Tiffany Chen ID: 998840686 10/9/15 STA 141

Assignment 1 Part 2

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

- Unusually high prices

There were several hundred prices that were unusually high for no good reason. For each observation, this is due to various reasons. For example, the maximum value in the data set is 600030000 which is posted twice by the same user and it is meant to be between $6,000 and $30,000 so I would correct this observation by replacing the current price with the mean of the two prices. This procedure differed between observations. The next highest observation, had price 30002500 so I would look at the observation and Google the specific car to see what a reasonable price would be. I believe this example also was a range of prices and found the median again. The next observation had price 9999999 so I read the post and they meant to sell it as $20 obo (or best offer) so I changed the value to 20. I continued this procedure for a few more prices. I think some prices were also typos in which they accidentally entered an extra digit. But, some cars are actually appropriately priced correctly because it actually is that expensive such as the 2006 Fort GT. I corrected a few observations but there are many more. I could identify them by subsetting the data for cars priced 50,000 or higher as that is what the average price of a car is about. Then, I would check each observation one by one for accuracy. I could also use grepl to search for if the "body" vector has a $ sign which could indicate a possible price. But, I believe it's best to go through the prices individually and determine on a case by case basis what to do. I think in analyzing the data with these corrections will have a positive impact in that it will reflect the true prices and will be more accurate for data analysis. The summary values, such as mean and plotting the values will make more sense and be more accurate.

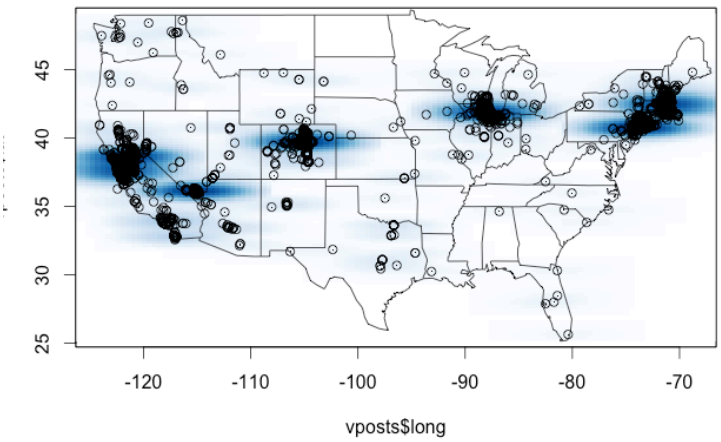
- Unusually high odometer readings

Similar to unusually high prices, there were also unusually high odometer readings that seemed unreasonable given that cars usually have a maximum odometer reading to about 6 digits. I looked through the high odometer readings and once again how I corrected it depended on the observation. For example, odometer reading 1234567890 was clearly wrong and the description had many typos so I would discard this observation. For most of the rest, if I could not determine from the description why they put the odometer value as they did, then I would change it to at most 999999 or simply delete it or delete a digit from their current one, assuming they made a typo. If I make these changes, I think similar to the prices; the values would make more sense and be more accurate. There won't be as much outliers or extreme values.

- Unusual years

There were a few odd years in the data set such as car with year 4 or 2022. But that was just from looking at just the years column in the data set. When looking through the whole observation, there would be a different year of the car. I would correct those cases by replacing the year with the correct one given in the headers. Depending on other cases, it was similar to the above anomalies in that there was a pattern. Some years were repeated many times because it was the same post. The impact on the data is removing the outliers and making the data more accurate.

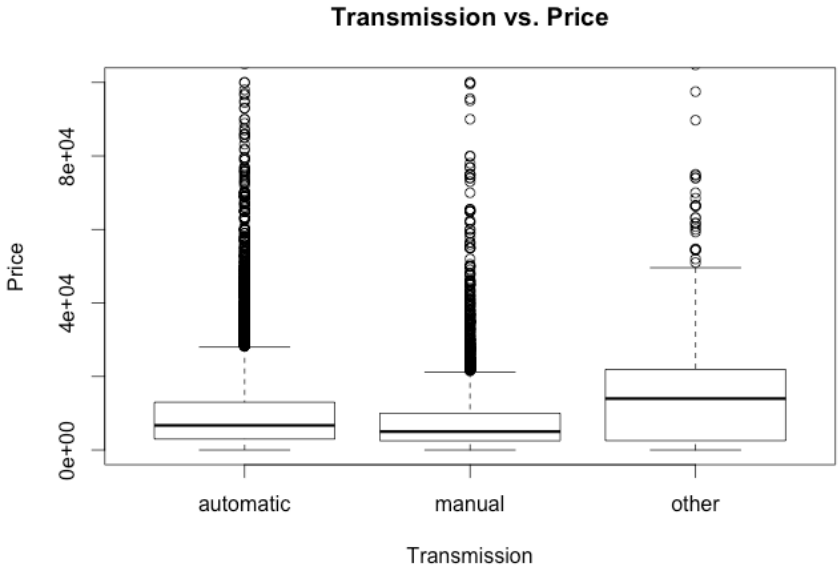
- Location in the ocean




After plotting the longitude and latitudes of the posts, one post was off the coast of California. It seems interesting that a car would be in the ocean. After looking at the specific observation, there just does not seem like any reasonable explanation. According to a post from Piazza, someone suggested that the user may have posted this on the plane, or some special circumstance happened. If not, I would delete this observation. Since this is just one observation, I don't think it would affect the overall data too much for analysis.

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

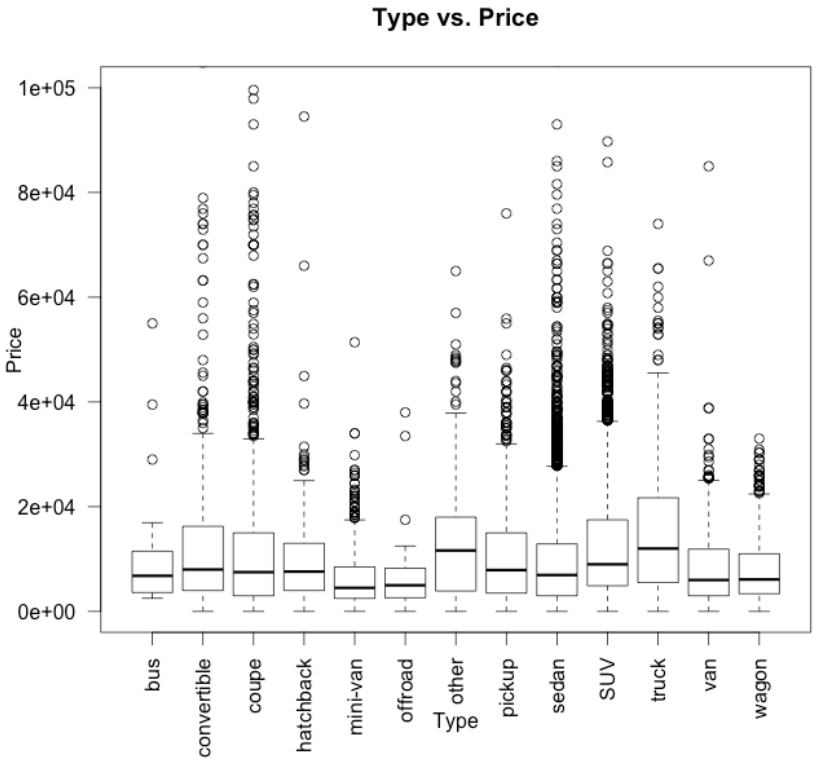
- Transmission versus price of car




Looking at a boxplot of transmission versus prices of the vehicles, "other" transmissions are the most pricey, then automatic and manual. In particular, the relationship between automatics and manuals is of most interest to buyers. In general, automatic vehicles are slightly more price than manual vehicles. This is very helpful to buyers to know which transmission is appropriately priced. I think this is fairly generalizable to other vehicle sales data according to (http://www.investopedia.com/university/newcar/what-to-look-for.asp) which states that "a manual transmission…likely to buy a sports car or a stripped down budget-model sedan or compact". I tried to plot transmission, type, and price to validate that statement but I could not fix the axis of the prices and it was hard to determine a difference. It would have been

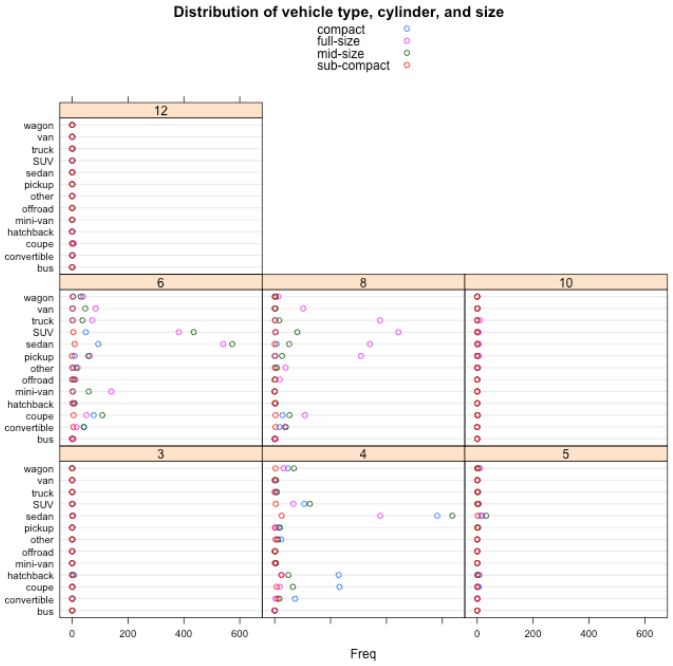
interesting to see if specific type of cars and transmission combos have different price points for just manual. Therefore, I displayed just the type and price in the next interesting insight.

- Type of car and price



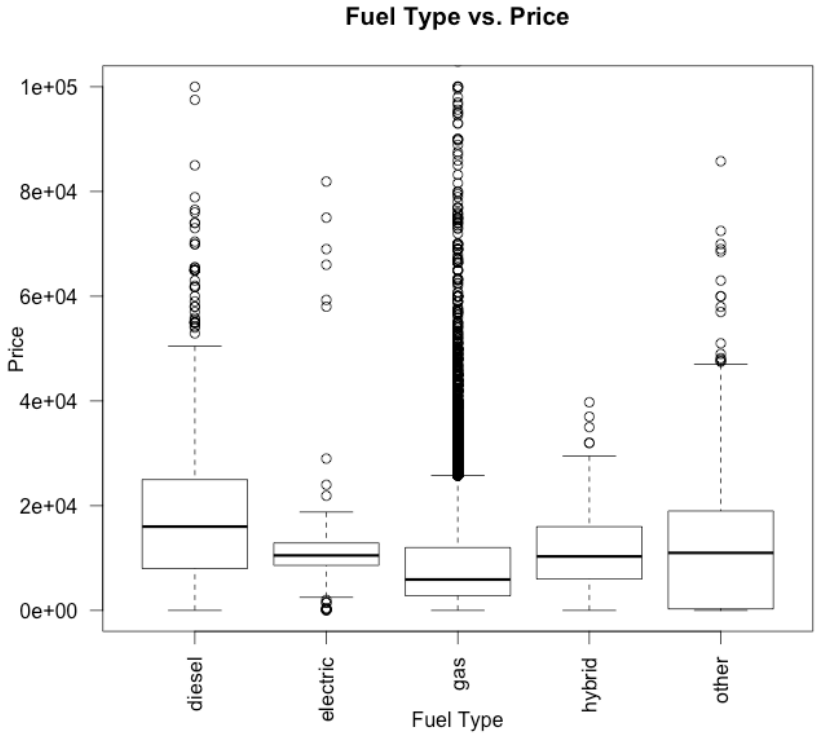

This is a boxplot of the type of cars and the prices. This is most interesting for a buyer so they can determine what type of car is most reasonable price relative to other types. This would be useful to know from this website or wherever the data is from, what is the general price range of the type of cars available. Based on this plot, “other” vehicles are the priciest. However, “other” vehicles are not just vehicles, which makes the conclusion not as accurate or generalizable. The distributions have many outliers and are spread out because the price ranges widely for the types. If we look at the next highest priced type of vehicle is truck and SUV. This probably isn’t very generalizable because of the many outliers and how inaccurate the data is.

- Distribution of vehicle type, cylinder, and size



I thought this would be interesting to explore the relationship between the vehicle type, cylinder, and size of car for the buyer. This way, the buyer knows the general picture of what kind of vehicles are being sold on this dataset. This would help if the buyer is looking for a specific car, perhaps this can show whether or not this online database would be a good site to buy it from. The most common vehicles seem to be 4, 6, and 8 cylinder vehicles, in particular sedans, SUVs, trucks, and pickups. For 8 cylinder vehicles, full-size vehicles truck, SUV, sedan, and pickups are most frequent. For 6 cylinder, most frequent are mid size and full sized SUV and sedans. For 4 cylinder vehicles, most frequently sold are compact, mid size, and full size sedans and compact hatchbacks and coupes. This is fairly generalizable to other vehicle sales data in terms of cylinder because 4 cylinder vehicles are most commonly sold. However, I am not sure how generalizable the specific type and size of car is because I couldn't find a link that displayed the top used cars sold on sites like Craigslist. There is, according to a site that has America's top 10 most popular used cars, and it seems most are mid size compact cars that are sedan. (http://www.cheatsheet.com/automobiles/americas-10-most-popular-used-cars.html/?a=viewall)

- Alternative fuel cars and price




This is a display of the fuel types and price. The priciest vehicle has fuel type diesel. This would be interesting to the buyer to know what kind of fuel type they should buy for the price. I don't think this is too generalizable because I would suggest looking at a case-by-case basis since many observations probably are listed incorrectly and there are many factors that effect the prices and fuel type. For example, an older hybrid car may be more expensive than a gas car and this plot doesn't show the age. I tried to plot age, fuel type, and price but was not successful.

Code Appendix

```
# Assignment 1 Part 2
# Tiffany Chen ID: 998840686

vehicles = load("/Users/tiffanychen/Desktop/STA 141/vehicles.rda")

# ANOMALY 1: large prices
## 1 HIGH PRICE
```

```
p = vposts$price
tail(sort(vposts$price), 50)
max(p,na.rm=TRUE)
ids = which(vposts$price == 600030000) # there are 2
ids
vposts[ids, ] # this posting for 600030000 is posted twice. Cost is supposed to be between 6000 & 30,000  sold byOwner
vposts[ids, "price"] = 18000 # let's change to the mean

ids = which(vposts$price == 30002500)
vposts[ids,] #2002 Cadillac Seville 8 cyl gas sold byOwner
vposts[ids, "price"] = 2750
# pretty sure this is a range of prices too

ids = which(vposts$price == 9999999)
vposts[ids, ] # just delete this one. "$20 obo. comes with complimentary Oboe. the description makes no sense. new used car, good bad condition. location is everywhere." contradictory information
vposts[-13937, ] # delete this one

ids = which(vposts$price == 569500)
vposts[ids, ] # post is byDealer, they have their own site, so credible source. the price doesnt seem right though. I went on their website www.lot1autosales.com and found the exact car on sale for $6,995
# http://www.lot1autosales.com/2007_Chevrolet_Monte%20Carlo_Melrose%20Park_IL_258261321.veh
vposts[ids,"price"] = 56950

ids = which(vposts$price == 400000)
vposts[ids, ]
# http://newyork.craigslist.org/wch/ctd/5215236700.html
# reasonable according to google
# 2006 Ford GT Canada
# http://www.autotrader.ca/cars/ford/gt/

ids = which(vposts$price == 359000)
vposts[ids, ]
vposts[ids, "price"] = 35900 # dropped a digit
# Chevrolet 2010
# http://www.cargurus.com/Cars/l-Used-2010-Chevrolet-Silverado-1500-LTZ-ev41
# i think theres a typo, extra 0

## 2 HIGH ODOMETER
tail(sort(vposts$odometer), 50)

od = which(vposts$odometer == 1234567890)
vposts[od, ]
vposts[-18161, ]
# delete this one. the odometer is clearly wrong, so many typos in the post.

od = which(vposts$odometer == 99999999)
vposts[od, ]
```

```
vposts[-4530, ]
# this owner also clearly typed a random odometer for his 1988 jeep comanche

od = which(vposts$odometer == 16000000)
vposts[od, ]
vposts[-19227, ]
vposts[-19537, ]
# posted twice, the maximum odometer is about 6 digits 999999
# if this car is like new, then the odometer reading is wrong
od = which(vposts$odometer == 9500000)
vposts[od, ]

od = which(vposts$odometer == 3000000)
vposts[od, ]

od = which(vposts$odometer == 2800000)
vposts[od, ]
## i dont get why its high. makes no sense.
# i can at most change it to 999999 or delete a digit.

vposts$price == 1 & grepl('\\$', vposts$body) # piazza
grepl('$', c('This has $ in it', 'This does not'))


## 3
head(sort(vposts$year), 50)
# there is a year 2022?
yr = which(vposts$year == 2022)
vposts[yr, ]
## based on the picture i think its 2012
yr = which(vposts$year == 1900) # delete. it was posted 7 times and has the same title CAR WON'T PASS
SMOG??WE'LL BUY TODAY!! - $750
vposts[yr, ]

yr = which(vposts$year == 1921) # ok
vposts[yr, ]

yr = which(vposts$year == 1922) # ok
vposts[yr, ]

yr = which(vposts$year == 1923) # ok
vposts[yr, ]

head(sort(vposts$year), 50)
# there is a year 4?
yr = which(vposts$year == 4)
vposts[yr, ]
## deleting this has too many typos. the link on craigslist, this post has been removed
# https://newyork.craigslist.org/mnh/cto/5233079816.html
```

```
# 4 location in the ocean
library(maps)
map('state')
smoothScatter(vposts$long, vposts$lat)
map('state', add = T)
points(vposts$long, vposts$lat, add = T)
locator(1)
identify(vposts$long, vposts$lat)
vposts[34388,]

## INSIGHT 1: transmission and price
plot(vposts$transmission, vposts$price, ylim = c(0, 100000), main = "Transmission vs. Price", xlab =
"Transmission", ylab = "Price")

# INSIGHT 2: type of car and price
plot(vposts$type, vposts$price, ylim = c(0, 100000), main = "Type vs. Price", xlab = "Type", ylab = "Price", las =
2)

head(vposts[other, ])
other = !is.na(vposts$transmission) & vposts$transmission == "other"

price_other = vposts[other, "price"]
tail(sort(price_other))

# INSIGHT 3: type, cylinder, size
x=table(vposts$type, vposts$cylinder, vposts$size)
library(lattice)
dotplot(x, besides=TRUE, breaks=seq(0,3000,10), main="Distribution of vehicle type, cylinder, and size",
auto.key=TRUE)

# alternative fuel cars vs. regular price
plot(vposts$fuel, vposts$price, ylim = c(0, 100000), main = "Fuel Type vs. Price", xlab = "Fuel Type", ylab =
"Price", las = 2)

# alternative fuel cars vs. regular price vs. age
vposts$age = 2015 - vposts$year # many negative values, so got to clean it
betterage = vposts$age[vposts$age>=-1 & vposts$age<=115]
table(vposts$age, vposts$fuel, vposts$price)

# this doesn't look good?
# i tried changing the axis??
library(lattice)
histogram( ~ vposts$price | vposts$transmission + vposts$type, vposts, main="Distribution of Price by
Transmission and type", type = "percent", xlab="Price of Car")
```