40K
30K
20K
10K
0K
good
excellent
like new
fair
new
salvage

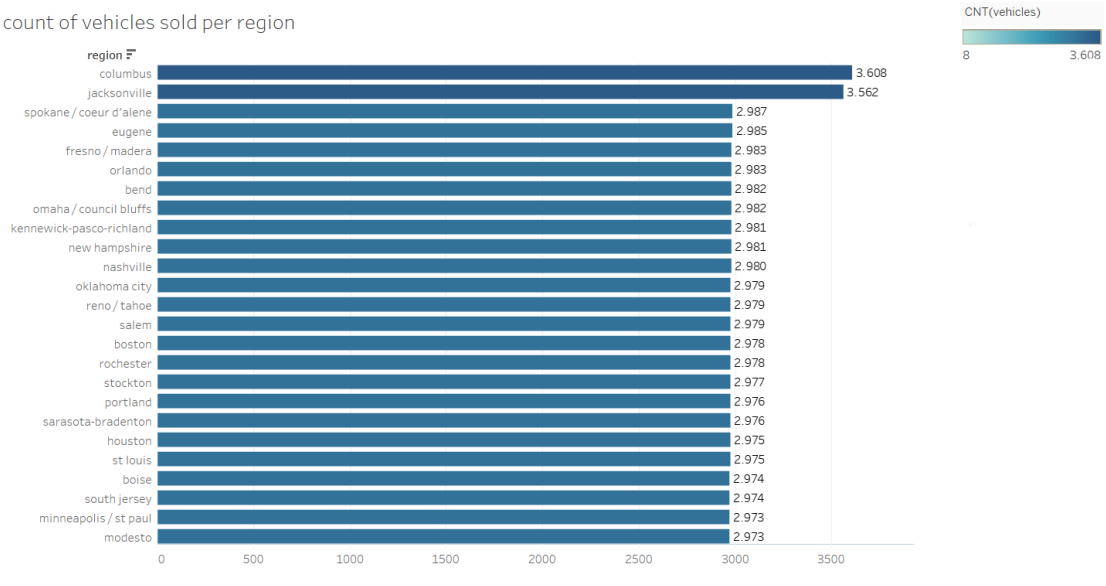

ount of vehicles and average price per manufacturer
manufacturer
audi
bmw
buick
cadillac
chevrolet
chrysler
datsun
dodge
ferrari
Fiat
ford
gmc
harley-davidson
honda
hyundai
infiniti
jaguar
jeep
kia
land rover
lexus
lincoln
mazda
mercedes-benz
mercury
mini
mitsubishi
morgan
nissan
not described
pontiac
porsche
ram
rover
saturn
subaru
tesla
toyota
volkswagen
volvo

count of vehicles per manufacturer

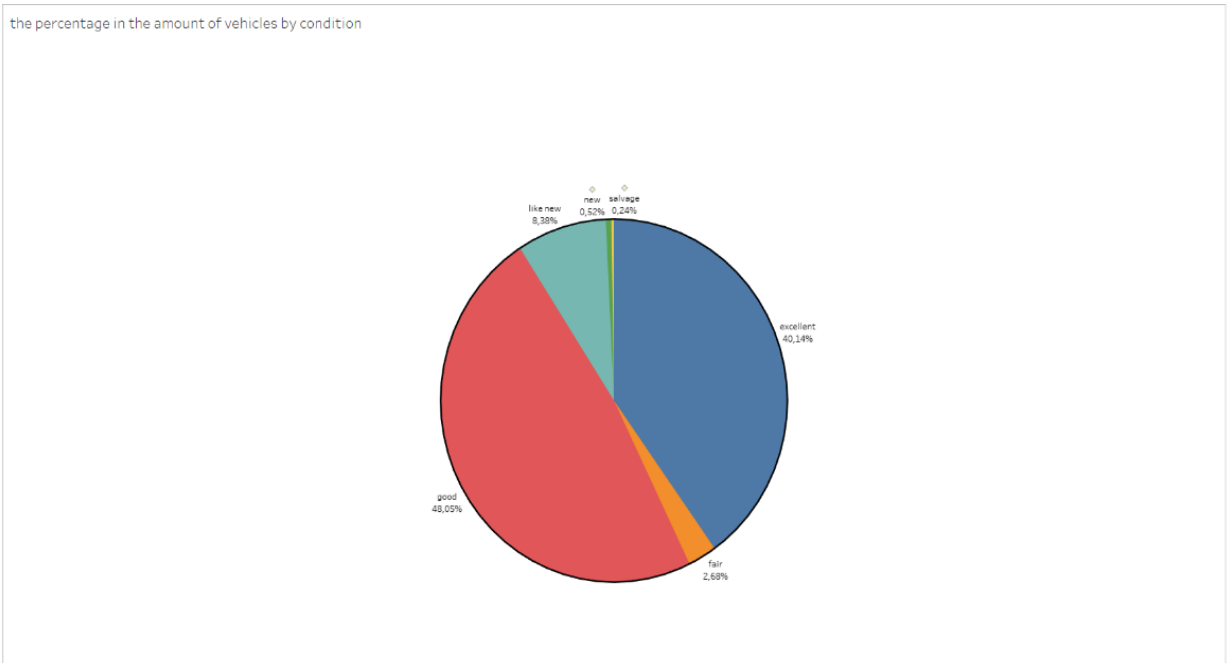


Median of price for each manufacturer. Color shows details about manufacturer. The marks are labeled by count of vehicles. The data is filtered on average of price, which keeps all values.

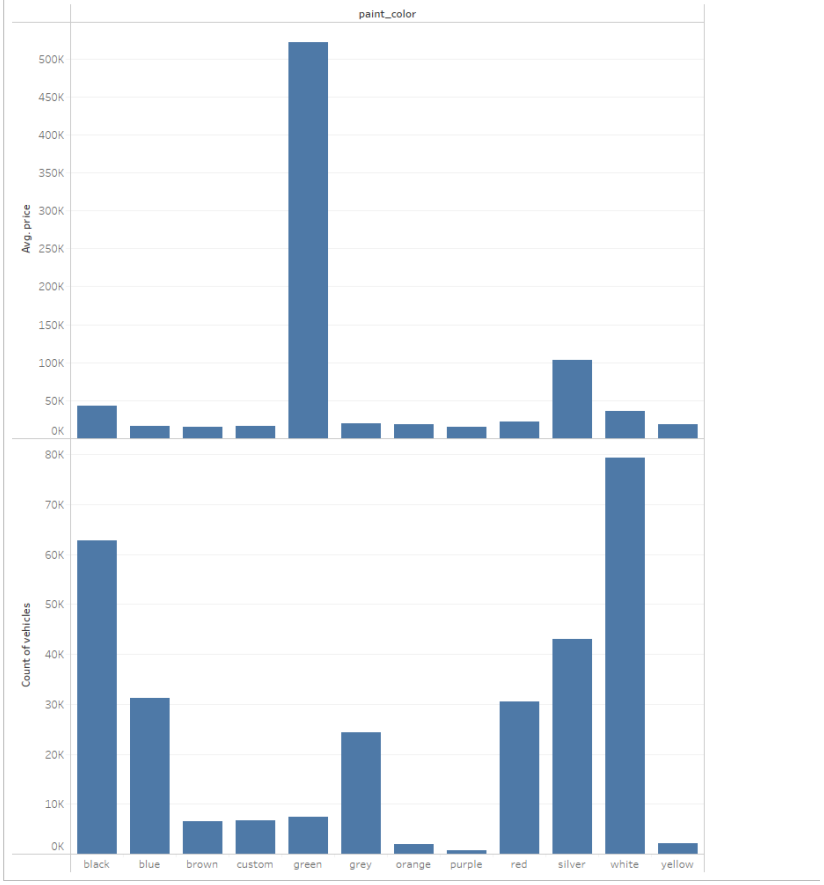
count of vehicles sold per region

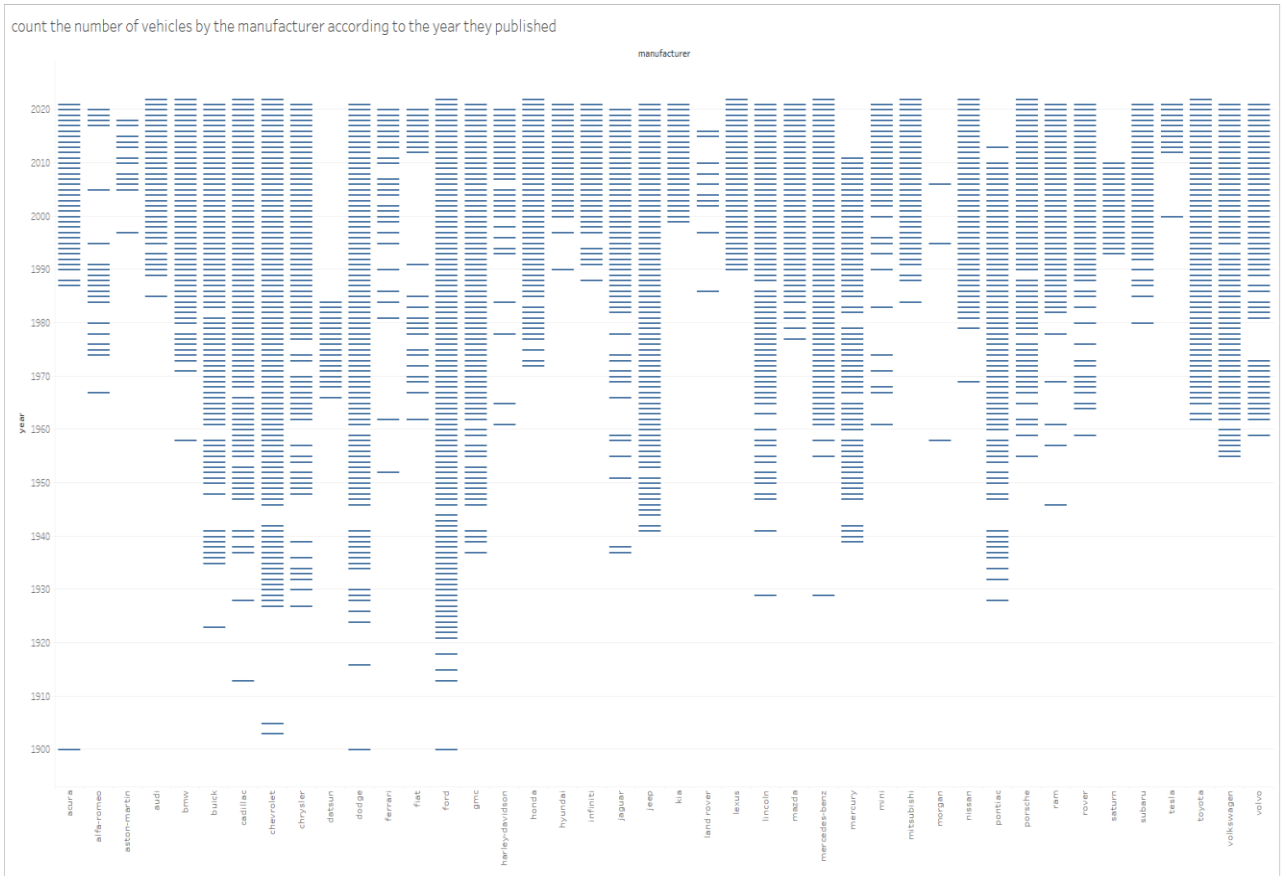


the percentage in the amount of vehicles by condition

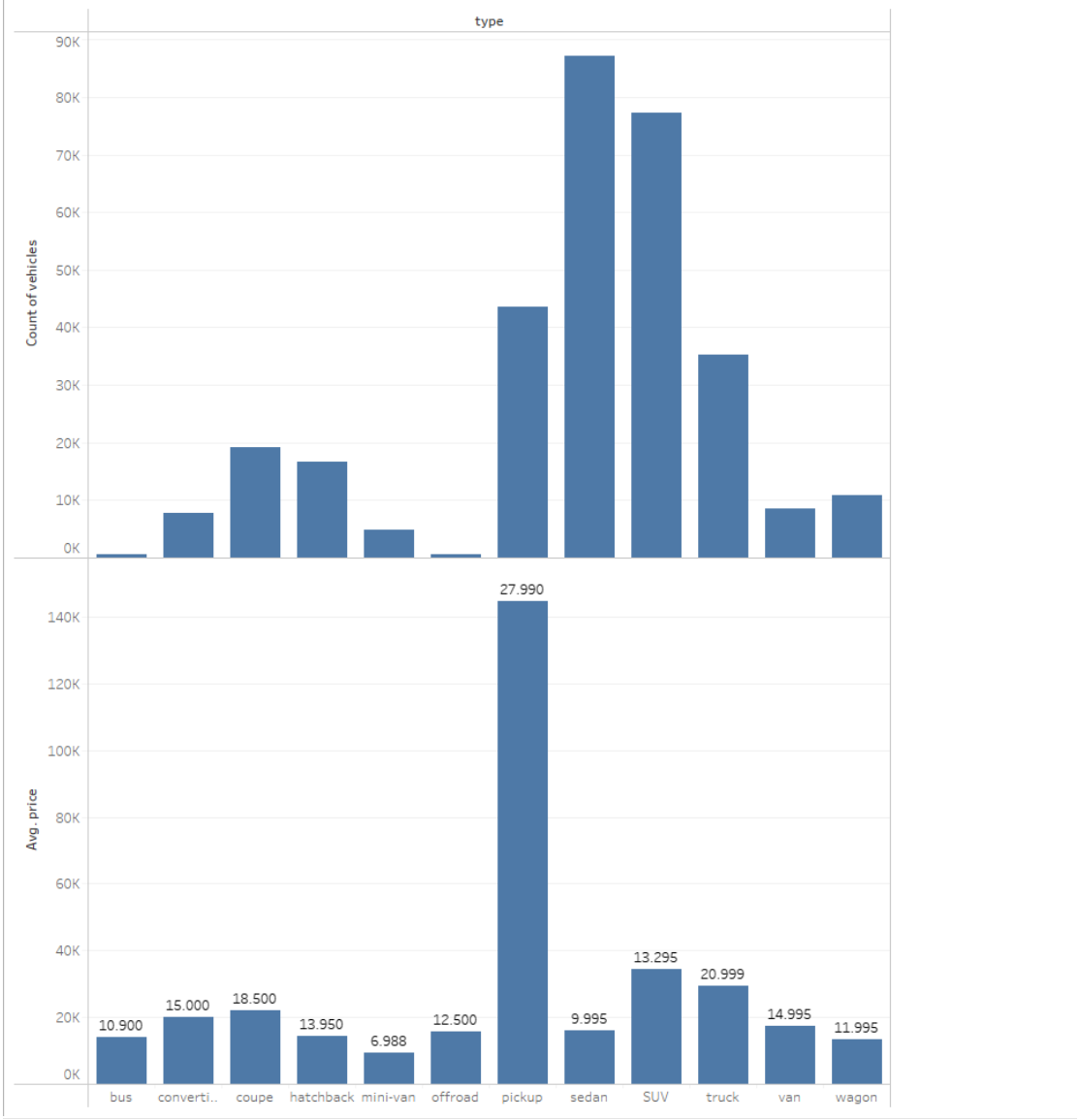


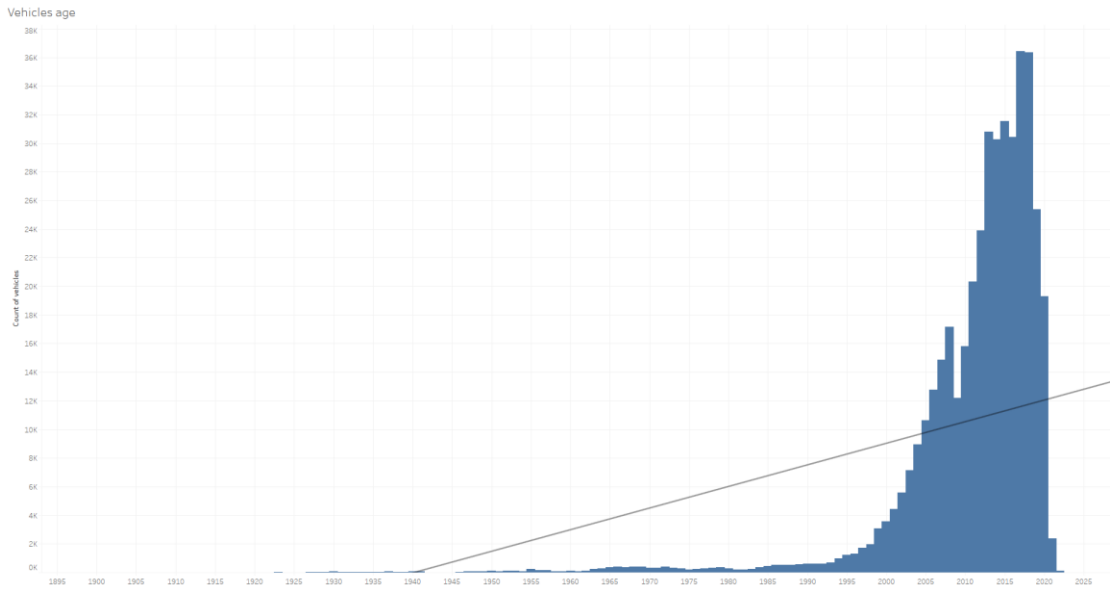
The influence of each color in the price

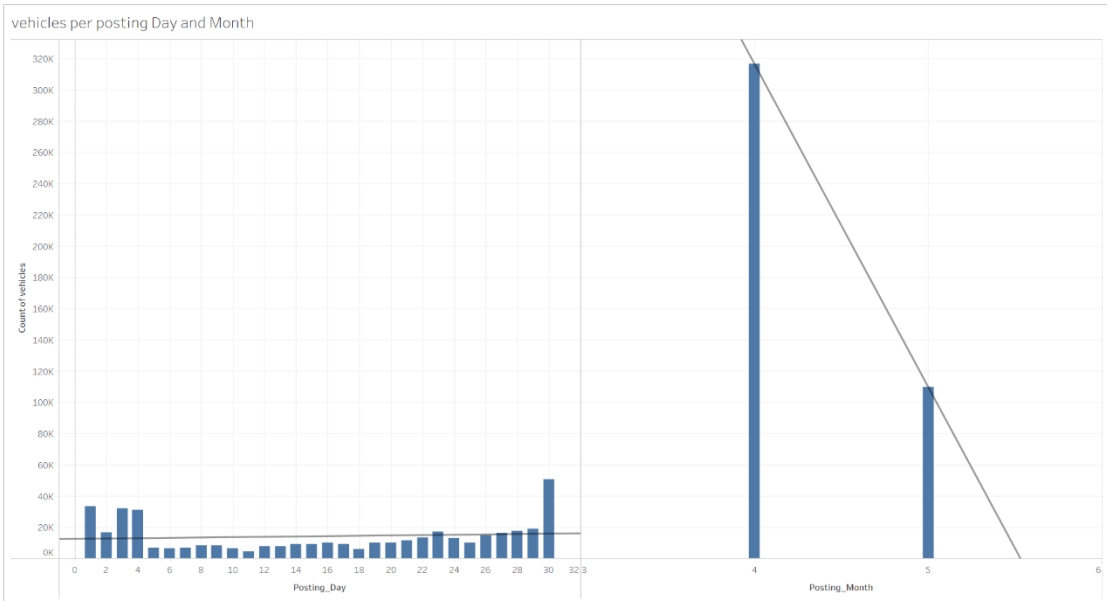



count the number of vehicles by the manufacturer according to the year they published
manufacturer
year
2020
2010
2000
1990
1980
1970
1960
1950
1940
1930
1920
1910
1900
acura
alfa-romeo
aston-martin
audi
bmw
buick
cadillac
chevrolet
chrysler
datsun
dodge
ferrari
fiat
ford
gmc
harley-davidson
honda
hyundai
infiniti
jaguar
jeep
kia
land rover
lexus
lincoln
mazda
mercedes-benz
mercury
mini
mitsubishi
morgan
nissan
pontiac
porsche
ram
rover
saturn
subaru
tesla
toyota
volkswagen
volvo

the influence of each type in the price of the vehicles





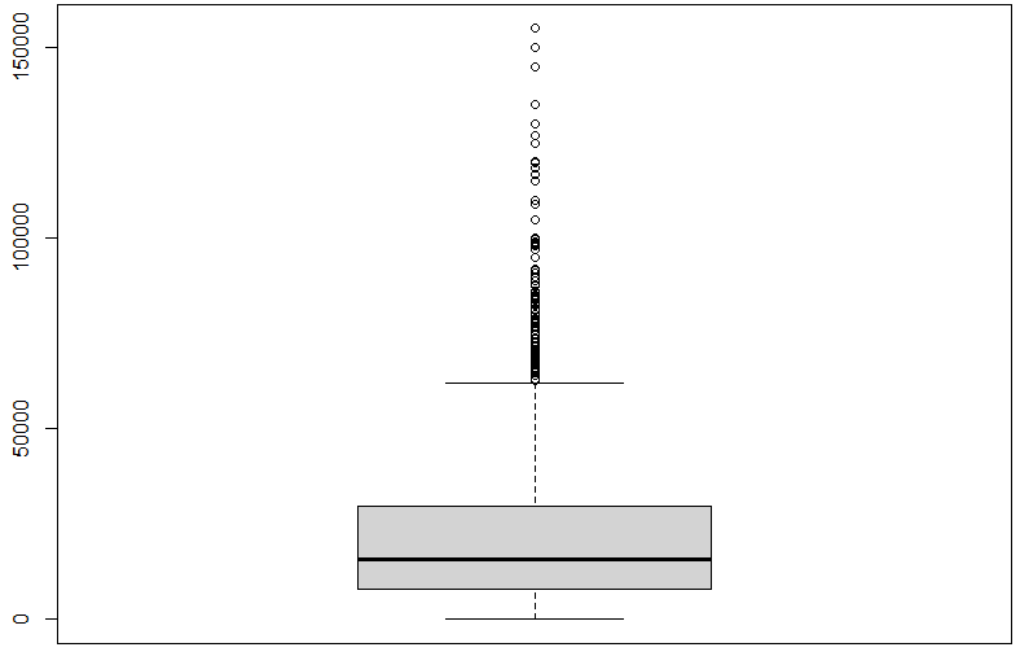

vehicles per posting Day and Month
Count of vehicles
320K
300K
280K
260K
240K
220K
200K
180K
160K
140K
120K
100K
80K
60K
40K
20K
0K
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
Posting_Day
3
4
5
6
Posting_Month

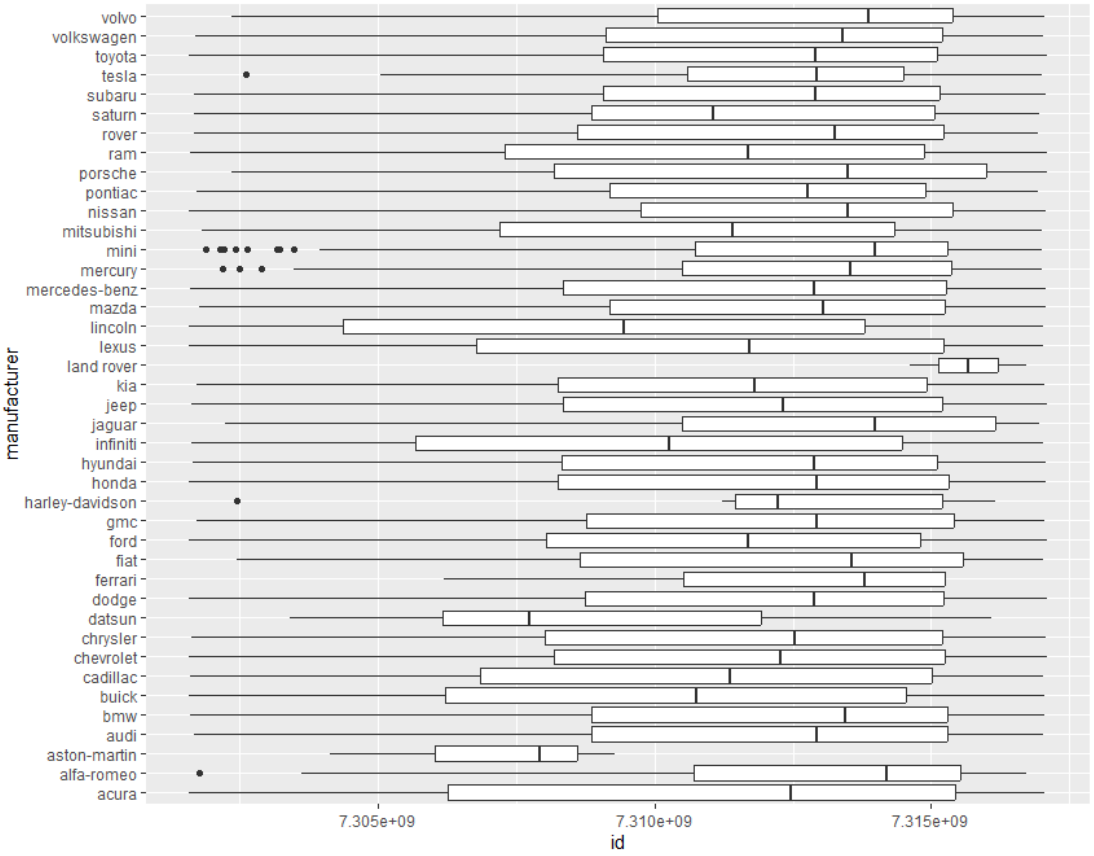
4. what is R

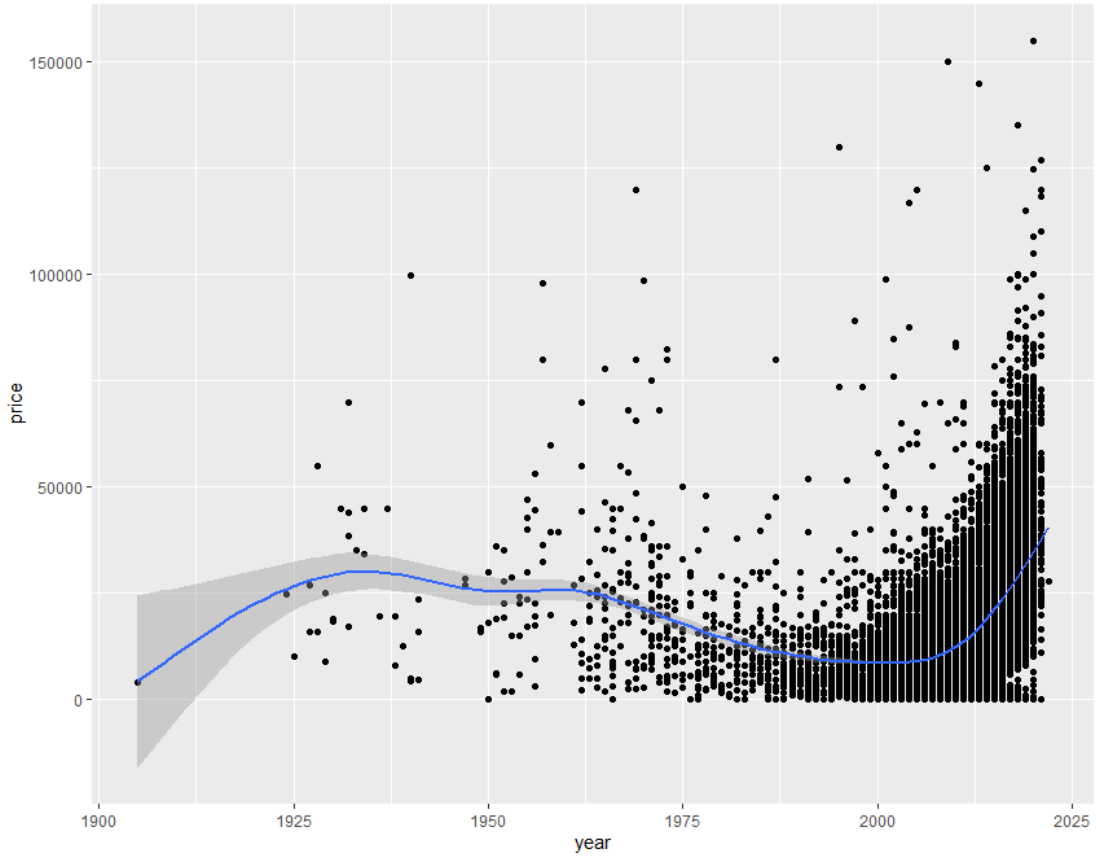
R is a programming language which is used ideally for statistical computing and graphics design. R provides many statistical and graphical techniques that assist every user in order to find the best match of these techniques according to the dataset they need to analyze. R can extend via the packages and libraries,these packages are helpful in order to make the analysis of the dataset better and faster.

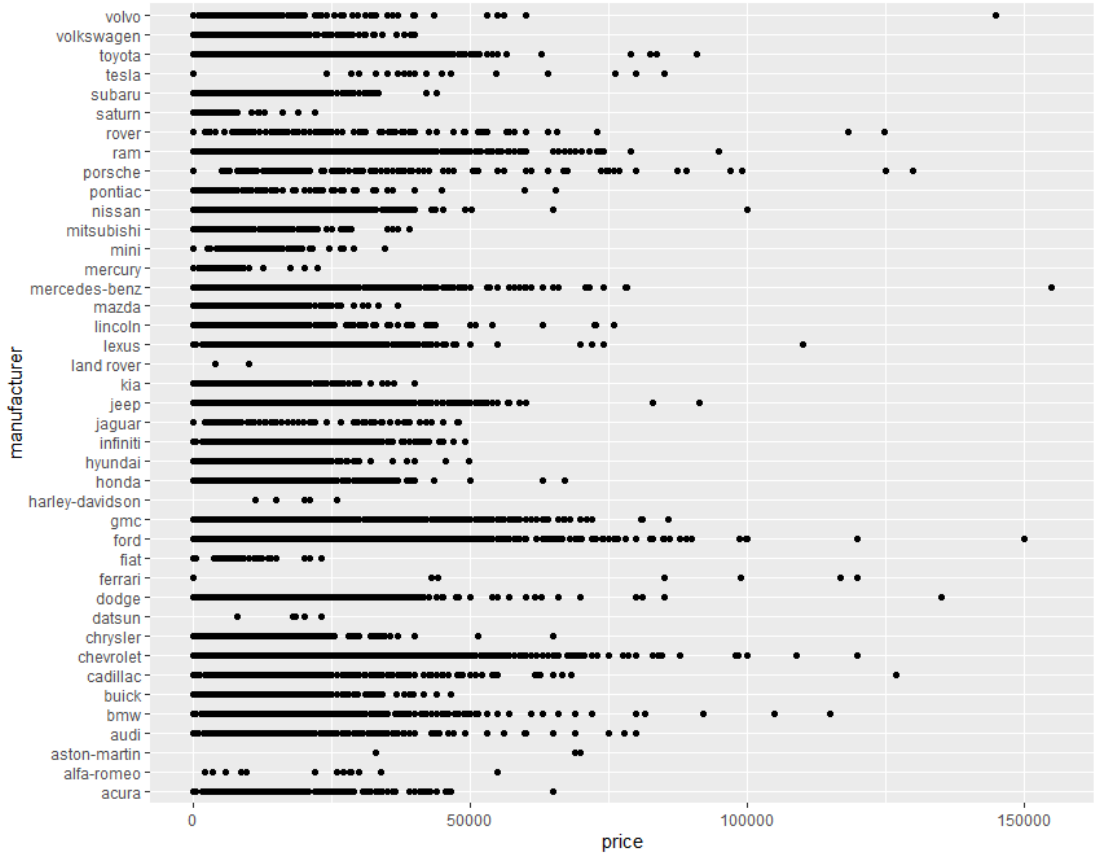
4.1 Visualizations via R in used cars vehicles dataset

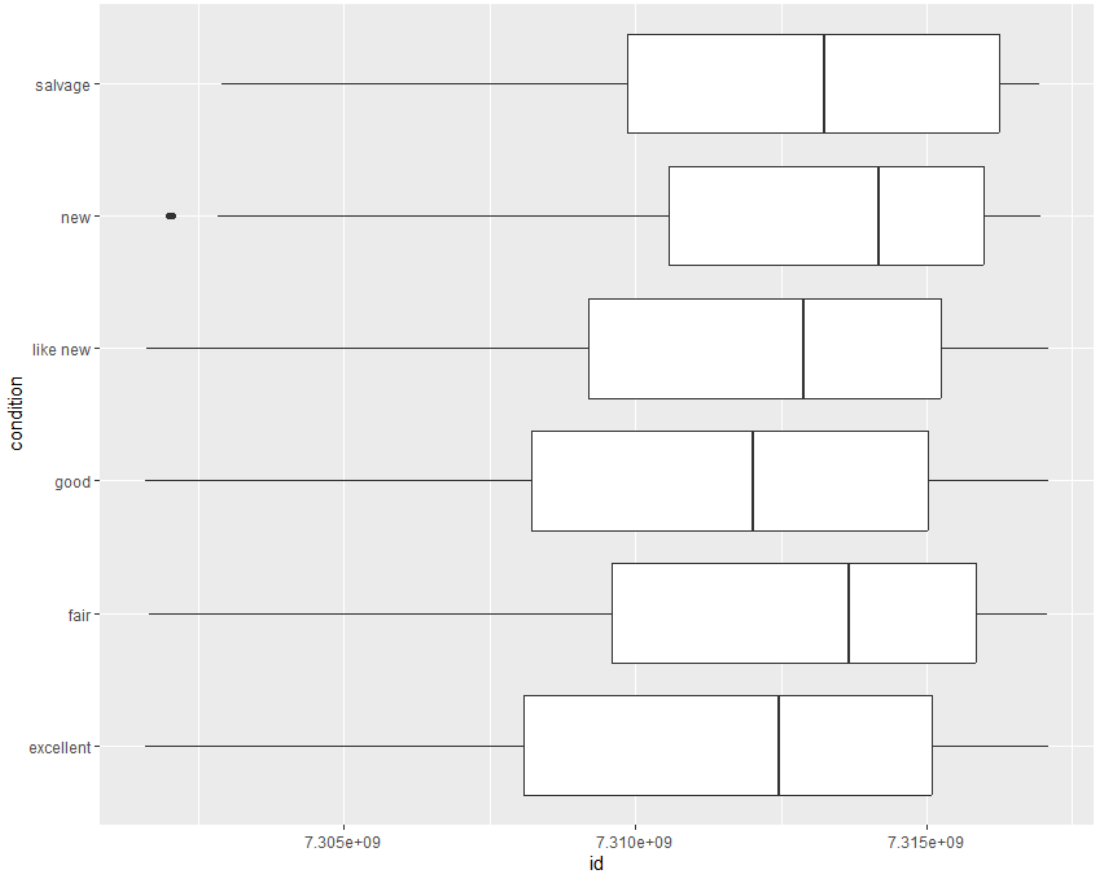
Boxplot in prices

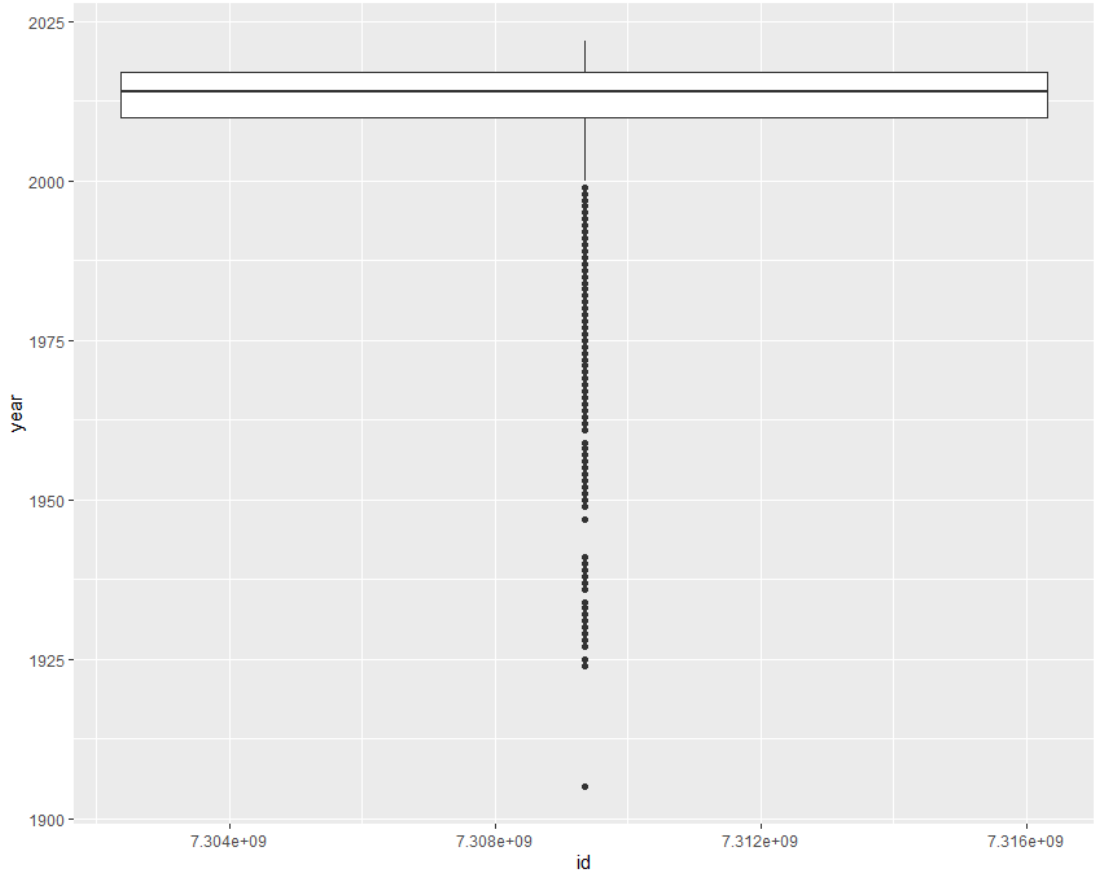


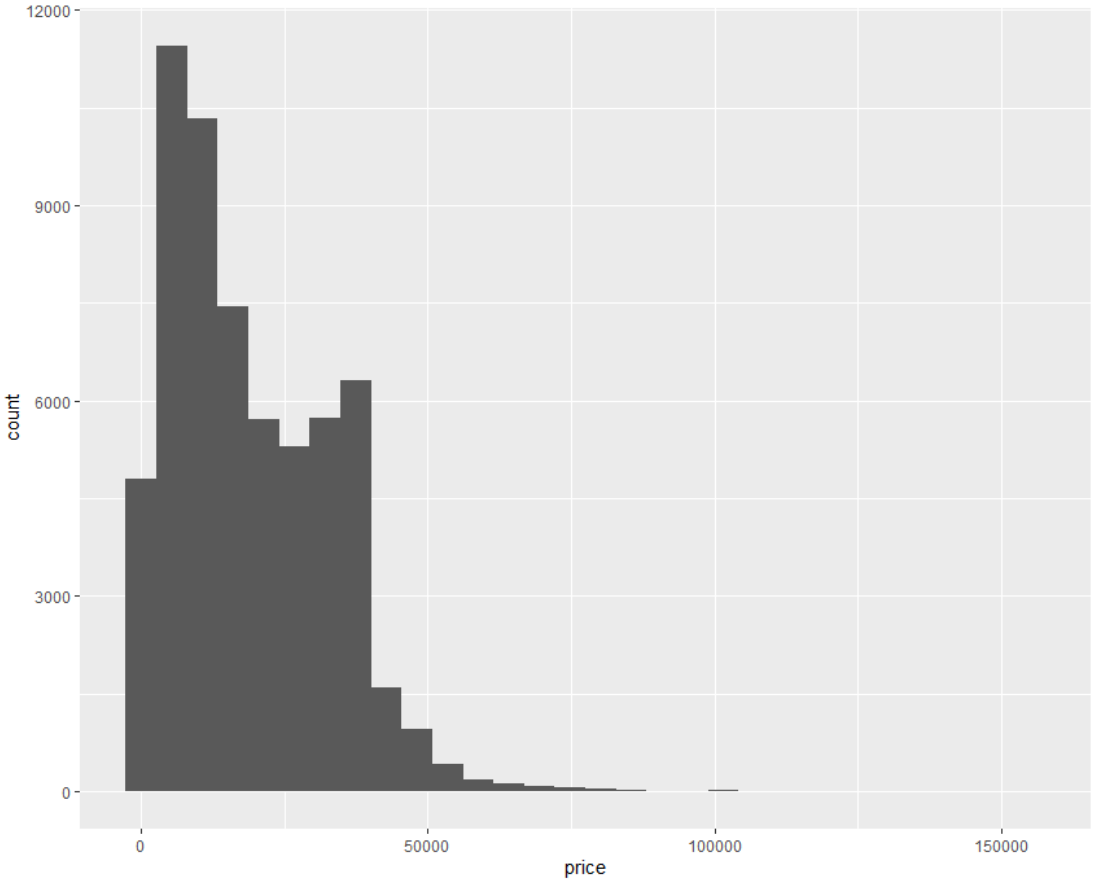


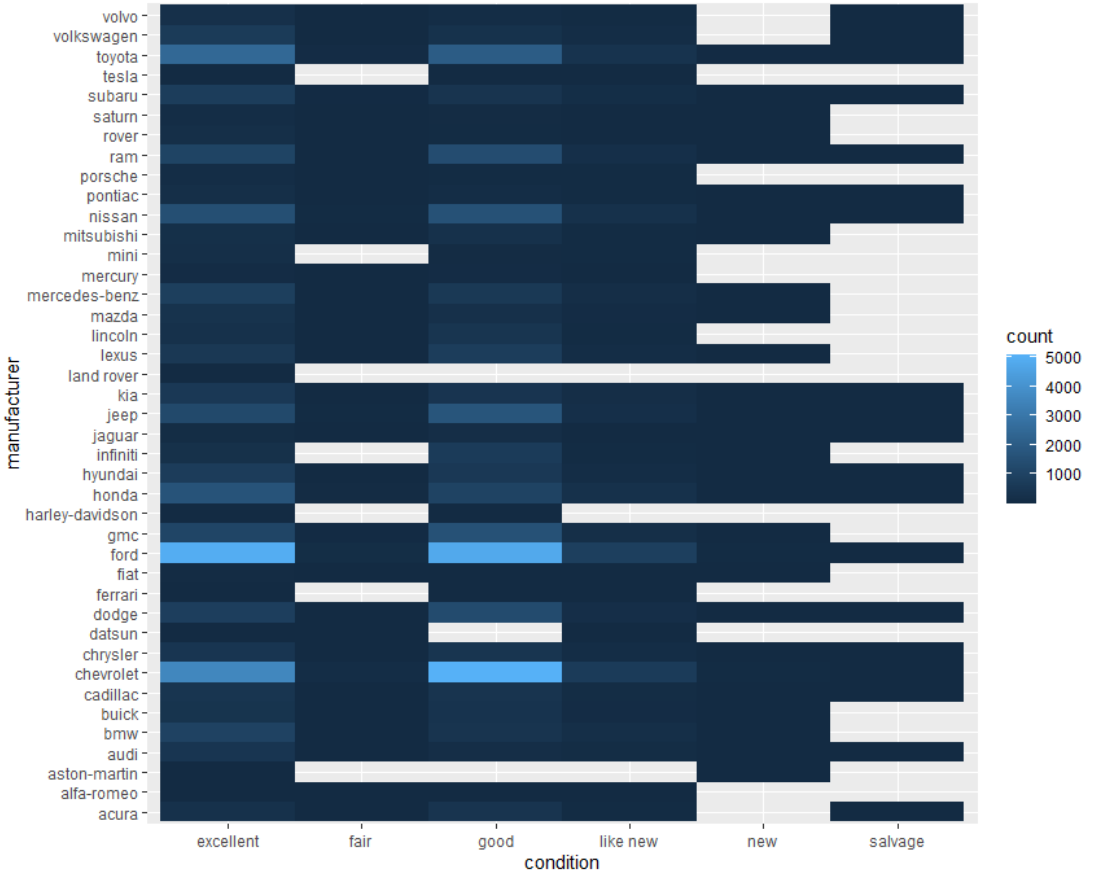

price
150000
100000
50000
0
1900
1925
1950
1975
2000
2025
year

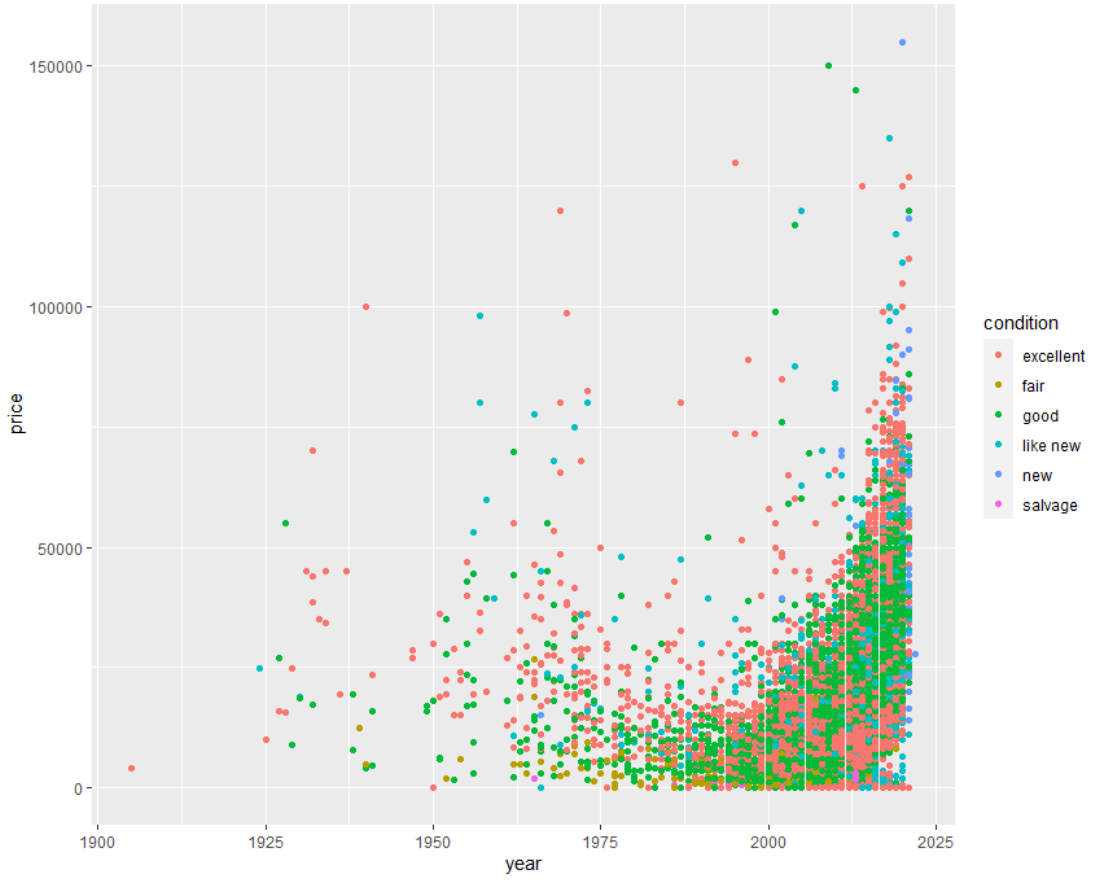

manufacturer
volvo
volkswagen
toyota
tesla
subaru
saturn
rover
ram
porsche
pontiac
nissan
mitsubishi
mini
mercury
mercedes-benz
mazda
lincoln
lexus
land rover
kia
jeep
jaguar
infiniti
hyundai
honda
harley-davidson
gmc
ford
fiat
ferrari
dodge
datsun
chrysler
chevrolet
cadillac
buick
bmw
audi
aston-martin
alfa-romeo
acura
0
50000
100000
150000
price

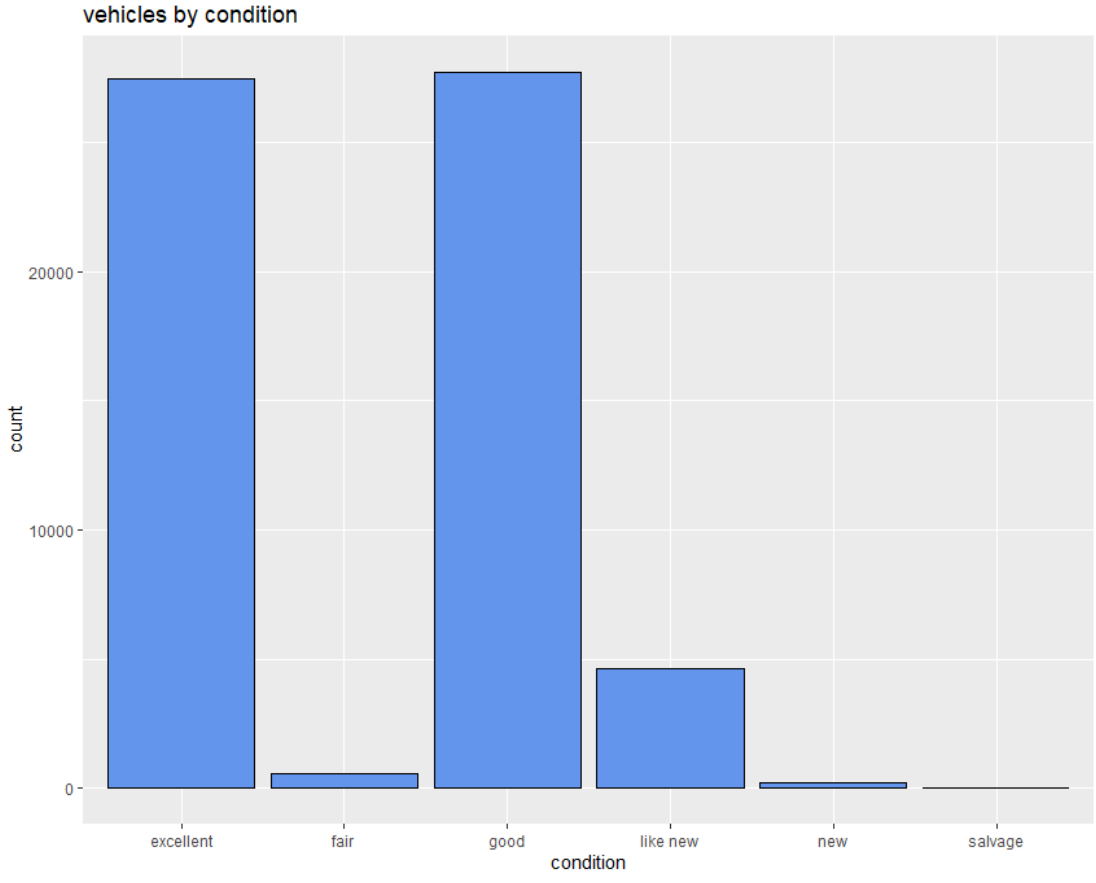

salvage
new
like new
good
fair
excellent
condition
7.305e+09
7.310e+09
7.315e+09
id

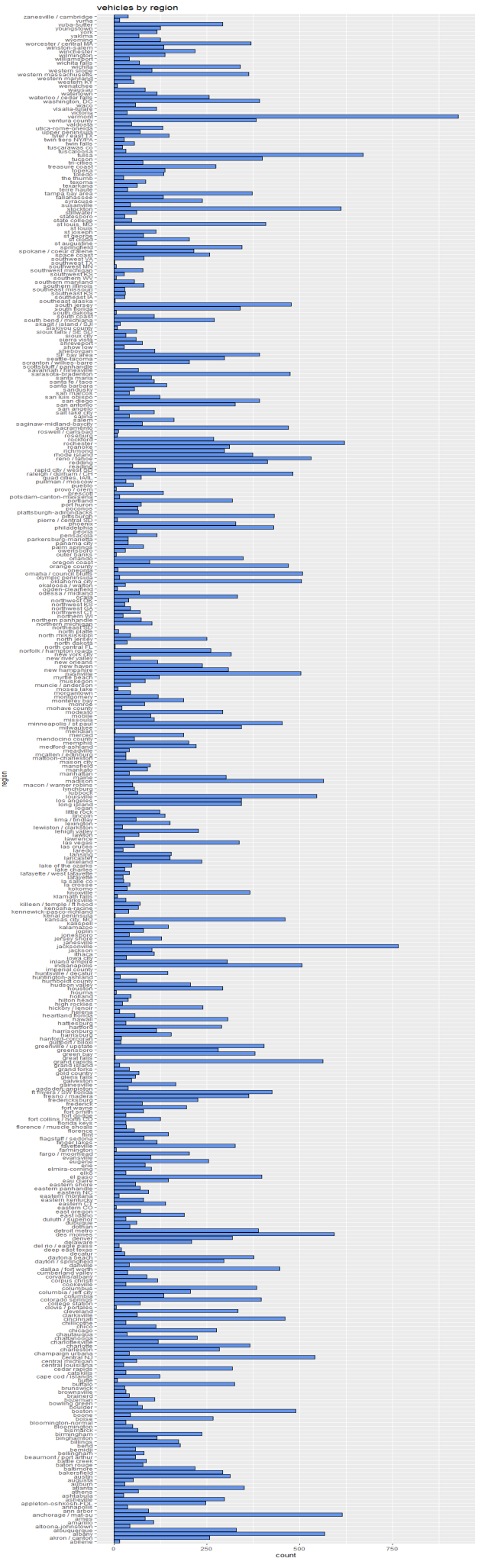

2025
2000
1975
1950
1925
1900
year
7.304e+09
7.308e+09
7.312e+09
7.316e+09
id


12000
9000
6000
3000
0
count
0
50000
100000
150000
price

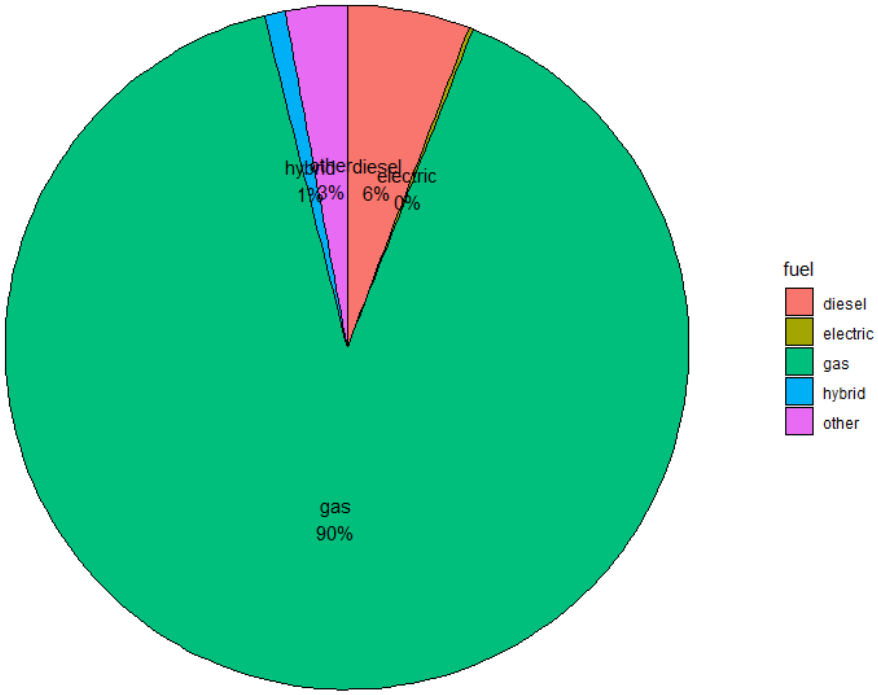

volvo
volkswagen
toyota
tesla
subaru
saturn
rover
ram
porsche
pontiac
nissan
mitsubishi
mini
mercury
mercedes-benz
mazda
lincoln
lexus
land rover
kia
jeep
jaguar
infiniti
hyundai
honda
harley-davidson
gmc
ford
fiat
ferrari
dodge
datsun
chrysler
chevrolet
cadillac
buick
bmw
audi
aston-martin
alfa-romeo
acura
manufacturer
excellent
fair
good
like new
new
salvage
condition
count
5000
4000
3000
2000
1000


150000
100000
50000
0
price
1900
1925
1950
1975
2000
2025
year
condition
excellent
fair
good
like new
new
salvage


vehicles by condition
20000
10000
0
count
excellent
fair
good
like new
new
salvage
condition


vehicles by region
region
count
0
250
500
750

vehicles by fuel



5. Analysis

First, in this case in order to view better and deeper this dataset we will analyze the dataset via R studio which is better for analysis.
The correlation between these numeric fields is low so it seems that the price is not correlated with the year that a vehicle is published as much as expected.

	id	price	year	odometer	lat	long
id	1.00	-0.06	-0.07	0.02	-0.07	-0.15
price	-0.06	1.00	0.38	-0.34	0.02	0.04
year	-0.07	0.38	1.00	-0.30	-0.03	0.03
odometer	0.02	-0.34	-0.30	1.00	0.03	-0.03
lat	-0.07	0.02	-0.03	0.03	1.00	-0.07
long	-0.15	0.04	0.03	-0.03	-0.07	1.00

Second, the Range of the price starts from zero which is an outlier because there is no possibility of selling a car for 0 price. The max. price has to do with a Mercedes-Benz car as we see in the graph from R studio where it is visualized the max price for every manufacturer.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
price	0	7950	15895	19069	29590	155000

Third, the range of the year that the cars were first published starts from 1905 but it is an outlier as we observe in the graph price per year in the group of years between 1905 and 1950 the prices are low and in 1905 the price is zero so we approve that this is an outlier and it is not a piece of evidence for further analysis.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
year	1905	2010	2014	2012	2017	2022

Fourth, comparing the prices of the manufacturer table and the mean prices of these vehicles we observe that there are some manufacturers that sell expensive their vehicles like Ferrari but the amount of cars they sell are not as much as Mercedes-Benz. Mercedes-Benz sells many cars and the range of the price of their vehicles is wider than Ferrari's.

Fifth, our observation is that the condition of the Ford vehicles that are for sale, they are from good to excellent in the majority of them which is remarkable and shows that this manufacturer focuses on the endurance of the auto parts in their vehicles.

Last, the most curious thing is that the majority of the used cars that are for sale use gas as fuel.

6. Conclusion

To sum up, both tools are useful for big data analysis. R is a software which is doing the data cleaning process fast, it needs high-level knowledge in order to make the visualizations you need and the statistical analysis. On the other hand, Tableau is a platform that is not so fast in the data cleaning process in high-volume datasets but it does not need so much statistical knowledge in order to do the statistical analysis and it is easier to make visualizations. As a result, big data analysts can use these two platforms but they should thing which one is better in every occasion.

7. Appendix

```
. spec(vehicles)
.   id = col_double(),
.   url = col_character(),
.   region = col_character(),
.   region_url = col_character(),
.   price = col_double(),
.   year = col_double(),
.   manufacturer = col_character(),
.   model = col_character(),
.   condition = col_character(),
.   cylinders = col_character(),
.   fuel = col_character(),
.   odometer = col_double(),
.   title_status = col_character(),
.   transmission = col_character(),
.   VIN = col_character(),
.   drive = col_character(),
.   size = col_character(),
.   type = col_character(),
.   paint_color = col_character(),
.   image_url = col_character(),
.   description = col_character(),
.   county = col_logical(),
.   state = col_character(),
.   lat = col_double(),
.   long = col_double(),
.   posting_date = col_datetime(format = "")
```

7.1 R code

```
library(readr)

library(dplyr)

library(data.table)

library(tidyverse)

library(parallel)

install.packages("conflicted")

library(conflicted)

conflict_prefer("dplyr", "data.table", warn = FALSE)

vehicles <- read_csv("C:/Users/user/Desktop/vehicles.csv")

vehicles_new <- select(vehicles, id,
region,price,year,manufacturer,model,condition,cylinders,fuel,odometer,title_status,
transmission,VIN,drive,type,paint_color,description,lat,long,posting_date)

spec(vehicles_new)

vehicles_new$price[is.na(vehicles_new$price)] <- mean(vehicles_new$price, na.rm
= TRUE)
```

```
vehicles_new$year[is.na(vehicles_new$year)] <- mean(vehicles_new$year, na.rm = TRUE)

vehicles_new$odometer[is.na(vehicles_new$odometer)] <- mean(vehicles_new$odometer, na.rm = TRUE)

vehicles_new <- na.omit(vehicles_new)

library(ggplot2)

ggplot(data = vehicles_new, aes(x=year, y=price)) +

 geom_point() +

 geom_smooth()

boxplot(vehicles_new$price)

ggplot(data = vehicles_new) +

 geom_point(mapping = aes(x=year, y=price , color= condition ))

ggplot(data = vehicles_new) +

 geom_boxplot(aes(x=id, y=condition))

ggplot(data = vehicles_new) +

 geom_histogram(aes(x=price, y=condition))

cor(vehicles_new$price,vehicles_new$year)

library(corrplot)

library(Hmisc)

install.packages("easystats")

library(correlation)

ggplot(data = vehicles_new) +

 geom_histogram(mapping = aes(x=price))

head(vehicles_new)

ggplot(data = vehicles_new) +

 geom_bin2d(mapping=aes(x=condition, y=manufacturer))

ggplot(vehicles_new,aes(x =condition))+

 geom_bar(fill ="cornflowerblue",color="black")+

 labs(x ="condition",y ="count",title ="vehicles by condition")

ggplot(vehicles_new,aes(y =region))+

 geom_bar(fill ="cornflowerblue",color="black")+

 labs(x ="count",y ="region",title ="vehicles by region")
```

```
plotdata <-vehicles_new %>%
  count(fuel) %>%
  arrange(desc(fuel)) %>%
  mutate(prop =round(n*100/sum(n),1),lab.ypos =cumsum(prop)-0.5*prop)
plotdata$label <-paste0(plotdata$fuel,"\n",
                        round(plotdata$prop),"%")
ggplot(plotdata,aes(x ="",y =prop,fill =fuel)) +
  geom_bar(width =1,stat ="identity",color ="black") +
  geom_text(aes(y =lab.ypos,label =label),color ="black") +
  coord_polar("y",start =0,direction =-1) +
  theme_void() +
  theme() +
  labs(title ="vehicles by fuel")
ggplot(Marriage,aes(x =age)) +
  geom_histogram() +
  labs(title ="Participants by age",x ="Age")
data(vehicles_new,package="mosaicData")
vn <-dplyr::select_if(vehicles_new,is.numeric)
r <-cor(vn,use="complete.obs")
round(r,2)
```

8.References

Anon., χ.χ. *Best Data Analytics Tools & Software (2023) – Forbes Advisor,* s.l.: s.n.

Anon., χ.χ. *Create Elegant Data Visualisations Using the Grammar of Graphics • ggplot2,* s.l.: s.n.

Anon., χ.χ. *R: Skim a data frame, getting useful summary statistics,* s.l.: s.n.

Anon., χ.χ. *R: What is R?,* s.l.: s.n.

Anon., χ.χ. *Tableau (version 2020.3),* s.l.: DOI: dx.doi.org/10.5195/jmla.2021.1135 jmla.mlanet.org.

Chen, C., Hao, L. & Xu, C., 2017. *Comparative analysis of used car price evaluation models,* s.l.: American Institute of Physics Inc..

C, O., χ.χ. *Making the decision on buying second-hand car market using data mining techniques,* s.l.: s.n.

C, O., Z, H., G, R. & S, P., χ.χ. *Multiple Linear Regression Applications Automobile Pricing,* s.l.: s.n.

Gelman, A. και συν., 2005. *Analysis of variance? Why it is more important than ever,* s.l.: Institute of Mathematical Statistics.

Hoelscher, J. & Mortimer, A., 2018. *Using Tableau to visualize data and drive decision-making,* s.l.: Pergamon.

Ihaka, R. & Gentleman, R., 1996. *R: A Language for Data Analysis and Graphics,* s.l.: s.n.

Kaya, E., Agca, M., Adiguzel, F. & Cetin, M., 2018. *Spatial data analysis with R programming for environment,* s.l.: Taylor & Francis.

Sallee, J. M., West, S. E. & Fan, W., 2016. *Do consumers recognize the value of fuel economy? Evidence from used car prices and gasoline price fluctuations,* s.l.: North-Holland.

Wu, J. D., Hsu, C. C. & Chen, H. C., 2009. *An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference,* s.l.: s.n.

Kabacoff, R., χ.χ. *Data Visualization with R,* s.l.: s.n.