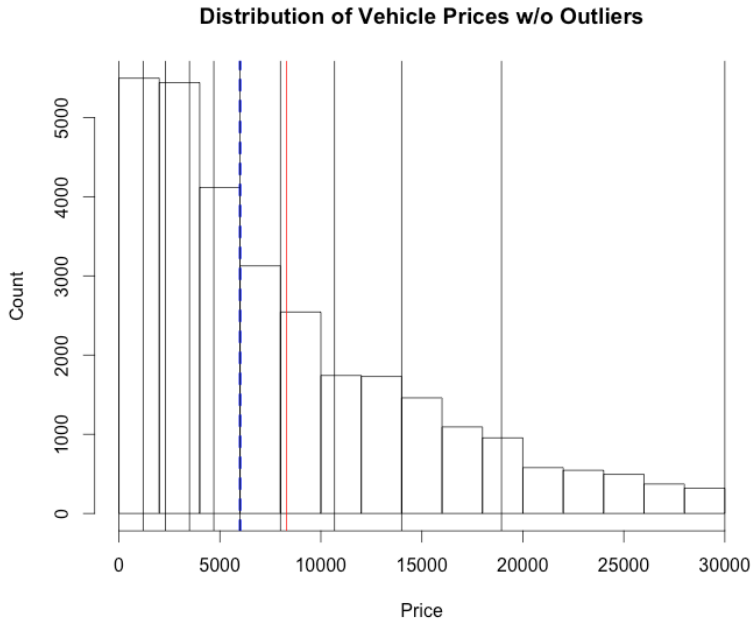
1. How many observations are there in the data set?
   There are 34677 rows and 26 columns, therefore there are 34677 observations in the data set.
2. What are the names of the variables? and what is the class of each variable?
   The names of the variables is id, title, body, lat, long, posted, updated, drive, odometer, type, header, condition, cylinders, fuel, size, transmission, byOwner, city, time, description, location, url, price, year, maker, and makerMethod. Id, title, body, header, description, location, url, maker have class character. Lat, long, price, makerMethod have class numeric. Drive, condition, fuel, size, transmission, city have class factor. Odometer, cylinders, year have class integer. byOwner is a logical. Posted, updated, and time are class POSIXt.
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.
   For the original data, the average price is $49449.9 and the median price is $6700. The deciles are

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

90%	100%
21997	600030000

   However, after graphing the vehicle prices, there were quite a few outliers. To determine which prices were outliers, I tried various methods from subsetting the data by 2 standard deviations from the mean to subsetting the data based on a price I felt was high. I finally settled on subsetting the data based on a USA Today article showing the average used vehicle transaction price for 2013 was about $29,967 for a car that was 1 year old, and used $30,000 for this data set. I decided to keep the lower price values such as $1 because it seems the seller purposely was trying to sell at that price, perhaps as a start bidding value. (http://www.usatoday.com/story/money/cars/2015/02/18/record-used-car-prices-in-2014/23637775/)
   Here is a plot of the distribution of vehicle prices for cars with prices $30,000 and lower. Red line is the mean ($8304.914) and the blue dotted line is median ($6000). The black lines are the deciles

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
1.0	1200.0	2300.0	3495.0	4700.0	6000.0	7999.0	10677.1	13998.0	18943.4	30000.0

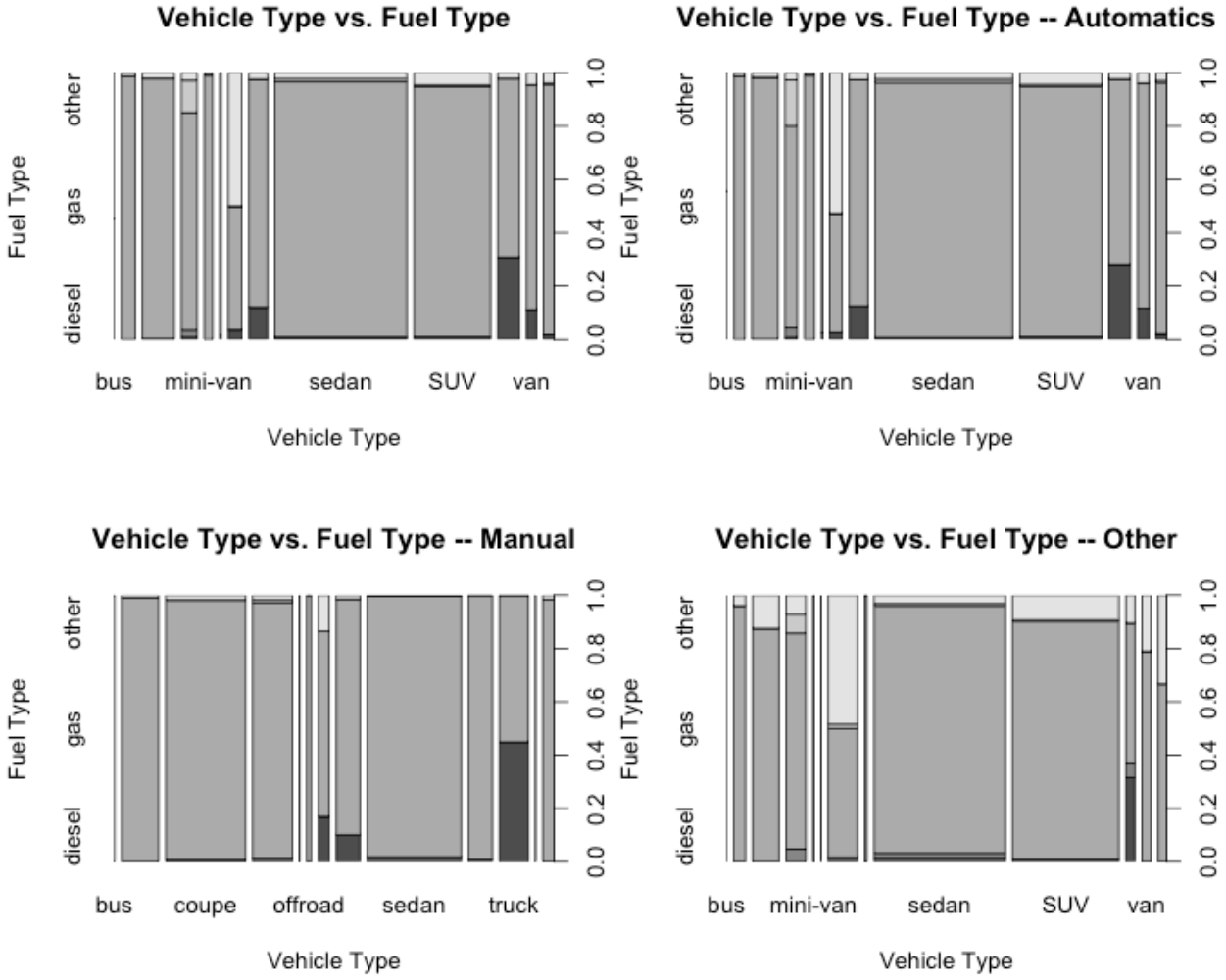


4. What are the different categories of vehicles, i.e. the **type** variable/column? What is the proportion for each category?

Categories of vehicles are bus, convertible, coupe, hatchback, minivan, offroad, other, pickup, sedan, SUV, truck, van, wagon.

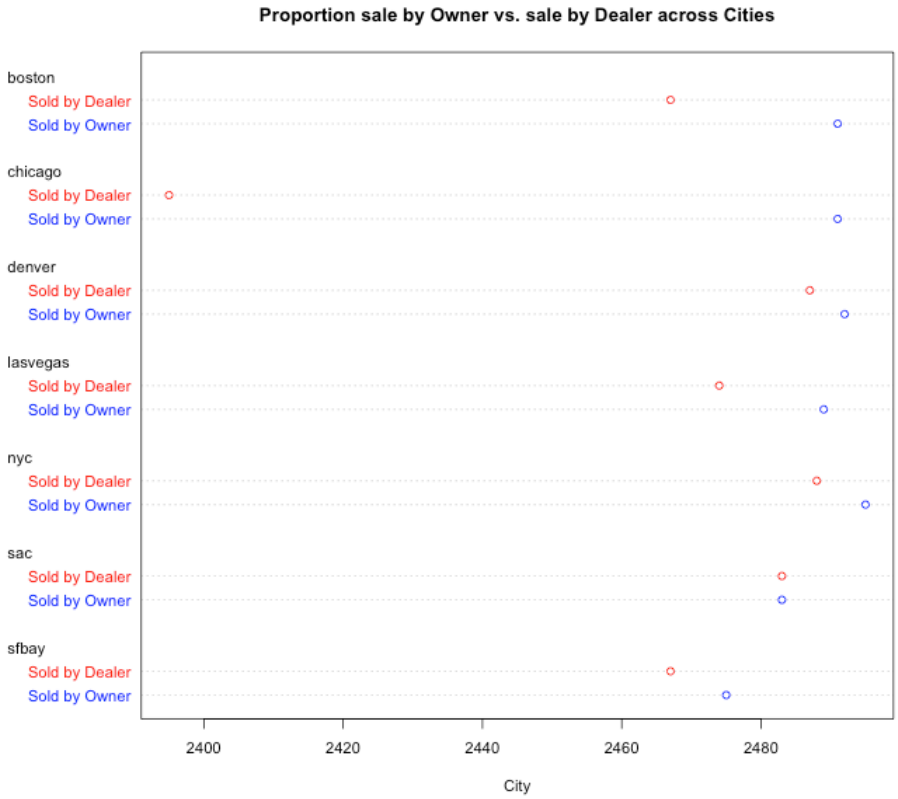
bus convertible coupe hatchback mini-van offroad other
0.001171147 0.037583178 0.086558424 0.043598616 0.024114985 0.003513442 0.035453820
pickup sedan SUV truck van wagon
0.048389673 0.374767101 0.224168219 0.063987224 0.026989619 0.029704552

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



Yes this depends on transmission type because different types of vehicles use different amounts of fuel types depending on the transmission type, notably automatics and manual has a drastic difference. It is evidenced by the sizes of the rectangular areas.

6. How many different cities are represented in the dataset?
   There are 7 different cities: Boston, Chicago, Denver, Las Vegas, NYC, Sac, SF Bay
7. Visually how the number/proportion of "for sale by owner" and "for sale by dealer" varies across city?



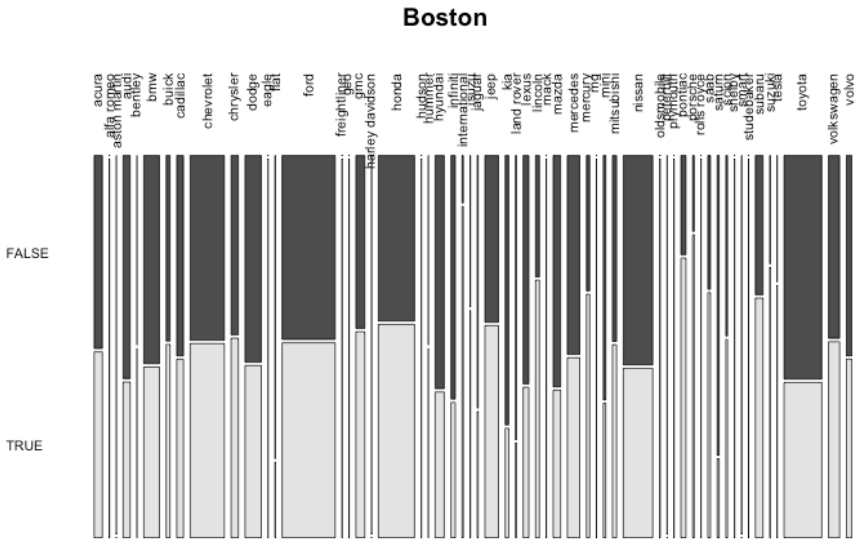
8. What is the largest price for a vehicle in this data set? Examine this and fix the value. Now examine the new highest value for price.
   The largest price for a vehicle is $600030000. I fixed it by replacing it with the value $21000 which is the average of the two prices because the description mentions it is a range of values. The new highest value is $30002500 which I followed a similar process, and replaced with $30000 which is about the average of a used car sale price (from the USA Today article). The new highest value is $9999999, which I again replaced with $30000.
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?
   The three most common makes of cars for sale by owner:

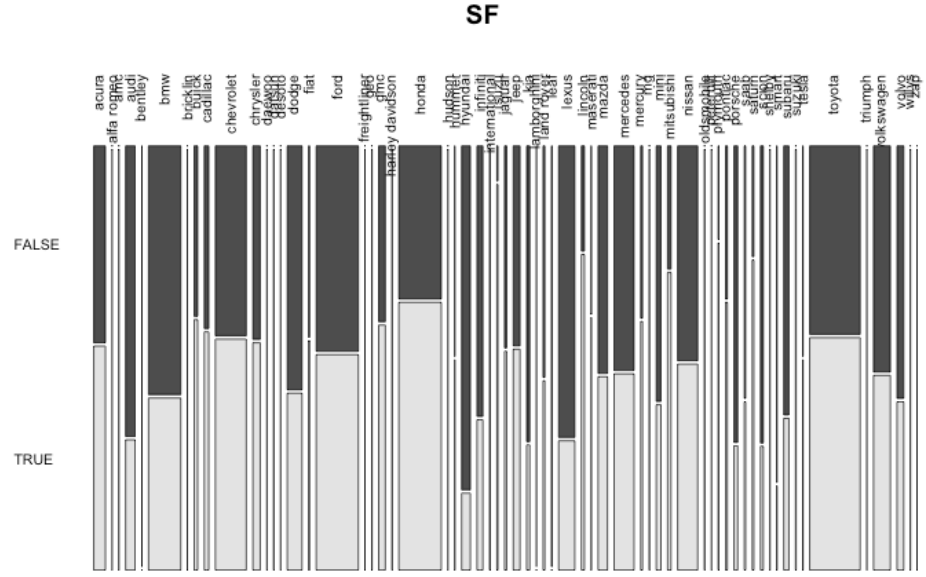
	boston	chicago	denver	lasvegas	nyc	sac	sfbay
[1,]	"ford"	"chevrolet"	"ford"	"ford"	"nissan"	"ford"	"toyota"
[2,]	"toyota"	"ford"	"chevrolet"	"chevrolet"	"toyota"	"toyota"	"honda"
[3,]	"honda"	"toyota"	"toyota"	"toyota"	"honda"	"chevrolet"	"ford"

   for sale by dealer:

	boston	chicago	denver	lasvegas	nyc	sac	sfbay
[1,]	"ford"	"chevrolet"	"ford"	"ford"	"nissan"	"ford"	"toyota"
[2,]	"toyota"	"ford"	"chevrolet"	"nissan"	"toyota"	"toyota"	"ford"
[3,]	"chevrolet"	"nissan"	"dodge"	"chevrolet"	"honda"	"chevrolet"	"bmw"

   They are quite similar. For each city between sale by owner and dealer, there is at least one top car make in common. For each city, I also made a mosaic plot to visually assess the relative sizes and they are about the same, which means the car makes sold either by owner or dealer is about similar. Below is example of what the mosaic plot looks like for Boston and SF Bay with False meaning sold by dealer and true meaning sold by owner.

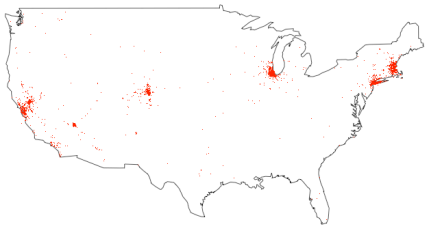




10. Visually compare the distribution of the age of cars for different cities and for "sale by owner" and "sale by dealer". Provide an interpretation of the plots, i.e., what are the key conclusions and insights?

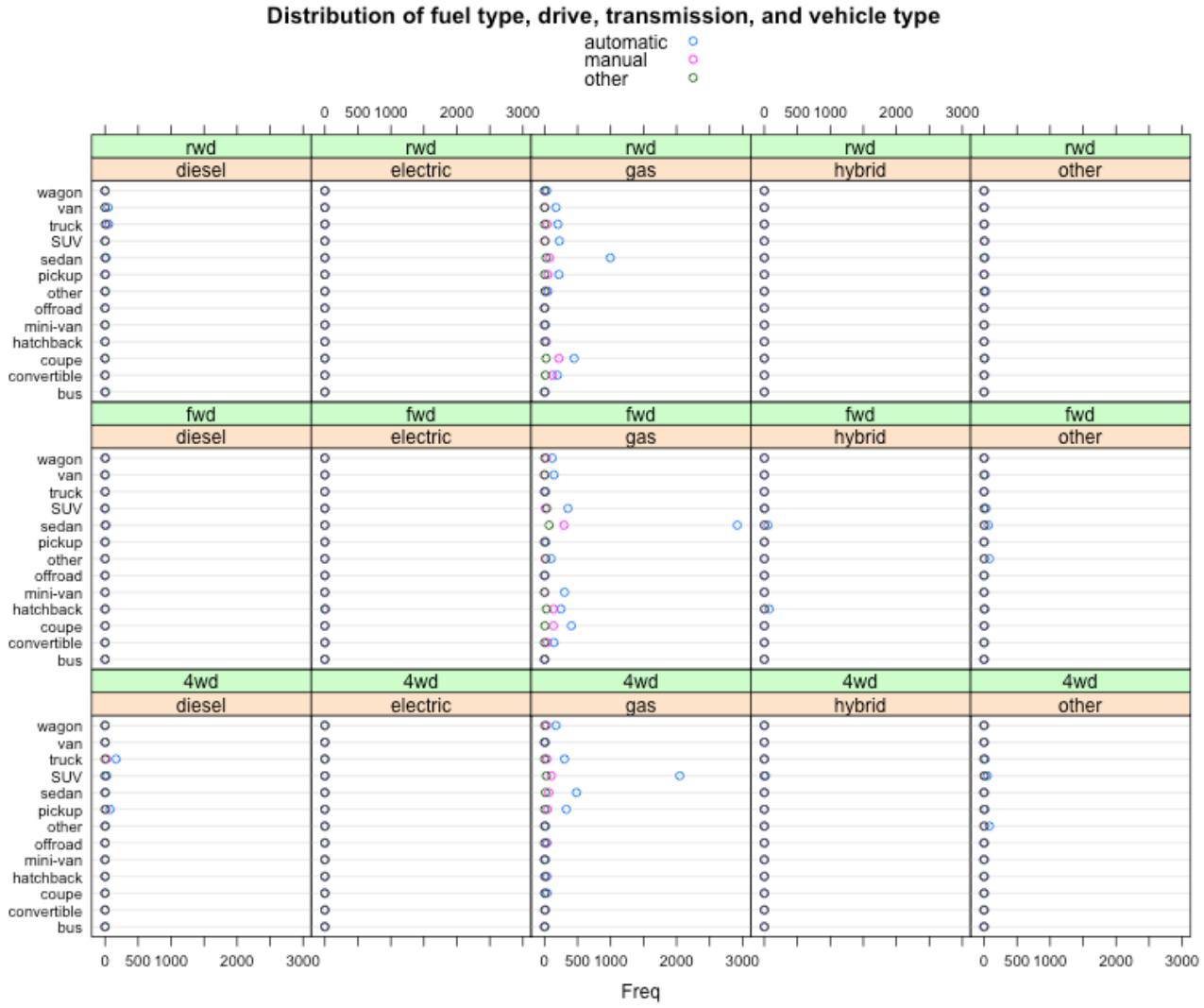
For each city, whether the car was sold by dealer (false) or owner (true), had similar distributions. Usually cars sold by dealer tend to be newer cars. Usually cars that were sold by owner were more varied in age of the car. In order to calculate the age of the car, I subtracted the year of the car from the current year 2015. There was a -7, and a couple hundred -1's and on the high end a couple hundred 115's year old cars and an extreme 2011 year old car. I looked into the description of those specific observations and found that the cars aged -1 were actually 2016 cars which is reasonable and cars aged 115 was also reasonable as 1900 year. Therefore, I subsetted the data to look at cars age from -1 to 115.

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



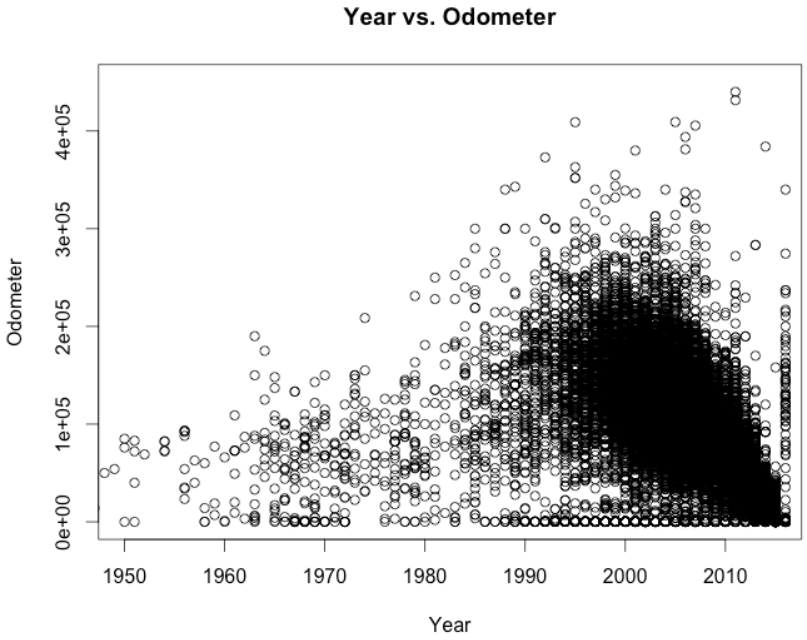
I noticed the posts come from the US, so I looked at only the US map. Then, I noticed the locations are particularly from major cities.

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

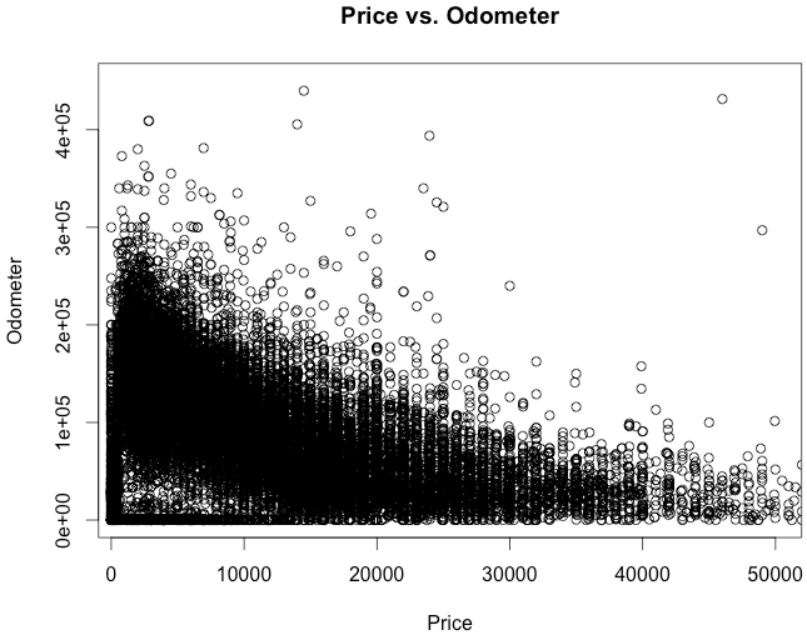


I used lattice dotplot of a table of the four variables to display the distribution.

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?

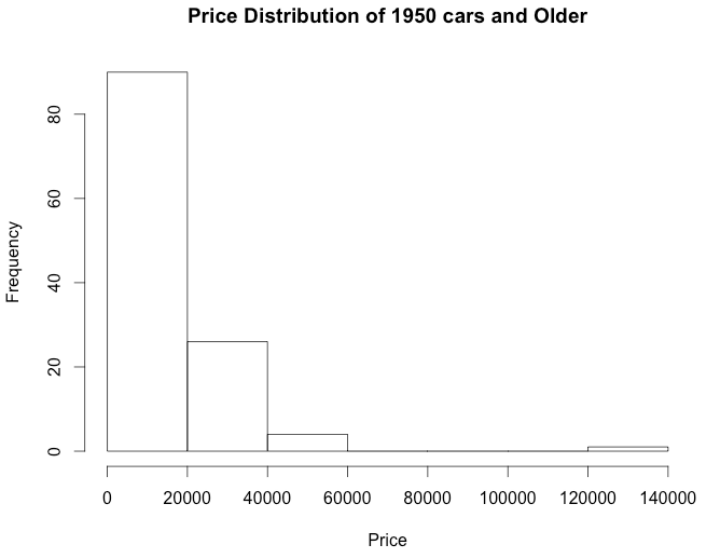


Yes there is a relationship between odometer reading and age of car. It seems that cars that are newer tend to have less miles traveled maybe because it hasn't been driven as much yet. Perhaps, cars that are newer have better technology that allows it to travel farther or more than older cars. For the plot, I looked at years 1950 to 2015 because I define "old" cars as 1950 or older in the next problem. I choose 2015 because that is our current year. I choose odometer readings from 0 to 450,000 miles because it yielded the best plot where I could see a relationship clearly, yet made logical sense.

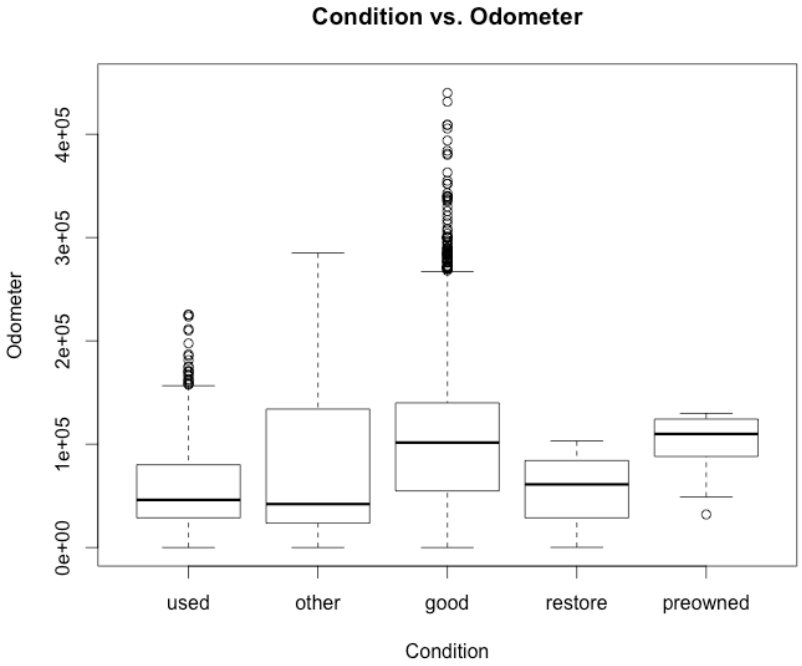


Yes there is also a relationship between odometer reading and price of car. Cars that have more mileage, tend to be valued less. This makes sense because the value of the car becomes more "used" as mileage increases, therefore it should be sold at a lower price. For this plot, I made the odometer ranges the same as the previous plot. For the price, I choose between 0 to $50,000 because it yielded the plot in which I could see the relationship quite clearly.

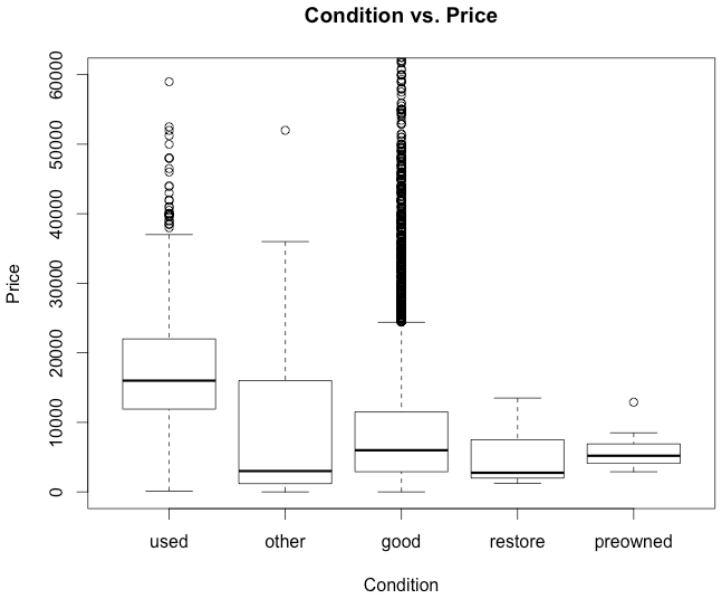
14. Identify the "old" cars. What manufacturers made these? What is the price distribution for these?
I will define "old" cars as at least 65 years old from the current year. So I am looking at car years from 1950 and older. The manufacturers that made these include: Bugatti, Buick, Cadillac, Chevrolet, Desoto, Dogde, Ford, Hudson, International, Jeep, Lincoln, Mercury, Oldsmobile, Plymouth, Pontiac, Studebaker, Willys.



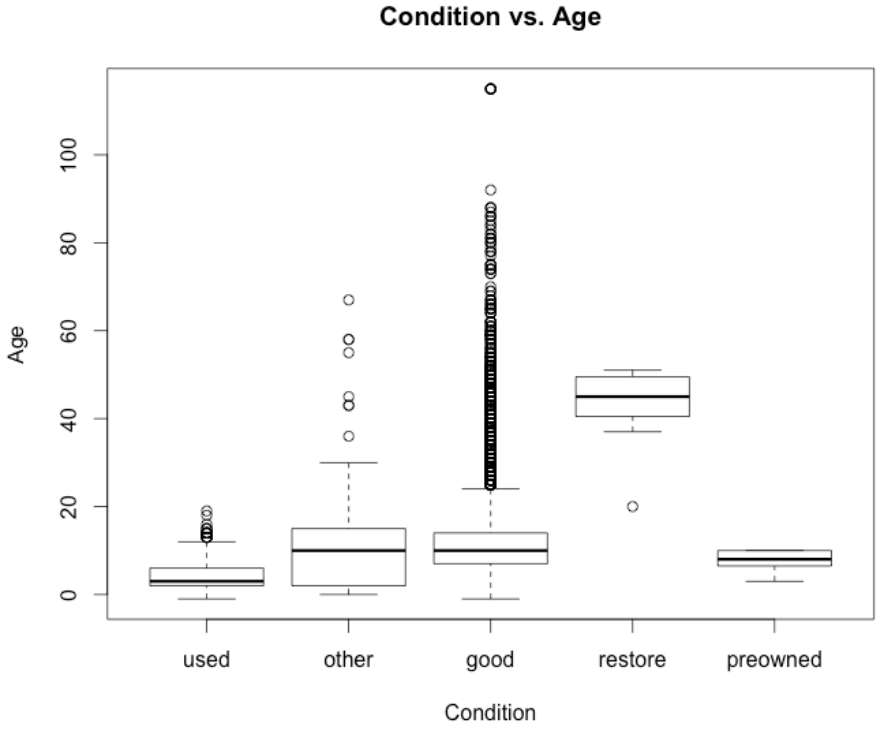
15. I have omitted one important variable in this data set. What do you think it is? Can we derive this from the other variables? If so, sketch possible ideas as to how we would compute this variable.
    I think the data set is missing the title status of the car. I think it may be determined from the body or description of the vehicles. Based on the description, we could categorize the variable into clean, salvage, rebuilt, lien, or other. I got this information from searching up used cars on Craigslist and noticed on the left side bar there is an option for “title status”. (https://sfbay.craigslist.org/search/cto)
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.
    Since there were about 43 different conditions, I grouped similar conditions into 5 categories: used, other, good, restore, and preowned. Then, I made a boxplot of each relationship. Below each plot, I will provide interpretation.



For condition vs. odometer, most of the vehicles were preowned and in good condition with the most amount of miles travelled. The good conditioned vehicles had many outliers, which might have just been how the data was. Used, other, and restored vehicles had similar odometer readings.



For condition vs price, majority of the used cars were most pricey compared to other, restored, and preowned. This makes intuitive sense, though not so much why the good conditioned vehicles weren't priced higher. However, there are many outliers even though I was looking at a subset of the prices up to $60,000. I previously looked up to $30,000 in the other question but wanted to get a better picture of the data.



For condition vs. age, I found that the cars in "worser" condition tender to be older. For example, restored cars were the oldest cars. For categories such as used, preowned, and other, the labels are really quite vague. I'm sure there might be some interaction between a few as in some vehicles can be counted as multiple conditions.

Code Appendix

```
##  Assignment 1 Part 1

# load the data
print(load(url("http://eeyore.ucdavis.edu/stat141/Data/vehicles.rda")))
vehicles = load("/Users/tiffanychen/Desktop/STA 141/vehicles.rda")

# q1
# there are 34677 rows and 26 columns
dim(vposts)
# number of observations
nrow(vposts)

# q2
names(vposts)
sapply(vposts,class)

# q3
mean(vposts$price, na.rm = TRUE)
```

```
sd(vposts$price, na.rm = T)
median(vposts$price, na.rm = TRUE)
quantile(vposts$price, probs = seq(0,1,0.1), na.rm = TRUE)

prices = subset(vposts$price, vposts$price <= 30000,) # from Piazza

# hist original data w/ outlier
hist(vposts$price)
rug(vposts$price)
summary(vposts$price)

# hist with no outlier
hist(prices, main = "Distribution of Vehicle Prices w/o Outliers", xlab = "Price", ylab = "Count")
m = mean(prices, na.rm = TRUE)
med = median(prices, na.rm = TRUE)
q = quantile(prices, probs = seq(0,1,0.1), na.rm = TRUE)
abline(v = c(m, med), col = c("red", "blue"), lty = c(1,2), lwd = c(1,3)) # mean and median
abline(v = q) # deciles

# from Piazza
price_boxplot = boxplot(vposts$price)
sort(price_boxplot$out)
## shows the outliers,
# i should subset the data
# excluding prices 29000 and higher, similar to online results

#q4
## this deleted the NAs
x = table(vposts$type)
prop.table(x)
# same as
table(vposts$type)/length(vposts$type)

#q5
levels(vposts$fuel) # give levels of the unique values

#relationship btwn type and fuel
plot(vposts$type,vposts$fuel,main='Vehicle Type vs. Fuel Type',xlab="Vehicle Type",ylab="Fuel Type")
with(vposts,plot(type,fuel))

# all three
x=table(vposts$type,vposts$fuel,vposts$transmission)
mosaicplot(x, las=2)
axis(las=2)

# does this depend on transmission type?
# yes
# piazza
par(mfrow=c(2,2))
```

```
auto = subset(vposts, transmission == "automatic")
plot(auto$type, auto$fuel, main = "Vehicle Type vs. Fuel Type -- Automatics", xlab = "Vehicle Type", ylab = "Fuel Type")
man = subset(vposts,transmission=="manual")
plot(man$type, man$fuel, main = "Vehicle Type vs. Fuel Type -- Manual", xlab= "Vehicle Type", ylab = "Fuel Type")

other = subset(vposts,transmission == "other")
plot(other$type, other$fuel, main="Vehicle Type vs. Fuel Type -- Other", xlab = "Vehicle Type", ylab = "Fuel Type")

#q6
# 7 different cities
table(vposts$city)

#q7
c=table(vposts$byOwner, vposts$city)
barplot(c, main="Proportion sale by Owner vs. sale by Dealer across Cities", xlab="City",ylab="Number of Cars",col=c("blue","red"),legend.text=c("Sold by Dealer","Sold by Owner"),beside=TRUE)
## dot chart from piazza
dotchart(c, main="Proportion sale by Owner vs. sale by Dealer across Cities",
xlab="City",col=c("blue","red"),labels=c("Sold by Owner", "Sold by Dealer"),cex=.7)

#q8
p = vposts$price
max(p,na.rm=TRUE)
ids=which(vposts$price==600030000)
vposts[ids,"price"]=21000
p = vposts$price
max(p,na.rm=TRUE)
ids=which(vposts$price==30002500)
vposts[ids,"price"]=30000
p = vposts$price
max(p,na.rm=TRUE)
ids=which(vposts$price==9999999)
vposts[ids,"price"]=30000
p = vposts$price
max(p,na.rm=TRUE)

# q9
# piazza Nick
owner = subset(vposts, byOwner == "TRUE") # subset byOwner first
counts = table(owner$city, owner$maker)
top3_makes = function(city_counts) {
  # Assume city_counts has counts for just one city.
  # Get indexes of "top 3" makes.
  # You could also use head(..., 3) instead of [1:3].
  top3 = order(city_counts, decreasing = TRUE)[1:3]
  # Convert indexes to names.
```

```
  # Use rownames() instead if "make" is the rows.
  colnames(counts)[top3]
}
apply(counts,1,top3_makes)
dealer=subset(vposts,byOwner=="FALSE") # then subset by dealer
counts=table(dealer$city,dealer$maker)
c=table(vposts$city,vposts$maker,vposts$byOwner)
c=table(vposts$byOwner,vposts$city,vposts$maker)

par(mfrow=c(2,2))
bos = subset(vposts,city=="boston")
x=table(bos$maker,bos$byOwner)
mosaicplot(x,color=TRUE,las=2,main="Boston")

chi = subset(vposts,city=="chicago")
x=table(chi$maker,chi$byOwner)
mosaicplot(x,color=TRUE,las=2,main="Chicago")

den = subset(vposts,city=="denver")
x=table(den$maker,den$byOwner)
mosaicplot(x,color=TRUE,las=2,main="Denver")

lv = subset(vposts,city=="lasvegas")
x=table(lv$maker,lv$byOwner)
mosaicplot(x,color=TRUE,las=2,main="Las Vegas")

nyc = subset(vposts,city=="nyc")
x=table(nyc$maker,nyc$byOwner)
mosaicplot(x,color=TRUE,las=2,main="NYC")

sac = subset(vposts,city=="sac")
x=table(sac$maker,sac$byOwner)
mosaicplot(x,color=TRUE,las=2,main="Sac")

sfbay = subset(vposts,city=="sfbay")
x=table(sfbay$maker,sfbay$byOwner)
mosaicplot(x,color=TRUE,las=2,main="SF")

#q10 ??
x=table(vposts$year,vposts$city,vposts$byOwner)
mosaicplot(x)
vposts$age = 2015 - vposts$year # many negative values, so got to clean it
head(sort(vposts$age), 250)

ids=which(vposts$age==-1)
vposts[1929,] # 2016 car
vposts[21975,] #2022 car, actually is a 2016 honda.
vposts[9361,] #1962
vposts[2078,] #2016
```

```
vposts[2785,] #2016 car

tail(sort(vposts$age),250)
ids=which(vposts$age==2011) # check out the aged 2011 car
vposts[8417,] # just delete this
ids=which(vposts$age==115)
vposts[27557,] # 1900  tires
vposts[27901,] # 1900 dodge

# subset by age -1 to 115
betterage = vposts$age[vposts$age>=-1 & vposts$age<=115]

library(lattice)
histogram( ~ betterage | byOwner + city, vposts, main="Distribution of Cars by City and Sale by
Owner/Dealer",xlab="Age of Car (2015 as current year)",breaks=seq(-5,120,5))

#q11
install.packages("maps")
library(maps)
map('usa')
map() # looks like its just the US
points(vposts$long, vposts$lat, col = "red", pch = ".")

#q12
x=table(vposts$type, vposts$fuel, vposts$drive, vposts$transmission)
library(lattice)
dotplot(x, besides=TRUE, breaks=seq(0,3000,10), main="Distribution of fuel type, drive, transmission, and
vehicle type", auto.key=TRUE)

## below is not good cuz the ranges on the axes are different
par(mfrow=c(2,2))
plot(vposts$fuel, main="Distribution of Fuel Type",xlab="Fuel Type",ylab="Frequency",ylim=c(0,50000))
plot(vposts$drive, main="Distribution of Drive Type",xlab="Drive Type",ylab="Frequency",ylim=c(0,50000))
plot(vposts$transmission, main="Distribution of Transmission Type",xlab="Transmission
Type",ylab="Frequency",ylim=c(0,50000))
plot(vposts$type,main="Distribution of Vehicle Type",xlab="Vehicle
Type",ylab="Frequency",ylim=c(0,50000),cex.names=.5)

# q13
# cars that are newer, have not been driven yet/less miles traveled
# cars that are newer with better technology can travel farther
# compared to older cars
plot(vposts$year, vposts$odometer) # clumped up

# year vs odometer
plot(vposts$year, vposts$odometer, xlab = "Year", ylab = "Odometer", main= "Year vs. Odometer", xlim =
c(1950,2015), ylim = c(0,450000))

# price vs odometer
```

```
plot(vposts$price, vposts$odometer, xlab="Price",ylab="Odometer", main="Price vs. Odometer", xlim = c(1000,50000), ylim = c(0,450000))
# cars that have more mileage, are valued less

#q14
# define old car as 25 years old. so 1990 and earlier
old = vposts[vposts$year <= 1950,]
unique(old$maker)
hist(old$price, main = "Price Distribution of 1950 cars and Older", xlab= "Price")

#q15
# colour, title status?

#q16
# groups conditions into good, restored, used, other, preowned
cond_cat = vposts # make a copy

levels(cond_cat$condition)[levels(cond_cat$condition) %in% c("excellent","like new", "nice", "nice teuck", "superb original", "very good", "good", "fair", "new") ] = "good"

levels(cond_cat$condition)[levels(cond_cat$condition) %in% c("muscle car restore", "needs restoration!", "needs total restore", "restoration", "restore","needs restored","needs work","nice rolling restoration","rebuildable project","restoration project","restored") ] = "restore"

levels(cond_cat$condition)[levels(cond_cat$condition)%in% c("0used", "used")] = "used"

levels(cond_cat$condition)[levels(cond_cat$condition)%in% c("ac/heater","certified","front side damage","hit and run :( gently","needs work/for parts","parts","pre-owned","preownes","project car","rough but runs","207,400","carfax guarantee!!","complete parts car, blown engine","honnda","mint","needs bodywork","needs work","not running","project","salvage")] = "other"

levels(cond_cat$condition)[levels(cond_cat$condition) %in% c("pre-owned","preownes","pre owned","preowned")] = "preowned"

# boxplots of condition vs odometer
plot(cond_cat$condition, cond_cat$odometer, ylim=c(0,450000), main = "Condition vs. Odometer", xlab = "Condition", ylab = "Odometer" )

# boxplot of condition vs price
plot(cond_cat$condition, cond_cat$price, main = "Condition vs. Price", xlab = "Condition", ylab = "Price", ylim=c(0,60000))

# boxplot of age and condition
cond_cat$age = 2015 - cond_cat$year
plot(cond_cat$condition, cond_cat$age, ylim = c(-1, 115), main = "Condition vs. Age", xlab = "Condition", ylab = "Age" )
```