Katherine Olson

Assignment One Part I

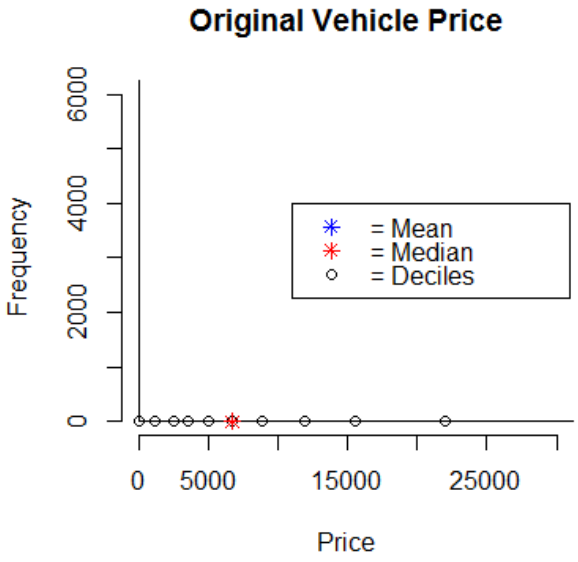
Question 1: How many observations are there in the data set?

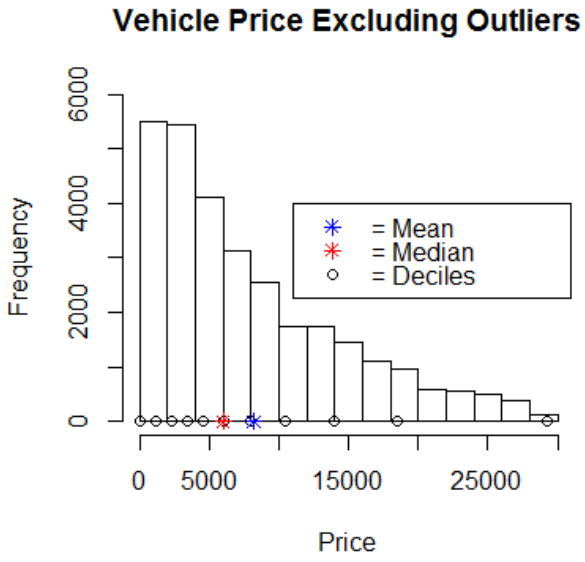
There are 34,677 observations.

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

Name	Class	Name	Class
id	character	fuel	factor
title	character	size	factor
body	character	transmission	factor
lat	numeric	byOwner	logical
long	numeric	city	factor
posted	POSIXct POSIXt	time	POSIXct POSIXt
updated	POSIXct POSIXt	description	character
drive	factor	location	character
odometer	integer	url	character
type	factor	price	integer
header	character	year	integer
condition	factor	maker	character
cylinders	integer	makerMethod	numeric

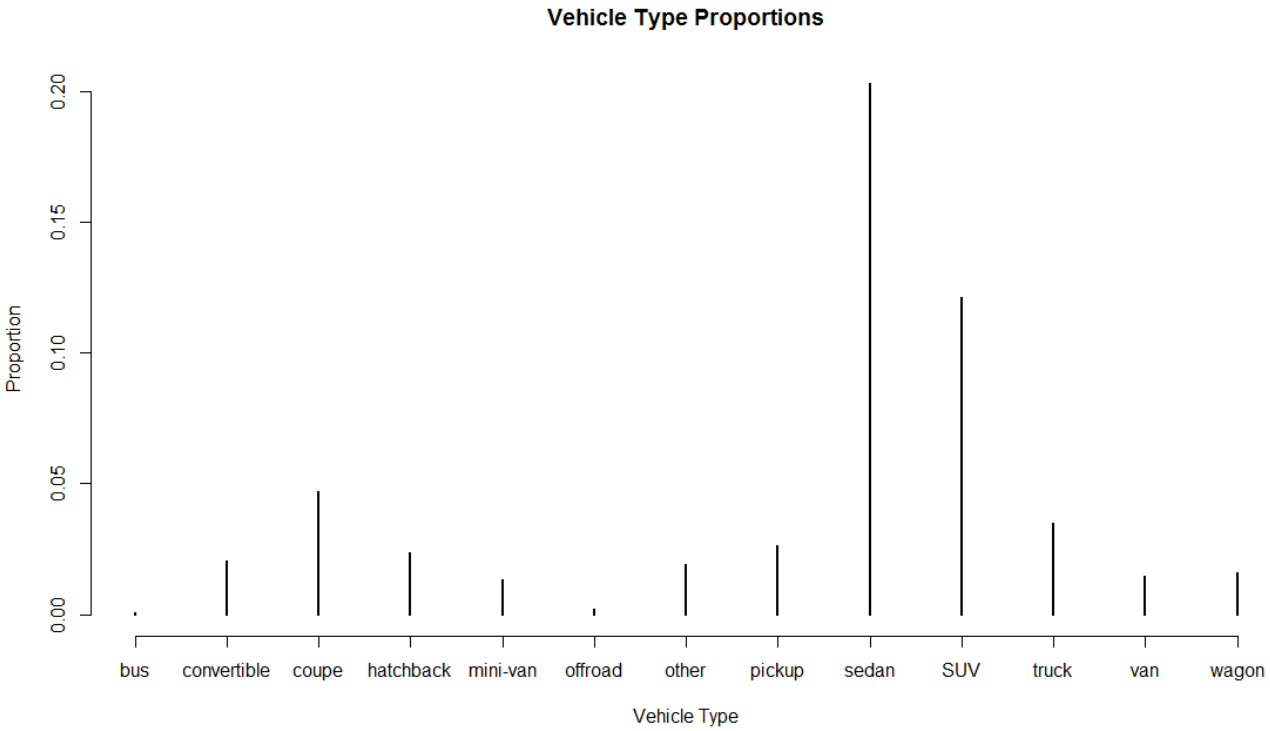
Question 3: What is the average price of all the vehicles? the median price? and the deciles? Displays these on a plot of the distribution of vehicle prices.





The original mean price is $49,450 and the original median price is $6,700. After plotting the original price, the graph does not look right. This is the left graph of the two graphs above, there is one vertical line at x = 0. Thinking that this strange graph might be due to outliers, I excluded outliers by using a lower bound of Q1 - 1.5 * IQR (2995) and an upper bound of Q3 + 1.5 * IQR (13500). The resulting graph is the right graph of the two graphs above. This graph gives a much better distribution of the vehicle price. The histogram is skewed right, meaning there are a lot of vehicles with low prices and fewer vehicles with high prices. The new mean price is $8,172 and the new median price is $6,000. This makes it clear that there were outliers that skewed the original mean. I just used this general rule of excluding outliers for the purpose of seeing a distribution, I understand that the actual process of changing outlier is more tedious.

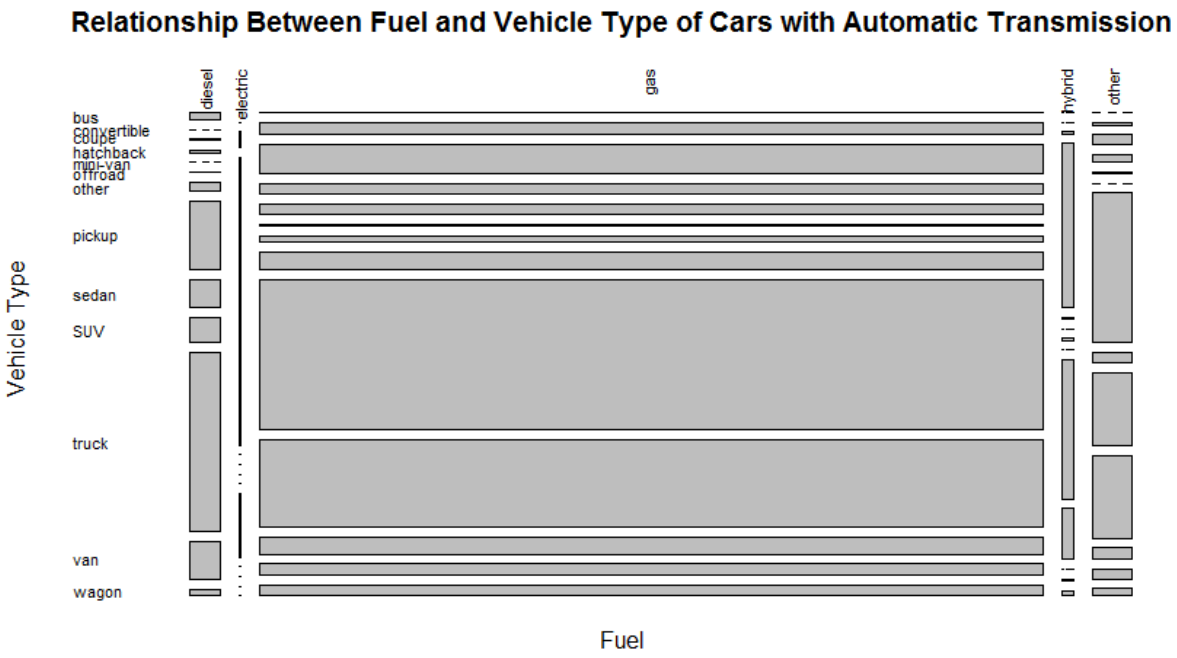
Question 4: What are the different categories of vehicles, i.e. the type variable/column? What is the proportion for each category ?



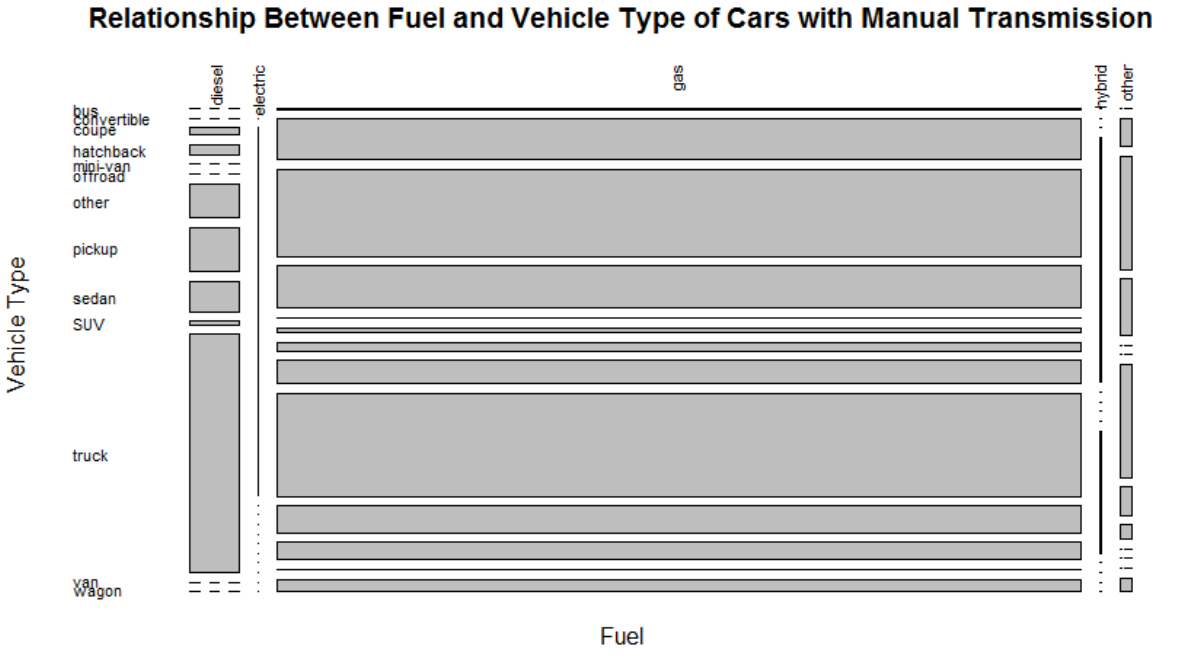
Looking at the graph above, there are 13 different categories of vehicles with sedans having the highest proportion, followed by SUVs and the rest of the vehicles close to or below 0.05.

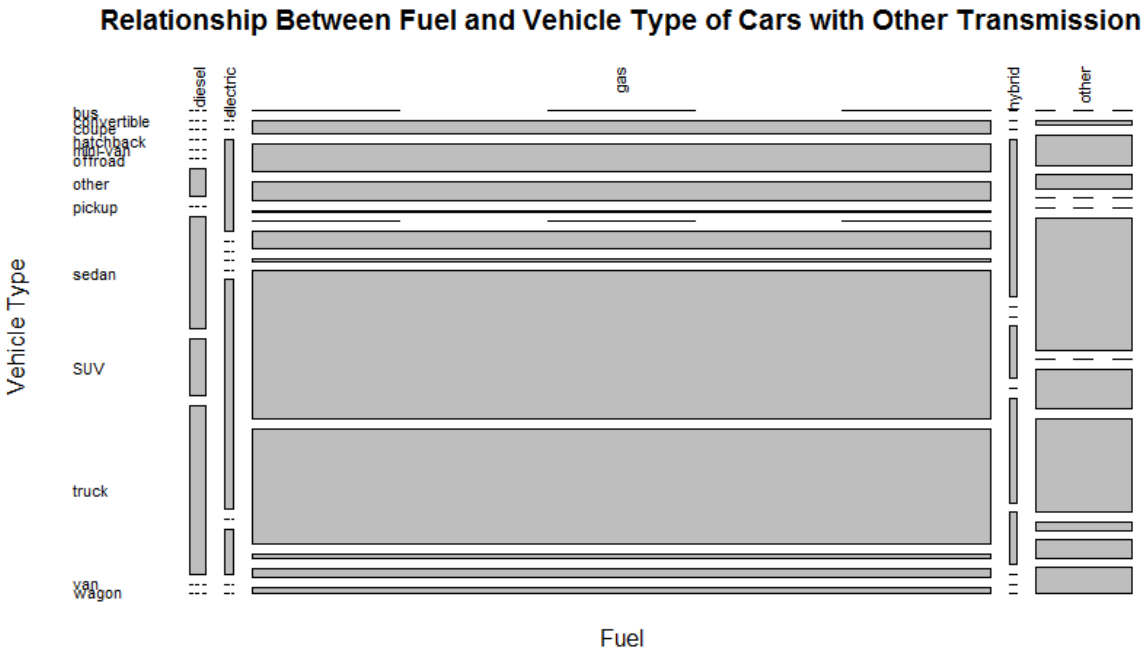
Question 5: Display the relationship between fuel type and vehicle type. Does this depend on transmission type?

The data was separated into three plots by sub-setting the data by transmission. The three data sets were then plotted to compare fuel type and vehicle type using the mosaicplot() function. These plots can be seen on the next page.



Looking at the vertical lengths of each type of vehicle gives us the proportion of that variable in each fuel category. So for diesel fuel, trucks have the highest proportion in all three graphs. For electric, the highest proportions are hatchback for automatic, coupe for manual, and sedan for other. For gas, sedans have the highest proportion in all three graphs. For hybrid, hatchbacks have the highest proportion in all three types of transmission. For other, the highest proportions are coupe for manual and other or automatic and other. These results for the highest proportions in each type of transmission are similar for some types of fuel, but different for others. We can also see that gas is by far the most common type of fuel for all three types of transmission.



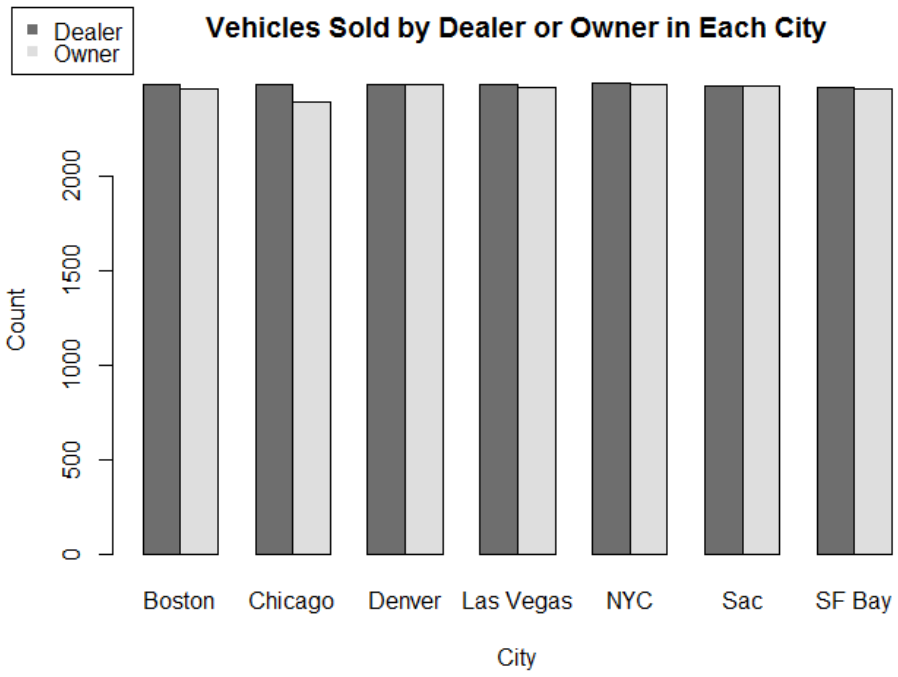


Since separating the data into three groups did not have any drastic changes to the three graphs. They are fairly similar, only with small differences. So we can conclude that the relationship between fuel and vehicle type does not depend on transmission.

Question 6: How many different cities are represented in the dataset?

There are seven cities.

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



Looking at the plot on the left, it is clear that the number of "for sale by owner" and "for sale by dealer" varies very little across city. Chicago has the greatest difference, followed by Boston, and then the rest have almost no difference. "Sale by dealer" appears to have a slightly higher count than "sale by owner" in each city.

Question 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.

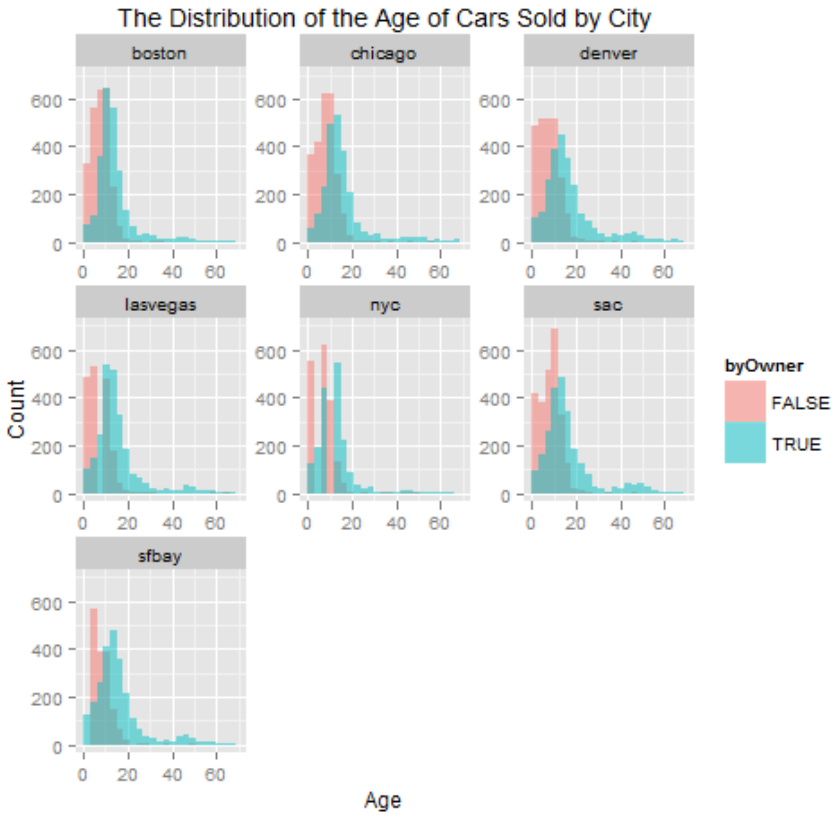
The largest price is $600,030,000. Looking at this row of data, the ad mentions that the price is between $6,000 and $30,000. There was a "-" that was not entered. To fix this, I took the mean of 6,000 and 30,000 which equals 18,000 and entered this as the price. Now the largest price is 30,002,500. Looking at that row, there are not any clues that the price is wrong or there was a typo. 30 million is just very large. I searched Google for a similar car and found one on craigslist for $10,000 dollars and decided to go with that. I repeated this process until I reached a price that was reasonable based on searches for that same car online. I ended up repeating the process seven times. I changed four prices by searching online, one by noticing a type, and deleted a duplicate row. The max is now $400,000 which seems correct given that most prices I found for this car were around $300,000.

Question 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?

City	Owner #1	Owner #2	Owner #3	Dealer #1	Dealer #2	Dealer #3
Boston	Ford	Honda	Chevrolet	Ford	Toyota	Chevrolet
Chicago	Chevrolet	Ford	Honda	Ford	Chevrolet	Nissan
Denver	Ford	Chevrolet	Toyota	Ford	Chevrolet	Dodge
Las Vegas	Ford	Chevrolet	Toyota	Ford	Nissan	Chevrolet
NYC	Nissan	Toyota	Honda	Nissan	Toyota	Honda
Sac	Toyota	Ford	Chevrolet	Ford	Toyota	Chevrolet
SF Bay	Toyota	Honda	Ford	Toyota	Ford	BMW

Looking at the table above, NYC and Sac have the same top three, but a different order for Sac. Boston, Chicago, Denver, Las Vegas, and SF Bay have two of the top three the same for owner and dealer. So they are very similar.

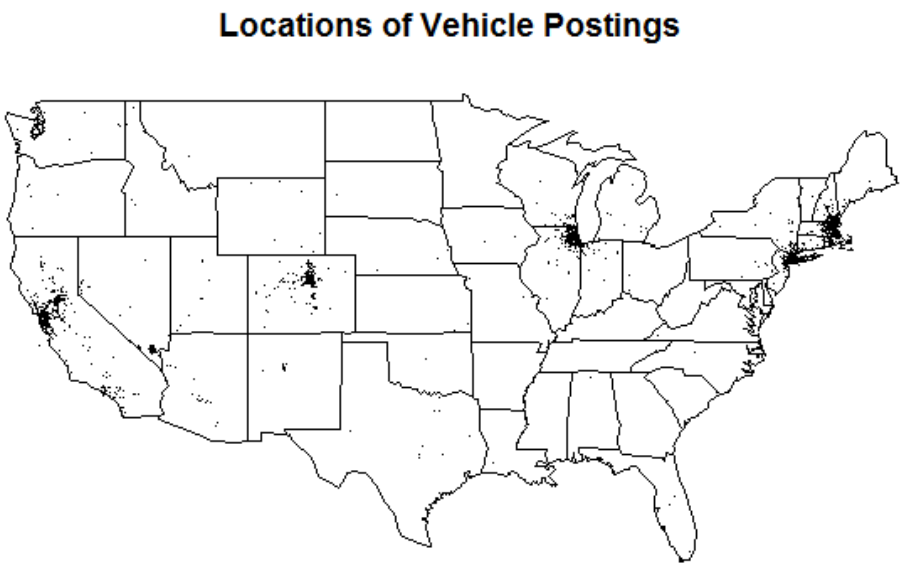
Question 10: Visually compare the distribution of the age of cars for different cities and for "sale by owner" and "sale by dealer". Provide an interpretation of the plots, i.e., what are the key conclusions and insights?



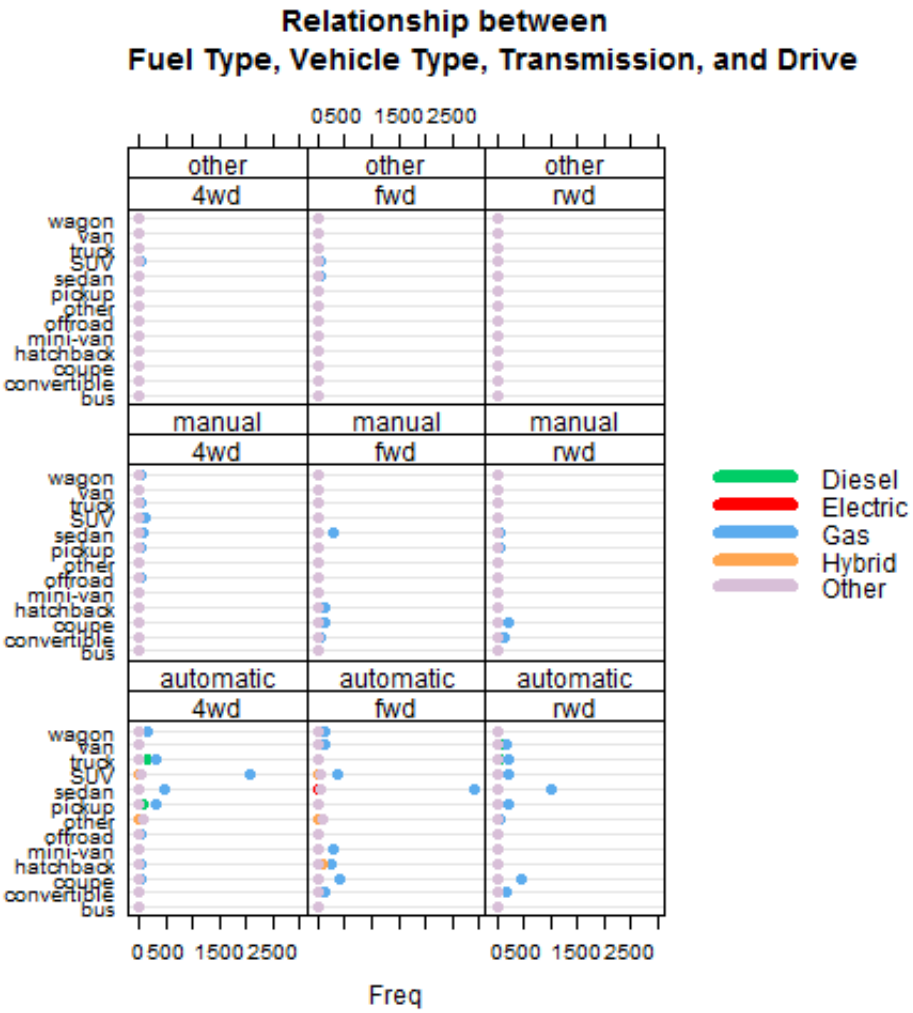
Looking at the plot on the right, cars sold by dealers are much younger than cars sold by owners. The distribution sold by owners is

skewed right while the distribution sold by dealers is more symmetric. This varies some by city, but the distributions are very similar from city to city.

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



Looking at the plot, there is a clear pattern. Most of the locations are in California, Colorado, Illinois, and along the east coast. Within the states, the postings are concentrated near big cities. It makes sense that people would want to sell their cars in big cities.

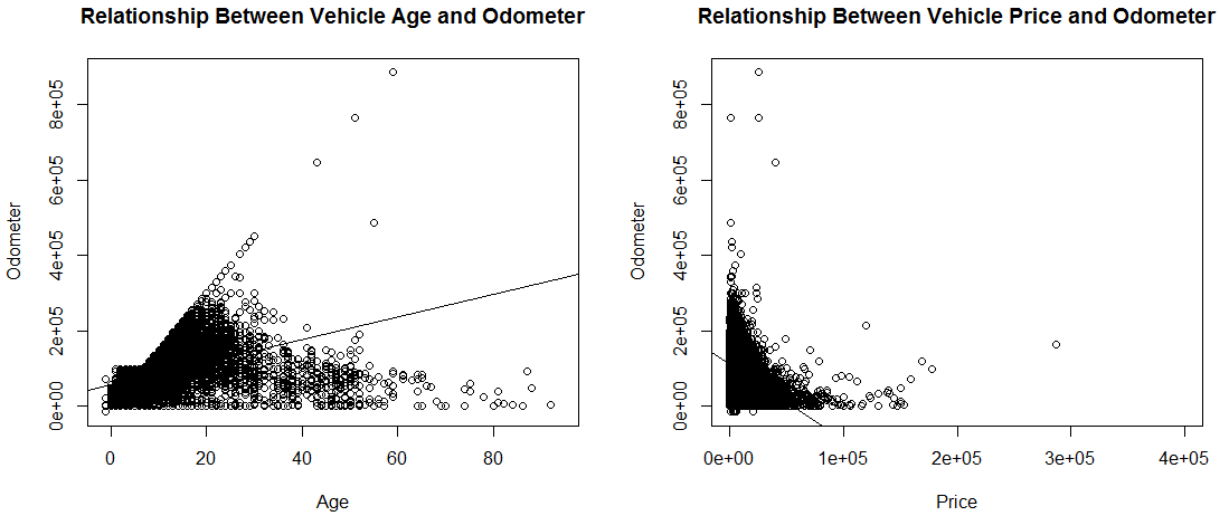


Question 12: Summarize the distribution of fuel type, drive, transmission, and vehicle type. Find a good way to display this information.

Looking at the plot to the left, the Vehicles with automatic transmission have the highest proportion, followed by manual and then other. In automatic, sedans with fwd are the most popular, followed by SUVs with 4wd and then sedans with rwd. Manual and other have more flat

distributions. Other is the most popular type of fuel, followed by gas, with the other three barely having a presence.

Question 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?



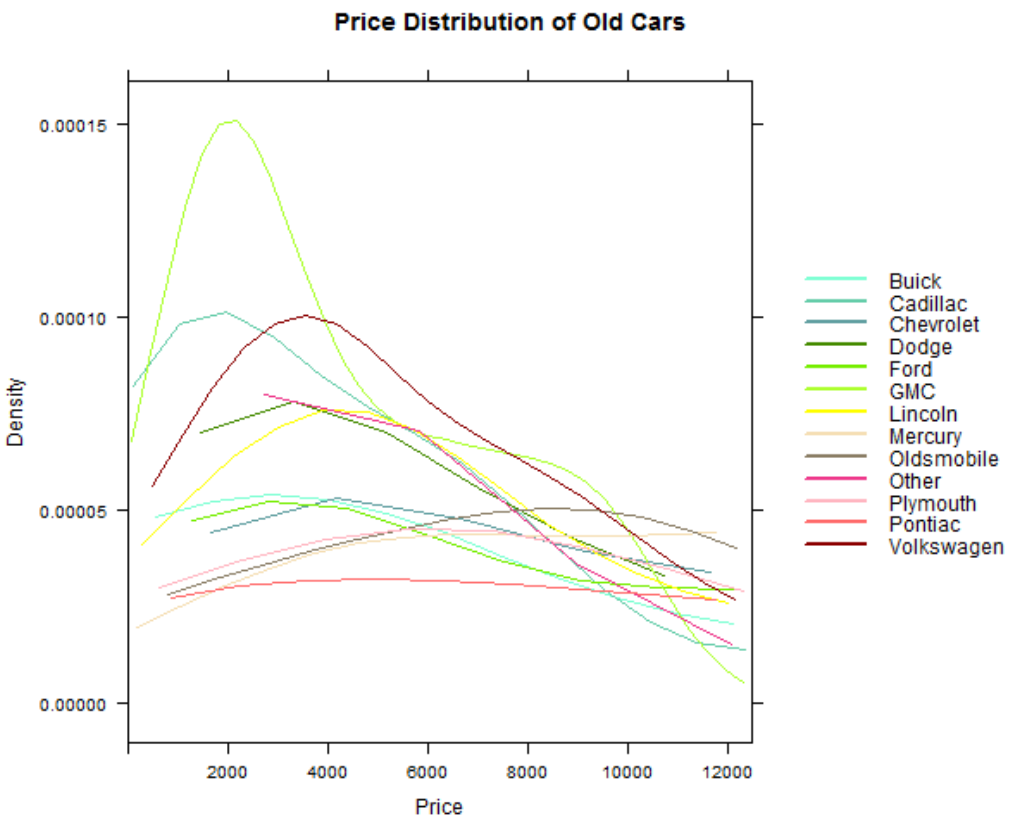
I started by plotting the original data which did not look right. Odometer had a bunch of very large outliers. Through researching online, I found out that a typical car is driven anywhere from 10,000 to 15,000 miles a year. Using this idea, I looped through each odometer, and if it was greater than 100,000 and greater than age * 15,000, I set that odometer to age * 15,000. This made the plots much easier to read.

Looking at the two plots above, there is a clear positive relationship between age and odometer. The line makes this clear. As one increases, so does the other. This makes sense because the older the vehicle is, the more miles it should have. Looking at the other plot for price and odometer, there is a negative relationship and most of the points are concentrated in the bottom left corner. As price increases, odometer goes down. This makes sense because people will charge more for a vehicle that has less miles on it.

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

I decided to classify old cars as cars older than 1980. I came to this conclusion by looking at various posts online and doing what felt right.

Looking at the plot below, the price distribution for most manufactures is fairly flat, and there are a few with more of a shape. GMC is skewed right. Volkswagen. Pontiac, Dodge, Lincoln and Cadillac are also skewed right, but a little less. The rest are very slightly skewed right, almost flat, but Oldsmobile is slightly skewed left. So many being skewed right tells us that for most



manufacturers of old cars, there are more low prices than high prices. As price increases, frequency decreases.

Question 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.

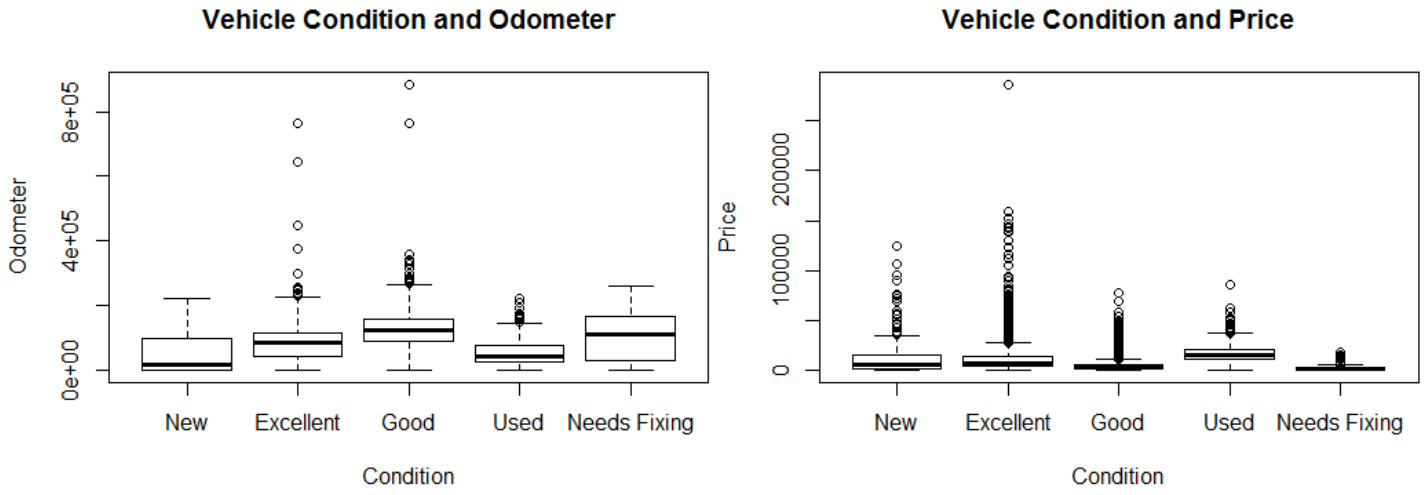
I think the omitted variable is engine size in liters. This cannot be derived from the other variables as far as I know. Looking through the body variable of vposts, I saw this pop up a lot. Based on seeing grep used a lot on Piazza, I would grep through each vposts$body looking for engine, liter, or a number followed by an L.

Question 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.

To deal with the 44 different conditions, I grouped them into five categories. The details of this can be found in the code appendix. For conditions that could not be put into any of the five groups, I looked at the rest of those rows to determine where they best fit. There was one row with a strange condition and upon further examination, the entire row looked empty, so I deleted that row.

I plotted two box plots that can be seen on the next page. The plot of condition and odometer almost has a bell shaped curve. The highest odometer is for vehicles in good condition, then vehicles that need fixing, then excellent, then used and then new. The plot of condition and price does not have much of a pattern. The condition with the highest mean price is used, followed by

excellent, new, good and needs fixing. This goes with the idea of paying more for a car in better condition.



Code Appendix

```
load("~/UC Davis/STA 141/HW 1/vehicles.rda")
# # # # # # #1:
nrow(vposts)
# # # # # # #2:
colnames(vposts)
print(sapply(vposts, class))
# # # # # # #3:
mean(vposts$price, na.rm = TRUE)
median(vposts$price, na.rm = TRUE)
quantile(vposts$price,c(.10,.20,.30,.40,.50,.60,.70,.80,.90,1.0), na.rm = TRUE) # Deciles
quantile(vposts$price,.25, na.rm = TRUE) # For outlier formula
q1 = 2995
quantile(vposts$price,.75, na.rm = TRUE)
q3 = 13500
iqr = q3 - q1
outLow = q1 - (1.5 * iqr)
outUpper = q3 + (1.5 * iqr)
noOutSubset = subset(vposts, (vposts$price > outLow) & (vposts$price < outUpper))
mean(noOutSubset$price, na.rm = TRUE) # New mean
```

```
median(noOutSubset$price, na.rm = TRUE) # New median
par(mfrow = c(1,2))
hist(vposts$price, main = "Original Vehicle Price", xlab = "Price", xlim = c(0,30000), ylim =
    c(0,6000))
points(quantile(vposts$price, prob = seq(0, 1, length = 11),na.rm = TRUE), rep(0,11)) #P iazza
points(mean(vposts$price, na.rm = TRUE), 0, col = "blue", pch = 8)
points(median(vposts$price, na.rm = TRUE), 0, col = "red", pch = 8)
leg.txt = c("= Mean", "= Median", "= Deciles")
legend(11000, 4000, legend = leg.txt, pch = c(8, 8, 1), col = c("blue", "red", "black"))
hist(noOutSubset$price, main = "Vehicle Price Excluding Outliers", xlab = "Price", xlim =
    c(0,30000), ylim = c(0,6000))
points(quantile(noOutSubset$price, prob=seq(0, 1, length = 11), na.rm = TRUE), rep(0, 11))
points(mean(noOutSubset$price, na.rm = TRUE), 0, col = "blue", pch = 8)
points(median(noOutSubset$price, na.rm = TRUE), 0, col = "red", pch = 8)
legend(11000, 4000, legend = leg.txt, pch = c(8, 8, 1), col = c("blue", "red", "black"))
# # # # # # #4:
type = table(vposts$type)
plot(type/length(vposts$type), main = "Vehicle Type Proportions", xlab = "Vehicle Type", ylab
= "Proportion", type = "h")
# # # # # # #5:
par(mar = c(3, 4.1, 3, 0.1)) # So labels don't overlap
autoTrans = subset(vposts, vposts$transmission == "automatic")
mosaicplot(autoTrans$fuel~autoTrans$type, las = 2, main = "Relationship Between Fuel and
    Vehicle Type of Cars with Automatic Transmission", ylab = "Vehicle Type", xlab =
    "Fuel")
manualTrans = subset(vposts, vposts$transmission == "manual")
mosaicplot(manualTrans$fuel~manualTrans$type, las = 2, main = "Relationship Between Fuel
    and Vehicle Type of Cars with Manual Transmission", ylab = "Vehicle Type", xlab =
    "Fuel")
otherTrans = subset(vposts, vposts$transmission == "other")
mosaicplot(otherTrans$fuel~otherTrans$type, las = 2,  main = "Relationship Between Fuel and
    Vehicle Type of Cars with Other Transmission", ylab = "Vehicle Type", xlab = "Fuel")
# # # # # # #6:
dim(table(vposts$city))
# # # # # # #7:
par(mar = c(4.1, 4.1, 4.1, 0.1))
par(xpd=TRUE)
barplot(table(vposts$byOwner, vposts$city), beside = TRUE, col = c("gray43", "gray87"), main
    = "Vehicles Sold by Dealer or Owner in Each City", names = c("Boston", "Chicago",
    "Denver", "Las Vegas", "NYC", "Sac", "SF Bay"), xlab = "City", ylab = "Count")
```

```
legend(-2.5, 2900, legend = c("Dealer", "Owner"), pch = 15, col = c("gray43", "gray87"))
# # # # # # #8:
maxPrice = print(max(vposts$price, na.rm = TRUE)) # 600030000
maxPriceRow = print(subset(vposts, vposts$price == maxPrice)) # Print entire row for reading
newPrice = (6000 + 30000) / 2 # Took mean of the given range for price: 6000 to 30000
maxLocation = which.max(vposts$price) # Row num of max
vposts[maxLocation,]$price = newPrice # Fixed

newMaxPrice = print(max(vposts$price, na.rm = TRUE)) # 30002500
newMaxPriceRow = print(subset(vposts, vposts$price == newMaxPrice))
newMaxLocation = which.max(vposts$price)
vposts[newMaxLocation,]$price = 10000 # Fixed - found similar car on craigslist

newMaxPrice2 = print(max(vposts$price, na.rm = TRUE)) # 9999999
newMaxPriceRow2 = print(subset(vposts, vposts$price == newMaxPrice2))
newMaxLocation2 = which.max(vposts$price)
vposts[newMaxLocation2,]$price = ((3495 + 2990 + 3695 + 2750 + 3500) / 5) # Fixed - found
        # similar cars online - mean price

newMaxPrice3 = print(max(vposts$price, na.rm = TRUE)) #569500
newMaxPriceRow3 = print(subset(vposts, vposts$price == newMaxPrice3))
newMaxLocation3 = which.max(vposts$price)
vposts[newMaxLocation3,]$price  = ((9383 + 6888 + 6991 + 5995 + 6900) / 5) # Fixed - Found
        # similar online

newMaxPrice4 = print(max(vposts$price, na.rm = TRUE)) #569500
newMaxPriceRow4 = print(subset(vposts, vposts$price == newMaxPrice4)) # Entered twice
newMaxLocation4 = which.max(vposts$price)
vposts = vposts[- newMaxLocation4, ] # Fixed - removed the duplicate row

newMaxPrice5 = print(max(vposts$price, na.rm = TRUE)) #569500
newMaxPriceRow5 = print(subset(vposts, vposts$price == newMaxPrice5))
newMaxLocation5 = which.max(vposts$price)
vposts[newMaxLocation5,]$price  = ((14995 + 15990 + 13995) / 3) # Fixed - Found similar
        # online, same seller as before

newMaxPrice6 = print(max(vposts$price, na.rm = TRUE)) #4e+05 = 400000
newMaxPriceRow6 = print(subset(vposts, vposts$price == newMaxPrice6)) #most online cost
        # 300000, so not unreasonable
# # # # # # #9:
```

```
byOwnSplit = split(vposts, vposts$byOwner) # Makes list
makeCityOwn = split(byOwnSplit$"TRUE"$maker, byOwnSplit$"TRUE"$city) # Piazza
makeCityDeal = split(byOwnSplit$"FALSE"$maker, byOwnSplit$"FALSE"$city)
lapply(makeCityOwn, function(m) head(sort(table(m), decreasing = TRUE), 3) ) # Piazza
lapply(makeCityDeal, function(m) head(sort(table(m), decreasing = TRUE), 3) )
# # # # # # #10:
# Looking for year = -1 and fixing:
vposts[ !is.na(vposts$year) & 2015 - vposts$year == -1, c("header", "odometer", "year",
        "maker", "title")] # Piazza
vposts[ !is.na(vposts$year) & 2015 - vposts$year == -1, c("year", "title")] # Piazza
vposts[vposts$header == "2016 2010 BMW X5 35d",]$year = 2010
vposts[vposts$header == "2016 1998 Ford Expedtion 1300",]$year = 1998
vposts[vposts$header == "2016 1978 Ford Mustang",]$year = 1978
vposts[vposts$header == "2016 1975",]$year = 1975
vposts[vposts$header == "2016 1991 ford bronco ii",]$year = 1991
vposts[vposts$header == "2016 Cobra mystic 1996",]$year = 1996
vposts[vposts$title == "Honda Civic Lx Coupe 2008 - $8000 (Everett)",]$year = 2008
vposts[vposts$title == "2002 Hyundai Elantra - $1400 (Tewksbury)",]$year = 2002
vposts[vposts$title == "NISSAN MAXIMA 1999 - $2000 (North shore area)",]$year = 1999
vposts[vposts$title == "1991 k5 blazer - $3200 (west bridgewater)",]$year = 1991
vposts[vposts$title == "2005 Hyundai Santa Fe - $4900 (chicago)",]$year = 2005
vposts[vposts$title == '2004 Nissan Quest SE "LEATHER" - $3000 (Grafton, Shrewsbury,
        Natick)',]$year = 2004
vposts[vposts$title == "01 LINCOLN TOWN CAR reduced!!!!! - $2500 (Walpole)",]$year =
        2001
vposts[vposts$title == "2012 Nissan Altima - $13500 (Tinley Park)",]$year = 2012
vposts[vposts$title == "selling 2002 cadillac escalade - $8500 (broadview)",]$year = 2002
vposts[vposts$title == "2012 F250 Super Duty FX4 Crew Cab Diesel. - $29500
        (Wakarusa)",]$year = 2012
vposts[vposts$title == "1962 chevy IMP. 2 DR HT. - $7500 (Oak lawn)",]$year = 1962
vposts[vposts$title == "2014 Buick Regal GS - $38500 (Oak Lawn)",]$year = 2014
vposts[vposts$title == "2000 saturn s1 - $1200 (Summit)",]$year = 2000
vposts[vposts$title == "2010 Chevy Camaro RS Super Clean!! Low Miles! - $19000
        (Johnstown)",]$year = 2010
vposts[vposts$title == "1990 HONDA CIVIC EX - $1600 (Thornton)",]$year = 1990
vposts[vposts$title == "1990 Chevy short box on 22s - $5000 (Denver)",]$year = 1990
vposts[vposts$title == "2010 Volkswagen Golf 2.5 - $14000 (denver)",]$year = 2010
vposts[vposts$title == "1997 Honda Civic LX - $2500 (western addition)",]$year = 1997
vposts[vposts$title == "2007 Dodge Ram 2500 with the 5.9L Cummins - $30 (treasure i
        sland)",]$year = 2007
```

```
vposts[vposts$title == "1996 chevy  2500 - $4000 (Byers)",]$year = 1996
vposts[vposts$title == "1996 Toyota Tacoma SR5 4WD - Cold AC - 119k miles - $8800
    (Denver)",]$year = 1996
vposts[vposts$title == "2000 Ford Mustang v6, 180,xxx miles - $2000 (Littleton,
    Colorado)",]$year = 2000
vposts[vposts$title == "1995 Chevy 2 door new engine and exhaust - $5400 (Arvada)",]$year =
    1995
vposts[vposts$title == "1973 el camino - $7000 (Roxborough)",]$year = 1973
vposts[vposts$title == "1990 1.8 swapped miata - $3400 (denver)",]$year = 1990
vposts[vposts$title == "2006 Lexus IS 250 AWD - $6500 (Aurora)",]$year = 2006
vposts[vposts$title == "1998 Toyota Camry xle. parts only - $1 (aurora)",]$year = 1998
vposts[vposts$title == "02 Expedition - $4500 (aurora)",]$year = 2002
vposts[vposts$title == "2000 Chevy Silverado - $2499 (Denver)",]$year = 2000
vposts[vposts$title == "2005 Acura tl - $4800 (Aurora)",]$year = 2005
vposts[vposts$title == "98 Chrysler Town and Country 7 passenger leather seats only 98k miles -
    $1800 (Denver)",]$year = 1998
vposts[vposts$title == "1987 - $5000 (Birmingham)",]$year = 1987
vposts[vposts$title == "2014 Kia Optima SX Limited Turbo!  Low miles! - $26000 (Castle
    Rock)",]$year = 2014
vposts[vposts$title == "2011 Chevrolet Shuttle /Party /Limo /Church Bus - $16950 (Oak Grove
    Missouri)",]$year = 2011
vposts[vposts$title == "2008 Toyota Tacoma PreRunner - $26000 (Las vegas)",]$year = 2008
vposts[vposts$title == "Mercedes Benz E320 1994 - $2000 (Northeast)",]$year = 1994
vposts[vposts$title == "White 2003 Chevy Silverad 3500 Duramax 4x4 6.6L Turbo Diesel A
    lisonT - $12900 (St. George)",]$year = 2003
vposts[vposts$title == "2011 Jeep Grand Cherokee rims with tires - $200 (Henderson)",]$year =
    2011
vposts[vposts$title == "2005 Dodge Magnum for sale - $3750 (Northeast)",]$year = 2005
vposts[vposts$title == "2003 Volkswagen Passat luxury - $3050 (Las Vegas)",]$year = 2003
vposts[vposts$title == "2015 Toyota Tundra monster truck 3k price drop - $43000",]$year =
    2015
vposts[vposts$title == "1999 forester - $4200 (Phoenix Arizona)",]$year = 1999
vposts[vposts$title == "1994 Honda accord LX automatic very nice - $2400 (last vegas)",]$year
    = 1994
vposts[vposts$title == "Brand New 2015/2016 Hyundais/Subarus for Sale (Cortlandt
    Manor)",]$year = 2015
vposts[vposts$title == "2015 NISSAN ALTIMA S - $12750 (BROOKLYN)",]$year = 2015
vposts[vposts$title == "2000 Nissan Maxima clean - $2700 (Bpt)",]$year = 2000
vposts[vposts$title == "2004 Toyota Sequoia - $5600 (Westchester County)",]$year = 2004
vposts[vposts$title == "2001 Pontiac am - $800 (Bpt)",]$year = 2001
```

```
vposts[vposts$title == "2002 subaru impreza  ts wagon - $2000 (Valley stream)",]$year = 2002
vposts[vposts$title == "1999 ford econoline custom - $2000 (nesconset)",]$year = 1999
vposts[vposts$title == "2000 CHEVY ASTRO 7 PASSENGER - $1800 (Midtown)",]$year =
    2000
vposts[vposts$title == "2009 Nissan Altima 128k miles Clean Title - $5700 (Rego Park,
    Queens)",]$year = 2009
vposts[vposts$title == "2001 maxima - $1700",]$year = 2001
vposts[vposts$title == "1988 Ford Mustang LX Notchback - $4500 (10308)",]$year = 1988
vposts[vposts$title == "2004 Mazda rx8 $5500 - $5500 (Hartsdale)",]$year = 2004
vposts[vposts$title == "1955 Chevy Belair 2dr Hardtop trades? - $10000 (Visalia)",]$year =
    1955
vposts[vposts$title == "2005 Audi A6 Quattro - $7900 (Citrus Heights)",]$year = 2005
vposts[vposts$title == "2007 335i twin turbo  low miles trade - $12000",]$year = 2007
vposts[vposts$title == '2004 Chevy Suburban Ls 1500 4WD "Clean Title" - $4500
    (Sacramento)',]$year = 2004
vposts[vposts$title == "1994 Chevy Silverado Z71 - $2300 (Stockton)",]$year = 1994
vposts[vposts$title == "2006 Acura TL - $8500 (Olivehurst)",]$year = 2006
vposts[vposts$title == "1999 Lexus RX300 CLEAN TITLE 116k ACTUAL MILES!!!! - $4800
    (Sacramento)",]$year = 1999
vposts[vposts$title == "1997 Honda Civic LX - $2500 (Modesto/Ceres)",]$year = 1997
vposts[vposts$title == "96 mustang gt,5 speed,4.6 - $2300 (Sacramento,antelope)",]$year = 1996
vposts[vposts$title == "MAZDA MPV LX 06 LOW MILES CLEAN TITLE/CLEAN CARFAX
    - $4800 (Sacramento)",]$year = 2006
vposts[vposts$title == "1964 nova 2door post - $3000",]$year = 1964
vposts[vposts$title == "***1931 MODEL A PROJECT*** - $1 (lake county)",]$year = 1931
vposts[vposts$title == "98 Honda civic DX - $3000 (Elk Grove)",]$year = 1998
vposts[vposts$title == "2006 Jeep Commander limited - $13999 (fremont / union city /
    newark)",]$year = 2006
vposts[vposts$title == "2006 Dodge Magnum $4700 - $1 (hayward / castro valley)",]$year =
    2006
vposts[vposts$title == "1999 ford ranger single cab - $3100 (dublin / pleasanton /
    livermore)",]$year = 1999
vposts[vposts$title == "1964 GMC 1500 3/4 Ton Pickup Truck - $1500 (fremont / union city / n
    ewark)",]$year = 1964
vposts[vposts$title == "95 PLYMOUTH VOYAGER - $1300 (fairfield / vacaville)",]$year =
    1995
vposts[vposts$title == "1969 mercedes 280s sedan classic - $2500 (santa rosa)",]$year = 2000
vposts[vposts$title == "03 Lexus es300 $3900 obo - $3900 (Sacramento)",]$year = 2003
vposts[vposts$title == "2011 Toyota Prius, Clean Tittle, 1 owner - $17500 (Fair oaks)",]$year =
    2011
```

```
vposts[vposts$title == "2001 6 speed Manual Cummins - $13 (Yolo County)",]$year = 2001
vposts[vposts$title == "2002 Toyota Avalon XL LOW MILAGE 35K - $7500
    (Sacramento)",]$year = 2002
vposts[vposts$title == "1998 Mitsubishi Galant mechanics special - $800 (Sacramento)",]$year =
    1998
vposts[vposts$title == "96 Mercedes Benz c280 $2000 FIRM - $2000",]$year = 2000
vposts[vposts$title == "Audi 2001 A6 - $2900 (Sacramento)",]$year = 2001
vposts[vposts$title == "2003 Toyota Corolla Needs work - $1800 (Roseville)",]$year = 2003
vposts[vposts$title == "1999 Nissan Pathfinder 4wd cold a/c daily driver - $2100
    (SAC.LOOMIS.LINCOLN.AUBURN.YUBA/SUTTER)",]$year = 1999
vposts[vposts$title == "2001 VW Jetta 120000 miles bad water pump - $1100 (San Jose
    Downtown)",]$year = 2001
vposts[vposts$title == "2006 Nissan Murano S - $6500 (Sacramento)",]$year = 2006
vposts[vposts$title == "2006 Harley sportster 883 - $4400 (santa rosa)",]$year = 2006
vposts[vposts$title == "1982 Chevy El camino - $4500 (santa rosa)",]$year = 1982
vposts[vposts$title == "1999 Ford Windstar Van - $1200 (mill valley)",]$year = 1999
vposts[vposts$title == "2006 Jeep Commander limited - $13999 (san mateo)",]$year = 2006
# Sketchy sounding titles
print(vposts[vposts$header == "2016 Variety",]) # Selling many cars on one post: delete
vposts = vposts[- (which(vposts$header == "2016 Variety")), ]
print(vposts[vposts$header == "2016 New",]) # Window cleaning: delete
vposts = vposts[- (which(vposts$header == "2016 New")), ]
print(vposts[vposts$header == "2016 2015",]) # Selling rims: delete (four)
vposts = vposts[- (which(vposts$header == "2016 2015")), ]
print(vposts[vposts$header == "2016 All years makes and models",]) # Want you to sell them
your car: delete (3)
vposts = vposts[- (which(vposts$header == "2016 All years makes and models")), ]
print(vposts[vposts$header == "2016 ALL CARS TRUCKS SUVS",]) # Ad for place that sells
    cars: delete (2)
vposts = vposts[- (which(vposts$header == "2016 ALL CARS TRUCKS SUVS")), ]
print(vposts[vposts$header == "2016 LOW DOWN PAYMENTS",]) # Ad: delete
vposts = vposts[- (which(vposts$header == "2016 LOW DOWN PAYMENTS")), ]
print(vposts[vposts$header == "2016 Caddilac",]) # Selling 44 cars: delete
vposts = vposts[- (which(vposts$header == "2016 Caddilac")), ]
print(vposts[vposts$header == "2016 Dodge, ford, Chevy",]) # Not a car: delete
vposts = vposts[- (which(vposts$header == "2016 Dodge, ford, Chevy")), ]
print(vposts[vposts$header == "2016 any",]) # delete
vposts = vposts[- (which(vposts$header == "2016 any")), ]
print(vposts[vposts$header == "2016 All",]) # delete
vposts = vposts[- (which(vposts$header == "2016 All")), ]
```

```
print(vposts[vposts$header == "2016 rr",]) # delete
vposts = vposts[- (which(vposts$header == "2016 rr")), ]
print(vposts[vposts$header == "2016 Acura ILX w/Technology Plus Pkg",]) # delete
vposts = vposts[- (which(vposts$header == "2016 Acura ILX w/Technology Plus Pkg")), ]
print(vposts[vposts$header == "2016 Acura ILX w/Premium A-SPEC Package",]) # delete
vposts = vposts[- (which(vposts$header == "2016 Acura ILX w/Premium A-SPEC Package")), ]
print(vposts[vposts$title == "1972 VW Super Beetle - $4000",]) # delte 2, change one's year to
    1972
locs = which(vposts$title == "1972 VW Super Beetle - $4000")
vposts = vposts[- locs[1:2], ]
vposts[vposts$title == "1972 VW Super Beetle - $4000",]$year = 1972
# Looking into the rest of the 2016 cars:
vposts[ !is.na(vposts$year) & 2015 - vposts$year == -1, c("header", "odometer", "year",
    "maker", "title")] # Piazza
vposts[ !is.na(vposts$year) & 2015 - vposts$year == -1, c("header", "description")] # Piazza
# The rest seem correct
# Looking at cars older than 100 years:
vposts[ !is.na(vposts$year) & 2015 - vposts$year > 99, c("header", "description", "title")] #
    Piazza
vposts[vposts$title == "argolic eni-04 JEeP wraNgler Clean lEATHeR - $2532 (chicago)",]$year
    = 2004
print(vposts[vposts$header == "1900 CAR",]) # delete
vposts = vposts[- (which(vposts$header == "1900 CAR")), ]
print(vposts[vposts$header == "1900 Wheels",]) # Wheels: Delete
vposts = vposts[- (which(vposts$header == "1900 Wheels")), ]
vposts[ !is.na(vposts$year) & 2015 - vposts$year > 80, c("header", "description", "title")] #
    Piazza
# Seems legit to me
# Looking at cars newer than 2015:
vposts[ !is.na(vposts$year) & 2015 - vposts$year < -1, c("title")] # Piazza
# One car: 2022 Honda Odyssey
print(vposts[vposts$title == "Check Out This Spotless 2022 Honda Odyssey with 117,102 Miles
    - $6999 (Jamaica)",])
# Found website, compared picture to different years of that car
vposts[vposts$title == "Check Out This Spotless 2022 Honda Odyssey with 117,102 Miles -
    $6999 (Jamaica)",]$year = 2015
# Data now ready to be plotted!
vposts$age = 2015 - vposts$year
install.packages("ggplot2")
library("ggplot2")
```

```
dealPlot = ggplot(vposts, aes(x = age, fill=byOwner)) + geom_histogram(binwidth = 3, alpha=.5,
        position="identity") + facet_wrap(~city, scales = "free") + scale_x_continuous(limit =
        c(0,70)) # Piazza
dealPlot + ggtitle("The Distribution of the Age of Cars Sold by City") + ylim(0, 700) +
        xlab("Age") + ylab("Count")
# # # # # # #11:
install.packages("maps")
par(mfrow = c(1,1))
map('state')
points(vposts$long, vposts$lat, pch = '.')
title(main = "Locations of Vehicle Postings")
# # # # # # #12:
install.packages("lattice")
library("lattice")
par(xpd=NA)
my.settings = list(strip.background=list(col = c("white", "white")), border="transparent")
dotplot(table(vposts$type, vposts$drive, vposts$transmission, vposts$fuel), pch = 19, col =
        c("springgreen3", "red", "steelblue2", "tan1", "thistle"), main = "Relationship between
        Fuel Type, Vehicle Type, Transmission, and Drive", par.settings = my.settings,
        key = list(space="right",lines =  list(col=c("springgreen3", "red", "steelblue2", "tan1",
        "thistle"), lty = c(1,5), lwd = 6), text = list(c("Diesel", "Electric", "Gas", "Hybrid",
        "Other"))))
## # # # # # #13:
plot(vposts$age,vposts$odometer) # Looks wrong
# Dealing with outliers:
# Found this online: # A typical car is driven anywhere from 10,000 to 15,000 miles a year.
# A ten-year old car should have roughly 100,000 to 150,000 miles on it
# Using this idea, I decided to go thought each odometer, and if it is greater than 100000
# and greater than age * 15000, set the odometer to age * 15000
vposts$odoNeedsChanging = (vposts$odometer > (15000 * vposts$age)) & (vposts$odometer >
        100000)
vposts$correctOdo = vposts$age * 15000
for ( i in 1:nrow(vposts))
{
        if((!is.na(vposts[i, ]$odometer)) & (!is.na(vposts[i, ]$odoNeedsChanging)) & (vposts[i,
        ]$odoNeedsChanging == TRUE))
        {
                vposts[i, ]$odometer = vposts[i, ]$correctOdo
         }
```

```
} # Slow, but the apply functions were giving me troubles, fast enough
# Plotting:
par(mfrow = c(1,2))
plot(vposts$age,vposts$odometer, main = "Relationship Between Vehicle Age and Odometer",
        xlab = "Age", ylab = "Odometer")
abline((lm(vposts$odometer~vposts$age)))
plot(vposts$price,vposts$odometer, main = "Relationship Between Vehicle Price and
        Odometer", xlab = "Price", ylab = "Odometer")
abline((lm(vposts$odometer~vposts$price)))
# # # # # # #14:
oldCars = subset(vposts, vposts$year < "1980") #Define old?
table(oldCars$maker)
# Grouping (cars with less than 20 were put in "other")
for ( i in 1:nrow(oldCars))
{
  if((!is.na(oldCars[i, ]$maker)) & ((oldCars[i, ]$maker == "alfa romeo") | (oldCars[i, ]$maker
        == "bentley") | (oldCars[i, ]$maker == "bmw") | (oldCars[i, ]$maker == "bugatti") |
        (oldCars[i, ]$maker == "desoto") | (oldCars[i, ]$maker == "eagle") | (oldCars[i, ]$maker
        == "geo") | (oldCars[i, ]$maker == "harley davidson") | (oldCars[i, ]$maker == "honda") |
        (oldCars[i, ]$maker == "hudson") | (oldCars[i, ]$maker == "jaguar") | (oldCars[i,
        ]$maker == "maserati") | (oldCars[i, ]$maker == "mazda") | (oldCars[i, ]$maker ==
        "mitsubishi") | (oldCars[i, ]$maker == "nissan") | (oldCars[i, ]$maker == "volvo") |
        (oldCars[i, ]$maker == "amc") | (oldCars[i, ]$maker == "fiat") | (oldCars[i, ]$maker ==
        "rolls royce") | (oldCars[i, ]$maker == "shelby") | (oldCars[i, ]$maker == "studebaker") |
        (oldCars[i, ]$maker == "triumph") | (oldCars[i, ]$maker == "chrysler") | (oldCars[i,
        ]$maker == "datsun") | (oldCars[i, ]$maker == "international") | (oldCars[i, ]$maker ==
        "jeep") | (oldCars[i, ]$maker == "mercedes") | (oldCars[i, ]$maker == "mg") | (oldCars[i,
        ]$maker == "porsche") | (oldCars[i, ]$maker == "toyota") | (oldCars[i, ]$maker == "
        willys")))
        {
                oldCars[i, ]$maker = "other"
        }
} # Slow, but the apply functions were giving me troubles, fast enough
library("lattice")
par(xpd=TRUE)
myCol = c("aquamarine", "aquamarine3", "cadetblue", "chartreuse4", "chartreuse2",
        "greenyellow", "yellow1", "wheat", "wheat4", "violetred2", "lightpink", "indianred1",
        "darkred")
densityplot(~ price, oldCars, group = maker, plot.points = FALSE, col.line = myCol, xlab =
        "Price", subset = !is.na(price), xlim=c(0,1.25e4), main = "Price Distribution of Old Cars",
```

```
        key=list(space="right",lines = list(col = myCol),lty=c(1,13), lwd=2,text=list(c("Buick",
        "Cadillac", "Chevrolet", "Dodge", "Ford", "GMC", "Lincoln", "Mercury", "Oldsmobile",
        "Other", "Plymouth", "Pontiac", "Volkswagen"))))
# # # # # # #15:
head(vposts[1000:1020,]$body)
# Going with engine size in liters
# # # # # # #16:
unique(vposts$condition) # 44 levels - need to group
table(vposts$condition)
# Splitting into five groups:
newSubset = subset(vposts, vposts$condition == "new")
excelSubset = subset(vposts, (vposts$condition == "excellent") | (vposts$condition == "superb
        original") | (vposts$condition == "very good") | (vposts$condition == "certified") |
        (vposts$condition == "like new") )
goodSubset = subset(vposts, (vposts$condition == "good") | (vposts$condition == "nice") |
        (vposts$condition == "nice teuck") | (vposts$condition =="ac/heater") | (vposts$condition
        == "fair"))
usedSubset = subset(vposts, (vposts$condition == "used") | (vposts$condition == "0used") |
        (vposts$condition == "pre-owned") | (vposts$condition == "pre owned") |
        (vposts$condition == "preowned") | (vposts$condition == "preowns") | (vposts$condition
        == "carfax guarantee") | (vposts$condition == "mint") | (vposts$condition == "honnda"))
needsFixingSubset = subset(vposts, (vposts$condition == "salvage") | (vposts$condition ==
        "complete parts car, blown engine") | (vposts$condition == "front side damage") |
        (vposts$condition == "hit and run :( gently") | (vposts$condition == "needs work/for
        parts") | (vposts$condition == "not running") | (vposts$condition == "needs total restore")
        | (vposts$condition == "needs bodywork") | (vposts$condition == "needs restoration") |
        (vposts$condition == "needs restored") | (vposts$condition == "needs work")|
        (vposts$condition == "parts")| (vposts$condition == "project car") | (vposts$condition ==
        "rebuildable project") | (vposts$condition == "restoration") | (vposts$condition ==
        "restoration project") | (vposts$condition == "project")| (vposts$condition == "nice
        rolling restoration") | (vposts$condition == "restore") | (vposts$condition == "restored") |
        (vposts$condition == "rough but runs") | (vposts$condition == "muscle car restore"))
# Plotting:
par(mfrow = c(1,2))
boxplot(newSubset$odometer, excelSubset$odometer, goodSubset$odometer,
        usedSubset$odometer, needsFixingSubset$odometer, main = "Vehicle Condition and
        Odometer", xlab = "Condition", ylab = "Odometer",
        names=c("New","Excellent","Good", "Used", "Needs Fixing"))
```

```
boxplot(newSubset$price, excelSubset$price, goodSubset$price, usedSubset$price,
        needsFixingSubset$price, main = "Vehicle Condition and Price", xlab = "Condition",
        ylab = "Price", names=c("New","Excellent","Good", "Used", "Needs Fixing"))
```

References

1. Used piazza for almost every question either for specifics, clarification or ideas

2. Referred to R assignments in past classes (STA 32 and 108) for some functions and reminders of how to do things

3. Used the ?help feature of R

4. Urls found with Google:

- http://www.jmp.com/support/help/Mosaic_Plot.shtml
- http://www.vicc.org/biostatistics/LuncheonTalks/DrTsai2.pdf
- http://www.cargurus.com/
- https://www.carsforsale.com/
- http://classiccars.com/
- https://mathematicaforprediction.wordpress.com/2014/03/17/mosaic-plots-for-data-visualization/
- https://onlinecourses.science.psu.edu/stat100/book/export/html/20
- http://stackoverflow.com/questions/3932038/plot-a-legend-outside-of-the-plotting-area-in-base-graphics
- http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf
- http://stackoverflow.com/questions/13449233/send-a-text-string-containing-double-quotes-to-function
- https://www.google.com/search?q=Honda+Odyssey&rlz=1C1CAFA_enUS630US630&es_sm=93&biw=1093&bih=534&source=lnms&tbm=isch&sa=X&ved=0CAYQ_AUoAWoVChMIh9DY7N2pyAIVUSuICh1g5gCv#imgrc=diBpBMG883Ys7M%3A
- http://www.autos.com/car-buying/odometer-reading-basic-when-buying-used-cars
- http://www.cookbook-r.com/Graphs/index.html
- http://stackoverflow.com/questions/25109196/r-lattice-package-add-legend-to-a-figure
- http://www.magesblog.com/2012/12/changing-colours-and-legends-in-lattice.html
- http://msemac.redwoods.edu/~darnold/math15/spring2013/R/Activities/BoxplotsII.html