Assignment 1 Ⅰ

Xinyi Hou

ID:913411456

Oct,1st,2015

Load the data:

```
load(url("http://eeyore.ucdavis.edu/stat141/Data/vehicles.rda"))
```

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

```
nrow(vposts)
```

```
## [1] 34677
```

or

```
length(vposts$id)
```

```
## [1] 34677
```

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

```
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"
```

```
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"
```

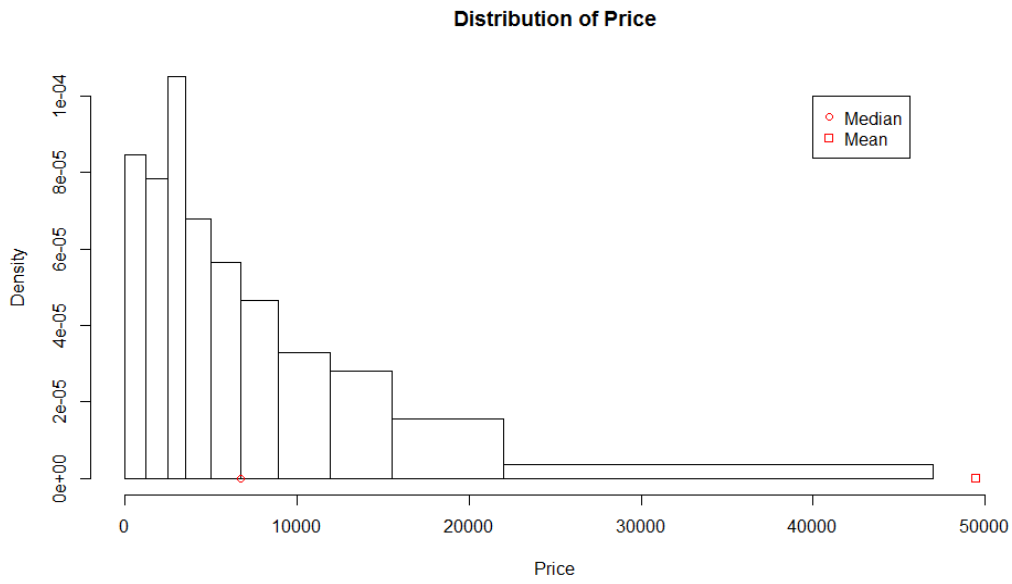
3. What is the average price of all the vehicles? The median price? And the deciles? Displays these on a plot of the distribution of vehicle prices.

```
avg=mean(vposts$price,na.rm=TRUE)
avg
```

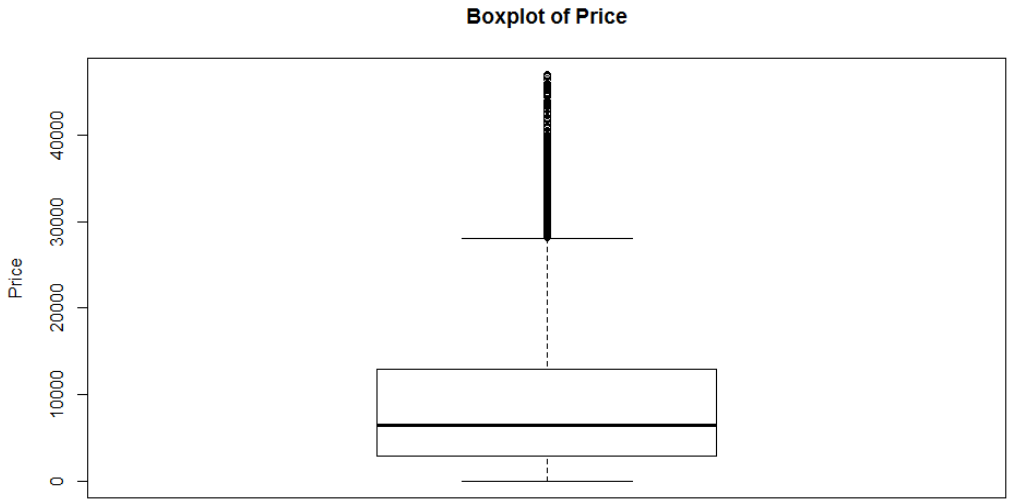
```
## [1] 49449.9

mid=median(vposts$price,na.rm=TRUE)
mid

## [1] 6700

q<-quantile(vposts$price,seq(0,0.9,0.1),na.rm=TRUE)
q

##    0%   10%   20%   30%   40%   50%   60%   70%   80%   90%
##     1  1200  2499  3500  4995  6700  8900 11888 15490 21997
```

As there are some extreme big values in price, I calculate the percentiles of 0.99 and omit prices bigger than it to get a more clear histogram. And I find that there is no obvious difference between the new and old median. So the abandonment of some data is reasonable.Furthermore, I plot deciles by using them as breaks of histogram.

```
quantile(vposts$price,c(0.01,0.99),na.rm=TRUE)

##    1%   99%
##     1 47000

#The 0.99 percentiles of price is 47000.
newprice<-vposts$price[vposts$price<47000]
hist(newprice,breaks=c(as.double(q),max(newprice,na.rm=TRUE)),xlim=c(0,
50000),main = "Distribution of Price",xlab='Price')
points(mid,0,col='red',pch=1)
points(avg,0,col='red',pch=0)
legend(40000,1e-04,c("Median", "Mean"), pch = c(1,0),col='red')
```



```
boxplot(newprice,ylab='Price',main='Boxplot of Price')
```



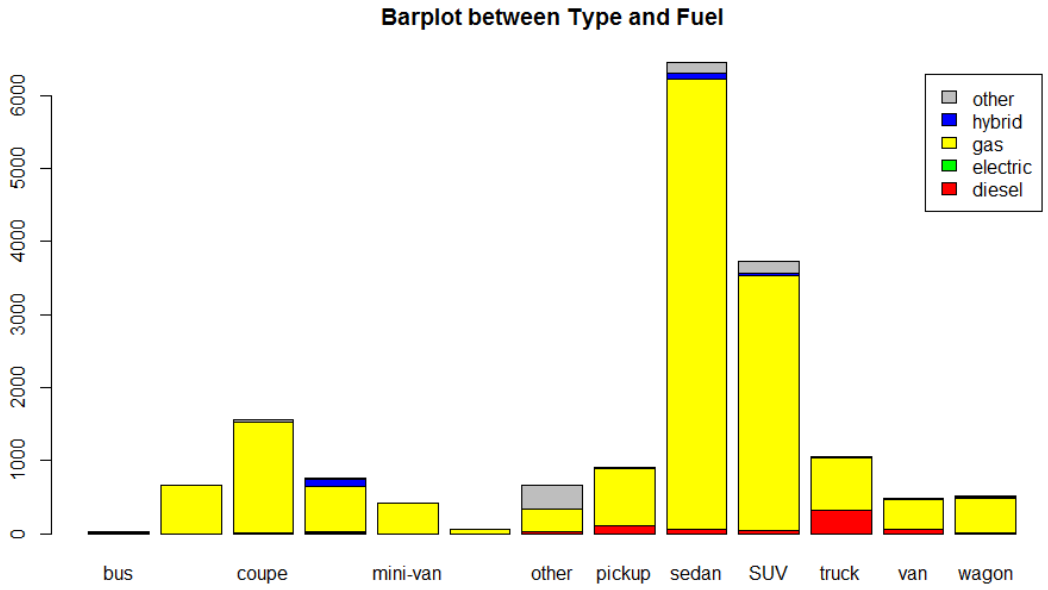
4. What are the different categories of vehicles, i.e. the type variable/column? What is the proportion for each category?

```
counts=table(vposts$type)
counts

##
##         bus convertible       coupe   hatchback    mini-van     offr
oad
##          22         706        1626         819         453
 66
##       other      pickup       sedan         SUV       truck
van
##         666         909        7040        4211        1202
507
##       wagon
##         558

propotion=counts/sum(!is.na(vposts$type))
print(propotion,digits=2)

##
##         bus convertible       coupe   hatchback    mini-van     offr
oad
##      0.0012      0.0376      0.0866      0.0436      0.0241      0.0
035
##       other      pickup       sedan         SUV       truck
van
```

```
##       0.0355       0.0484       0.3748       0.2242       0.0640       0.0
270
##        wagon
##       0.0297
```

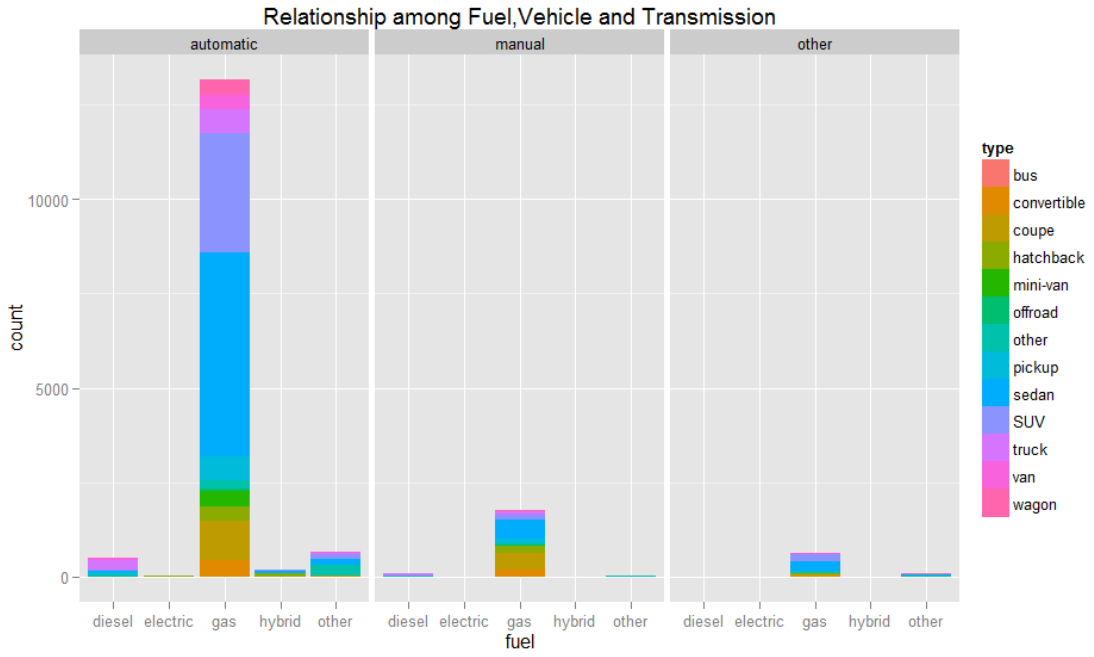
5. Display the relationship between fuel type and vehicle type. Does this depend on transmission type?

```
library(ggplot2)
counts=table(vposts$fuel,vposts$type)
barplot(counts,legend=rownames(counts),col=c('red','green','yellow','bl
ue','grey'),main='Barplot between Type and Fuel')
```




```
ggplot(vposts[!is.na(vposts$type)&!is.na(vposts$fuel)&!is.na(vposts$tra
nsmission),],aes(fuel,fill=type))+geom_bar()+facet_wrap(~transmission)+
ggtitle("Relationship among Fuel,Vehicle and Transmission")
```

According to the first bar plot, most of vehicles use gas as fuel. What is more, the percentages of truck, pickup and van using diesel are higher than others. This finding is the same as our common knowledge that these vehicles need more power. Hatchback uses hybrid and electric as fuel more that this is an environment friendly vehicle. Last, about half of the 'other' vehicle use 'other' fuel and these vehicles may be really special.



In the second graph, I take transmission type into considering.Automatic vehicles try all kinds of fuel, yet other transmissions are more classical which only use gas, diesel and other. When I focus on gas using vehicles, sedan and SUV take most percentages in automatic vehicles part,but there is less difference among vehicle type of manual. In other words, most sedan and SUV are automatic, yet other kinds do not show such obvious bias.

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

```
unique(vposts$city)
```

```
## [1] boston   chicago  denver   lasvegas nyc      sac      sfbay
## Levels: boston chicago denver lasvegas nyc sac sfbay
```

```
length(unique(vposts$city))
```

```
## [1] 7
```

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

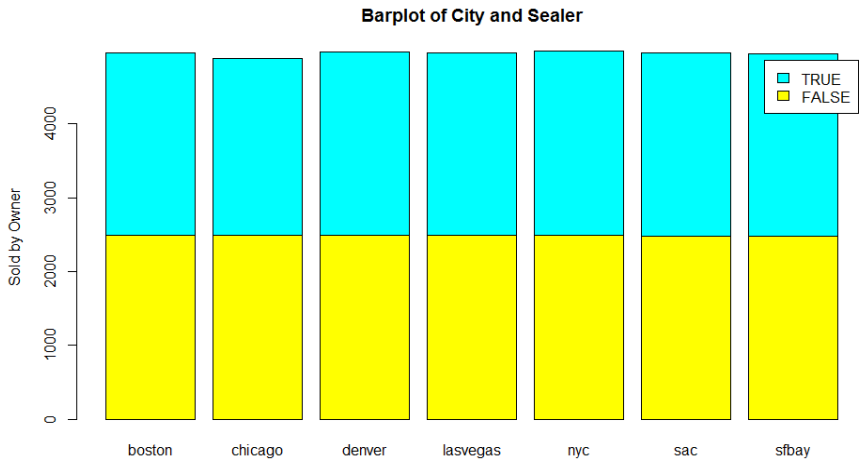
```
library(vcd)
```

```
## Loading required package: grid
```

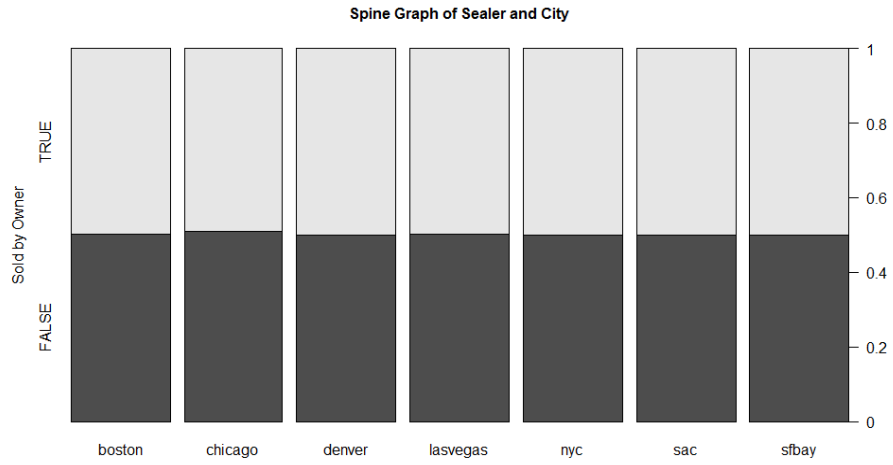
```
counts<-table(vposts$byOwner,vposts$city)
counts
```

```
##
##         boston chicago denver lasvegas  nyc  sac sfbay
##   FALSE   2491    2491   2492     2489 2495 2483  2475
##   TRUE    2467    2395   2487     2474 2488 2483  2467
```

```
barplot(counts,legend=rownames(counts),col=c(87,525),main='Barplot of City and Sealer',ylab='Sold by Owner')
```




```
counts<-table(vposts$city,vposts$byOwner)
spine(counts,ylab='Sold by Owner',main='Spine Graph of Sealer and City')
```



According to the table and graphs, we can see that there is no significant difference between numbers of vehicles sold by dealer or owner among the cities. Roughly, half of the vehicles are sold by owner and the others by dealer and no matter what the city is.

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.**

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

```
## [1] 600030000
```

```
#hist(vposts$price)
#rug(vposts$price,lwd=4)
#hist(vposts$price[vposts$price<8e4])
#rug(vposts$price[vposts$price<8e4],lwd=4)
vposts$price[vposts$price>8e4]=NA
max(vposts$price,na.rm=TRUE)
```

```
## [1] 79998
```

I use the function hist() and rug() several times to help me find extreme values of price. And I use 8e4 as the standard and fix the value by replacing all price bigger than 8e4 as NA.

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?**

```
library(dplyr)
```

```
library(data.table)
```

```
vposts1=count(vposts,maker,city,byOwner)
vposts1=vposts1 %>%
  filter(!maker=='NA')
#The data is ordered based on n, the number of every group we counted a
bove.
d <- data.table(vposts1, key='n')
#.SD means subset the data, 'by' gives the method of grouping.
d<-d[, tail(.SD, 3), by=c('byOwner','city')]
arrange(d,city,byOwner)
```

```
##     byOwner     city     maker   n
##  1:   FALSE   boston chevrolet 215
##  2:   FALSE   boston    toyota 288
##  3:   FALSE   boston      ford 333
##  4:    TRUE   boston chevrolet 226
##  5:    TRUE   boston     honda 263
##  6:    TRUE   boston      ford 353
##  7:   FALSE  chicago    nissan 208
##  8:   FALSE  chicago chevrolet 305
##  9:   FALSE  chicago      ford 305
## 10:    TRUE  chicago     honda 180
## 11:    TRUE  chicago      ford 331
## 12:    TRUE  chicago chevrolet 365
```

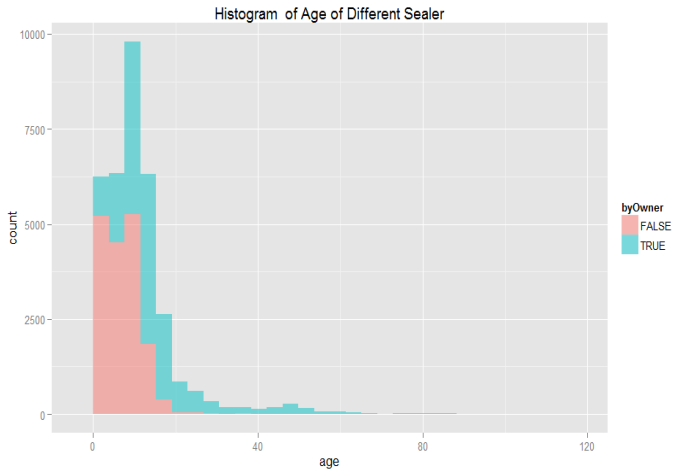
```
## 13:   FALSE   denver     dodge 210
## 14:   FALSE   denver chevrolet 291
## 15:   FALSE   denver      ford 313
## 16:    TRUE   denver    toyota 191
## 17:    TRUE   denver chevrolet 313
## 18:    TRUE   denver      ford 378
## 19:   FALSE lasvegas chevrolet 238
## 20:   FALSE lasvegas    nissan 249
## 21:   FALSE lasvegas      ford 307
## 22:    TRUE lasvegas    toyota 193
## 23:    TRUE lasvegas chevrolet 306
## 24:    TRUE lasvegas      ford 394
## 25:   FALSE      nyc     honda 220
## 26:   FALSE      nyc    toyota 238
## 27:   FALSE      nyc    nissan 328
## 28:    TRUE      nyc     honda 260
## 29:    TRUE      nyc    toyota 274
## 30:    TRUE      nyc    nissan 308
## 31:   FALSE      sac chevrolet 206
## 32:   FALSE      sac    toyota 273
## 33:   FALSE      sac      ford 337
## 34:    TRUE      sac chevrolet 299
## 35:    TRUE      sac      ford 305
## 36:    TRUE      sac    toyota 340
## 37:   FALSE    sfbay       bmw 227
## 38:   FALSE    sfbay      ford 245
## 39:   FALSE    sfbay    toyota 269
## 40:    TRUE    sfbay      ford 257
## 41:    TRUE    sfbay     honda 322
## 42:    TRUE    sfbay    toyota 332
##     byOwner     city     maker   n

#Another Method
#bymakercity=split(vposts1,list(vposts1$city,vposts1$byOwner))
#sapply(bymakercity,function(x)x[(x$n)>(sort(x$n,decreasing=TRUE)[4]),'
maker'])
```

The three most popular makers are similar among different cities. However, there are some other interesting findings. Such as, Ford seals great in most cities except NYC and Chevrolet is quiet common except NYC and SFbay. What is more, Honda is sold by owners more than dealers. Dodge is welcomed in Denver and BMW is common in SFbay,yet they are not three top popular makers in other cities.

Reference: http://stackoverflow.com/questions/14800161/how-to-find-the-top-n-values-by-group-or-within-category-groupwise-in-an-r-dat

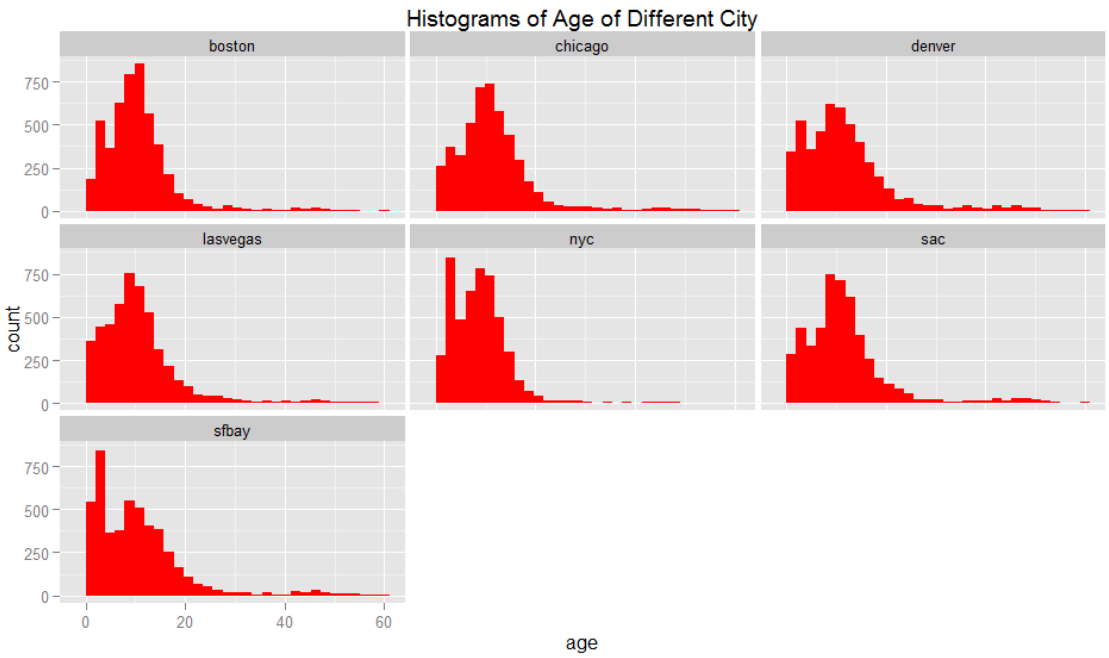
10. Visually compare the distribution of the age of cars for different cities and for "sale by owner" and "sale by dealer". Provide an interpretation of the plots, i.e., what are the key conclusions and insights?

```
library(dplyr)
library(ggplot2)
#Vposts2:Add a variable, age, to original dataframe vposts
vposts2=
  vposts %>%
  filter(year<=2015&year>1800) %>%
  mutate(age=2015-year)
ggplot(vposts2,aes(age, fill = byOwner)) + geom_histogram(alpha = 0.5)+
ggtitle("Histogram  of Age of Different Sealer")
```




```
ggplot(vposts2,aes(x=city,y=age,fill=city))+geom_boxplot()+ggtitle("Box
plot of Age of Different City")
```

```
vposts2 %>%
  filter(age<60) %>%
  ggplot(aes(age))+geom_histogram(fill='red')+facet_wrap(~city)+ggtitle
('Histograms of Age of Different City')
```

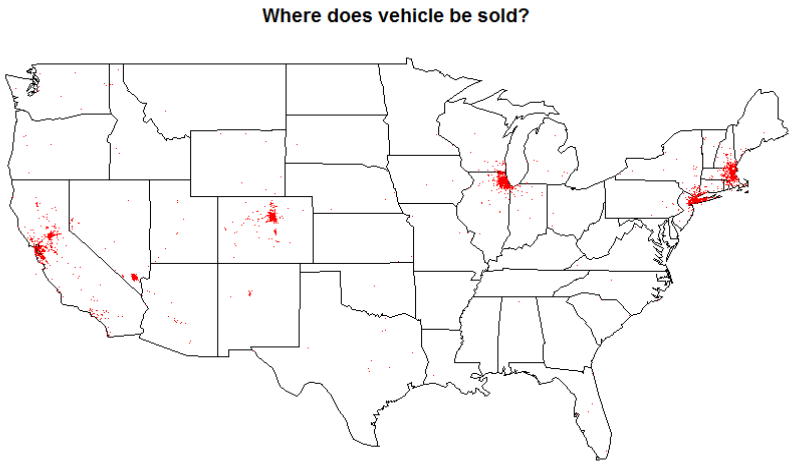


Based on the histogram of age of different sealers, the age of vehicles sold by owner is longer than it sold by dealer. What we can get from the second graph is that the age medians of Denver,SAC and SFbay are bigger than others. And the ranges of age of Denver and SFbay are wider. This finding can also be gotten from the histograms of age of different city.

Reference: http://stackoverflow.com/questions/3541713/how-to-plot-two-histograms-together-in-r

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

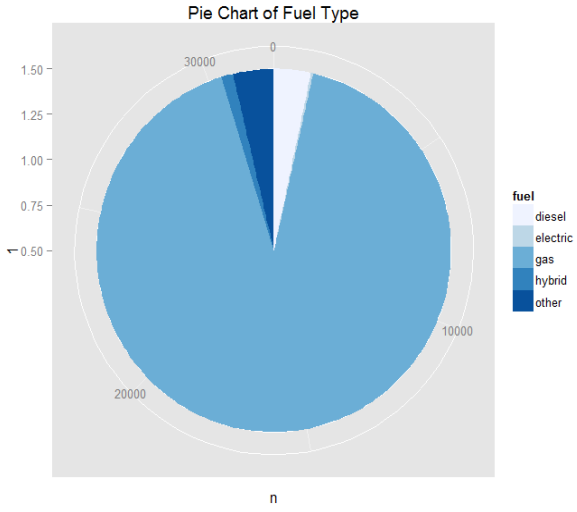
```
library(maps)
map(database='state')
with(vposts,points(long,lat,pch='.',col='red'))
title(main="Where does vehicle be sold?")
```



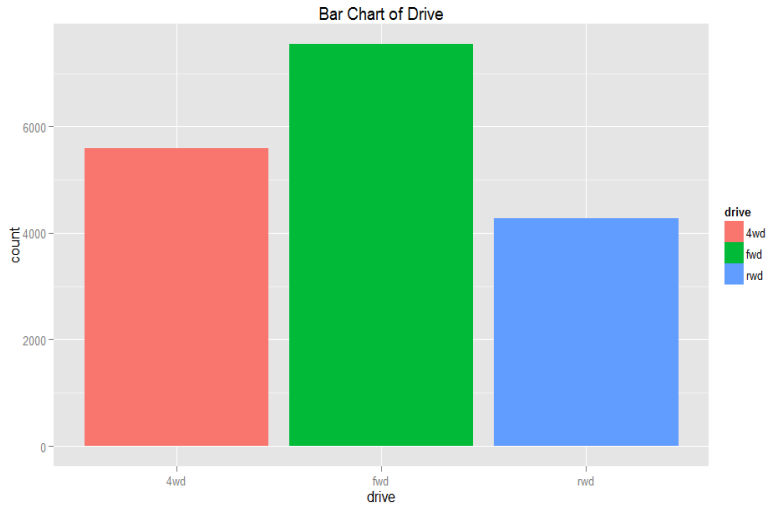
I have two findings from the map.Firstly, fewest vehicles are sold in Las Vegas compared to other cities. What is more, the vehicles are not sold only in the listed 7 cities and there are vehicles sold in other places, although just a few.

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

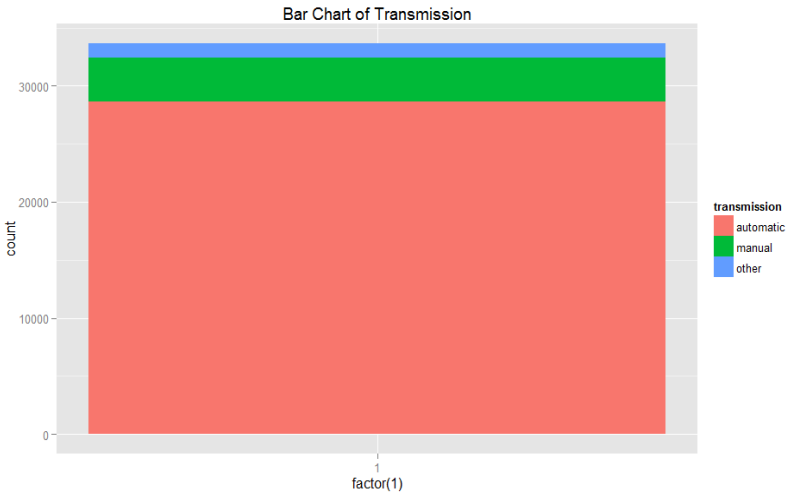
```
library(ggplot2)
library(dplyr)
vposts %>%
  filter(!is.na(fuel))%>%
  count(fuel) %>%
  ggplot(aes(x=1,y=n,fill=fuel))+geom_bar(stat='identity',width=1)+coor
d_polar("y",start=0)   +scale_fill_brewer(palette="Blues")+ggtitle('Pie
 Chart of Fuel Type')
```



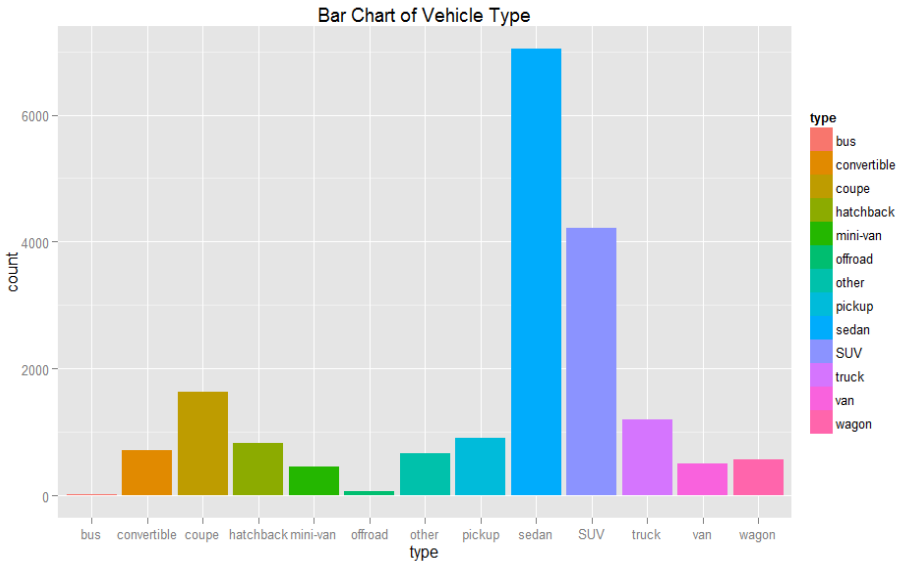
```
#Another kind of code
#vposts %>%
#  filter(!is.na(fuel))%>%
#  ggplot(aes(x=factor(1),fill=fuel))+geom_bar(width=1)+coord_polar("y",
start=0)   #+scale_fill_brewer(palette="Blues")

vposts %>%
  filter(!is.na(drive))%>%
  ggplot(aes(drive))+geom_bar(aes(fill=drive))+ggtitle('Bar Chart of Dr
ive')
```



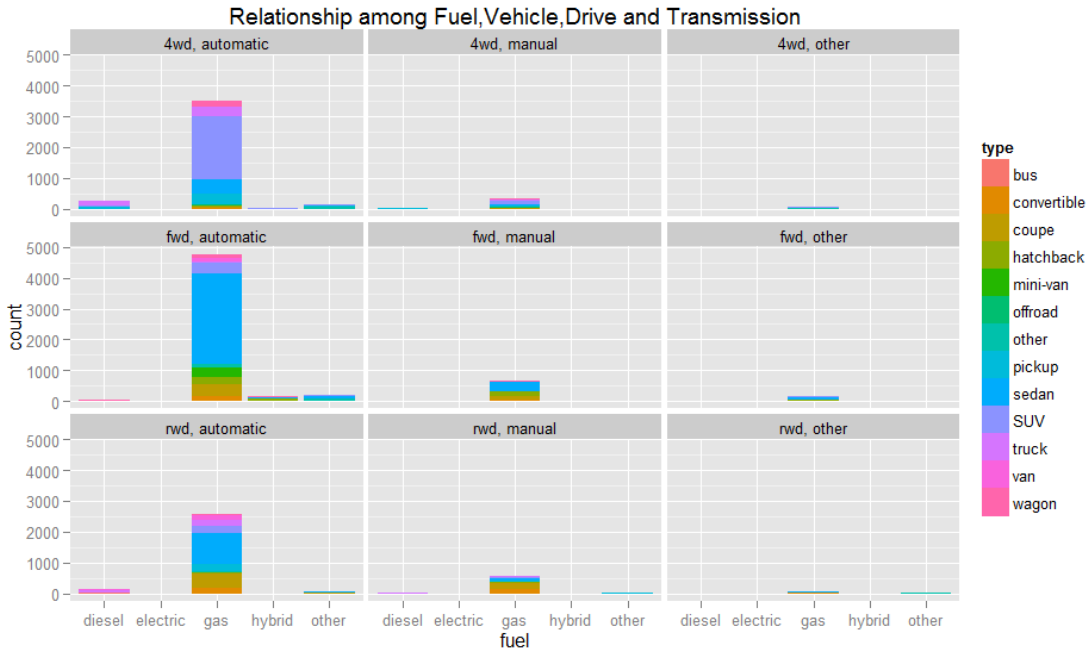

```
vposts %>%
  filter(!is.na(transmission))%>%
  ggplot(aes(x=factor(1),fill=transmission))+geom_bar()+ggtitle('Bar Ch
art of Transmission')
```



```
vposts %>%
  filter(!is.na(type))%>%
  ggplot(aes(type))+geom_bar(aes(fill=type))+ggtitle('Bar Chart of Vehi
cle Type')
```




```
vposts %>%
  filter(!is.na(type)&!is.na(fuel)&!is.na(transmission)&!is.na(driv
e))%>%
  ggplot(aes(fuel,fill=type))+geom_bar()+facet_wrap(drive~transmission)
+ggtitle("Relationship among Fuel,Vehicle,Drive and Transmission")
```



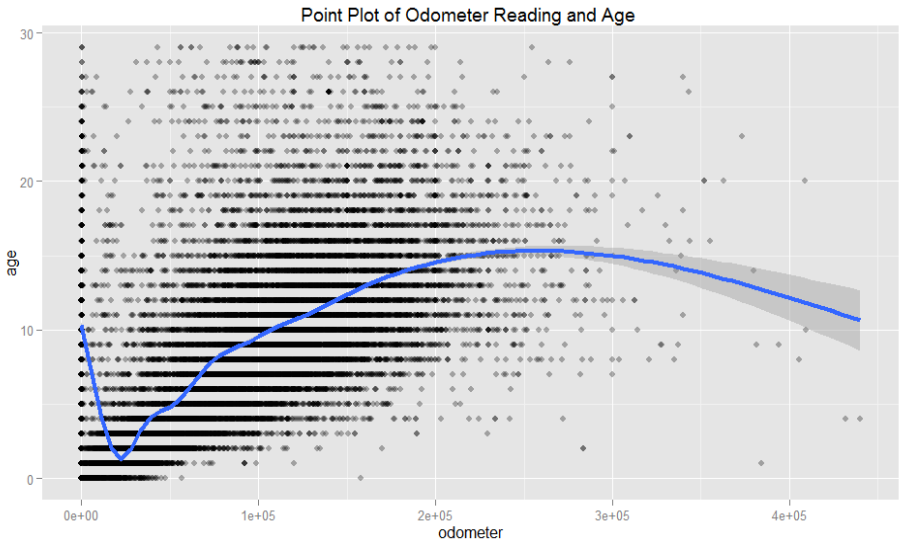
As we know, most of vehicles are automatic, front-wheel drive and use gas as fuel. And I notice that most suv is 4wd, yet most sedan is fwd.

Reference: http://www.sthda.com/english/wiki/ggplot2-pie-chart-quick-start-guide-r-software-and-data-visualization

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?

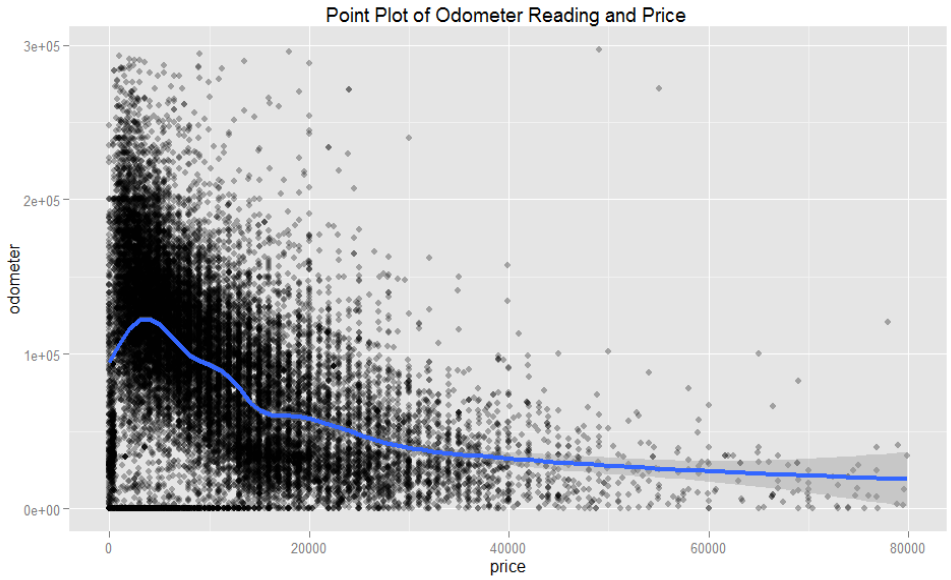
```
library(ggplot2)
library(dplyr)
#To get a good graph, I omit some obeservations whose odometers are ext
reme big and age are old.
vposts2 %>%
  filter(!is.na(odometer)&(odometer<5e5)&(age<30)) %>%
  ggplot(aes(x=odometer,y=age))+geom_point(alpha=0.3)+geom_smooth(size=
1.5)+ggtitle('Point Plot of Odometer Reading and Age ')
```

```
## geom_smooth: method="auto" and size of largest group is >=1000, so u
sing gam with formula: y ~ s(x, bs = "cs"). Use 'method = x' to change
the smoothing method.
```




```
vposts2 %>%
  filter(!is.na(odometer)&(odometer<3e5)&!is.na(price)&(price<8e4)) %>%

  ggplot(aes(x=price,y=odometer))+geom_point(alpha=0.3)+geom_smooth(siz
e=1.5)+ggtitle('Point Plot of Odometer Reading and Price ')
```

```
## geom_smooth: method="auto" and size of largest group is >=1000, so u
sing gam with formula: y ~ s(x, bs = "cs"). Use 'method = x' to change
the smoothing method.
```



According to the plot between age and odometer reading, there is a positive relationship between them in the concentrated area. Instead, there is a negative relationship between price and odometer reading in the concentrated area. Actually, in my opoinion, these relationships are not very obvious.

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

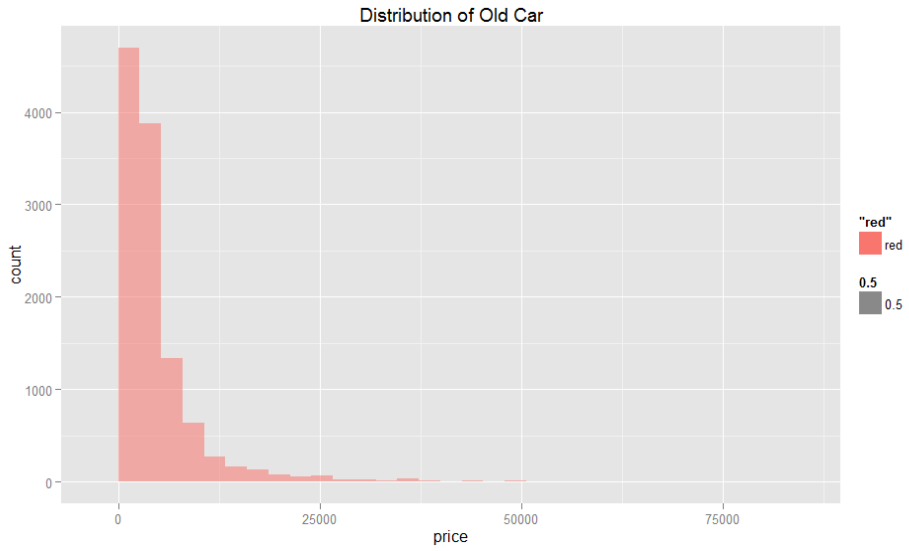
Atomakers say they are building the best cars ever, and there's proof on the road. The average age of vehicles in the U.S. has climbed to a record 11.5 years, according to research firm IHS Automotive. So I define the old car as the car whose age is bigger than or equals to 12 years.

```
library(dplyr)
vposts2 %>%
  filter(age>=12) %>%
  count(maker) %>%
  arrange(desc(n))
```

```
## Source: local data frame [67 x 2]
##
##         maker     n
##         (chr) (int)
## 1        ford  1803
## 2   chevrolet  1497
## 3      toyota  1086
## 4       honda  1060
## 5       dodge   561
## 6      nissan   513
## 7    mercedes   445
```

```
## 8          bmw   418
## 9   volkswagen   418
## 10        jeep   359
## ..         ...   ...
```

```
#Use the same threshhold of price as queation 8.
vposts2 %>%
  filter(age>=12&price<8e4) %>%
  ggplot(aes(price,alpha=0.5))+geom_histogram(aes(fill='red'))+ggtitle
('Distribution of Old Car')
```



Reference: http://www.latimes.com/business/autos/la-fi-hy-ihs-average-car-age-20150729-story.html

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.

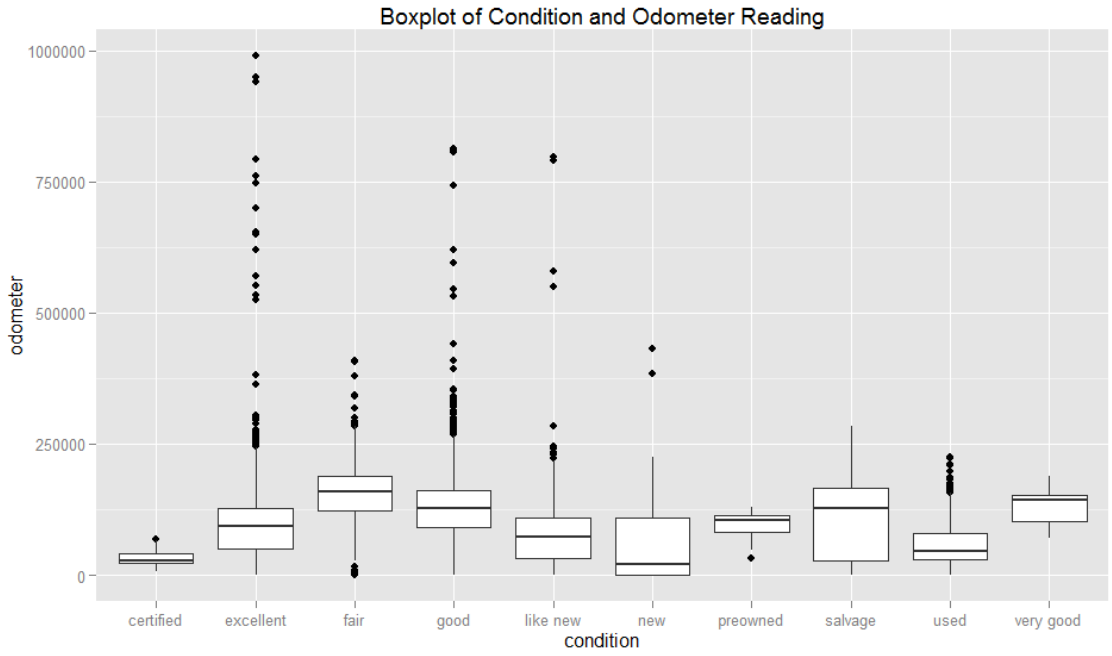
As far as I see, MPG (miles per gallon) is missed. However,professor Duncan said on piazza that only about 1500 observations contain the MPG in the 'body' and it is not enough to construct another variable. In addition, the color is also missed. Someone posted on piazza that a lot of observations contain this information in the 'body' and it is possible to extract out and build another variable.

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.

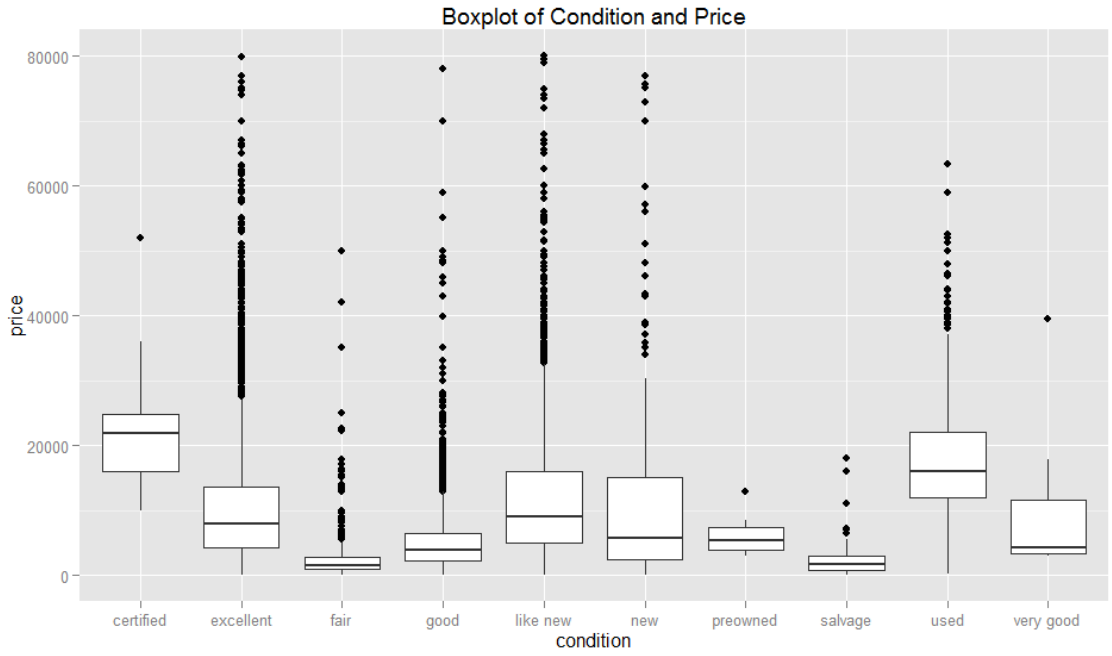
- If you display results in a table, consider a more appropriate plot.
- If you do use a table, consider the number of digits in the numbers displayed.

- If you use a single plot, consider using a legend to identify the different sub-elements.
- If you use multiple panels/plots, consider if the scales are the same and ensure they are if appropriate.
- Make certain plots have the approriate labels on the axes.

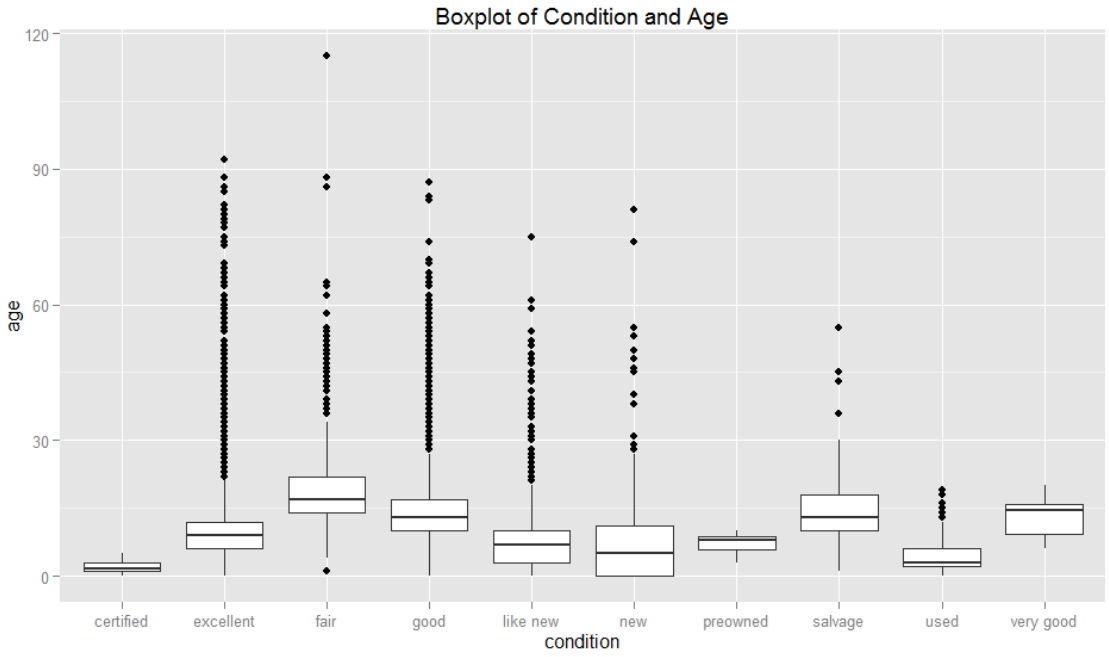
```
library(ggplot2)
library(dplyr)

#I only plot the conditions which show more than 5 times in the datafra
me as the drop part is not representative enough. After, we can get a m
ore clear graph.
commoncondition=
  vposts %>%
  filter(!is.na(condition)) %>%
  count(condition) %>%
  filter(n>5)

#To compare easily,omit some large odometer cars
vposts %>%
  filter(!is.na(odometer)&(odometer<1e6)&!is.na(condition)& (condition %
in%     commoncondition$condition)) %>%
  ggplot(aes(x=condition,y=odometer))+geom_boxplot()+ggtitle("Boxplot o
f Condition and Odometer Reading")
```



```
vposts %>%
  filter(!is.na(price)&!is.na(condition)& (condition %in% commonconditi
on$condition)) %>%
  ggplot(aes(x=condition,y=price))+geom_boxplot()+ggtitle("Boxplot of C
ondition and Price")
```




```
vposts2 %>%
  filter(!is.na(condition)& (condition %in% commoncondition$condition))
 %>%
  ggplot(aes(x=condition,y=age))+geom_boxplot()+ggtitle("Boxplot of Con
dition and Age")
```



We can get some interesting results from three graphs. Mostly,the vehicles in positive conditions such as 'like new' or 'excellent' have higher price,lower odometer reading and shorter age comapred to vehicles in the general or negative conditions such as 'fair' or 'salvage'. In addition, certified and used vehicles have highest price, lowest age and almost lowest odometer readings. However, there are a lot of extreme big values of odometer readings on ‘excellent’, ‘good’ and ‘like new’ condition. There must be some sealers using better words to describe cars and it is a little faked. Besides, the condition 'very good' and 'fair' have similar odometer reading and age, yet the price of 'very good' vehicles are higher. Thus, some vehicles in 'very good' condition may not work so well actually.