Probability Theory and Statistics

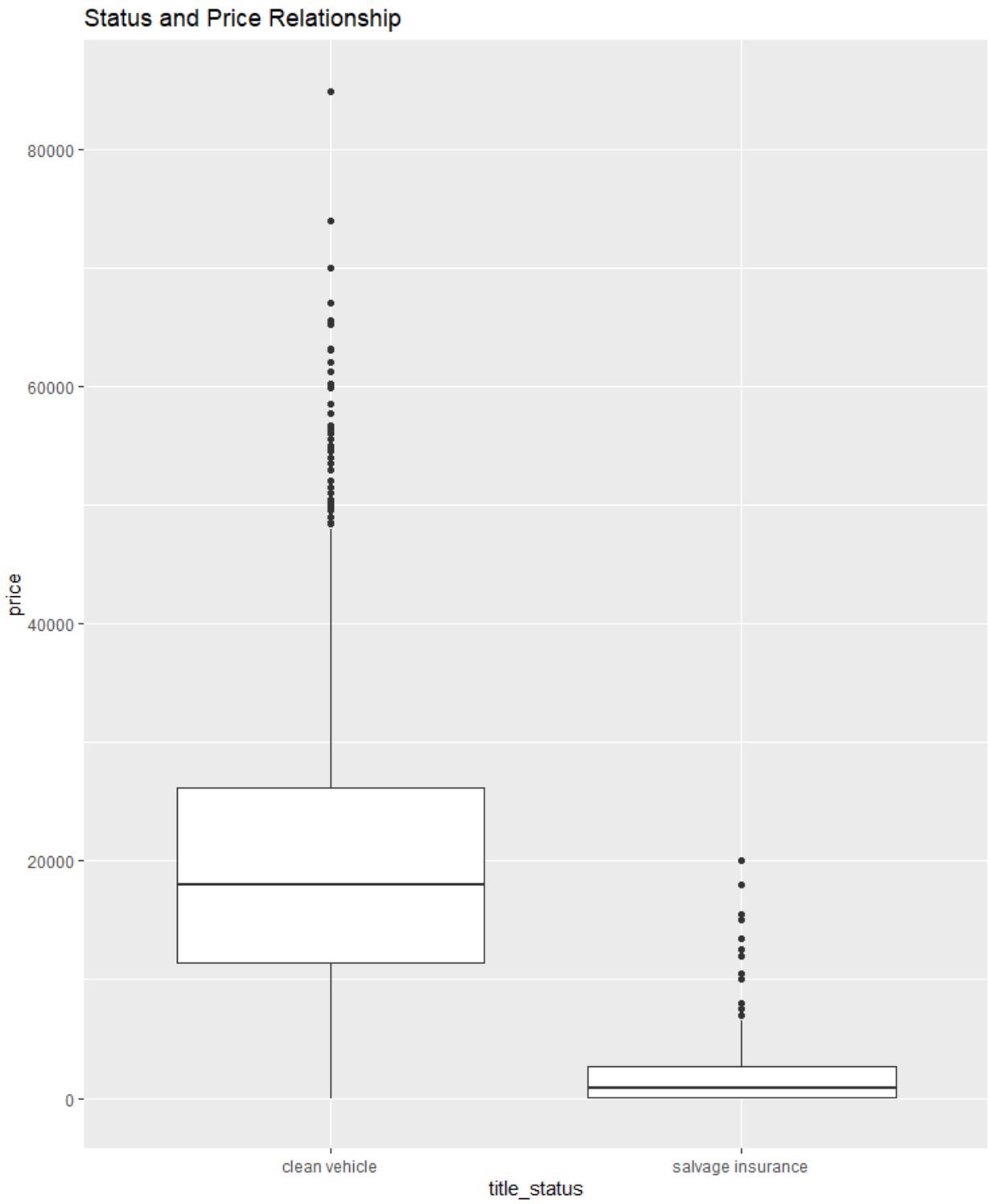
Rashad Gafarli 2018IVSB

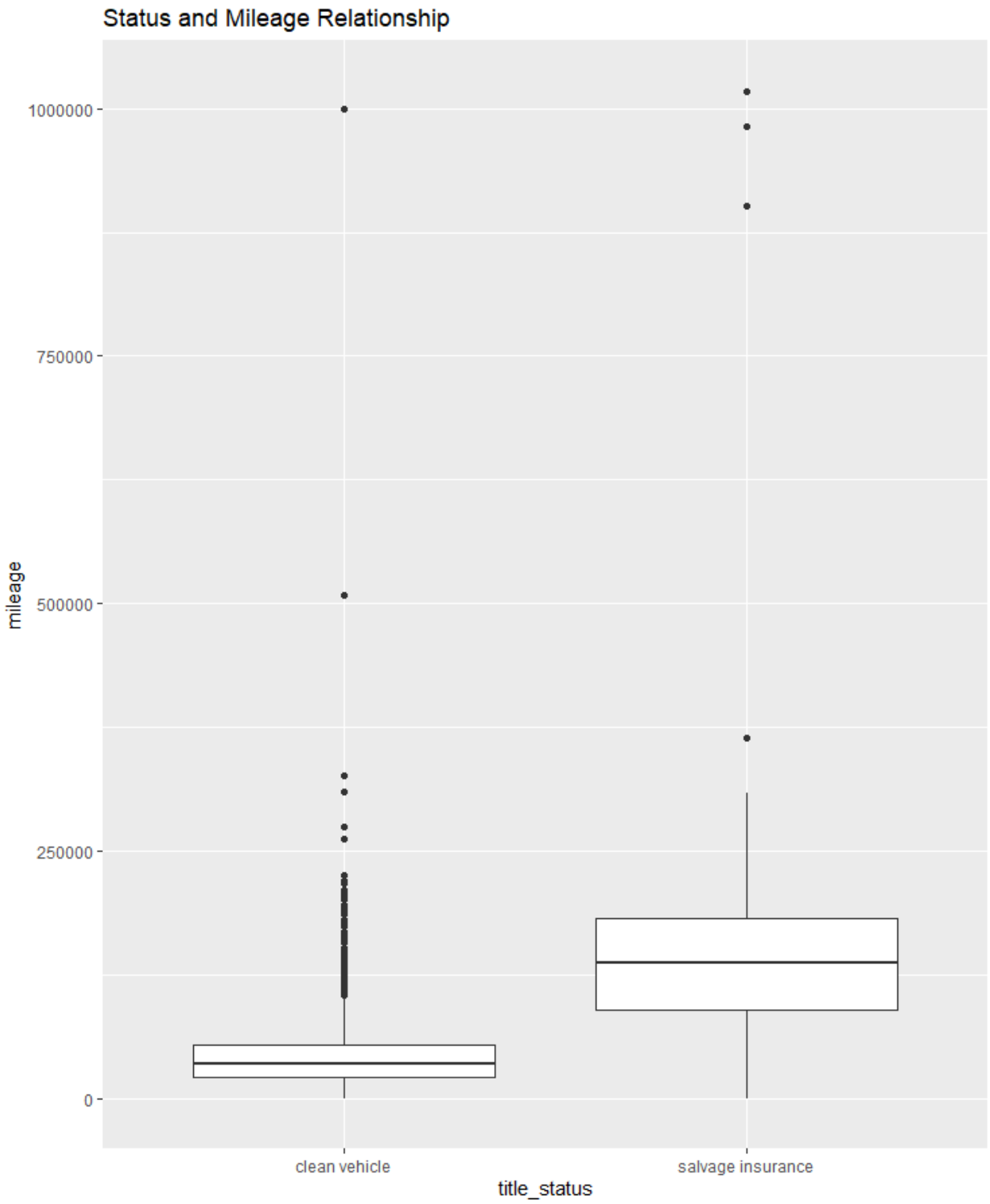
https://www.kaggle.com/doaaalsenani/usa-cers-dataset

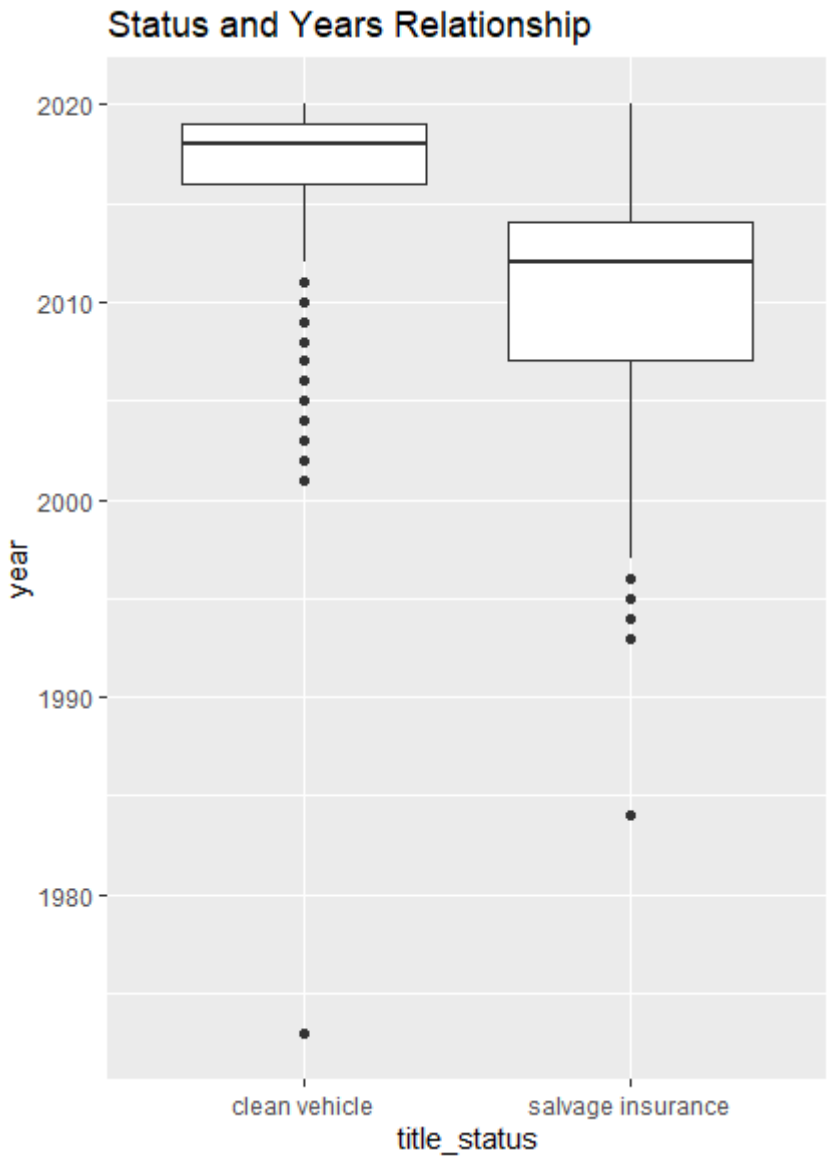
A dataset scraped from AUCTION EXPORT.com on 28 brands of clean and used vehicles for sale In the United States.
Acquired from Kaggle.com with fair use law provided by the site and poster of data.
The dataset possesses 11 variables as follows:
Price – The sale price of the vehicle in the add. Integer
Year – The vehicle registration year. Integer
Brand – The brand of the vehicle. String
Model – Model of the vehicle. String
Color – Color of the vehicle. String
State/City – The location where the car is available for purchase String
Mileage - miles traveled by vehicle. Float
Vin – Unique vehicle identification numbers as a collection of 17 characters (digits and capital letters). String
Title Status – Status indicating if a vehicle is a clean vehicle or salvage insurance in the add. String
Lot – 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. Integer
Condition – Time left for the sale of the car on the add. String
Summary of Data:

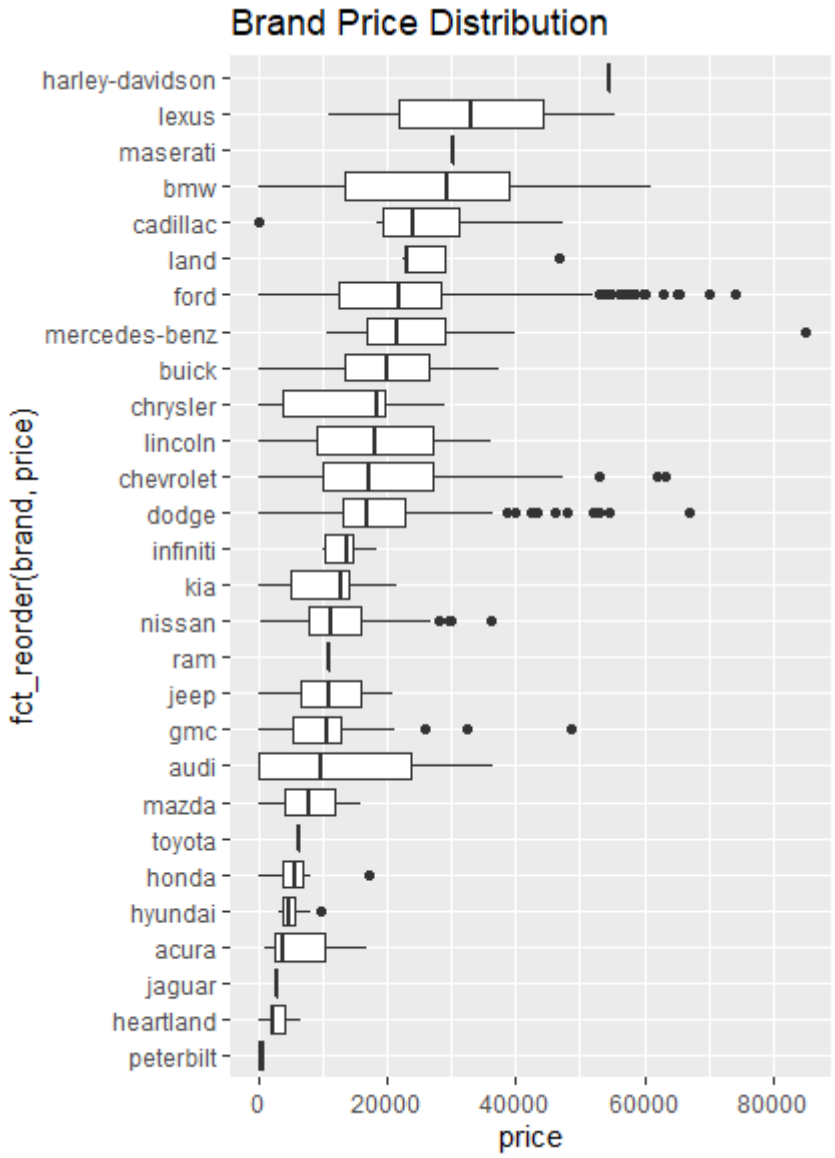
```
       X              price          brand              model               year      title_status
 Min.   :   0.0   Min.   :    0   Length:2499        Length:2499        Min.   :1973   Length:2499
 1st Qu.: 624.5   1st Qu.:10200   Class :character   Class :character   1st Qu.:2016   Class :character
 Median :1249.0   Median :16900   Mode  :character   Mode  :character   Median :2018   Mode  :character
 Mean   :1249.0   Mean   :18768                                         Mean   :2017
 3rd Qu.:1873.5   3rd Qu.:25556                                         3rd Qu.:2019
 Max.   :2498.0   Max.   :84900                                         Max.   :2020
    mileage           color               vin                 lot                state              country
 Min.   :      0   Length:2499        Length:2499        Min.   :159348797   Length:2499        Length:2499
 1st Qu.:  21467   Class :character   Class :character   1st Qu.:167625331   Class :character   Class :character
 Median :  35365   Mode  :character   Mode  :character   Median :167745058   Mode  :character   Mode  :character
 Mean   :  52299                                         Mean   :167691389
 3rd Qu.:  63473                                         3rd Qu.:167779772
 Max.   :1017936                                         Max.   :167805500
  condition
 Length:2499
 Class :character
 Mode  :character
```

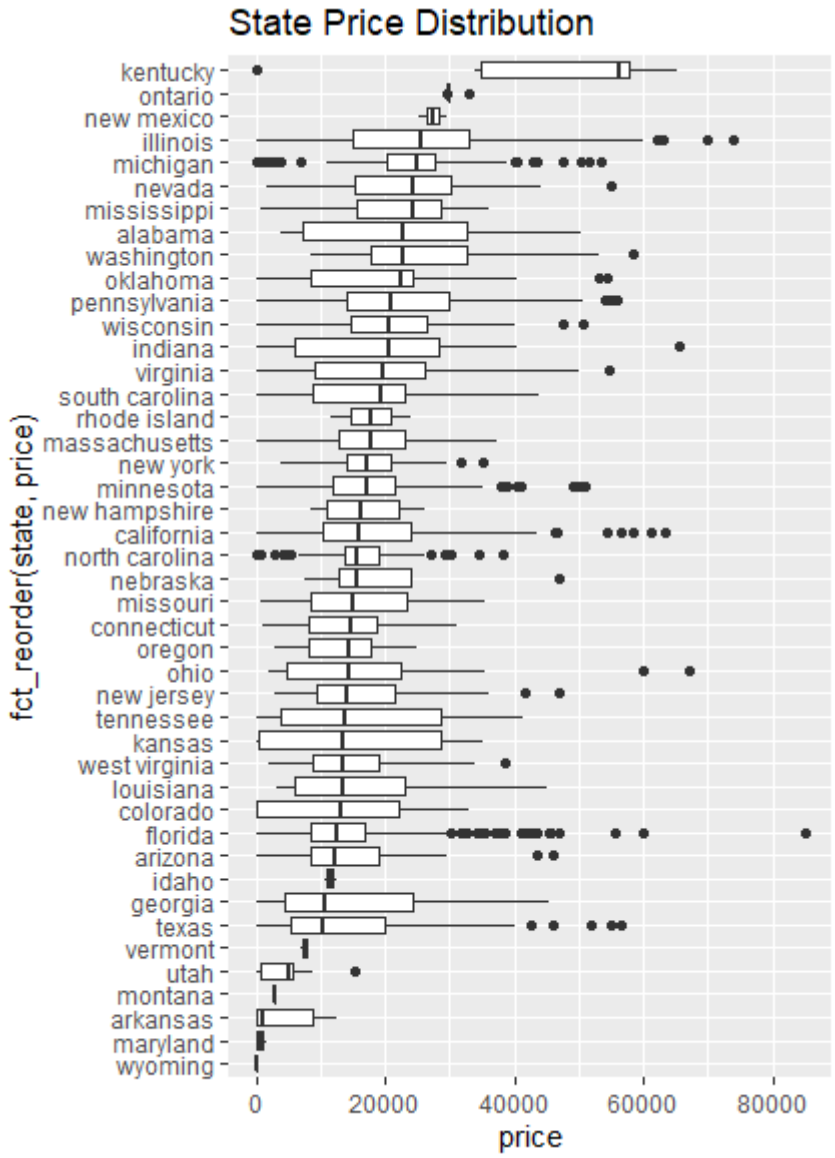
1. Boxplot Status and Years shows that there is an inclination for cars made more recently to be sold as clean vehicles while older cars are more likely to be put on a as salvage insurance.
2. Boxplot Status and Price relationship clearly displays that vehicles which are classified as salvage insurance are put on sale more cheaply compared to clean vehicles.
3. Boxplot Status and Mileage display that cars that have put on more miles are more likely to be salvage insurance vehicles. Remove outliers to verify.
4. Brand price distribution provides insight that most offers no matter the brand lie between the 7500 to 47500 range. Ford brand vehicles display the widest range as well as having some of the priciest single vehicles.
5. State price distribution provides insight that most prices in lie between 0 to 30000 with Kentucky having the most expensive vehicles.
6. Car Price histogram provides a display of the prices being prevalent between the 0 – 32500 range.
7. Car Model Years histogram displays that vehicles between 2014 and 2020 are much more frequently put-on sale compared the previous year.
8. The Price Year histogram shows an upward trend for prices the more modern the vehicle is.
9. The Most Popular Colors Display that White colored vehicles are on auction most frequently follows by Black, Gray, and Silver.
10. The Most Popular Colors by brand display that Toyota, Jaguar, Heartland, and Harley-Davison provide the least variety with only 1 option a color pallet, while Dodge and Chevrolet have the most variety of colors.
11. The Correlation Matrix provides a strong relationship between price and year of manufacture, but a weak one between price and miles travelled, and very little correlation between manufacture of a vehicle and mileage.
12. The Price and Mileage relationship regression displays that the more the car has been driven the lower in value it becomes.
13. The test and training sets display a similar regression to the total sum of data.

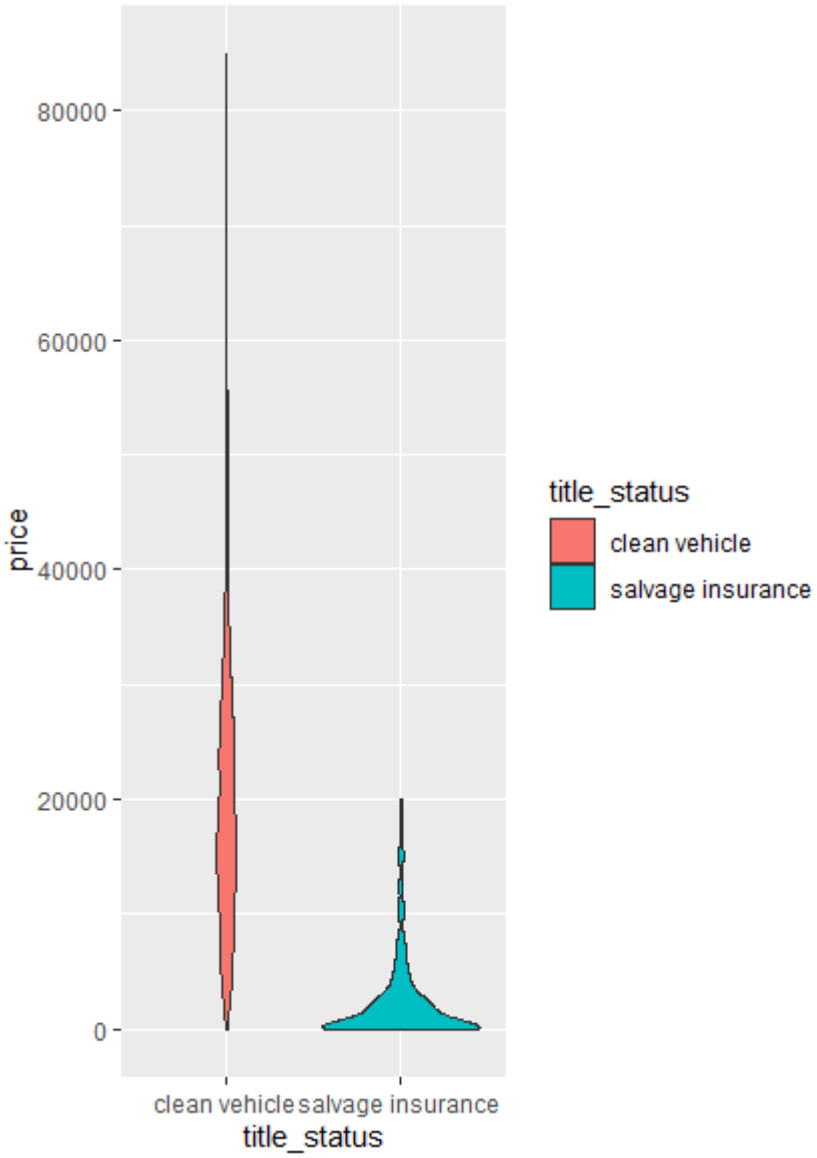

Status and Price Relationship
80000
60000
40000
20000
0
price
clean vehicle
salvage insurance
title_status

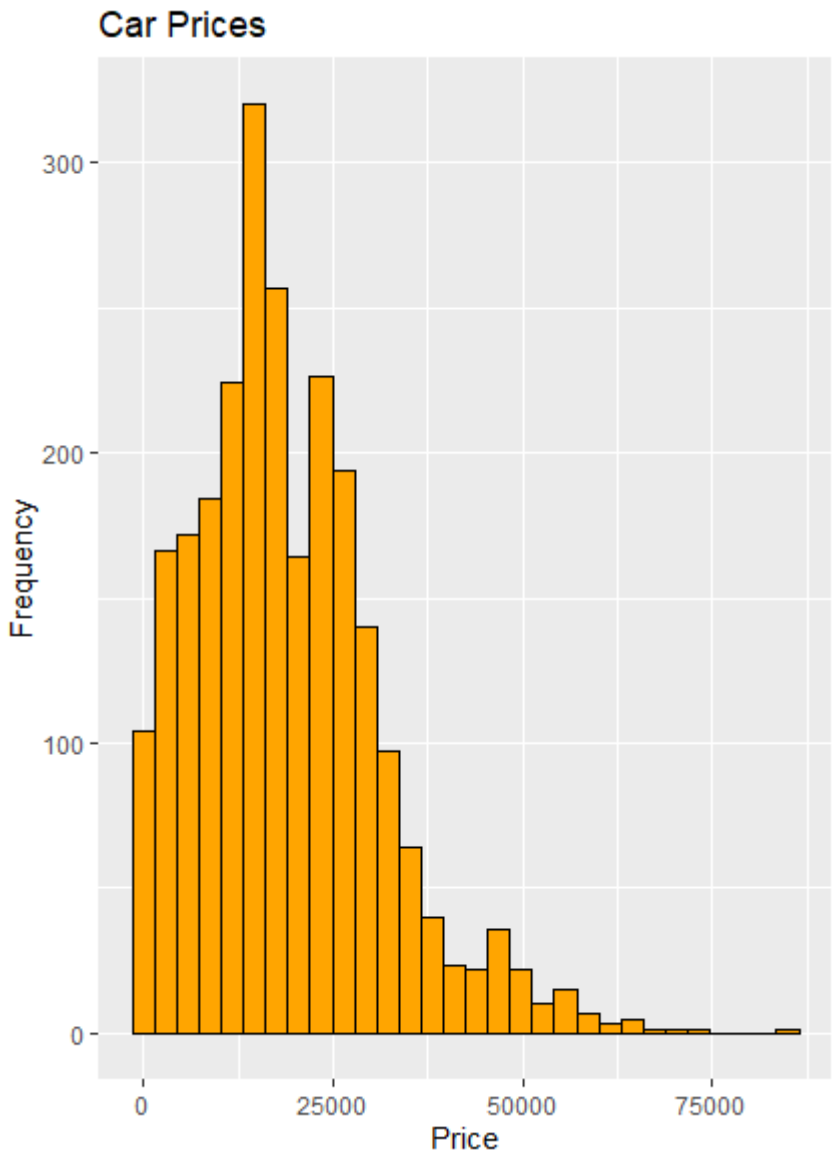

Status and Mileage Relationship
1000000
750000
500000
250000
0
mileage
clean vehicle
salvage insurance
title_status

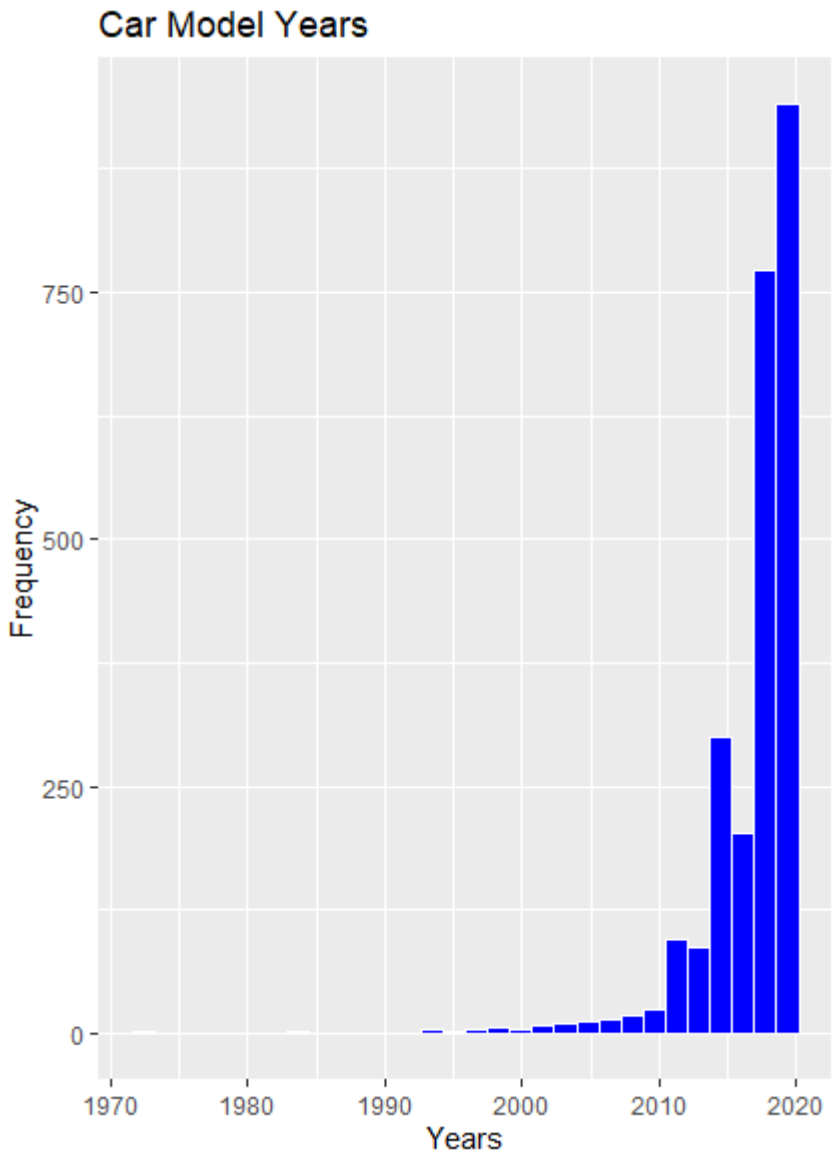

Status and Years Relationship
2020
2010
2000
1990
1980
year
clean vehicle
salvage insurance
title_status

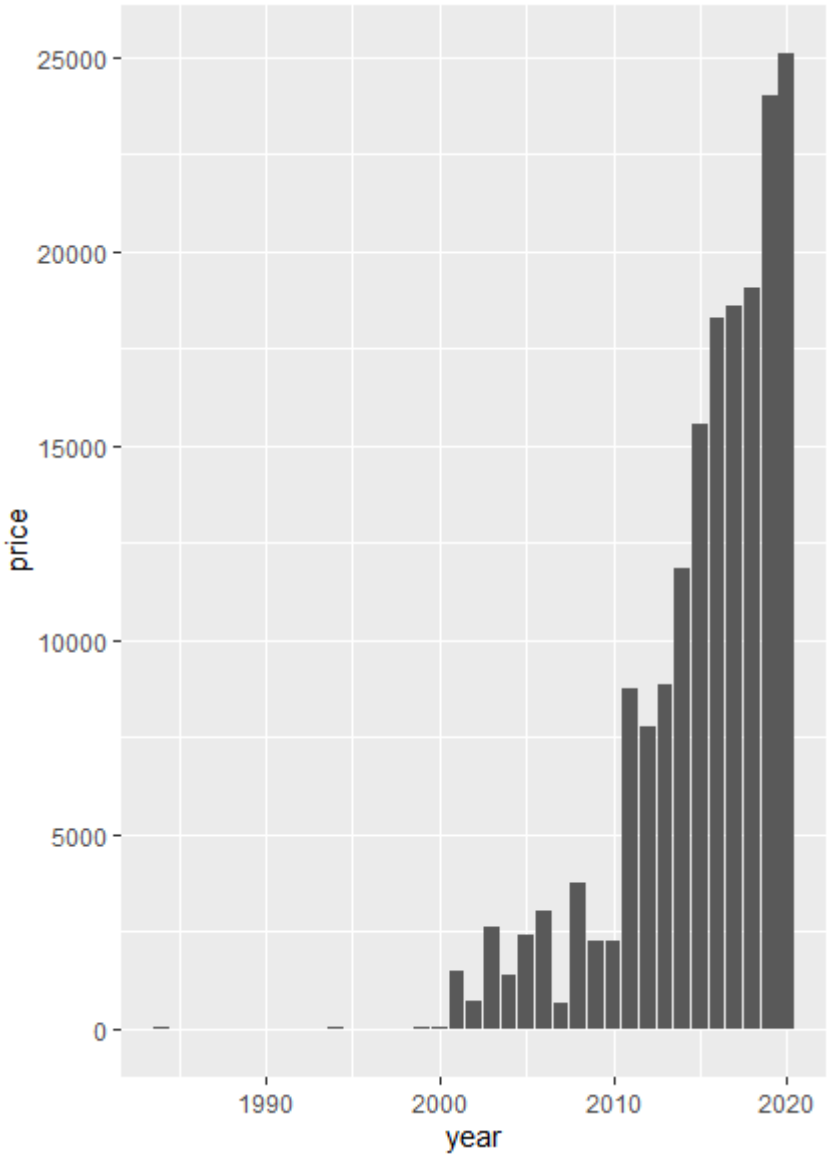

Brand Price Distribution
harley-davidson
lexus
maserati
bmw
cadillac
land
ford
mercedes-benz
buick
chrysler
lincoln
chevrolet
dodge
infiniti
kia
nissan
ram
jeep
gmc
audi
mazda
toyota
honda
hyundai
acura
jaguar
heartland
peterbilt
fct_reorder(brand, price)
0
20000
40000
60000
80000
price

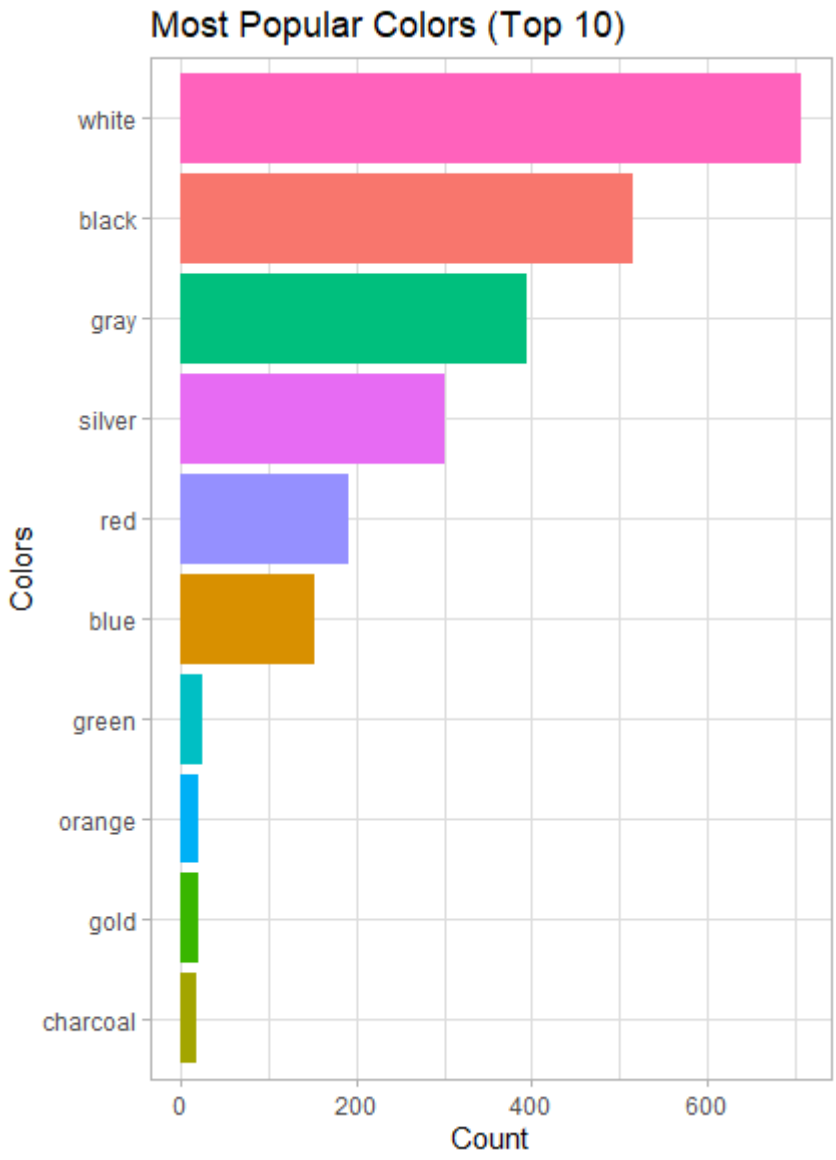

State Price Distribution
fct_reorder(state, price)
kentucky
ontario
new mexico
illinois
michigan
nevada
mississippi
alabama
washington
oklahoma
pennsylvania
wisconsin
indiana
virginia
south carolina
rhode island
massachusetts
new york
minnesota
new hampshire
california
north carolina
nebraska
missouri
connecticut
oregon
ohio
new jersey
tennessee
kansas
west virginia
louisiana
colorado
florida
arizona
idaho
georgia
texas
vermont
utah
montana
arkansas
maryland
wyoming
0
20000
40000
60000
80000
price

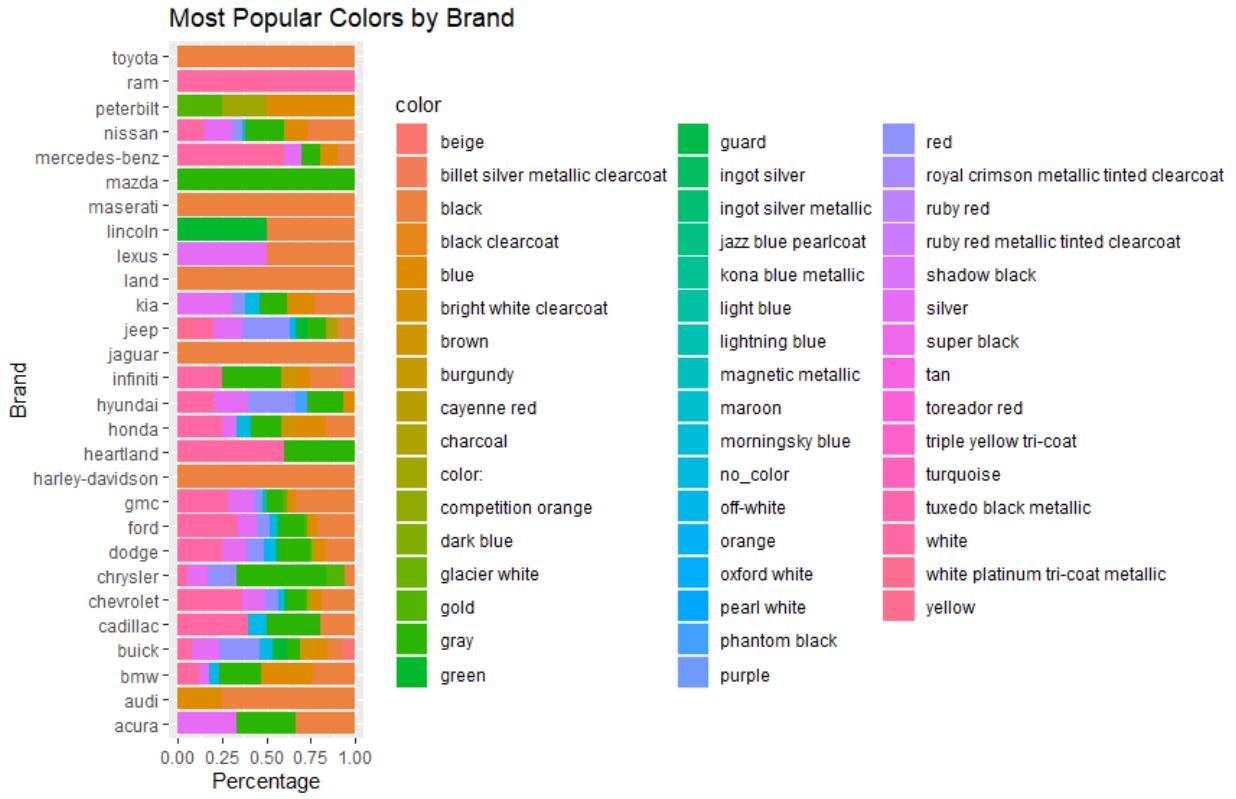

80000
60000
40000
20000
0
price
clean vehicle
salvage insurance
title_status
title_status
clean vehicle
salvage insurance

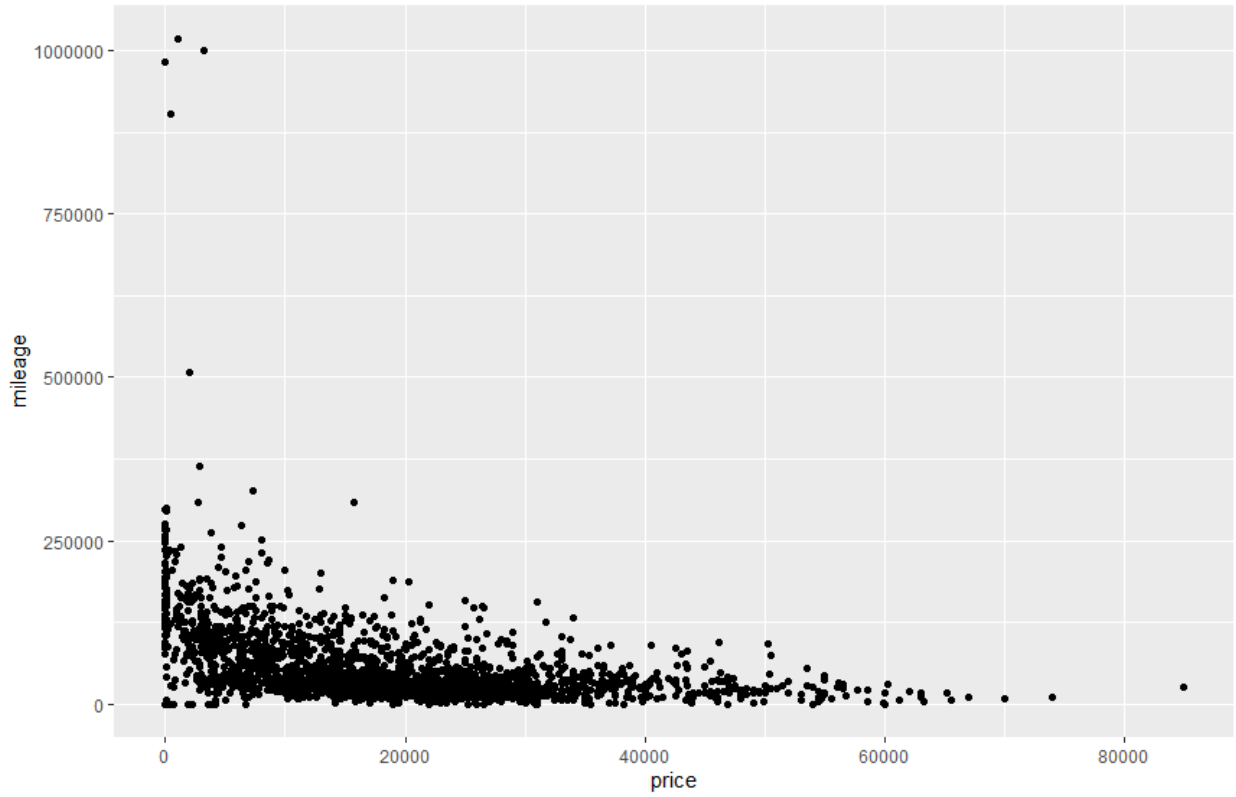

Car Prices
300
200
100
0
Frequency
0
25000
50000
75000
Price

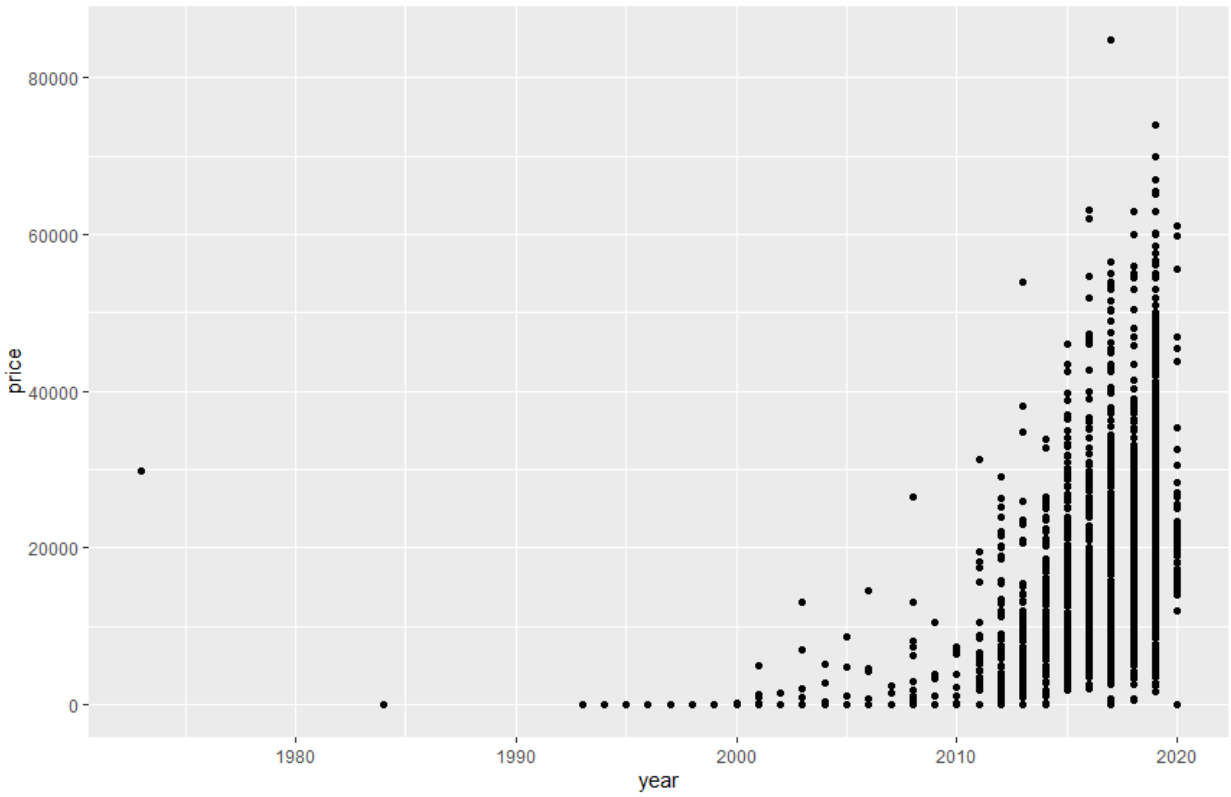

Car Model Years
Frequency
750
500
250
0
1970
1980
1990
2000
2010
2020
Years

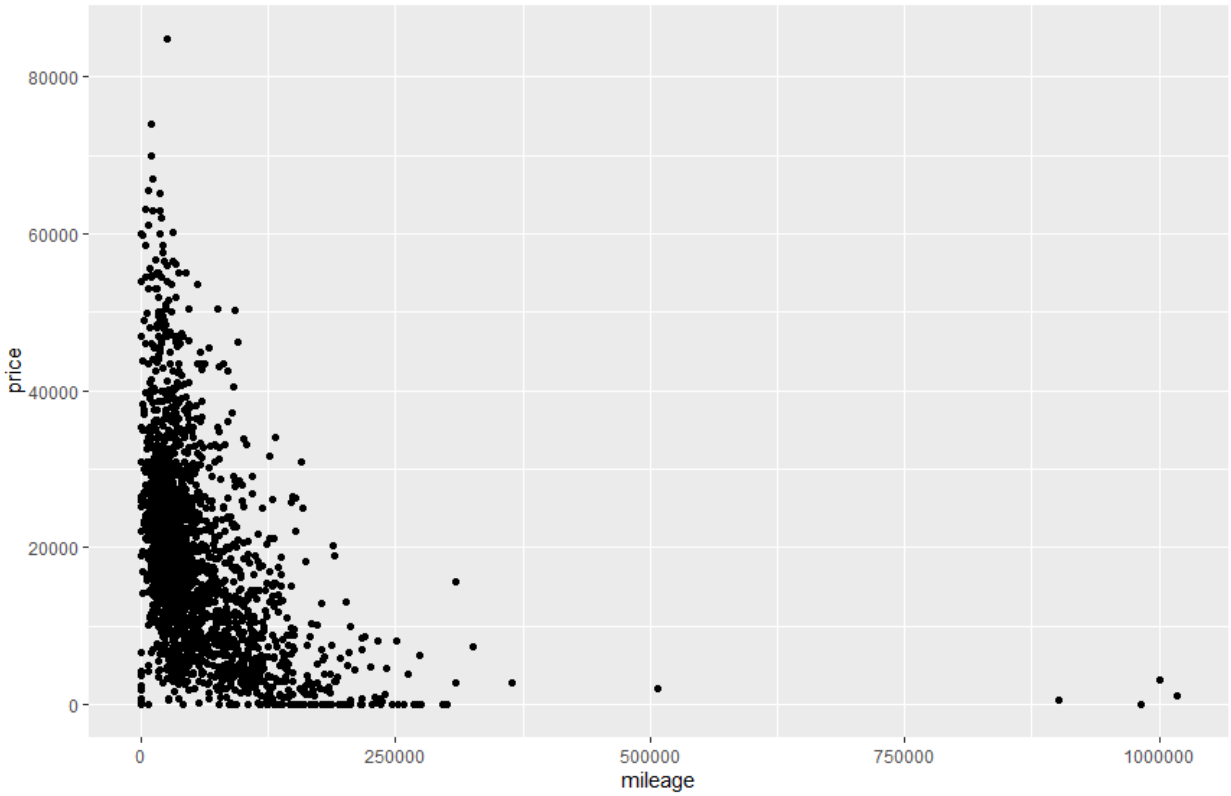

25000
20000
15000
10000
5000
0
price
1990
2000
2010
2020
year

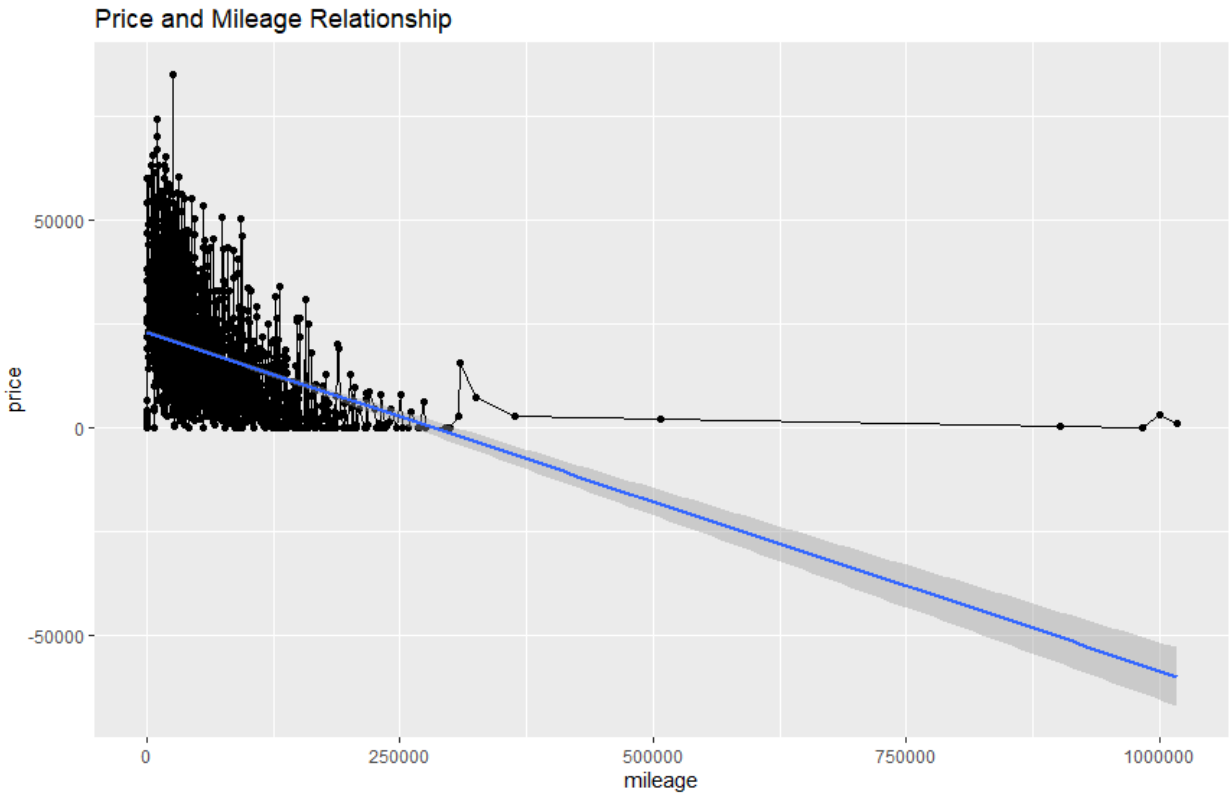

Most Popular Colors (Top 10)
white
black
gray
silver
red
blue
green
orange
gold
charcoal
Colors
0
200
400
600
Count

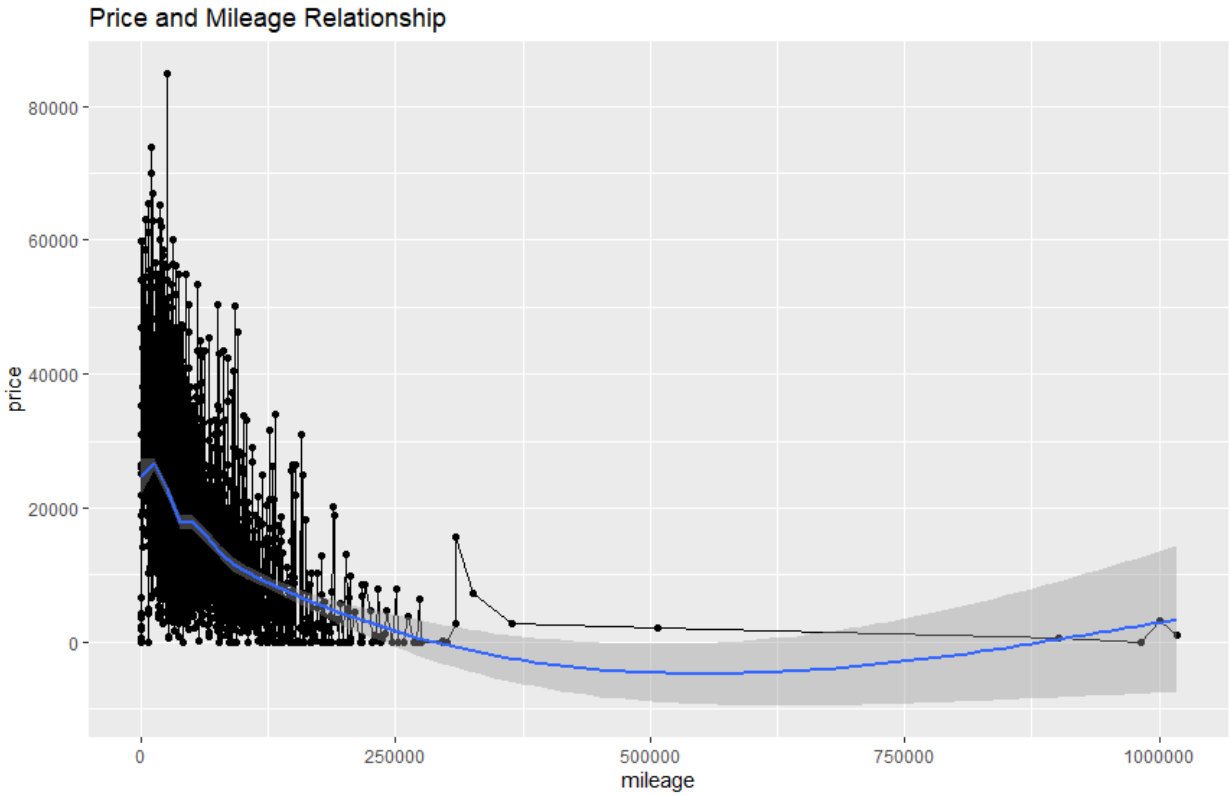

Most Popular Colors by Brand
Brand
toyota
ram
peterbilt
nissan
mercedes-benz
mazda
maserati
lincoln
lexus
land
kia
jeep
jaguar
infiniti
hyundai
honda
heartland
harley-davidson
gmc
ford
dodge
chrysler
chevrolet
cadillac
buick
bmw
audi
acura
0.00 0.25 0.50 0.75 1.00
Percentage
color
beige
billet silver metallic clearcoat
black
black clearcoat
blue
bright white clearcoat
brown
burgundy
cayenne red
charcoal
color:
competition orange
dark blue
glacier white
gold
gray
green
guard
ingot silver
ingot silver metallic
jazz blue pearlcoat
kona blue metallic
light blue
lightning blue
magnetic metallic
maroon
morningsky blue
no_color
off-white
orange
oxford white
pearl white
phantom black
purple
red
royal crimson metallic tinted clearcoat
ruby red
ruby red metallic tinted clearcoat
shadow black
silver
super black
tan
toreador red
triple yellow tri-coat
turquoise
tuxedo black metallic
white
white platinum tri-coat metallic
yellow

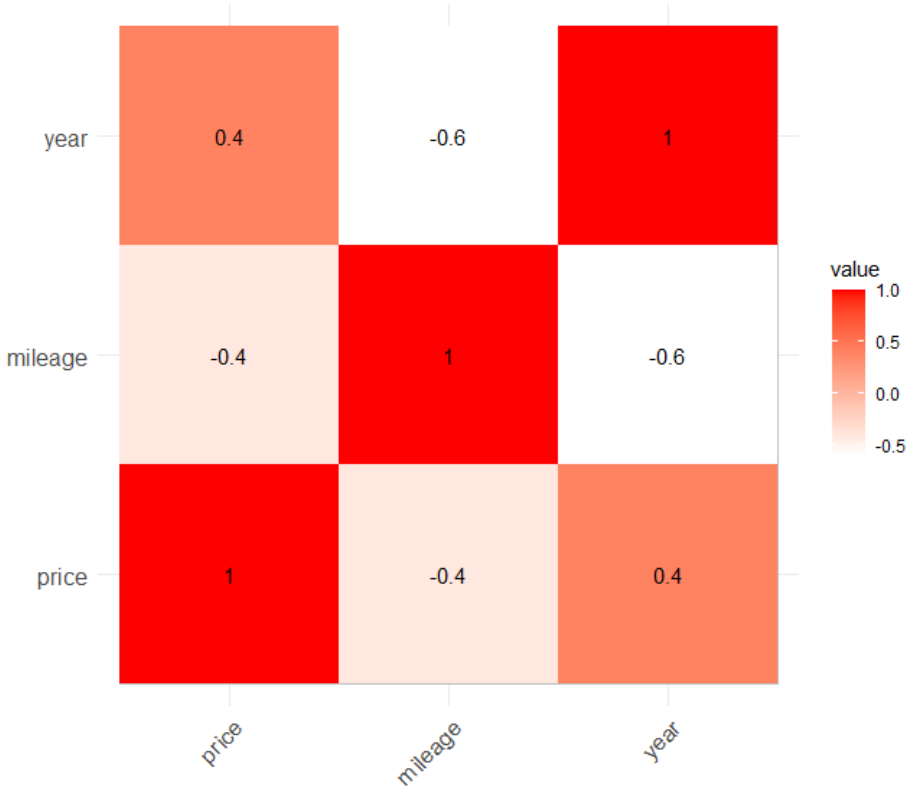

mileage
1000000
750000
500000
250000
0
0
20000
40000
60000
80000
price

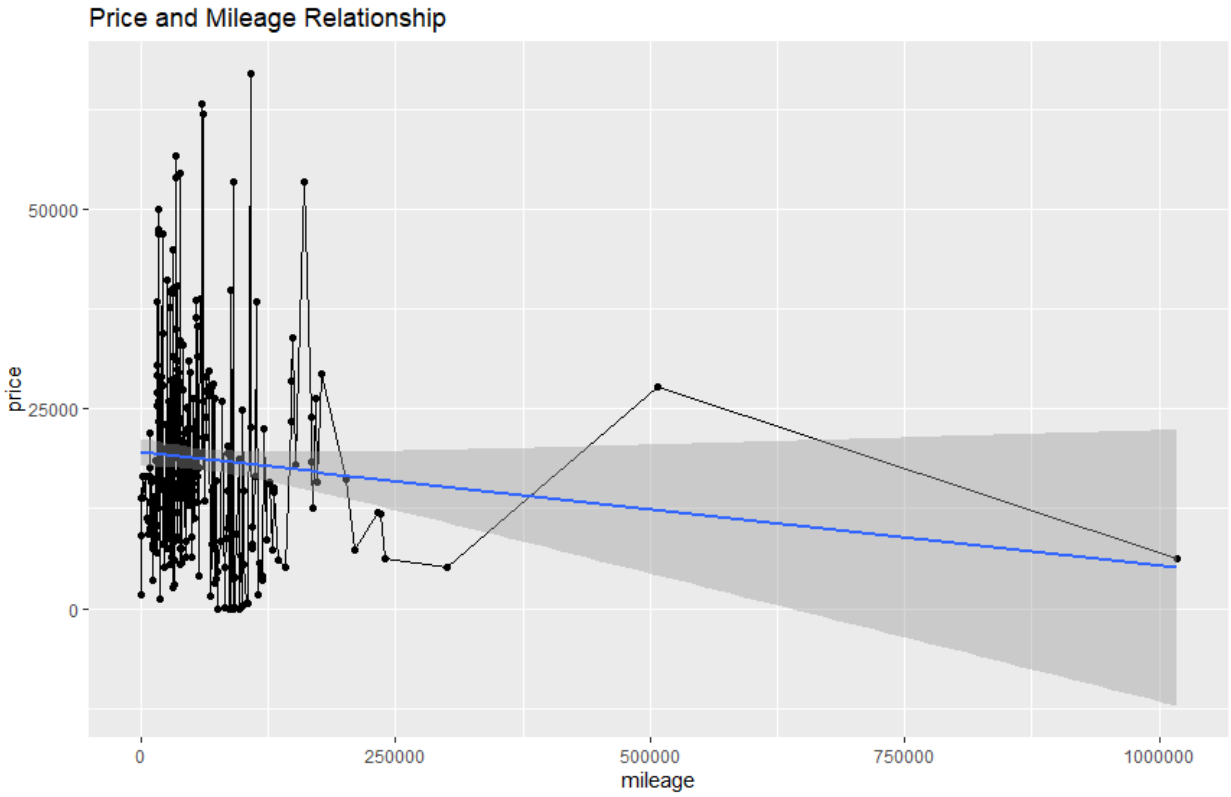

80000
60000
40000
20000
0
price
1980
1990
2000
2010
2020
year

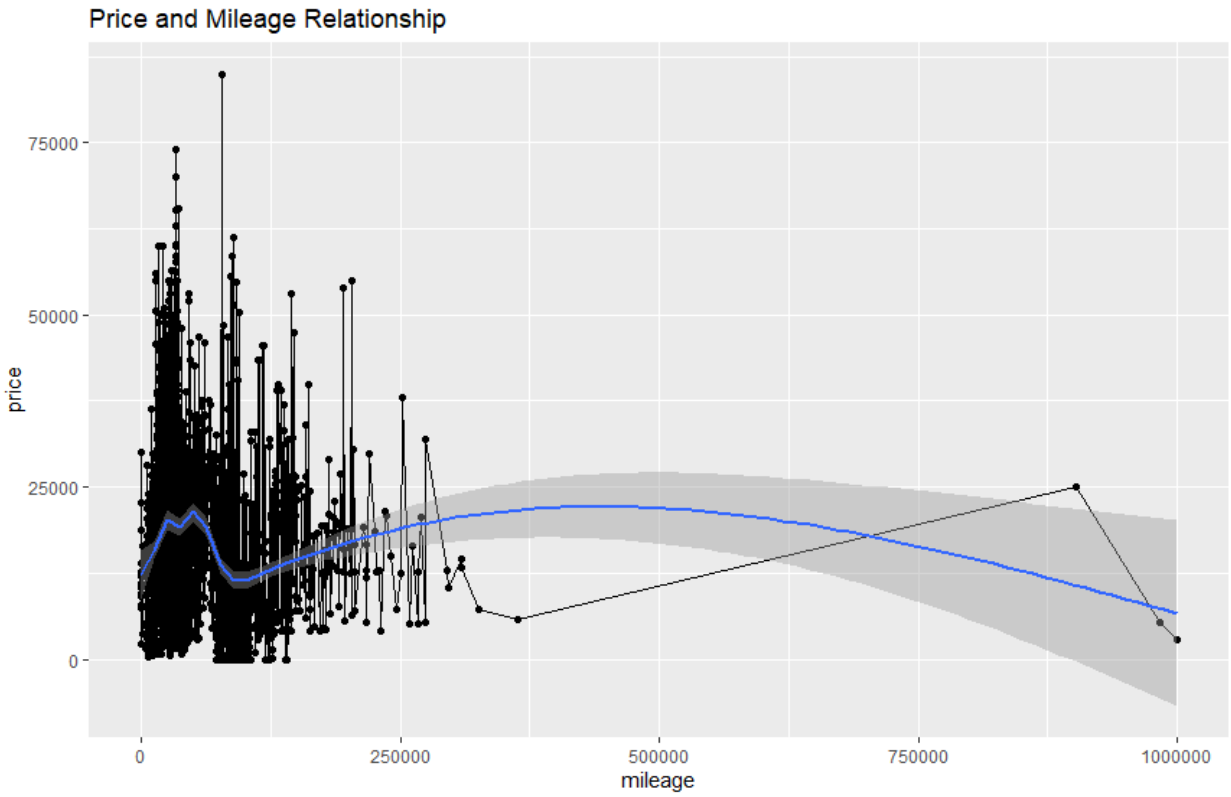

80000
60000
40000
20000
0
price
0
250000
500000
750000
1000000
mileage

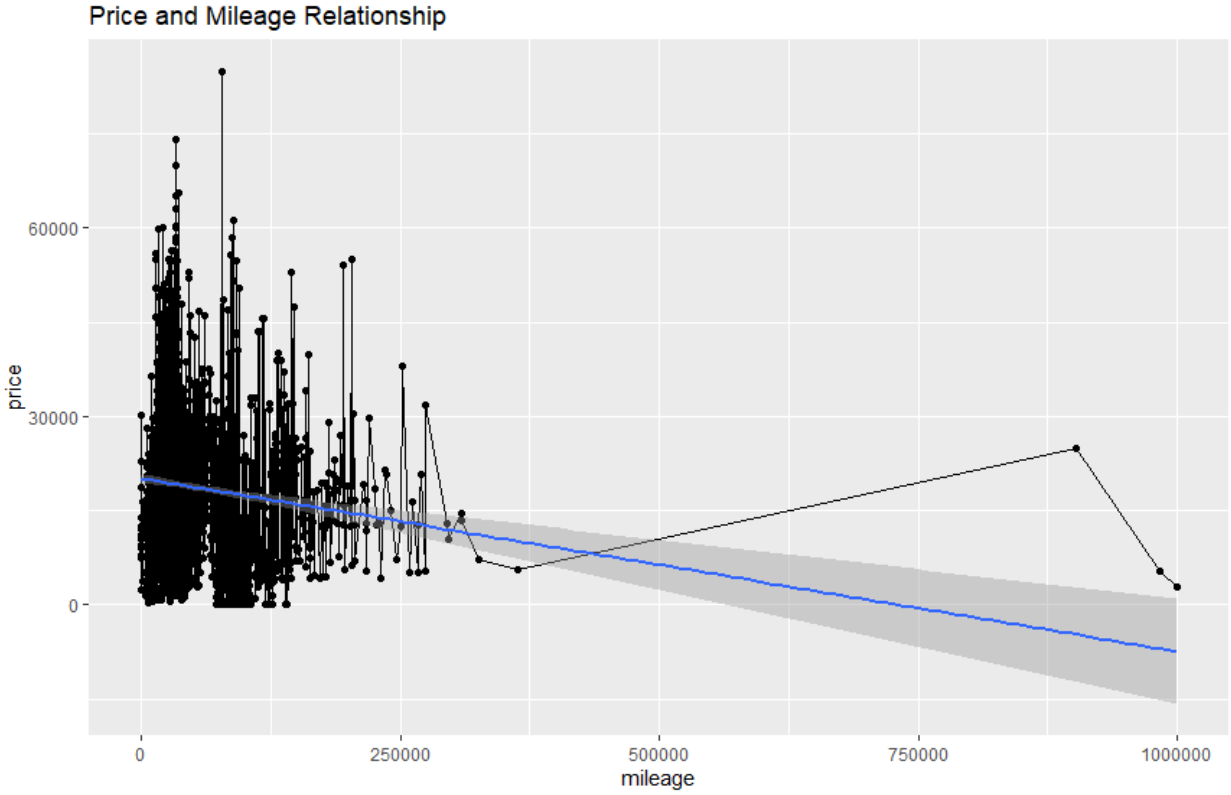

Price and Mileage Relationship
price
50000
0
-50000
0
250000
500000
750000
1000000
mileage

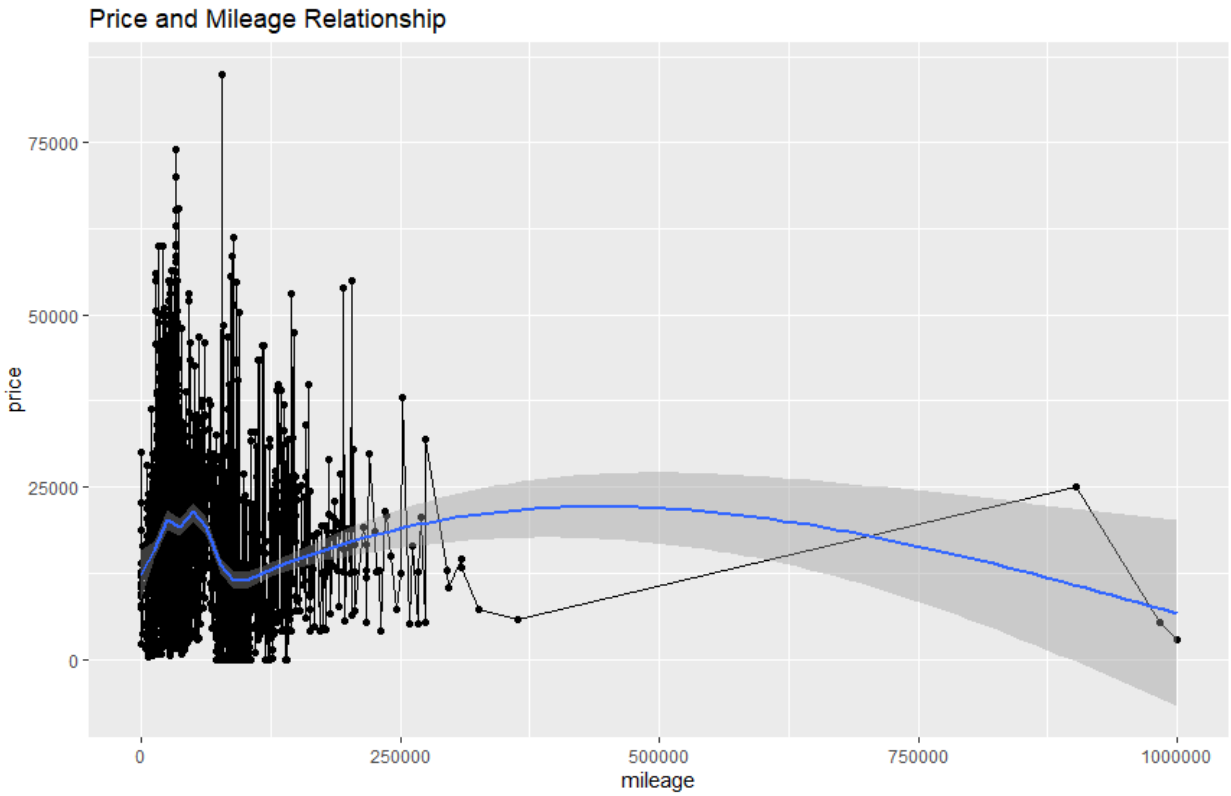

Price and Mileage Relationship
price
80000
60000
40000
20000
0
0
250000
500000
750000
1000000
mileage


year
mileage
price
0.4
-0.6
1
-0.4
1
-0.6
1
-0.4
0.4
value
1.0
0.5
0.0
-0.5
price
mileage
year


Price and Mileage Relationship
price
50000
25000
0
0
250000
500000
750000
1000000
mileage


Price and Mileage Relationship
75000
50000
25000
0
price
0
250000
500000
750000
1000000
mileage


Price and Mileage Relationship
60000
30000
0
price
0
250000
500000
750000
1000000
mileage


Price and Mileage Relationship
75000
50000
25000
0
price
0
250000
500000
750000
1000000
mileage