STA141 Assignment 1 I

Weitong(Jessie) Lin

Download the dataset *vehicles.rda* from website

```
setwd('~/Desktop/UC Davis/141/Day1')
download.file('http://eeyore.ucdavis.edu/stat141/Data/vehicles.rda',des
tfile="vehicles.rda")
print(load("vehicles.rda"))

## [1] "vposts"

library(ggplot2)
```

1. How many observations are there in the data set?

```
str(vposts)

## 'data.frame':    34677 obs. of  26 variables:
##  $ id          : chr  "5228397709" "5228437424" "5228475701" "522850
6948" ...
##  $ title       : chr  "2012 Chevrolet Camaro SS - All Credit Accepte
d - $29896 (Automax Preowned of Framingham)" "2013 Chevrolet Equinox LT
 - All Credit Accepted - $18797 (Automax Preowned of Framingham)" "2013
 Nissan Altima 2.5 SV - All Credit Accepted - $15792 (Automax Preowned
of Framingham)" "2009 Infiniti M35x X - All Credit Accepted - $18288 (A
utomax Preowned of Framingham)" ...
##  $ body        : chr  "2012 Chevrolet Camaro SS\nOffered by: Automax
 Preowned of Framingham â\u0080\u0094 (508) 205-1046 â\u0080\u0094 $29,
896\nVIN: 2"| __truncated__ "2013 Chevrolet Equinox LT\nOffered by: Aut
omax Preowned of Framingham â\u0080\u0094 (508) 205-1046 â\u0080\u0094
$18,797\nVIN: "| __truncated__ "2013 Nissan Altima 2.5 SV\nOffered by:
Automax Preowned of Framingham â\u0080\u0094 (508) 205-1046 â\u0080\u00
94 $15,792\nVIN: "| __truncated__ "2009 Infiniti M35x X\nOffered by: Au
tomax Preowned of Framingham â\u0080\u0094 (508) 205-1046 â\u0080\u0094
 $18,288\nVIN: JNKCY"| __truncated__ ...
##  $ lat         : num  42.3 42.3 42.3 42.3 42.3 ...
##  $ long        : num  -71.4 -71.4 -71.4 -71.4 -71.4 ...
##  $ posted      : POSIXct, format: "2015-09-18 15:50:15" "2015-09-18
16:19:48" ...
##  $ updated     : POSIXct, format: NA NA ...
##  $ drive       : Factor w/ 3 levels "4wd","fwd","rwd": 3 NA 2 NA NA
NA 2 NA 2 NA ...
##  $ odometer    : int  16324 61095 40880 76108 14942 35230 94227 3664
1 7914 81136 ...
##  $ type        : Factor w/ 13 levels "bus","convertible",..: 3 10 9
9 9 10 9 10 9 9 ...
```

```
##  $ header      : chr  "2012 Chevrolet Camaro SS" "2013 Chevrolet Equ
inox LT" "2013 Nissan Altima 2.5 SV" "2009 Infiniti M35x X" ...
##  $ condition   : Factor w/ 43 levels "0used","207,400",..: NA NA NA
NA NA NA NA NA NA NA ...
##  $ cylinders   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ fuel        : Factor w/ 5 levels "diesel","electric",..: NA NA NA
 NA NA NA NA NA NA NA ...
##  $ size        : Factor w/ 4 levels "compact","full-size",..: NA NA
NA NA NA NA NA NA NA NA ...
##  $ transmission: Factor w/ 3 levels "automatic","manual",..: NA NA N
A 1 1 1 1 1 1 1 ...
##  $ byOwner     : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ city        : Factor w/ 7 levels "boston","chicago",..: 1 1 1 1 1
 1 1 1 1 1 ...
##  $ time        : POSIXct, format: "2015-09-18 18:50:00" "2015-09-18
19:19:00" ...
##  $ description : chr  "2012 Chevrolet Camaro SS - All Credit Accepte
d" "2013 Chevrolet Equinox LT - All Credit Accepted" "2013 Nissan Altim
a 2.5 SV - All Credit Accepted" "2009 Infiniti M35x X - All Credit Acce
pted" ...
##  $ location    : chr  "  (Automax Preowned of Framingham)   pic map
" "  (Automax Preowned of Framingham)   pic map " "  (Automax Preowned
of Framingham)   pic map " "  (Automax Preowned of Framingham)   pic ma
p " ...
##  $ url         : chr  "/bmw/ctd/5228397709.html" "/bmw/ctd/522843742
4.html" "/bmw/ctd/5228475701.html" "/bmw/ctd/5228506948.html" ...
##  $ price       : int  29896 18797 15792 18288 26389 28996 NA 24995 1
5995 NA ...
##  $ year        : int  2012 2013 2013 2009 2013 2012 2010 2012 2014 2
009 ...
##  $ maker       : chr  "chevrolet" "chevrolet" "nissan" "infiniti" ...
##  $ makerMethod : num  1.5 1.5 1.5 1.5 1.5 1.5 1.5 1.5 1.5 1.5 ...
```

From the above output from R, we can find that there are **34667** observations in the data set.

2. What are the names of the variables? and what is the class of each variable?

From the output from the fisrt question, we can know the **name** (after $) and **class**(after :) for each variables.

Or,

- We can also get the names for variables by:

```
names(vposts)

##  [1] "id"           "title"        "body"         "lat"
##  [5] "long"         "posted"       "updated"      "drive"
##  [9] "odometer"     "type"         "header"       "condition"   
## [13] "cylinders"    "fuel"         "size"         "transmission"
## [17] "byOwner"      "city"         "time"         "description" 
## [21] "location"     "url"          "price"        "year"        
## [25] "maker"        "makerMethod"
```

- The type for each variable by:

```
sapply(vposts, class)

## $id
## [1] "character"
## 
## $title
## [1] "character"
## 
## $body
## [1] "character"
## 
## $lat
## [1] "numeric"
## 
## $long
## [1] "numeric"
## 
## $posted
## [1] "POSIXct" "POSIXt" 
## 
## $updated
## [1] "POSIXct" "POSIXt" 
## 
## $drive
## [1] "factor"
## 
## $odometer
## [1] "integer"
## 
## $type
## [1] "factor"
## 
## $header
## [1] "character"
## 
## $condition
## [1] "factor"
## 
## $cylinders
## [1] "integer"
## 
## $fuel
## [1] "factor"
```

```
## 
## $size
## [1] "factor"
## 
## $transmission
## [1] "factor"
## 
## $byOwner
## [1] "logical"
## 
## $city
## [1] "factor"
## 
## $time
## [1] "POSIXct" "POSIXt" 
## 
## $description
## [1] "character"
## 
## $location
## [1] "character"
## 
## $url
## [1] "character"
## 
## $price
## [1] "integer"
## 
## $year
## [1] "integer"
## 
## $maker
## [1] "character"
## 
## $makerMethod
## [1] "numeric"
```

3. What is the average price of all the vehicles? the median price? and the deciles? Displays these on a plot of the distribution of vehicle prices.

I will answer the question after Question 8.

4. What are the different categories of vehicles, i.e. the type variable/column? What is the proportion for each category ?

- Categories of vehicles:

```
levels(vposts$type)
```

```
##  [1] "bus"          "convertible" "coupe"        "hatchback"    "mini-v
an"
##  [6] "offroad"      "other"        "pickup"       "sedan"        "SUV"

## [11] "truck"        "van"          "wagon"
```

- The proportion for each category:

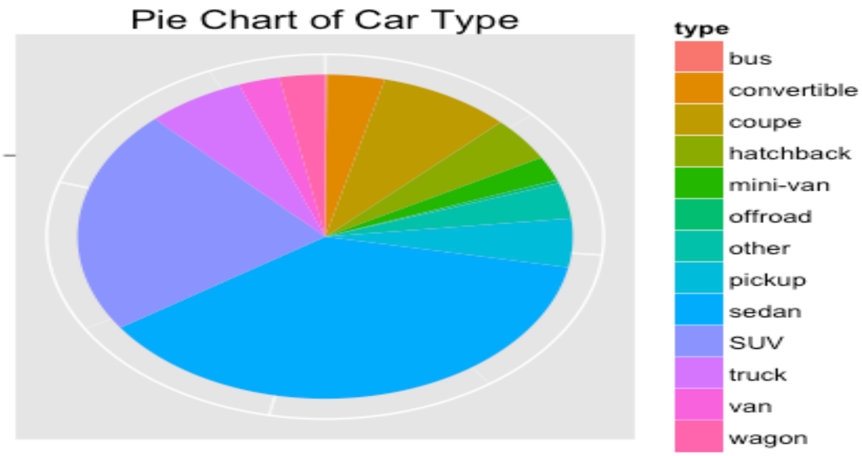
We need to move the data where type=NA,

```
Type_clean=vposts[which(vposts$type != "NA"),]
prop_type=round(prop.table(summary(Type_clean$type)),digits=3)
prop_type
```

```
##         bus convertible       coupe   hatchback    mini-van     offr
oad
##       0.001       0.038       0.087       0.044       0.024       0.
004
##       other      pickup       sedan         SUV       truck
van
##       0.035       0.048       0.375       0.224       0.064       0.
027
##       wagon
##       0.030
```

Here is a pie chart which can show the the proportion of car types visually.

```
count_type_table=as.data.frame(table(Type_clean$type))
colnames(count_type_table)=c("type","counts")
ggplot(count_type_table, aes(x="", y=count_type_table$counts,
fill=type))+
geom_bar(width = 1, stat = "identity") +
coord_polar("y", start=0)+
theme(axis.text.x=element_blank())+
labs(list(title = "Pie Chart of Car Type", x = "", y = ""))
```



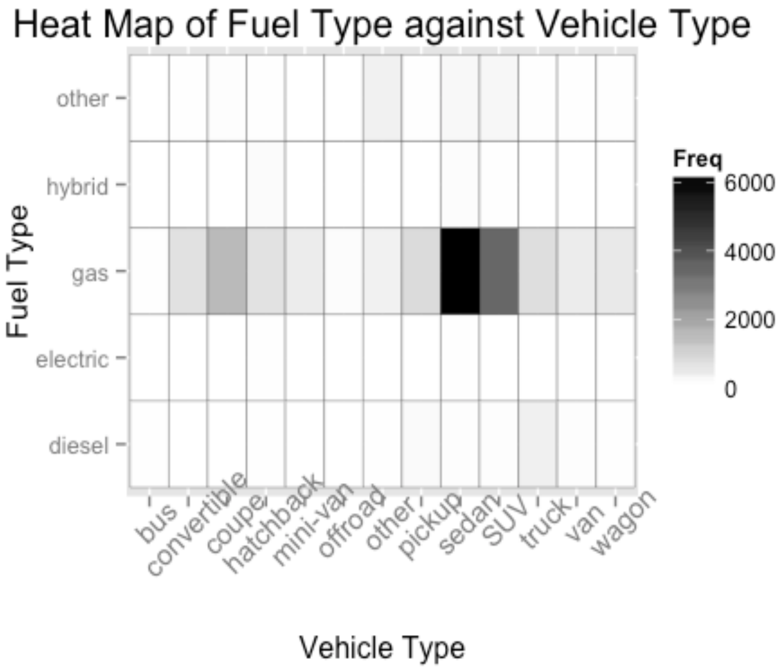
5. Display the relationship between fuel type and vehicle type. Does this depend on transmission type?

- First, all data whose "type" and "fuel are "NA" should be removed.

```
Fuel_Type_clean=vposts[which(vposts$type != "NA" & vposts$fuel != "NA"
& vposts$transmission != "NA"),]
```

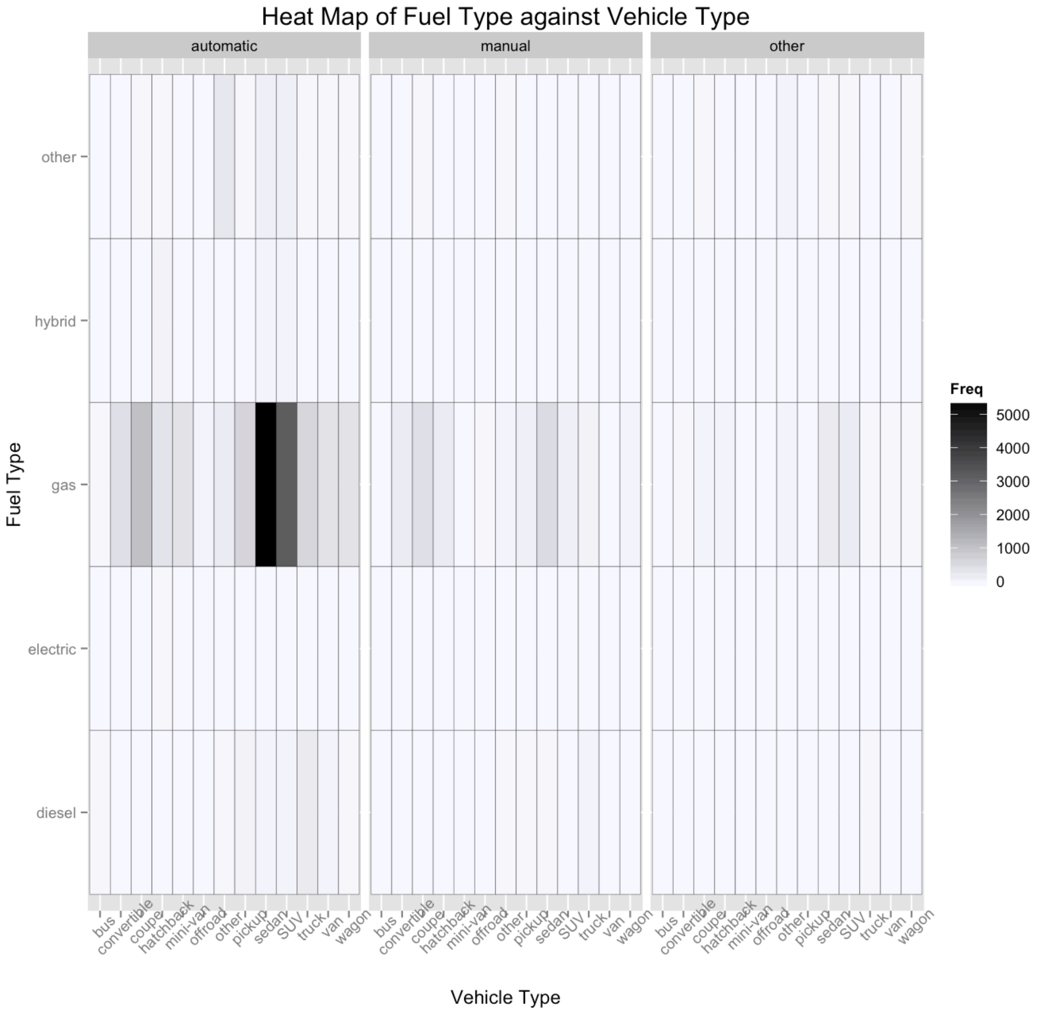
Now let's see the relationship between fuel type and vehicle type. In piazza, most students use mosaic plot to show their relationship. However, when I plot it, I think it's a little bit messy. So I also try the heat map, which might be more clear.

```
Fuel_Type_clean_df = with(Fuel_Type_clean, as.data.frame(table(type, fu
el)))
ggplot(Fuel_Type_clean_df, aes(type, fuel)) +
   geom_tile(aes(fill = Freq), colour = "black")+
   scale_fill_gradient(low = "white", high = "black") +
   xlab("Vehicle Type")+
   ylab("Fuel Type")+
   labs(title="Heat Map of Fuel Type against Vehicle Type")+
   theme(axis.text.x = element_text(size = rel(1.2),angle=45),
         plot.title = element_text(size = rel(1.3)))
```



- Then we draw another plot to see whether the relationship depends on transmission type.

```
Fuel_Type_trans_df = with(Fuel_Type_clean, as.data.frame(table(type, fu
el,transmission)))
ggplot(Fuel_Type_trans_df, aes(type, fuel)) +
   geom_tile(aes(fill = Freq), colour = "black")+
   scale_fill_gradient(low = "ghostwhite", high = "black") +
   xlab("Vehicle Type")+
   ylab("Fuel Type")+
   labs(title="Heat Map of Fuel Type against Vehicle Type")+
   theme(axis.text.x = element_text(size = rel(1),angle=45),
         plot.title = element_text(size = rel(1.3))) +
   facet_wrap(~transmission)
```



From the above plot, we can find no matter which type of transmission, the number of gas-used sedan is always a larger one.

6. How many different cities are represented in the dataset?

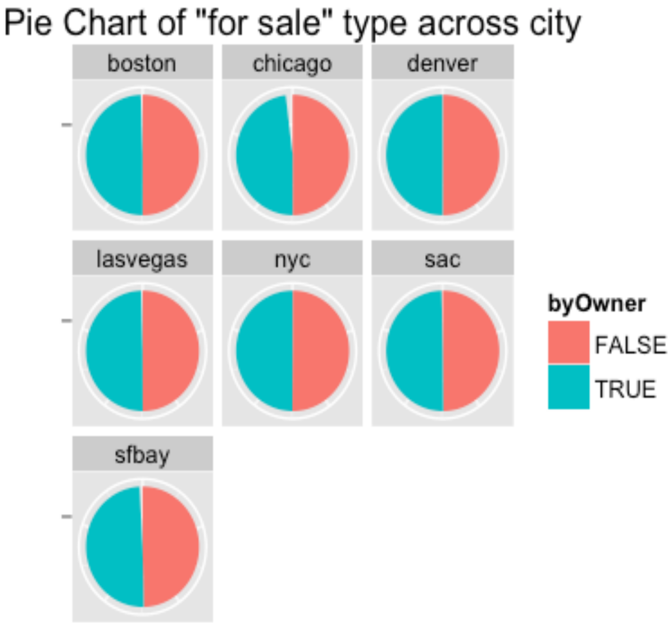
```
levels(vposts$city)

## [1] "boston"   "chicago"  "denver"   "lasvegas" "nyc"      "sac"

## [7] "sfbay"
```

7. Visually display how the number/proportion of "for sale by owner" and "for sale by dealer" varies across city?

```
for_sale_clean=subset(vposts, !is.na(byOwner) & !is.na(city))
for_sale_clean_count=with(for_sale_clean, as.data.frame(table(byOwner,
city)))
ggplot(for_sale_clean_count, aes(x='', y=Freq, fill=byOwner))+
geom_bar(width = 1, stat = "identity") +
facet_wrap(~city) +
coord_polar("y", start=0)+
theme(axis.text.x=element_blank())+
labs(list(title = "Pie Chart of \"for sale\" type across city", x = "",
 y = ""))
```



From the above chart, we can find that the proportion of "for sale by owner" and "for sale by dealer" across city are almost the same, which is around 50%.

8. What is the largest price for a vehicle in this data set? Examine this and fix the value. Now examine the new highest value for price.

First, let's find the highest price for a vehicle in data set.

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

## [1] 600030000
```

It is incredible large for this value that we need to figure out what really happen in this data.

```
vposts[which.max(vposts$price),]$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    "
```

From the information showed above, 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

vposts[which(vposts$price==max(vposts$price,na.rm=T)),]$price=median(c
(6000,30000))
```

So this car is on longer with the highest price. Now let's see the top 6 largest price.

```
TopSixPrice=head(sort(vposts$price,decreasing=T))
TopSixPrice

## [1] 30002500  9999999   569500   559500   400000   359000
```

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

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

## [1] 30002500
```

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

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

## [1] "2002 Caddy Seville sls"
```

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

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

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

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

```
## [1] 9999999
```

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

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

## [1] "2001 Honda Accord"

vposts[which.max(vposts$price),]$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. Thus, we just fix it as $20.

```
vposts[which(vposts$price==max(vposts$price,na.rm=T)),]$price=20
```

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

```
TopSixPrice[c(3,4)]

## [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.

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

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

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

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

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

## [1] "2006 FORD GT"

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

## [1] "\n        *CANADIAN CAR NO ACCIDENTS*RARE LOW KM*Less than 2,00
0 kms!!! You don't have to worry about depreciation on this superb 2006
 Ford GT!!!!** This vehicle has its original front wind shield stickers
 from factory. Safety equipment includes: ABS, Xenon headlights, Passen
ger Airbag - Cancellable, Front fog/driving lights...Other features inc
lude: Leather seats, Power locks, Manual Transmission,\nFeatures and Sp
ecifications\nOther Features\nAir Conditioning\nCD Player\nKeyless Entr
y\nLeather Interior\nCruise Control\nCup Holder\n5.4L DOHC MPFI superch
arged handbuilt all-aluminum V8 engine\nElectronic ignition system w/pu
sh-button start\nDry sump lubrication system\nTwin disc self-adjusting
hydraulic clutch\nMid-engine/rear wheel drive\n48-AH maintenance-free b
attery w/battery saver feature\nFront/rear independent unequal length
(SLA) aluminum suspension w/steel coil springs\nFront/rear non-adjustab
le forged aluminum shock absorbers w/forged aluminum housings\nFront/re
ar tubular stabilizer bars\nTire inflation kit-no spare tire available\
nPwr rack & pinion steering\nBrembo front & rear vented 4-piston disc b
rakes w/black painted calipers\n66.2 litre fuel tank\nStainless steel d
ual exhaust\n1-306-525-1555 MORGAN\n    "
```

From the data showed in cars.com, $40000 sounds a appropriate price for a 2006 Ford GT.

Thus, the new highest price for the revised dataset should be:

```
options(scipen=3)
max(vposts$price,na.rm=T)

## [1] 400000
```

3. What is the average price of all the vehicles? the median price? and the deciles? Displays these on a plot of the distribution of vehicle prices.

```
Price=subset(vposts, ! is.na(vposts$price) & vposts$price>5000 )
```

- The #average# price of all the vehicles:

```
mean(Price$price)

## [1] 15173.47
```

- The #median# price of all the vehicles:

```
median(Price$price)

## [1] 11995
```

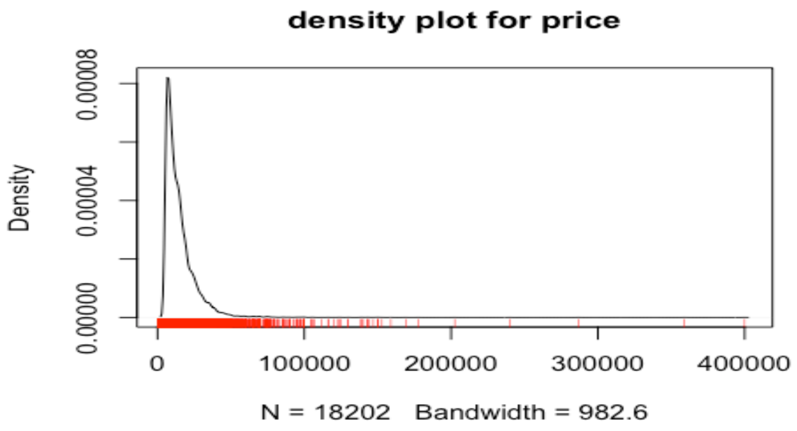
- The #declies# for price of all the vehicles:

```
quantile(Price$price ,seq(0, 1, length = 11))

##       0%       10%       20%       30%       40%       50%       60%       
70% 
##   5049.0    6000.0    7464.6    8500.0    9998.0   11995.0   13999.0   1680
0.0 
##      80%       90%      100% 
##  20000.0   27267.9 400000.0
```

- a plot of the distribution of vehicle prices

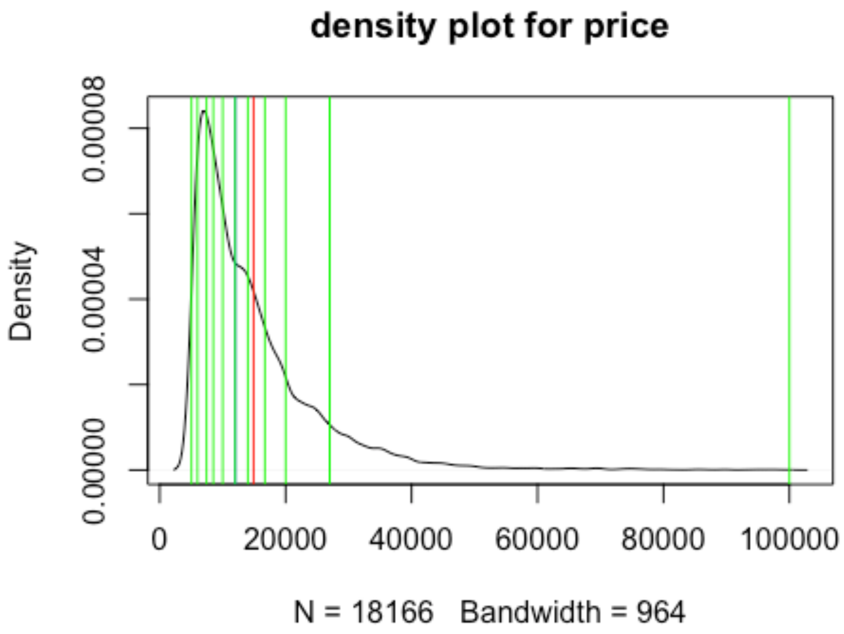
```
plot(density(Price$price),main="density plot for price")
rug(vposts$price,col='red')
```



d

From the rug function, I will reset the limit for price.

```
Price=subset(vposts, vposts$price>5000 & vposts$price< 100000)
plot(density(Price$price),main="density plot for price")
abline(v=mean(Price$price),col="red")
abline(v=median(Price$price),col="blue")
abline(v=quantile(Price$price ,seq(0, 1, length = 11)),col="green")
```



9. What are the three most common makes of cars in each city for "sale by owner" and for "sale by dealer"? Are they similar or quite different?

```
sortFreq1=function(Freq){
  sort_top3=sort(Freq, decreasing = TRUE);
  order_top3=order(Freq, decreasing = TRUE);
  Maker=maker_clean_count_O$maker[order_top3];
  data.frame(Maker=Maker[1:3], Top3=sort_top3[1:3])
}

For_Sale_Owner=with(maker_clean_count_O,tapply(Freq,city,sortFreq1))
#######
sortFreq2=function(Freq){
  sort_top3=sort(Freq, decreasing = TRUE);
  order_top3=order(Freq, decreasing = TRUE);
  Maker=maker_clean_count_D$maker[order_top3];
  data.frame(Maker=Maker[1:3], Top3=sort_top3[1:3])
}

For_Sale_Dealer=with(maker_clean_count_D,tapply(Freq,city,sortFreq2))
#######
Top3_Maker=mapply(function(Owner, Dealer) merge(Owner, Dealer, by = 0),
 Owner = For_Sale_Owner, Dealer = For_Sale_Dealer, SIMPLIFY = F)
Top3_Maker
```

```
## $boston
##   Row.names   Maker.x Top3.x   Maker.y Top3.y
## 1         1      ford    353      ford    333
## 2         2     honda    263    toyota    288
## 3         3 chevrolet    226 chevrolet    215
## 
## $chicago
##   Row.names   Maker.x Top3.x   Maker.y Top3.y
## 1         1 chevrolet    365 chevrolet    305
## 2         2      ford    331      ford    305
## 3         3     honda    180    nissan    208
## 
## $denver
##   Row.names   Maker.x Top3.x   Maker.y Top3.y
## 1         1      ford    378      ford    313
## 2         2 chevrolet    313 chevrolet    291
## 3         3    toyota    191     dodge    210
## 
## $lasvegas
##   Row.names   Maker.x Top3.x   Maker.y Top3.y
## 1         1      ford    394      ford    307
## 2         2 chevrolet    306    nissan    249
## 3         3    toyota    193 chevrolet    238
## 
## $nyc
##   Row.names Maker.x Top3.x Maker.y Top3.y
## 1         1  nissan    308  nissan    328
## 2         2  toyota    274  toyota    238
## 3          3   honda     260   honda     220
## 
## $sac
##   Row.names   Maker.x Top3.x   Maker.y Top3.y
## 1          1    toyota    340      ford    337
## 2          2      ford    305    toyota    273
## 3          3 chevrolet    299 chevrolet    206
## 
## $sfbay
##   Row.names Maker.x Top3.x Maker.y Top3.y
## 1          1  toyota    332  toyota    269
## 2          2   honda    322    ford    245
## 3          3    ford    257     bmw    227
```

The above result shows the three most common makes of cars in each city for "sale by owner"(left) and for "sale by dealer"(right). They are similar The skill that merge two lists into one list is found on Stackoverflow

10. Visually compare the distribution of the age of cars for different cities and for "sale by owner" and "sale by dealer". Provide an interpretation of the plots, i.e., what are the key conclusions and insights?

First let's see the "year for this data:

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

```
##  [1]    4 1900 1921 1922 1923 1925 1926 1927 1928 1929 1930 1931 193
2 1933
## [15] 1934 1935 1936 1937 1938 1939 1940 1941 1942 1945 1946 1947 194
8 1949
## [29] 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960 1961 196
2 1963
## [43] 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 197
6 1977
## [57] 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 199
0 1991
## [71] 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 200
4 2005
## [85] 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2022
```

Then we find that there are "4", "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
```

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

After watching the "title" for this car, I change the year"4" into "2004"

- **"2022" year**:

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

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

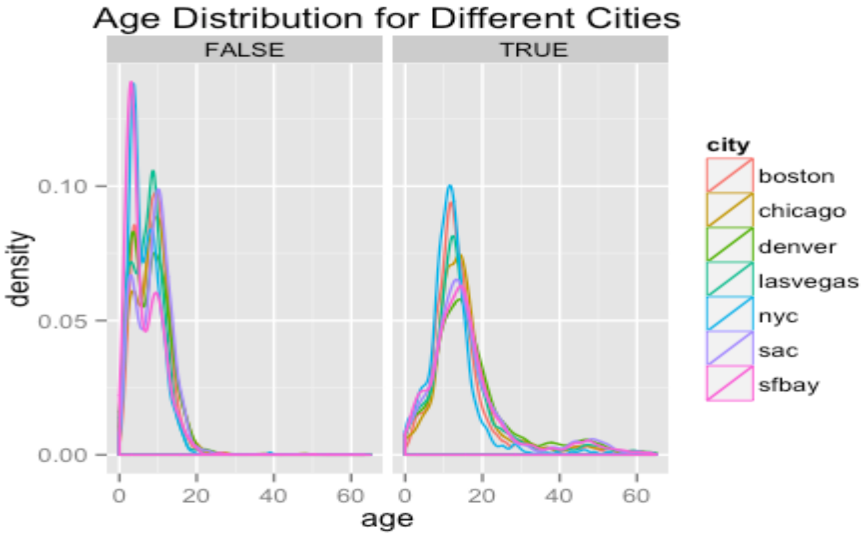
```
                                     body
## posted21888 2022 Honda Odyssey LX AT Automatic Gray Cloth on Silver
Silver Pearl Metallic 104208\nTake a look at this 2022 Honda Odyssey LX
 AT. It has only 117102 miles.\nColor: Silver Cloth on Silver Silver Pe
arl Metallic\nEngine: 3.5 V6 Cylinder Engine\nStock 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.

Now let's draw density plots.

```
city_byOwner_clean=subset(vposts, !is.na(byOwner) & !is.na(city) & !is.
na(year))
city_byOwner_clean$age = 2016 - city_byOwner_clean$year
ggplot(city_byOwner_clean, aes(x=age,col=city)) +
   geom_density() +
   xlim(c(0,65)) +
    facet_wrap(~byOwner) +
  labs(title = "Age Distribution for Different Cities")
```



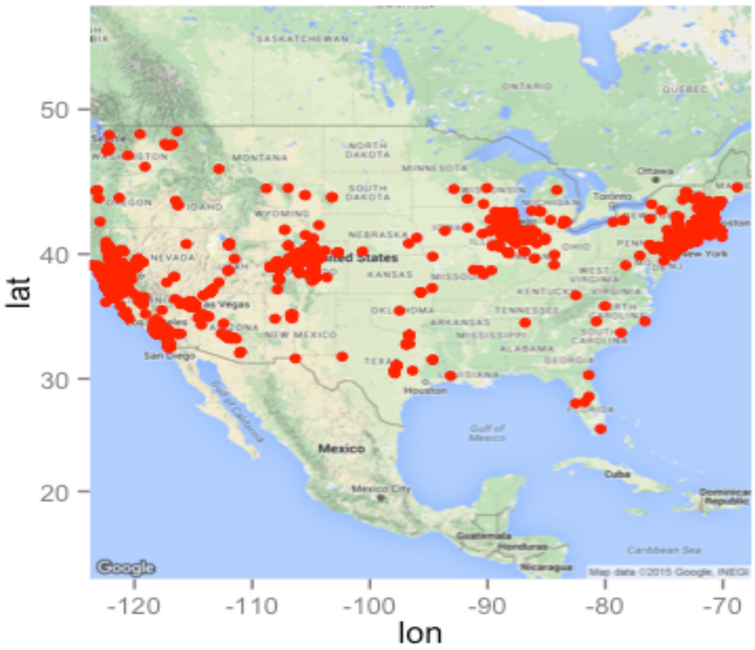
In the above plot, which "Sale by Dealer" on the left and "Sale by Owner" on the right, we can find that for sale by dealer, most cars in nyc are aged around 1 to 5 years.For sale by owner, most cars in nyc are aged around 5 to 7 years. Also, some very old car can also be found when the cars are sold by owner. Comparing between different cities, nyc always has the younger cars.

11. Plot the locations of the posts on a map? What do you notice?

```
map_clean=subset(vposts, !is.na(long) & !is.na(lat))
library(ggmap)
USAmap = get_map(location="United States", zoom = 4)

## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=
United+States&zoom=4&size=640x640&scale=2&maptype=terrain&language=en-E
N&sensor=false
## Information from URL : http://maps.googleapis.com/maps/api/geocode/j
son?address=United%20States&sensor=false

ggmap(USAmap) +
 geom_point(aes(x = long, y = lat), col='red', map_clean)
```



From the map, we can see that most of the posts are located in major 4 areas.

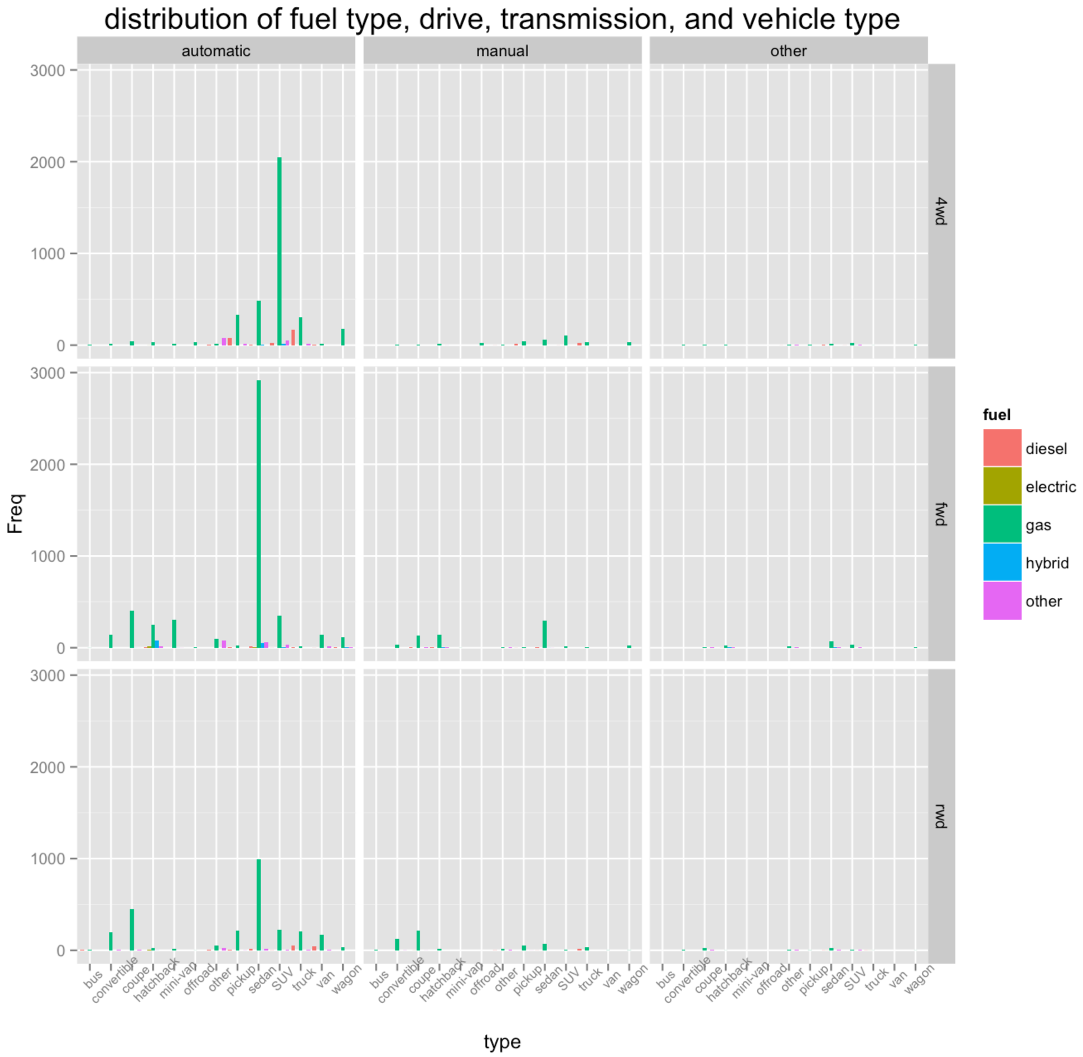
12. Summarize the distribution of fuel type, drive, transmission, and vehicle type. Find a good way to display this information.

Here are the distribution table of fuel type, drive, transmission, and vehicle type

```
FTDV=subset(vposts,!is.na(fuel) & !is.na(drive) & !is.na(transmission)
& !is.na(type))
FTDV_count=with(FTDV, as.data.frame(table(fuel,drive,transmission,typ
e)))
```

When we display this table into a plot:

```
ggplot(FTDV_count, aes(x=type, y=Freq))+
geom_bar(stat="identity",aes(fill = fuel), position = "dodge")+
  facet_grid(drive ~transmission) +
  labs(title="distribution of fuel type, drive, transmission, and vehic
le type")+
   theme(axis.text.x = element_text(size = rel(0.8),angle=45),
         plot.title = element_text(size = rel(1)))
```



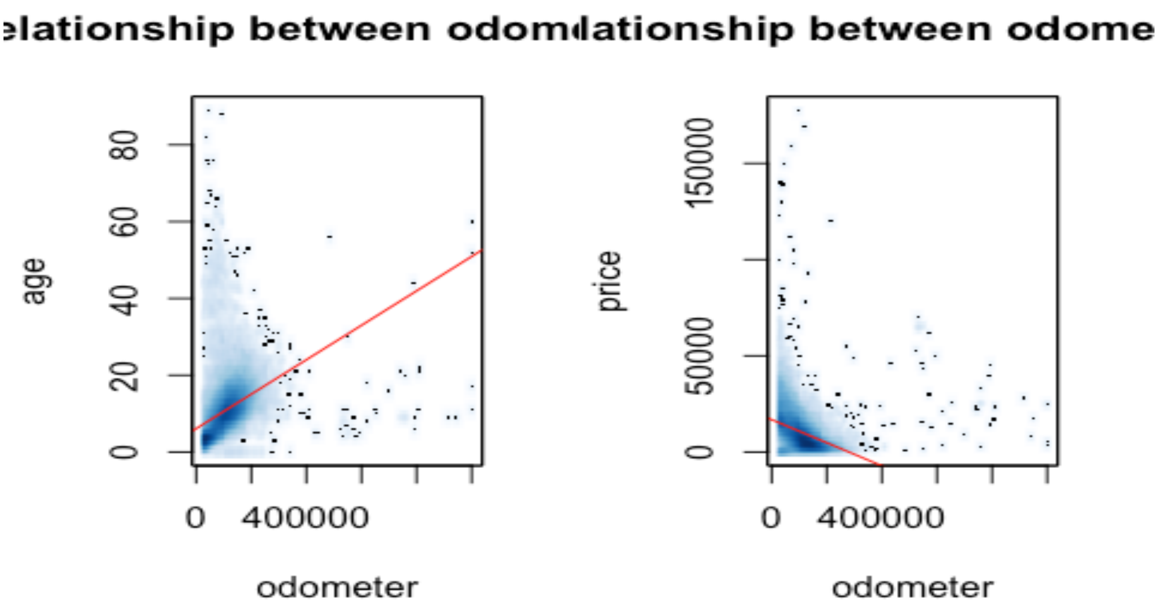
From the plot, we can conclude that most of cars are "automatic" and "gas-used". Also, most of "4wd" is Jeep. Most of "fwd" and 'rwd' is sedan.

13. Plot odometer reading and age of car? Is there a relationship? Similarly, plot odometer reading and price? Interpret the result(s). Are odometer reading and age of car related?

I only consider the cars which odometers are between 25000 miles and 1000000 miles.

```
odometer_clean=subset(odometer_clean, odometer>=25000 & odometer<=10000
00 & price<=200000)
op=par(mfrow=c(1,2))
with(odometer_clean,smoothScatter(odometer,age,main="The relationship b
etween odometer & age"))
```

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



```
par(op)
```

From the plot, we can see an approximate trend that odometer and age are positively related. The larger the odometer, the larger the age. Also, there is an approximate trend that odometer and price are negatively related. The larger the odometer, the lower the age.

14. Identify the "old" cars. What manufacturers made these? What is the price distribution for these?

In my opinion, I'd like to define those cars which were manufactured before 2005 or odometer were larger than 150000 miles as "old cars".

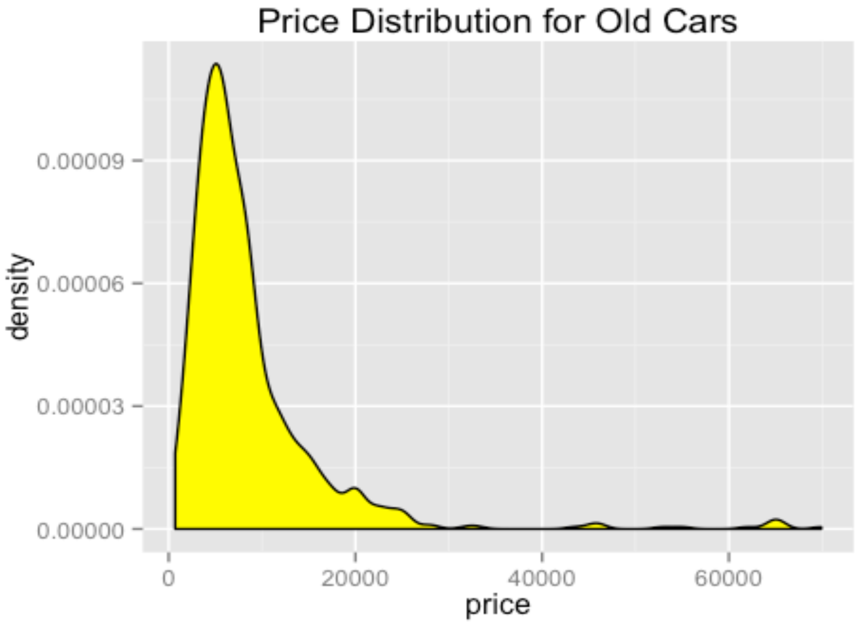
Here are the manufacturers:

```
unique(oldcar$maker)
```

```
##  [1] "ford"         "peterbilt"    "subaru"       "bmw"
##  [5] "honda"        "scion"        "toyota"       "mercedes"
##  [9] "kia"          "mitsubishi"   "saab"         "acura"
## [13] "pontiac"      "chrysler"     "hyundai"      "nissan"
## [17] "chevrolet"    "gmc"          "jeep"         "mazda"
## [21] "lincoln"      "dodge"        "infiniti"     "porsche"
## [25] "volvo"          "cadillac"       "mercury"        "international"
## [29] "saturn"         "audi"           "freightliner"   "hummer"
## [33] "buick"          "lexus"          "volkswagen"     "land rover"
## [37] "jaguar"
```

Here is the price distribution:

```
ggplot(oldcar, aes(x=price)) +
  geom_density(fill="yellow") +
  labs(title = "Price Distribution for Old Cars")
```



15.I have omitted one important variable in this data set. What do you think it is? Can we derive this from the other variables? If so, sketch possible ideas as to how we would compute this variable.

From the "body" part, I guess it's Engine information.

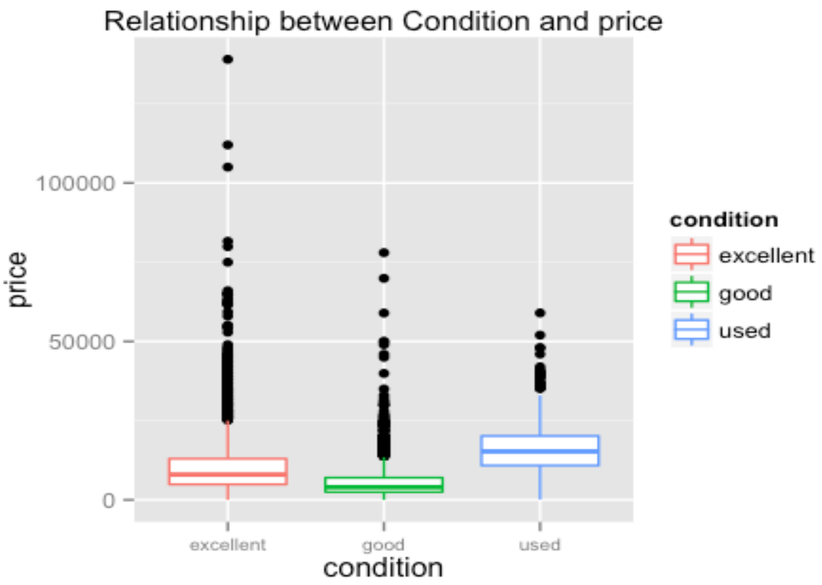
```
table(grepl('Engine', vposts$body))

##
## FALSE  TRUE
## 23695 10982
```

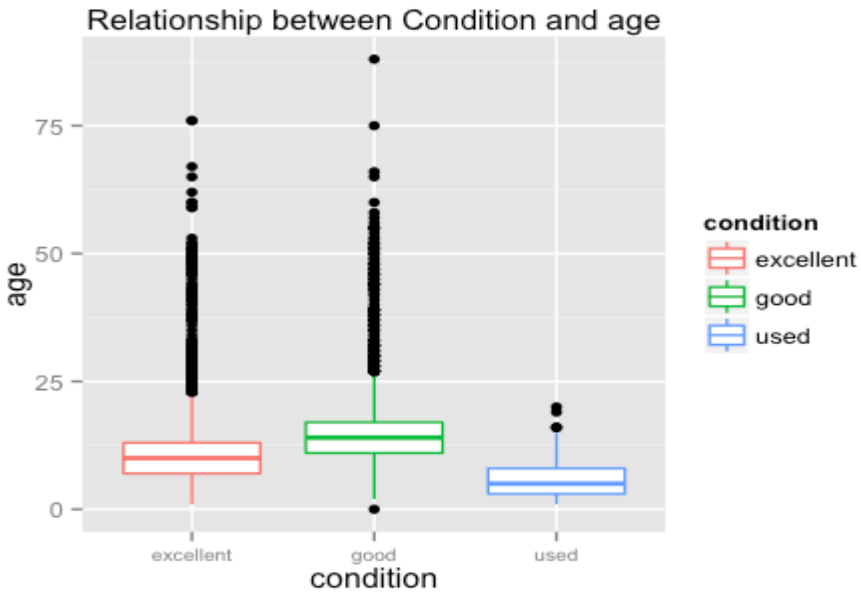
16. Display how condition and odometer are related. Also how condition and price are related. And condition and age of the car. Provide a brief interpretation of what you find.

Here we only consider three conditions: 'good', 'excellent' and 'used'

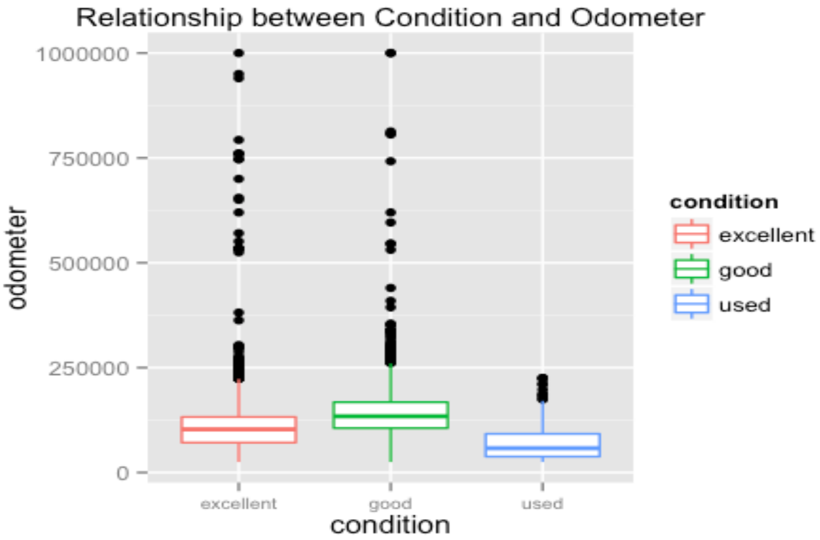
```
ggplot(odometer_clean, aes(x=condition, y=price,col=condition))+
geom_boxplot()+
  labs(title="Relationship between Condition and price")+
   theme(axis.text.x = element_text(size = rel(0.8)),
         plot.title = element_text(size = rel(1)))
```



```
ggplot(odometer_clean, aes(x=condition, y=age,col=condition))+
geom_boxplot()+
  labs(title="Relationship between Condition and age")+
   theme(axis.text.x = element_text(size = rel(0.8)),
         plot.title = element_text(size = rel(1)))
```



```
ggplot(odometer_clean, aes(x=condition, y=odometer,col=condition))+
geom_boxplot()+
  labs(title="Relationship between Condition and Odometer")+
   theme(axis.text.x = element_text(size = rel(0.8)),
         plot.title = element_text(size = rel(1)))
```



From the above plots, I find that the 'used' cars have the highest mean price. "good" cars have both the highest age and odometer.