STA141 Assignment 1 II

Weitong(Jessie) Lin

ID: 913513998

Download the dataset *vehicles.rda* from website

```
setwd('~/Desktop/UC Davis/141/STA141 Assignment 1 II')
download.file('http://eeyore.ucdavis.edu/stat141/Data/vehicles.rda',destfile="vehicles.rda")
print(load("vehicles.rda"))
```

```
## [1] "vposts"
```

1. find at least 3 types of anomalies in the data. Provide succinct justification for identifying them as anomalies. Then correct the corresponding observations appropriately, again providing justification. What impact does this have on analyzing the data?

What's wrong with "year"?

Now let's see the "year"

```
sort(unique(vposts$year))
```

```
##  [1]    4 1900 1921 1922 1923 1925 1926 1927 1928 1929 1930 1931 1932 1933
## [15] 1934 1935 1936 1937 1938 1939 1940 1941 1942 1945 1946 1947 1948 1949
## [29] 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963
## [43] 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977
## [57] 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991
## [71] 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005
## [85] 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2022
```

I find that there are "4", "1900",2016", "2022"years existing in the data which is not reasonable.

- **"4" year**:

```
vposts[which(vposts$year == 4),]
```

```
##                     id
## posted9673 5233798193
##                                                                      ti
tle
## posted9673 argolic eni-04 JEeP wraNgler Clean lEATHeR - $2532 (chica
go)
##


                                                                       b
ody
## posted9673 \n        and passengeranwig Please do not low ball, and
no dealers please mlkzxv AM/FM cassette player-muli CD player\nPlease d
o not low ball, and no dealers please  and passenger\nAM/FM cassette pl
ayer-muli CD player Please do not low ball, and no dealers please louwt
bwl
##                lat    long              posted updated drive odomete
r type
## posted9673 42.1458 -88.023 2015-09-22 09:23:17    <NA>  <NA>       N
A <NA>
##                   header condition cylinders fuel size transmission b
yOwner
## posted9673 04 vctvhmfdk      good        NA  gas <NA>    automatic
  TRUE
##               city                time
## posted9673 chicago 2015-09-22 11:35:00
##                                          description
## posted9673 argolic eni-04 JEeP wraNgler Clean lEATHeR
##                           location                      url price yea
r
## posted9673   (chicago)   pic map  /chc/cto/5233798193.html  2532
4
##            maker makerMethod
## posted9673  jeep           1
```

```
vposts[which(vposts$year == 4),]$year=2004
```

After watching the "title" for this car, which shows"eni-04 JEeP wraNgler Clean lEATHeR", I change the year"4" into "2004"

- **"1900" year**:

```
vposts[which(vposts$year == 1900),]
```

```
##                     id
## posted30411  5190770818
## posted64811  5203617469
## posted64911  5203619345
## posted80511  5207480246
## posted80611  5207481444
## posted84711  5208175430
```

```
## posted112011 5213391719
##
      title
## posted30411                                    Miata enkei wheels - $3
00 (Folsom)
## posted64811  CAR WON'T PASS SMOG??WE'LL BUY TODAY!! - $750 (GREATER
SACRAMENTO)
## posted64911  CAR WON'T PASS SMOG??WE'LL BUY TODAY!! - $750 (GREATER
SACRAMENTO)
## posted80511  CAR WON'T PASS SMOG??WE'LL BUY TODAY!! - $750 (GREATER
SACRAMENTO)
## posted80611  CAR WON'T PASS SMOG??WE'LL BUY TODAY!! - $750 (GREATER
SACRAMENTO)
## posted84711  CAR WON'T PASS SMOG??WE'LL BUY TODAY!! - $750 (GREATER
SACRAMENTO)
## posted112011    YOUR CAR WON'T PASS SMOG??WE'LL BUY TODAY!! - $750
(SACRAMENTO)
##

              body
## posted30411       \n        Hey I have my stock miata enkei wheels f
or sale. They all have tires that still have a good amount of life left.
 4x100 15x6\n
## posted64811  \n        IF YOU HAVE A RUNNING CAR THAT WON'T PASS SMO
G WE WILL GIVE YOU UP TO $750 FOR YOUR CAR TODAY!! GIVE STEVE A CALL AT
 show contact info
## posted64911  \n        IF YOU HAVE A RUNNING CAR THAT WON'T PASS SMO
G WE WILL GIVE YOU UP TO $750 FOR YOUR CAR TODAY!! GIVE STEVE A CALL AT
 show contact info
## posted80511  \n        IF YOU HAVE A RUNNING CAR THAT WON'T PASS SMO
G WE WILL GIVE YOU UP TO $750 FOR YOUR CAR TODAY!! GIVE STEVE A CALL AT
 show contact info
## posted80611  \n        IF YOU HAVE A RUNNING CAR THAT WON'T PASS SMO
G WE WILL GIVE YOU UP TO $750 FOR YOUR CAR TODAY!! GIVE STEVE A CALL AT
 show contact info
## posted84711  \n        IF YOU HAVE A RUNNING CAR THAT WON'T PASS SMO
G WE WILL GIVE YOU UP TO $750 FOR YOUR CAR TODAY!! GIVE STEVE A CALL AT
 show contact info
## posted112011 \n        IF YOU HAVE A RUNNING CAR THAT WON'T PASS SMO
G WE WILL GIVE YOU UP TO $750 FOR YOUR CAR TODAY!! GIVE STEVE A CALL AT
 show contact info
##                   lat      long              posted             updat
ed
## posted30411       NA        NA 2015-08-25 17:59:55 2015-09-12 00:41:
02
## posted64811  38.4797 -121.4438 2015-09-02 18:11:30 2015-09-11 23:34:
35
## posted64911  38.4797 -121.4438 2015-09-02 18:13:04 2015-09-12 00:06:
51
## posted80511  38.4797 -121.4438 2015-09-05 09:10:36 2015-09-11 23:47:
```

```
03
## posted80611  38.4797 -121.4438 2015-09-05 09:11:16 2015-09-11 23:30:
55
## posted84711  38.4797 -121.4438 2015-09-05 16:28:57 2015-09-12 01:38:
11
## posted112011      NA        NA 2015-09-09 07:53:11 2015-09-12 00:38:
47
##              drive odometer type      header condition cylinders  fu
el
## posted30411   <NA>       NA <NA> 1900 Wheels      <NA>        NA oth
er
## posted64811   <NA>       NA <NA>    1900 CAR      fair        NA   g
as
## posted64911   <NA>       NA <NA>    1900 CAR      fair        NA   g
as
## posted80511   <NA>       NA <NA>    1900 CAR      fair        NA   g
as
## posted80611   <NA>       NA <NA>    1900 CAR      fair        NA   g
as
## posted84711   <NA>       NA <NA>    1900 CAR      fair        NA   g
as
## posted112011  <NA>       NA <NA>    1900 CAR      fair        NA   g
as
##              size transmission byOwner city                time
## posted30411  <NA>        other    TRUE  sac 2015-09-12 00:46:00
## posted64811  <NA>    automatic    TRUE  sac 2015-09-11 23:39:00
## posted64911  <NA>    automatic    TRUE  sac 2015-09-12 00:11:00
## posted80511  <NA>    automatic    TRUE  sac 2015-09-11 23:52:00
## posted80611  <NA>    automatic    TRUE  sac 2015-09-11 23:35:00
## posted84711  <NA>    automatic    TRUE  sac 2015-09-12 01:43:00
## posted112011 <NA>    automatic    TRUE  sac 2015-09-12 00:43:00
##                                          description
## posted30411                           Miata enkei wheels
## posted64811       CAR WON'T PASS SMOG??WE'LL BUY TODAY!!
## posted64911       CAR WON'T PASS SMOG??WE'LL BUY TODAY!!
## posted80511       CAR WON'T PASS SMOG??WE'LL BUY TODAY!!
## posted80611       CAR WON'T PASS SMOG??WE'LL BUY TODAY!!
## posted84711       CAR WON'T PASS SMOG??WE'LL BUY TODAY!!
## posted112011 YOUR CAR WON'T PASS SMOG??WE'LL BUY TODAY!!
##                                       location                 url
price
## posted30411                     (Folsom)   pic  /cto/5190770818.html
  300
## posted64811        (GREATER SACRAMENTO)   map  /cto/5203617469.html
  750
## posted64911    (GREATER SACRAMENTO)   pic map  /cto/5203619345.html
  750
## posted80511        (GREATER SACRAMENTO)   map  /cto/5207480246.html
  750
## posted80611    (GREATER SACRAMENTO)   pic map  /cto/5207481444.html
```

```
  750
## posted84711        (GREATER SACRAMENTO)   map  /cto/5208175430.html
  750
## posted112011               (SACRAMENTO)   pic  /cto/5213391719.html
  750
##               year maker makerMethod
## posted30411   1900  <NA>           0
## posted64811   1900 dodge           3
## posted64911   1900 dodge           3
## posted80511   1900 dodge           3
## posted80611   1900 dodge           3
## posted84711   1900 dodge           3
## posted112011 1900 dodge           3
```

When I look at these records, I find that it's "we will buy the cars which won't pass smog test". This is not about selling cars. 1900 is not the car year. So I remove these data.

```
# remove data
vposts = vposts[-which(vposts$year == 1900),]
```

- **"2022" year**:

```
vposts[which(vposts$year == 2022),]
```

```
##                       id
## posted21888 5218261938
##
                   title
## posted21888 Check Out This Spotless 2022 Honda Odyssey with 117,102
Miles - $6999 (Jamaica)
##
```

```
                                      body
## posted21888 2022 Honda Odyssey LX AT Automatic Gray Cloth on Silver
Silver Pearl Metallic 104208\nTake a look at this 2022 Honda Odyssey LX
 AT. It has only 117102 miles.\nColor: Silver Cloth on Silver Silver Pe
arl Metallic\nEngine: 3.5 V6 Cylinder Engine\nStock number: 104208\nTra
nsmission: Automatic\nMiles: 117,102\nQueens Best Auto, Inc.\n179-18, H
illside Ave. Jamaica, New York 11432\nPLEASE REPLY TO THIS AD TO GET MO
RE INFORMATION ABOUT THIS VEHICLE\nOR 718    297   2900\nCARFAX REPORT
IS AVAILABLE ON DEMANDFINANCING AVAILABLE FOR ALL CUSTOMERS.\n641e3384-
5b99-4cbd-91e6-75885952a684\n 3.1.7\n
##              lat long              posted             updated drive
## posted21888  NA   NA 2015-09-12 08:24:38 2015-09-12 08:24:40  <NA>
```

```
##              odometer type              header condition cylinders fue
l size
## posted21888   117102 <NA> 2022 Honda Odyssey excellent        NA  ga
s <NA>
##              transmission byOwner city                time
## posted21888    automatic   FALSE  nyc 2015-09-12 11:24:00
##                                                                descri
ption
## posted21888 Check Out This Spotless 2022 Honda Odyssey with 117,102
Miles
##                       location                      url price year m
aker
## posted21888   (Jamaica)   pic  /que/ctd/5218261938.html  6999 2022 h
onda
##              makerMethod
## posted21888         1.5
```

```
vposts[which(vposts$year == 2022),]$year=2012
```

After watching the detail, I find the "odometer" for this car is a little bit large considering it's a Japaness car, so "2002" is more reasonable.

- **"2016" year**:

```
length(which(vposts$year == 2016))
```

```
## [1] 206
```

There are 206 records show that the year is 2016. We can not see their details one by one which may spend too much time. So we will use grepl to extract information from "title","body" and "description"

```
# search whether it's around 20xx year which shows in "body"
Year2016 = which(!is.na(vposts$year) & vposts$year == 2016)
grab_Year2016 = grepl("20[0-9][0-9]", vposts$body[Year2016])
vposts$year[Year2016[grab_Year2016]] = as.integer(gsub(".*(20[0-9][0-
9]).*", "\\1", vposts$body[Year2016[grab_Year2016]]))
# search whether it's around 20xx year which shows in "title"
Year2016 = which(!is.na(vposts$year) & vposts$year == 2016)
grab_Year2016 = grepl("20[0-9][0-9]", vposts$title[Year2016])
vposts$year[Year2016[grab_Year2016]] = as.integer(gsub(".*(20[0-9][0-
9]).*", "\\1",vposts$title[Year2016[grab_Year2016]]))
# search whether it's around 20xx year which shows in "description"
Year2016 = which(!is.na(vposts$year) & vposts$year == 2016)
grab_Year2016 = grepl("20[0-9][0-9]", vposts$description[Year2016])
vposts$year[Year2016[grab_Year2016]] = as.integer(gsub(".*(20[0-9][0-
9]).*", "\\1",vposts$description[Year2016[grab_Year2016]]))
# search whether it's around 19xx year which shows in "body"
Year2016 = which(!is.na(vposts$year) & vposts$year == 2016)
grab_Year2016 = grepl("19[0-9][0-9]", vposts$body[Year2016])
vposts$year[Year2016[grab_Year2016]] = as.integer(gsub(".*(19[0-9][0-
```

```
9]).*", "\\1", vposts$body[Year2016[grab_Year2016]]))
# search whether it's around 19xx year which shows in "title"
Year2016 = which(!is.na(vposts$year) & vposts$year == 2016)
grab_Year2016 = grepl("19[0-9][0-9]", vposts$title[Year2016])
vposts$year[Year2016[grab_Year2016]] = as.integer(gsub(".*(19[0-9][0-
9]).*", "\\1",vposts$title[Year2016[grab_Year2016]]))
# search whether it's around 19xx year which shows in "description"
Year2016 = which(!is.na(vposts$year) & vposts$year == 2016)
grab_Year2016 = grepl("19[0-9][0-9]", vposts$description[Year2016])
vposts$year[Year2016[grab_Year2016]] = as.integer(gsub(".*(19[0-9][0-
9]).*", "\\1", vposts$description[Year2016[grab_Year2016]]))

length(which(vposts$year == 2016))

## [1] 92
```

However, some records still can not find any information from other variables, so I remove these points.

```
vposts=vposts[-which(vposts$year == 2016),]
```

To avoid some wrong message that we get from the detail, I remove those wrong data which caused by "grepl"

```
vposts=subset(vposts, vposts$year <2016)
```

This year can really cause some misunderstanding to users. When a user want to search cars with an age limit, this would cause users to miss a lot of cars.

Outliers in "Price"

First, let's take a summary of "price" in data set.

```
summary(vposts$price)

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA'
s
##         1      2995      6700     49550     13500 600000000      330
8
```

From the result, I found that there are some "NA"s existing. So I remove all the dataset where the 'price' is "NA":

```
Price_data=with(vposts,subset(vposts,!is.na(price)))
```

Also, we can see that the maximum price and minimum price, which are $600030000 and $1 are incredible unreasonable. So the variable "price" is a kind of anomaly that we need to correct it.

Firstly I will go through the "price"" which is large unreasonably.

```
TopTenPrice=sort(Price_data$price,decreasing=T)[1:10]
TopTenPrice
```

```
##  [1] 600030000 600030000  30002500   9999999    569500    559500
400000
##  [8]    359000    286763    240000
```

It seems like somethings wrong because it's too experience for a car. Also we can briefly see some hidden patterns which reveal somethings wrong.

Now let's see the description for the cars which have the highest value.

```
Price_data[which(Price_data$price==TopTenPrice[1]),]$body
```

```
## [1] "\n        We have 1968 & 1969 Pontiac GTO's.\nCurrently we are
working on a 1968 end a 1969 Gto project is almost complete.\nOur Inten
tion is the custom to specification by owner.\nCost will be between $60
00 & $30,000. This will be depending on the car in the condition and th
e Owner financial capabilities. \nSerious inquires only inquiries only..
 please call Tony at \n show contact info\n\n    "
## [2] "\n        We have 1968 & 1969 Pontiac GTO's.\nCurrently we are
working on a 1968 end a 1969 Gto project is almost complete.\nOur Inten
tion is the custom to specification by owner.\nCost will be between $60
00 & $30,000. This will be depending on the car in the condition and th
e Owner financial capabilities. \nSerious inquires only inquiries only..
 please call Tony at \n show contact info\n\n    "
```

From the information showed above, we get two exactly similar data which means that there are two duplicated data. So I remove one of them.

```
Price_data=Price_data[-(which(Price_data$price==TopTenPrice[2])[1]),]
```

Also, we can know that the price actually should be between $6000 and $30000, not $600030000. Thus, I take a median of the 6000 and 30000 to be as the price for this car, which is:

```
median(c(6000,30000))
```

```
## [1] 18000
```

```
Price_data[which(Price_data$price==max(Price_data$price)),]$price=media
n(c(6000,30000))
```

Now let's move on to the highest value for the revised dataset, which is:

```
max(Price_data$price)
```

```
## [1] 30002500
```

Now let's see the description and the maker for this car:

```
Price_data[which.max(Price_data$price),c("header","body")]
```

```
##                              header
## posted6903 2002 Caddy Seville sls
##
```

```
                                                                    body
## posted6903 \n        clean, fully loaded, nice shine, good running e
ngine and trans, willing to trade for old school or truc
k???????????????????? Mounted on 22 inch rims new tires no bends no crack
s\n
```

Then we search it in the Google. From the data showed in cars.com, it should be around $2500 to $3000. So we take the median to assign this typo.

```
median(c(2500,3000))

## [1] 2750

Price_data[which(Price_data$price==max(Price_data$price,na.rm=T)),]$pri
ce=median(c(2500,3000))
```

Now let's move on to the next highest value for the revised dataset, which is:

```
max(Price_data$price,na.rm=T)

## [1] 9999999
```

$9999999 seems like a really typo. Now let's see the description and the maker for this car:

```
Price_data[which.max(Price_data$price),]$header

## [1] "2001 Honda Accord"

Price_data[which.max(Price_data$price),]$body

## [1] "\n        Selling my car for some lunch money. $20 OBO. Comes w
ith complimentary Oboe.\n    "
```

Owner acually said that it would be $20 obo. In my option, It's more like a joke. So, I remove the whole record of this car.

```
Price_data=Price_data[-(which(Price_data$price==max(Price_data$pric
e))),]
```

Now let's keep moving to next two larger value.

```
TopTenPrice[c(5,6)]

## [1] 569500 559500
```

These two cars are the same type of car. From the data showed in cars.com, it should be around $9500. So I correct these two typo as $9500.

```
Price_data[which(Price_data$price==TopTenPrice[5]),]$price=9500
Price_data[which(Price_data$price==TopTenPrice[6]),]$price=9500
```

Let's move on to the highest value in the revised dataset.

```
max(Price_data$price)
```

```
## [1] 4e+05
```

Now let's see the description and the maker for this car:

```
Price_data[which.max(Price_data$price),c('header','maker','body')]
```

```
##                    header maker
## posted23788 2006 FORD GT  ford
##
```

```
                                                 body
## posted23788 \n        *CANADIAN CAR NO ACCIDENTS*RARE LOW KM*Less th
an 2,000 kms!!! You don't have to worry about depreciation on this supe
rb 2006 Ford GT!!!!** This vehicle has its original front wind shield s
tickers from factory. Safety equipment includes: ABS, Xenon headlights,
 Passenger Airbag - Cancellable, Front fog/driving lights...Other featu
res include: Leather seats, Power locks, Manual Transmission,\nFeatures
 and Specifications\nOther Features\nAir Conditioning\nCD Player\nKeyle
ss Entry\nLeather Interior\nCruise Control\nCup Holder\n5.4L DOHC MPFI
supercharged handbuilt all-aluminum V8 engine\nElectronic ignition syst
em w/push-button start\nDry sump lubrication system\nTwin disc self-adj
usting hydraulic clutch\nMid-engine/rear wheel drive\n48-AH maintenance
-free battery w/battery saver feature\nFront/rear independent unequal l
ength (SLA) aluminum suspension w/steel coil springs\nFront/rear non-ad
justable forged aluminum shock absorbers w/forged aluminum housings\nFr
ont/rear tubular stabilizer bars\nTire inflation kit-no spare tire avai
lable\nPwr rack & pinion steering\nBrembo front & rear vented 4-piston
disc brakes w/black painted calipers\n66.2 litre fuel tank\nStainless s
teel dual exhaust\n1-306-525-1555 MORGAN\n
```

From the data showed in cars.com, $40000 sounds an appropriate price for a 2006 Ford GT. So I won't change this record.

Now let's move to next price:

```
Price_data[which(Price_data$price==TopTenPrice[8]), c("header", "price",
"body")]
```

```
##                               header  price
## posted1460 2010 CHEVROLET SILVERADO 359000
##
```

```
       body
## posted1460 \n        2010 CHEVROLET SILVERADO LTZ 2500HD, 4X4, 6.6L
LMM DURAMAX DIESEL ENGINE, ALLISON 6 SPEED AUTO TRANSMISSION\n HEATED L
EATHER SEATS WITH FULL CENTER CONSOLE, TOWING PACKAGE, POWER WINDOWS/LO
CKS/MIRRORS/SEATS, CLIMATE CONTROL, CD/MP3 PLAYER, PUSH BUTTON 4X4, FOG
 LIGHTS, WOODGRAIN DASH, BOSE, TOW MIRRORS\n TRUCK INCLUDES 33" PRO COM
P XTERRAIN TIRES AND XD WHEELS DPF BACK 5" EXHAUST.\n TRUCK SALE COMES
WITH A BRAND NEW INSPECTION STICKER AND A 30 DAY WARRANTY.\n FRONT END
IS TIGHT WITH MOOG PARTS OIL AND FUEL FILTER CHANGE.\n 103,000 MILES\n
APLUS DIESEL SALES\n 143 PORTLAND RD\n GRAY, ME 04039\nWE SPECIALIZE IN
 DURAMAX TRUCKS AND HAVE MANY TO CHOOSE FROM, IF A DURAMAX IS WHAT YOUR
 LOOKING FOR DO NOT HESITATE TO CONTACT US FOR BOTH SALES AND PERFORMAN
CE. OUR SHOP IS LOCATED 20 MINUTES FROM PORTLAND MAINE BUT WE DO OFFER
DELIVERY ON AN AS NEEDED BASIS FOR THOSE CUSTOMERS UNABLE TO MAKE THE D
RIVE!\n
```

For a 2010 CHEVROLET SILVERADO, I believe that there is an extra zero in the end of price. The real price should be $ 35900

```
Price_data[which(Price_data$price==TopTenPrice[8]),]$price=35900
```

For the next price,

```
Price_data[which(Price_data$price==TopTenPrice[9]), c("header", "price",
"body")]
```

```
##                           header  price
## posted9976 2004 Toyota Corolla 286763
##
```

body
posted9976 La Joya Auto Sales & Lease\n2520 Fremont St. Las Vegas, NV 89104\nHave a question about this vehicle? \nCall at (702) 385-9505\n \n2004 Toyota Corolla 4dr Sdn CE Auto - $286,763\n\n\n\tVIN: 1NXBR32E64Z286763Type: 1Body: SedanTransmission: AutomaticColor: BlackInterior: BlackEngine: 4 CylindersDrivetrain: Front Wheel DriveStock ID: 286763City MPG: 30.50*Hwy MPG: 39.00*\n   *Actual rating may vary   \n   \n\n   Call Lourdes @ (702) 385-9505 for an appointment today.\n   \tWebsite: www.lajoyaauto.net\n   \n \n â\u0080¢ Side-impact door beamsâ\u0080¢ Variable pwr rack & pinion steeringâ\u0080¢ 4-way adjustable front seatsâ\u0080¢ Pwr front vented disc/rear drum brakesâ\u0080¢ Tilt steering wheelâ\u0080¢ Dual front/rear cup holdersâ\u0080¢ Temporary spare tireâ\u0080¢ Trunk entrapment releaseâ\u0080¢ Color-keyed door handlesâ\u0080¢ HD rear window defoggerâ\u0080¢ Digital clockâ\u0080¢ Driver & front passenger airbags Supplemental Restraint System (SRS)â\u0080¢ Front wheel driveâ\u0080¢ 60/40 split fold-down rear seatâ\u0080¢ Daytime running lightsâ\u0080¢ ETR AM/FM stereo w/CD player-inc: (4) speakersâ\u0080¢ Dual 12V aux pwr outletsâ\u0080¢ 1.8L DOHC EFI 16-valve 4-cyl aluminum engine w/VVT-i variable valve timingâ\u0080¢ 13.2 gallon fuel tankâ\u0080¢ Fabric-trimmed interiorâ\u0080¢ 15\\" steel wheels w/full wheel coversâ\u0080¢ Defroster-linked CFC-free air conditioning w/air filterâ\u0080¢ Independent MacPherson strut front/torsion beam rear suspensionâ\u0080¢ Remote releases-inc: fuel-filler door, hood, trunk w/cancel featureâ\u0080¢ Center console w/storageâ\u0080¢ Child restraint system lower anchors & top tether anchorsPhone: (702) 385-9505   \tWebsite: www.lajoyaauto.net\n   \n   2520 Fremont St. Las Vegas, NV 89104\t\t\t\n\n

From the detail, we can see that people who post this car is pretty sure that he will sell a 2004 Toyota Corolla for $286863. Let's find the similar cars in the dataset.(The code here is inspired by Duncan's idea.)

```
car_2004_Corolla=Price_data[ which(Price_data$year %in% c(2004)& Price_
data$price > 100 &Price_data$price < 286763 & grepl(pattern = "Corolla",
 x = Price_data$header, ignore.case = TRUE)), c("header", "price", "mak
er", "year") ]
car_2004_Corolla
```

```
##                                      header price  maker year
## posted770              2004 TOYOTA COROLLA LE  7500 toyota 2004
## posted890                 2004 Toyota corolla  3999 toyota 2004
## posted1081                2004 toyota corolla  5500 toyota 2004
## posted3341                2004 toyota corolla  3500 toyota 2004
## posted4971                2004 Toyota corolla  4200 toyota 2004
## posted8201                2004 Toyota corolla  2900 toyota 2004
## posted21582               2004 Toyota Corolla  5500 toyota 2004
## posted9015                       2004 Corolla  1100 toyota 2004
## posted13863               2004 Toyota Corolla  3100 toyota 2004
## posted13654               2004 TOYOTA COROLLA  5200 toyota 2004
## posted16924            2004 Toyota Corolla LE 10000 toyota 2004
## posted16265             2004 Toyota Corolla S  2800 toyota 2004
## posted18520               2004 Toyota Corolla   499 toyota 2004
## posted8236                2004 Toyota Corolla  4998 toyota 2004
## posted10456               2004 Toyota Corolla   499 toyota 2004
## posted18256               2004 Toyota Corolla  4998 toyota 2004
## posted21646               2004 Toyota Corolla   895 toyota 2004
## posted21656               2004 Toyota Corolla   895 toyota 2004
## posted5447                2004 toyota corolla  5300 toyota 2004
## posted12048               2004 TOYOTA COROLLA  8300 toyota 2004
## posted6189                2004 Toyota Corolla  4290 toyota 2004
## posted7989                2004 toyota corolla  1800 toyota 2004
## posted188113              2004 toyota corolla  4200 toyota 2004
## posted20599         2004 Toyota/Corolla wagon  2100 toyota 2004
## posted21389        2004 toyota corolla matrix  2100 toyota 2004
## posted8350   2004 2004 Toyota Corolla Matrix  4950 toyota 2004
## posted144412              2004 Toyota Corolla  7000 toyota 2004
## posted177012              2004 toyota corolla  3400 toyota 2004
## posted70813         2004 toyota corolla sport  3700 toyota 2004
```

Here we can find that people who sell it may overestimate this car. So let's take the average price of other "2004 TOYOTA COROLLA" and assign it as the price of this car.

```
mean(car_2004_Corolla$price)
```

```
## [1] 3973.207
```

```
Price_data[which(Price_data$price==TopTenPrice[9]),]$price=mean(car_200
4_Corolla$price)
```

For the last top 10 price car,

```
Price_data[which(Price_data$price==TopTenPrice[10]), c("header", "price
","body")]
```

```
##                                header  price
## posted12630 2014 ferrari 458 italia 240000
##


                                body
## posted12630 \n        Selling my 2014 Ferrari 458 Italia F1 Coupe\nH
as 11,936 Miles\nClean Title, Clean Car Fax\nRuns and drives perfect\nS
erious offers only\ncall / text if you have questions or want to check
the car out.\nthanks\npaul\n
```

This price for a ferrari is quite reasonable.

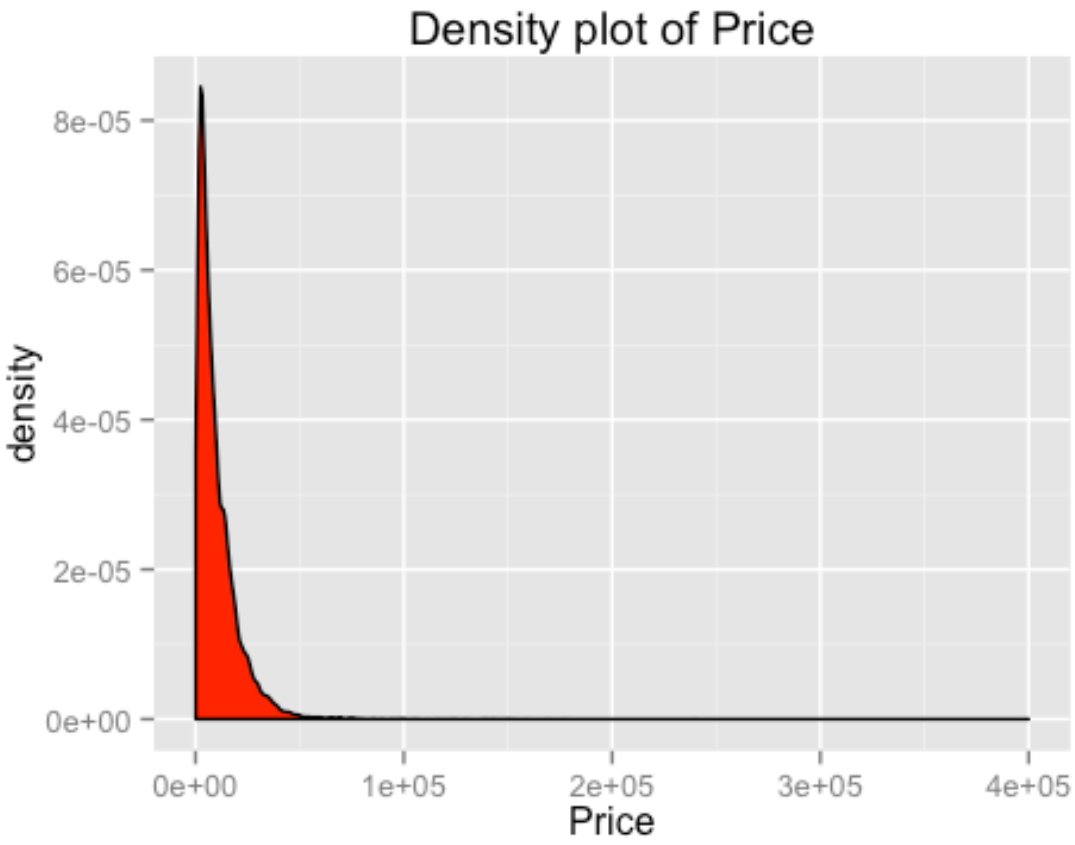
Also, when we look at the data, we can find that there are a lot of cars which are sold with a price under $500. This is quite abnormal.

```
dim(Price_data[Price_data$price<500,c("header", "price", "maker", "year
")])[1]
```

```
## [1] 1611
```

There are 847 records. So I decide to remove them.

```
Price_data=Price_data[-which(Price_data$price<100),]

library(ggplot2)
ggplot(Price_data, aes(x=Price_data$price)) +
  geom_density(fill="red") +
  xlab("Price") +
  labs(title="Density plot of Price")
```



Those extremely large and small prices (i.e. $600030000 and $1) can cause misunderstanding to users. Especially when we calculate mean of price.

Odometer

Now Let's see odometer

```
Price_data$age = 2016 - Price_data$year
Odometer_Price_data=subset(Price_data,!is.na(Price_data$odometer))
summary(Odometer_Price_data$odometer)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.000e+00 4.219e+04 9.300e+04 1.580e+05 1.319e+05 1.235e+09
```

According to the official record, The higghest-vehicle milage is around 3100000 miles. Thus, for those cars who have odometers above 3100000 are not right, we need to correct or remove them.

```
Odometer_Price_data=subset(Odometer_Price_data, Odometer_Price_data$odo
meter<3100000)
```

However, this is still an extreme case. Thus, let's see the quantitle for this data.

```
quantile(x = Odometer_Price_data$odometer, probs = c(0,0.99))
```

```
##       0%      99% 
##      0.0 261915.3
```

99% of data are below 261915 miles. Thus, we remove those data are above 261915 miles, which may be typos.

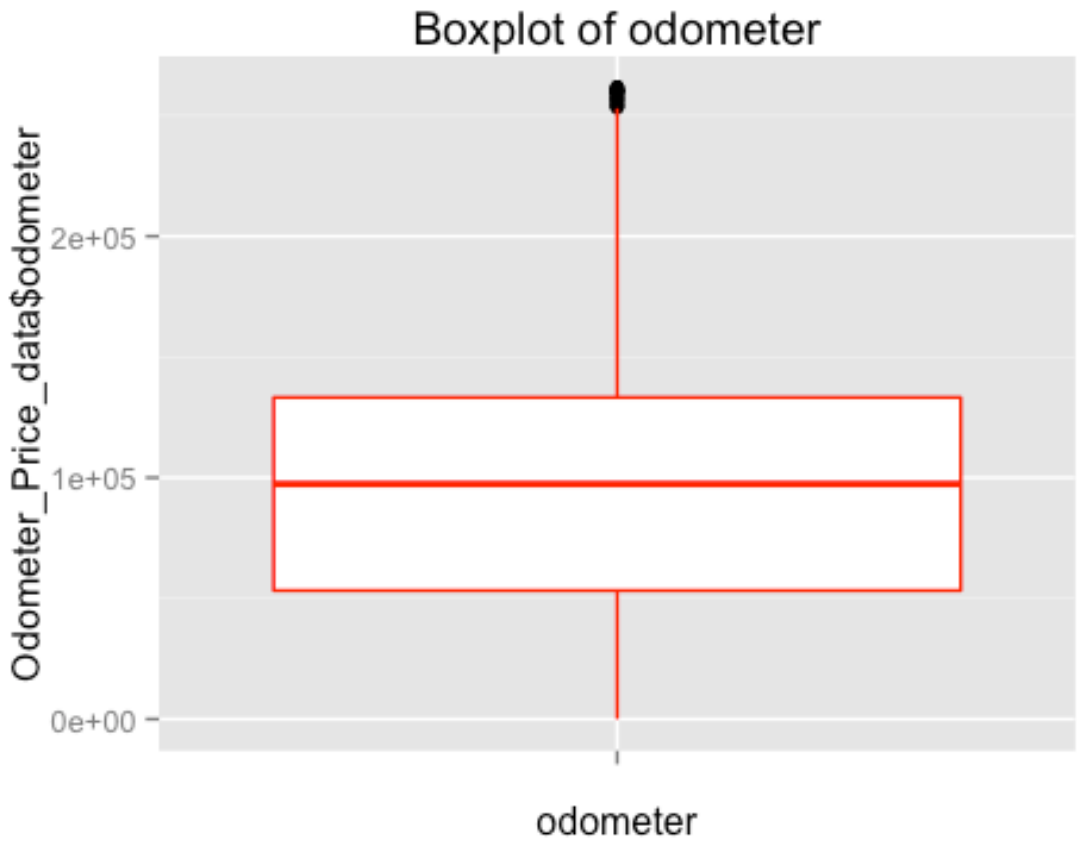
```
Odometer_Price_data=subset(Odometer_Price_data, Odometer_Price_data$odo
meter<261915)

length(which(Odometer_Price_data$odometer<1000))

## [1] 1573
```

Also, some odometers are very small, or we can say "smaller than 1000 miles". This situation can happen because those cars may just be bought and the sellers want to change another car. However, for those cars' ages are very large, this situation is quite rare. Thus, I will remove those cars whose odometer is smaller than 1000 miles and car age is larger than 2.

```
# remove those cars
Odometer_Price_data=Odometer_Price_data[-which(Odometer_Price_data$odom
eter < 1000 & Odometer_Price_data$age >2),]
```

Now let's draw a box plot of odometer.

```
ggplot(Odometer_Price_data, aes(x='',y=Odometer_Price_data$odometer)) +
  geom_boxplot(col="red") +
  xlab("odometer") +
  labs(title="Boxplot of odometer")
```



From the plot, we can see that the mean odometer is around 100000 miles.

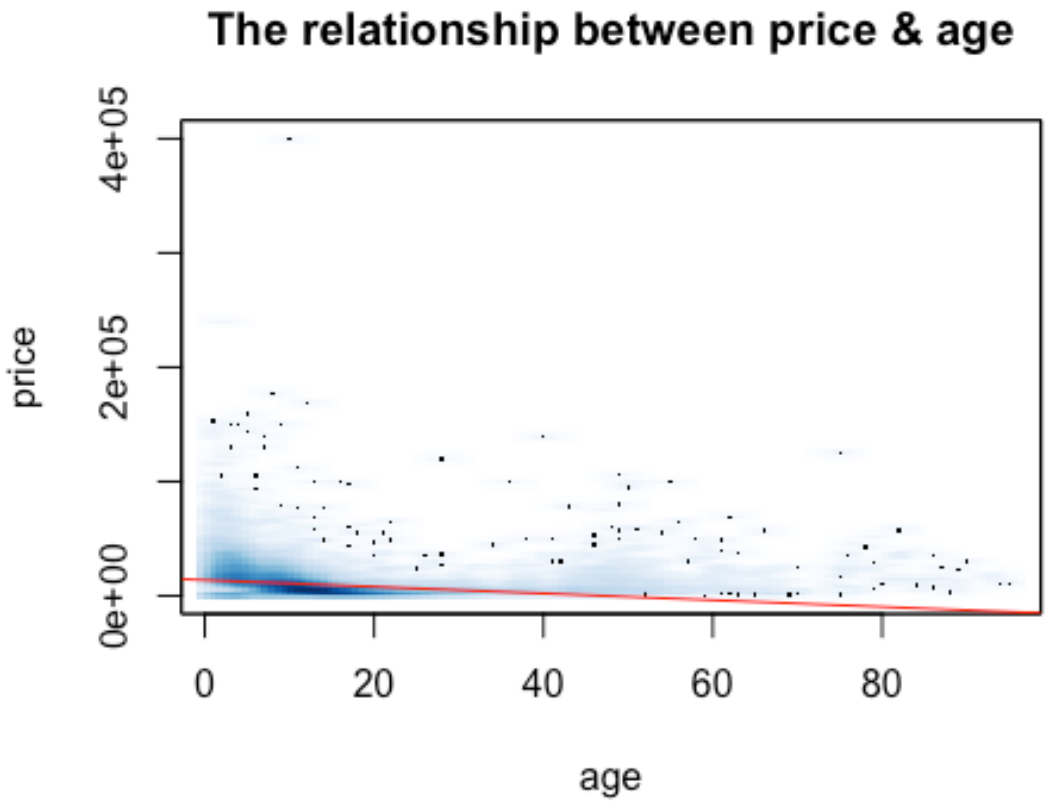
This anomaly would cause a big problem. Most buyers will judge the car based on their odometer. For those cars with extremely large and small odometer, it will mislead buyers a lot.

2.Find at least 3 interesting insights/characteristics/features illustrated by the data. Explain in what way these insights are interesting (to whom? why?) and provide evidence for any inference/conclusions you draw. How generalizable are these insights to other vehicle sales data?

Price vs. Age

Here we can also draw a plot about price and age

```
with(Price_data,smoothScatter(age,price,main="The relationship between price & age"))
abline(lm(Price_data$price~Price_data$age),col="red")
```



From the above plot, we can see that there is an approciamate trend that age and price have a negative relationship. The older the car, the lower the price. However, there is still some old car have high price. Let's take a such point as example.

```
Price_data[which(Price_data$age>65 & Price_data$price > 100000),c("titl
e", 'price','age')]
```

```
##                                                 title  price age
## posted212613 willys  coupe 1941 blown - $125000 125000  75
```

"willys coupe 1941 blown" is a really fancy car that althought it's old, it's quite expensive.

Thus, we can conclude that although the older the car, the lower the price, some fancy cars can still be found in "older car" list. This message is quite useful for those buyers who have enough money as well as want to buy fancy cars. When these kind of buyers search car online, don't limit the age in case of missing some really fancy vintage cars. I believe this insight can also apply to other vehicle sales data.

Price vs. Fuel type

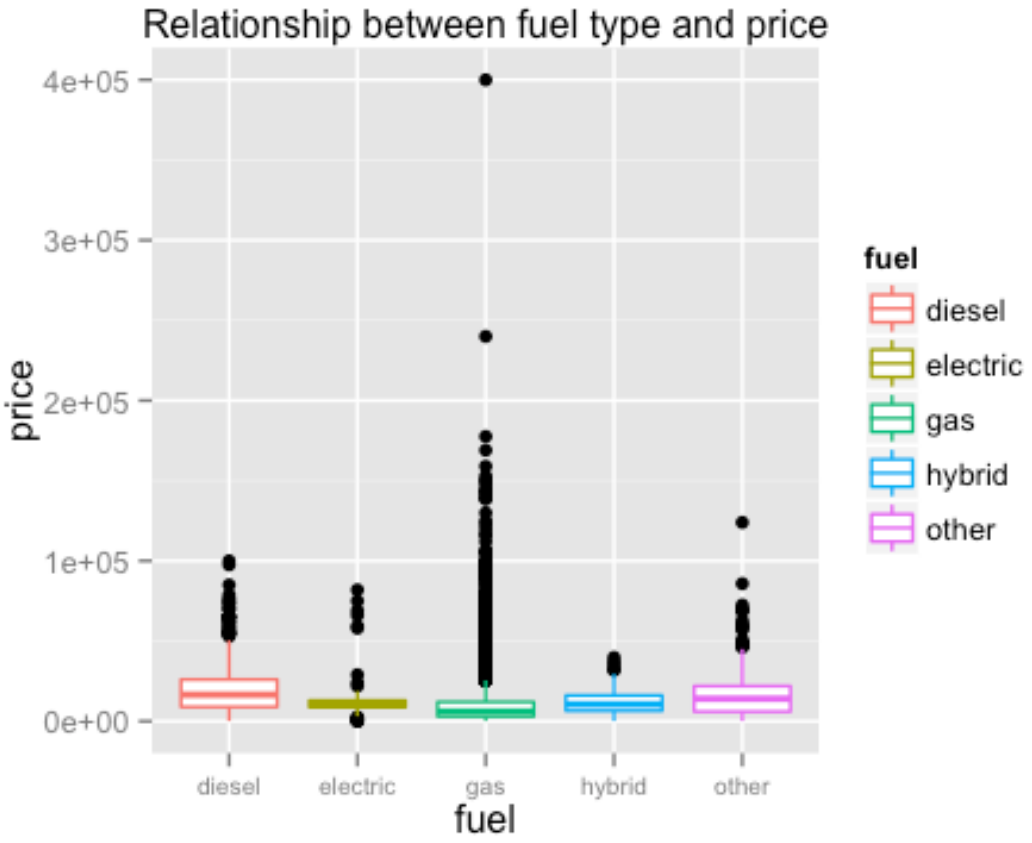
Now, there is another issue that buyers may need to decide: fuel type of car.

```
# remove the "fuel" is NA
Fuel_Price_data=subset(Price_data,!is.na(Price_data$fuel))
summary(Fuel_Price_data$fuel)
```

```
##   diesel electric      gas   hybrid    other
##      925       64    26223      325      813
```

Now, let's see the relationship between price and fuel type:

```
ggplot(Fuel_Price_data, aes(x=fuel, y=price,col=fuel))+
geom_boxplot()+
  labs(title="Relationship between fuel type and price")+
   theme(axis.text.x = element_text(size = rel(0.8)),
         plot.title = element_text(size = rel(1)))
```



According to the boxplot, we can conclude that the mean price of "gas car" is the lowest. "Diesel car" is relatively more expensive. The reason is that the residual value of "Diesel car" is higher than gas car (see the relative link). Also, the cost of diesel is lower than gas. Thus, in the future, diesel car will cost less than gas car in the fuel fee. It leads American who want to buy diesel car to increase. This situation may make the price of "diesel car" rises. This situation may provide buyers an insight that gas cars may cost less than other fuel-type cars.

Size vs. price

Let's see the relationship between size and price.

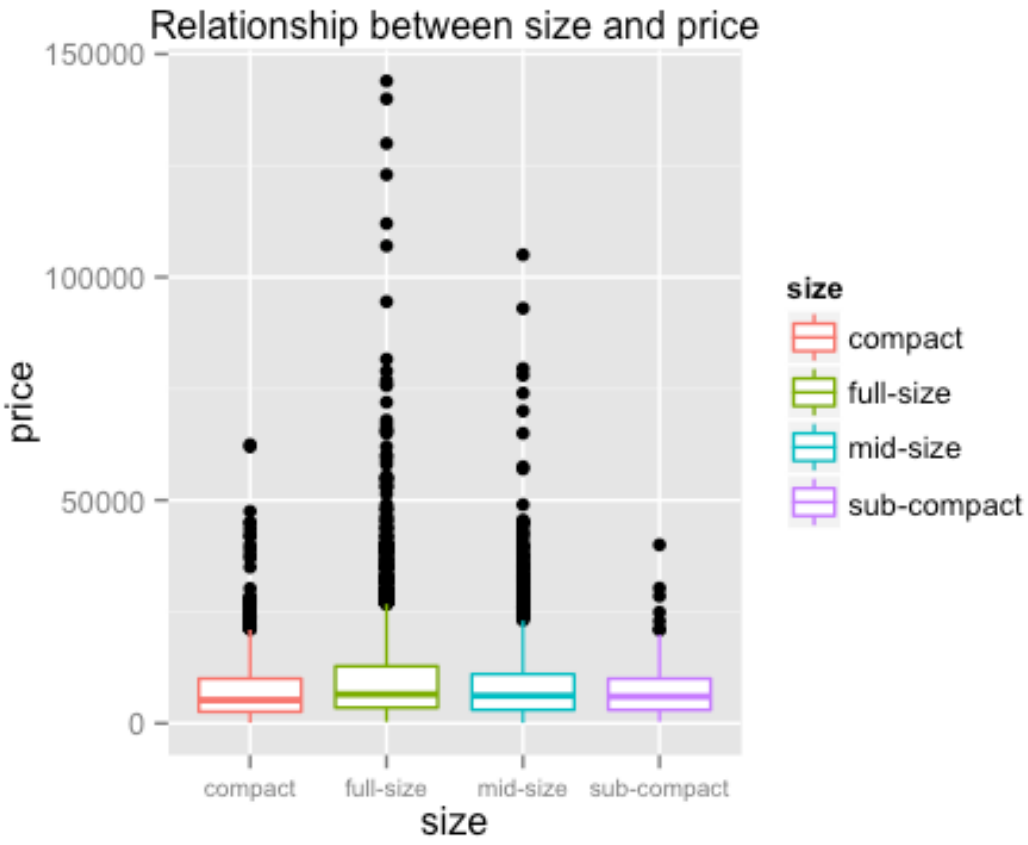
```
Size_Price_data=subset(Price_data,!is.na(Price_data$size))
summary(Size_Price_data$size)
```

```
##     compact   full-size    mid-size sub-compact
##        1843        4081        2942         199
```

There are 4 size of cars in this dataset.

```
ggplot(Size_Price_data, aes(x=size, y=price,col=size))+
geom_boxplot()+
  labs(title="Relationship between size and price")+
```

```
  theme(axis.text.x = element_text(size = rel(0.8)),
        plot.title = element_text(size = rel(1)))
```



From the plot above, we can see the mean prices for different size. "Full-size car" holds the highest mean price. "sub-compact car" holds the lowest mean price. This situation meets our common knowledge. However, we can also find that there is not too much difference for the mean price between "compact car" and "sub-compact car":

```
# mean price for each size
tapply(Size_Price_data$price,Size_Price_data$size,mean)
```

```
##     compact   full-size    mid-size sub-compact
##    7337.071    9868.363    8534.209    7274.673
```

Here we can conclude that the cost for "compact car" and "sub-compact car" is quite close. If buyers have a tight budget and want to buy a car, which have enough space to carry passengers and cargos, "compact car" is a better choice than "sub-compact car". This situation may also apply to other dataset.