Assignment 1 Ⅱ

Xinyi Hou

ID： 913411456

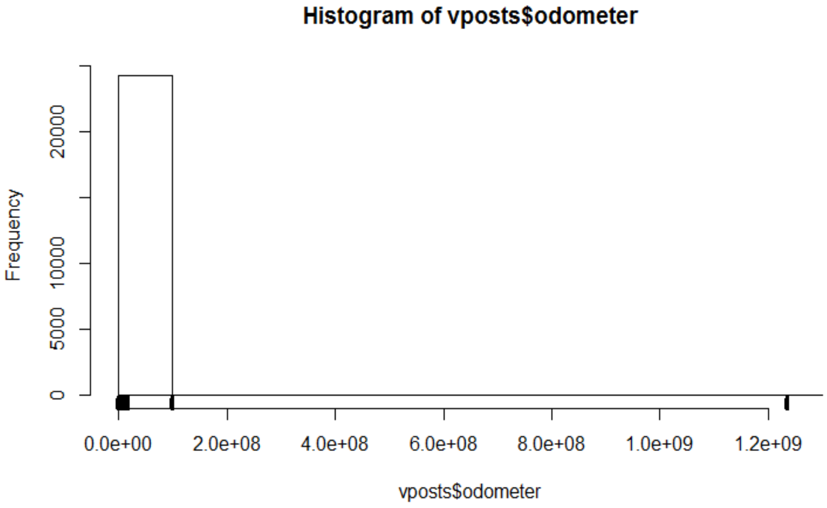
Oct,8th,2015

Load the data:

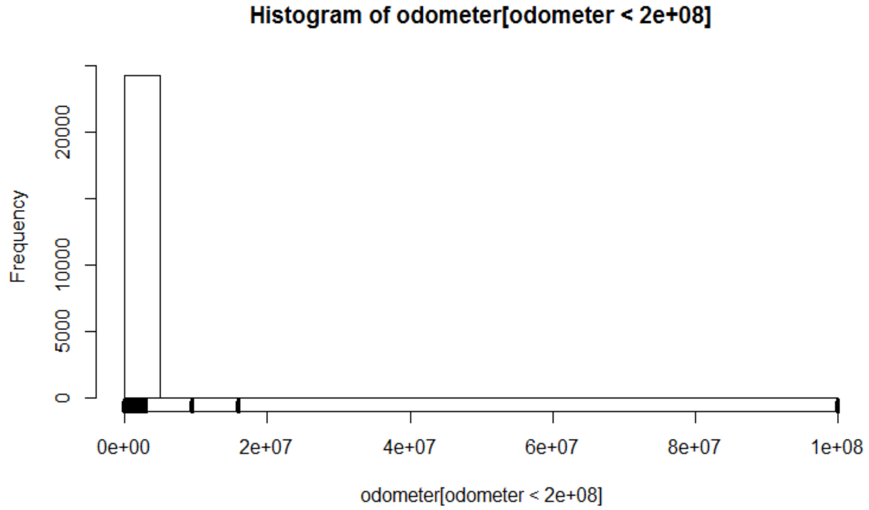
```
load(url("http://eeyore.ucdavis.edu/stat141/Data/vehicles.rda"))
```

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?**

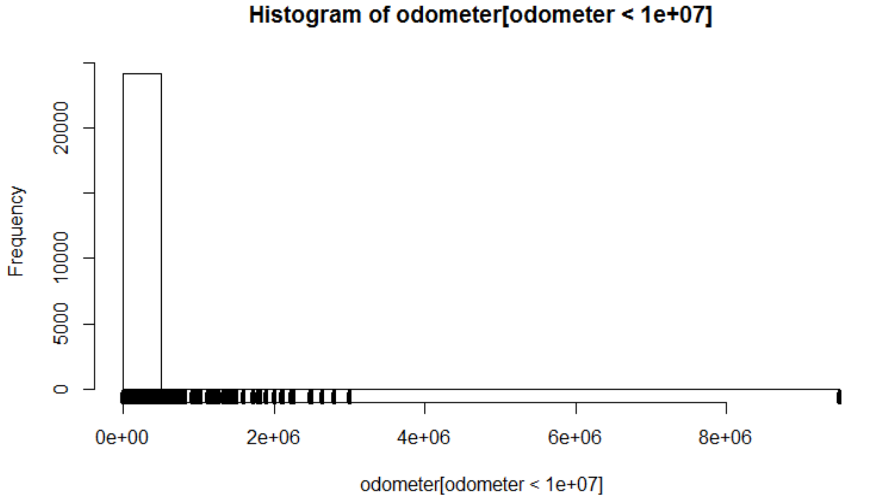
```
#odometer
hist(vposts$odometer)
rug(vposts$odometer,lwd=4)
```



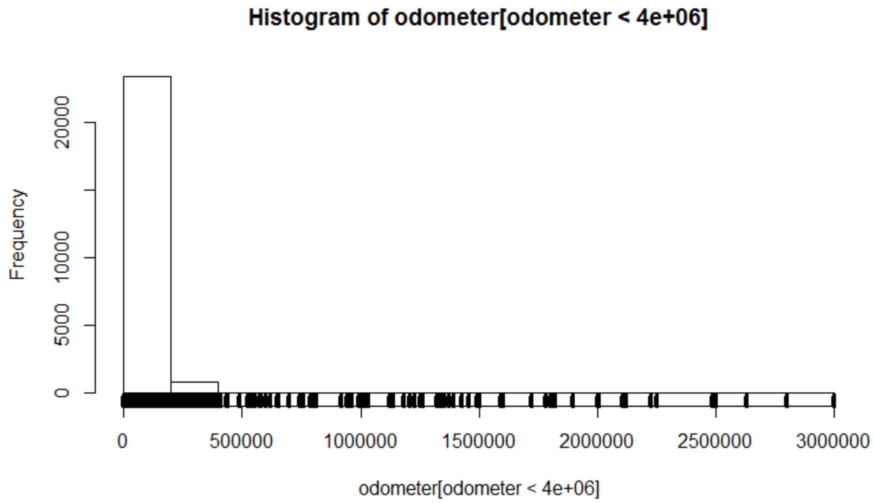
```
#Omit extreme value bigger then 2e8
with(vposts,hist(odometer[odometer<2e8]))
rug(vposts$odometer,lwd=4)
```



```
#Omit extreme value bigger then 1e7
with(vposts,hist(odometer[odometer<1e7]))
rug(vposts$odometer,lwd=4)
```



```
#Omit extreme value bigger then 4e6
with(vposts,hist(odometer[odometer<4e6]))
rug(vposts$odometer,lwd=4)
```



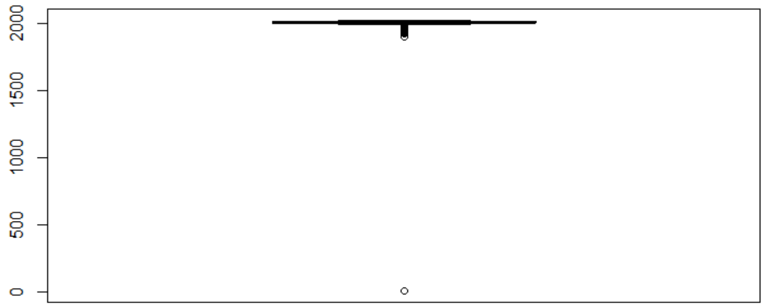
```
#Fix value
vposts$odometer[vposts$odometer>4e6]=NA
max(vposts$odometer,na.rm=TRUE)

## [1] 3000000
```

These anomalies would bring biased mean and median, so we should exclude them.

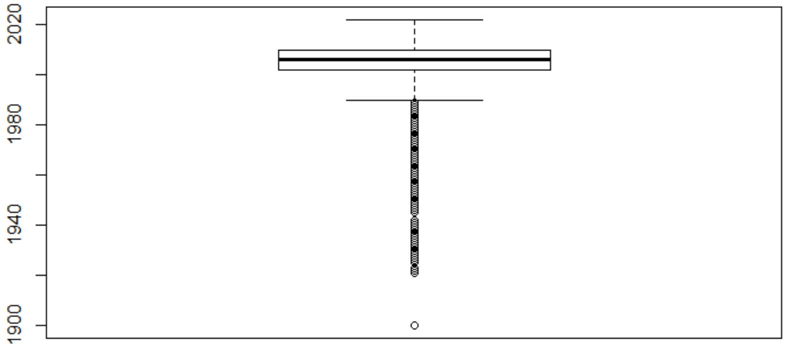
```
#condition
t<-table(vposts$condition)
t=t[t<=3]
head(t)

##
##                               0used                             207,400
##                                   1                                   1
##                           ac/heater                   carfax guarantee!!
##                                   1                                   2
## complete parts car, blown engine                   front side damage
##                                   1                                   1

#The condition "207,400" is really weird
vposts$condition[vposts$condition=='207,400']=NA

#Drop the level:"207,400"
vposts$condition=droplevels(vposts$condition)

#Have fixed it!
head(table(vposts$condition))

##
##                               0used                           ac/heater
##                                1                                1
##                carfax guarantee!!                       certified
##                                2                               54
## complete parts car, blown engine                        excellent
##                                1                             7543
```

Reference: http://stackoverflow.com/questions/1195826/drop-factor-levels-in-a-subsetted-data-frame

```
#year
boxplot(vposts$year)
```



```
head(sort(vposts$year,decreasing=TRUE),10)

##  [1] 2022 2016 2016 2016 2016 2016 2016 2016 2016 2016

#Omit year smaller than 1500
boxplot(vposts$year[vposts$year>1500])
```



```
#Fix the value
vposts$year[vposts$year<1500]=NA
min(vposts$year,na.rm=TRUE)

## [1] 1900

vposts$year[vposts$year>2016]=NA
max(vposts$year,na.rm=TRUE)

## [1] 2016

#How many year equaling to 2016?
i=!is.na(vposts$year)&vposts$year==2016
table(i)

## i
## FALSE  TRUE
## 34471   206

#Does the body conatin another year information?
table(grepl("20[0-1][0-9]",vposts$body[i]))

##
## FALSE  TRUE
##   111    95

table(grepl("19[0-9][0-9]",vposts$body[i]))

##
## FALSE  TRUE
##   168    38

w=grepl("20[0-1][0-9]",vposts$body[i])
x=grepl("19[0-9][0-9]",vposts$body[i])

#Replace the 'year' with the year information provided in 'body'.
i=which(i)
vposts$year[i[w]]=as.integer(gsub(".*(20[0-1][0-9]).*",'\\1',vposts$bod
y[i[w]]))
vposts$year[i[x]]=as.integer(gsub(".*(19[0-9][0-9]).*",'\\1',vposts$bod
y[i[x]]))

#Let the remaining 2016 be NA.
vposts$year[!is.na(vposts$year)&vposts$year==2016]=NA

#Have fixed it!
vposts$year[!is.na(vposts$year)&vposts$year==2016]

## integer(0)
```

These anomalies year would bring some confusion and bad graph during analysising data.

```
#location
t<-table(vposts$location)
t=t[t<=3]
head(t)

## 
##          (鈽\x8620% DOWN - GUARANTEED APPROVAL!!! 鈽\x86)   pic map 
##                                                                   1 
##           (鈽匘IESEL DUALLY 4X4 鈽匳IVA AUTO SALES 鈽\x85)   pic map 

##                                                                   3 
##                                    (you'll drive anyway 鈽\xba) 
##                                                                   1 
##             (鈽\x86 20% DOWN - GUARANTEED APPROVAL!!! 鈽\x86)   pic 
##                                                                   2 
##           (鈽\x86 $4,995 - GUARANTEED APPROVAL!!! 鈽\x86)   pic map 
##                                                                   1 
##   (鉁擥OOD/BAD/NO CREDIT OK 鈽\x8e CALL US NOW!! 鈽\x8e)   pic map 
##                                                                   3

#Fix:contain "WE ".
head(vposts$location[grepl('WE ', vposts$location)])## [1] "  (HUDSON, NH >>WE FINANCE<<)   pic map "
## [2] "  (WE FINANCE - TEWKSBURY)   pic "
## [3] "  (WE FINANCE ANY CREDIT)   pic map "
## [4] "  (WE FINANCE ANYONE (BAD CREDIT, NO CREDIT)   pic map "
## [5] "  (WWW.ADVANCEDAUTOSALESINC.COM WE FINANCE)   pic map "
## [6] "  (WE FINANCE - TEWKSBURY)   pic map "

#I find some locations contain exact city or town name and "We " patter
n at the same time. So, I want to leave these a little bit weird locati
ons there without fixing.

vposts$location[grepl('WE ', vposts$location)&!grepl('Chicago', vposts
$location)&!grepl('Sacramento', vposts$location)]=NA
head(vposts$location[grepl('WE ', vposts$location)])

## [1] "  (Chicago,IL / WE DON'T BELIEVE IN BAD CREDIT SCORES!)   pic m
ap "
## [2] "  (Chicago,IL / WE DON'T BELIEVE IN BAD CREDIT SCORES!)   pic m
ap "
## [3] "  (Chicago,IL / WE DON'T BELIEVE IN BAD CREDIT SCORES!)   pic m
ap "
## [4] "  (Chicago,IL / WE DON'T BELIEVE IN BAD CREDIT SCORES!)   pic m
ap "
## [5] "  (Chicago,IL / WE DON'T BELIEVE IN BAD CREDIT SCORES!)   pic m
```

```
ap "
## [6] "  (Chicago,IL / WE DON'T BELIEVE IN BAD CREDIT SCORES!)   pic m
ap "

#Fix:contain:"You "
head(vposts$location[grepl('YOU ', vposts$location)])

## [1] "  (YOU ARE APPROVED - TEWKSBURY MA)   pic map "
## [2] "  (CREDIT ISN'T A PROBLEM FOR US YOU HERE!!)   pic map "
## [3] "  (ALL YOU NEED IS A CHECK STUB TO DRIVE!!!)   pic map "
## [4] "  (YOU WILL GET APPROVED WITH US CALL NOW!!)   pic map "
## [5] "  (YOUR INCOME WILL GET YOU DRIVING HERE!!!)   pic map "
## [6] "  (YOUR INCOME WILL GET YOU DRIVING HERE!!!)   pic map "

vposts$location[grepl('YOU ', vposts$location)]=NA
vposts$location[grepl('YOU ', vposts$location)]

## character(0)

#Fix:contain:"www."or "WWW."
head(vposts$location[grepl('www.', vposts$location)|grepl('WWW.', vpost
s$location)])

## [1] "  (www.HughesMotorGroup.Com)   pic map "
## [2] "  (WWW.HughesMotorGroup.Com)   pic map "
## [3] "  (Worcester www.massmotorcars.com)   pic map "
## [4] "  (worcester www.massmotorcars.com)   pic map "
## [5] "  (worcester www.massmotorcars.com)   pic map "
## [6] "  (worcester www.massmotorcars.com)   pic map "

#I find some locations contain exact city or town name and "www" patter
n at the same time. So, I want to leave these a little bit weird locati
ons there without fixing.
vposts$location[grepl('WWW.', vposts$location)|grepl('www.', vposts$loc
ation)& !grepl('Brooklyn', vposts$location)&!grepl('Lowell', vposts$loc
ation)&!grepl('worcester', vposts$location)]=NA
vposts$location[grepl('www.', vposts$location)|grepl('WWW.', vposts$loc
ation)]

##  [1] "  (worcester www.massmotorcars.com)   pic map "
##  [2] "  (worcester www.massmotorcars.com)   pic map "
##  [3] "  (worcester www.massmotorcars.com)   pic map "
##  [4] "  (worcester http://www.massmotorcars.com)   pic map "
##  [5] "  (Lowell, MA *www.daveducharmesauto.com*)   pic map "
##  [6] "  (Lowell, MA *www.daveducharmesauto.com*)   pic map "
##  [7] "  (Lowell, MA *www.daveducharmesauto.com*)   pic map "
##  [8] "  (Lowell www.cstreetauto.com)   pic map "
##  [9] "  (Lowell www.cstreetauto.com)   pic map "
## [10] "  (Brooklyn - www.WBAUTOINC.com)   pic "
## [11] "  (Brooklyn - www.WBAUTOINC.com)   pic "
## [12] "  (Brooklyn - www.WBAUTOINC.com)   pic "
## [13] "  (Brooklyn - www.WBAUTOINC.com)   pic "
## [14] "  (Brooklyn - www.WBAUTOINC.com)   pic "
## [15] "  (Brooklyn - www.WBAUTOINC.com)   pic "
## [16] "  (Brooklyn - www.WBAUTOINC.com)   pic "
## [17] "  (Brooklyn - www.WBAUTOINC.com)   pic "
## [18] "  (Brooklyn - www.WBAUTOINC.com)   pic "
## [19] "  (Brooklyn - www.WBAUTOINC.com)   pic "
```

There are a lot of weird locations in the data. Some ones contain tipical patterns such as *we, your* or *www*. Then it would be able to fix them conveniently. However, there is also a quantity of weird locations containging no typical patterns and I do not know how to fix them currently.

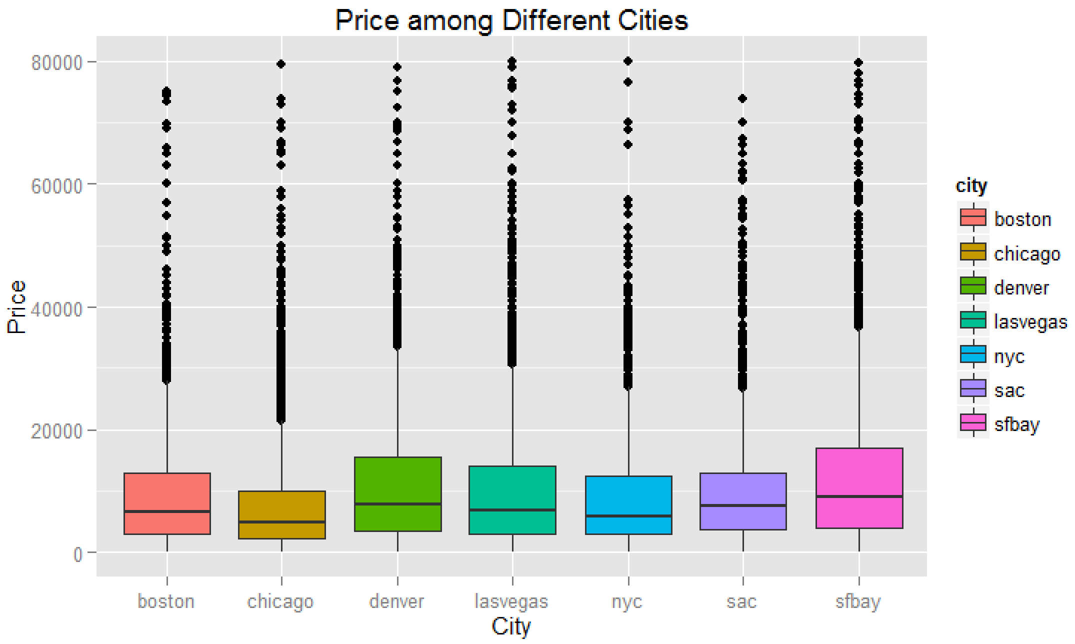
```
#Price
#As far as I see, price smaller then 0.05 percentiles is anomalies.
quantile(vposts$price,0.05,na.rm=TRUE)

##  5%
## 499

#Fix the value samller then 500

#How many prices smaller than 500?
i=!is.na(vposts$price)&vposts$price<500
table(i)

## i
## FALSE  TRUE
## 33046  1631

#Does the 'body' include the price?
table(grepl("\\$",vposts$body[i]))

##
## FALSE  TRUE
##   737   894

#Select observations whose 'body' does not contain $.
w=!grepl("\\$",vposts$body[i])

#Then change these price to be NA and leave the obeservations whose pri
ce smaller than 500 yet contain $ in 'body' there.
i=which(i)
vposts$price[i[w]]=NA

i=!is.na(vposts$price)&vposts$price<500
table(i)
```

```
## i
## FALSE  TRUE
## 33783   894

#Fix the extreme big value by same way used in part 1 Q8.
vposts$price[vposts$price>8e4]=NA
max(vposts$price,na.rm=TRUE)

## [1] 79998
```

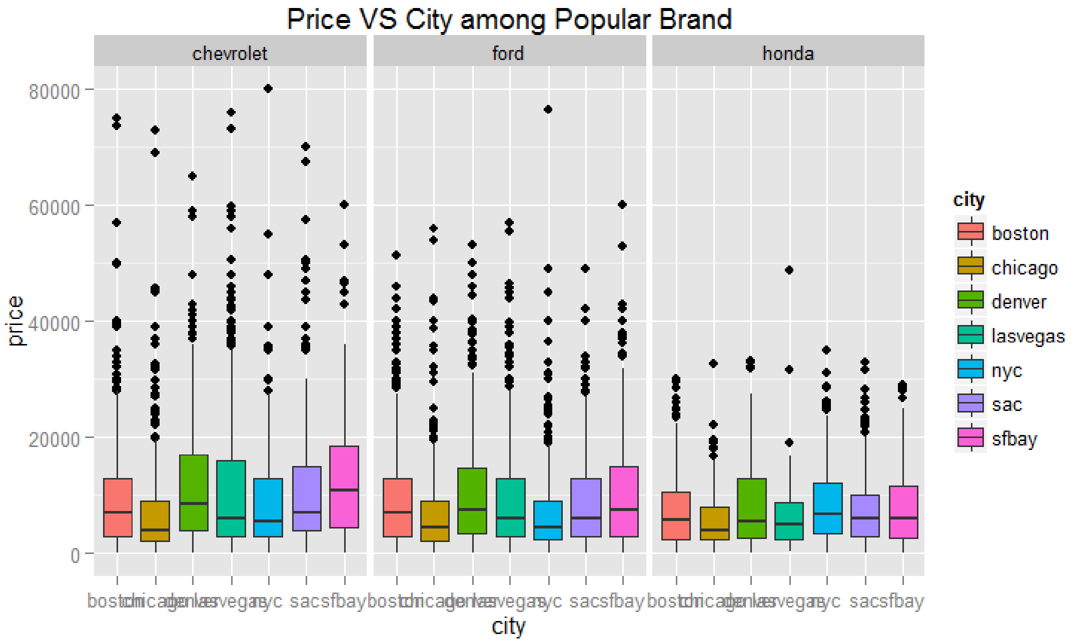
We find that there are 1631 prices smaller than 500 at begining, yet only 894 in the end. These extreme small price would affect the mean price calculation as well as the distribution of price.

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?**

```
library(ggplot2)
library(dplyr)

#Prive VS City
ggplot(vposts,aes(x=vposts$city,y=vposts$price,fill=city))+geom_boxplot
()+xlab('City')+ylab('Price')+ggtitle("Price among Different Cities")
```



```
#Does it depends on brand?
vposts %>%
  filter(maker=='chevrolet'|maker=='ford'|maker=='honda') %>%
 ggplot(aes(x=city,y=price,fill=city))+geom_boxplot()+facet_grid(~maker)+ggtitle("Price VS City among Popular Brand")
```



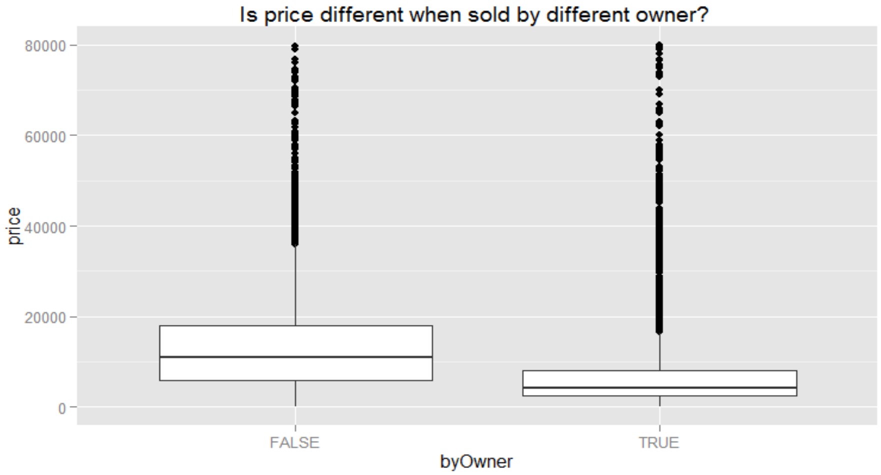
Based on the graph, the average price is highest in SFbay and vehicles sold in NYC and Denver are also expensive, while it is lowest in Chicago. When I focus on different brands,the price orders are similar. I think it is really interesting. For the buyer, if it is possible, they can choose Chiago market to save money. And for the saler, they may prefer SFbay. To test this finding, I choose two car models randomly and search for their prices at TRUECar(website). And the prices of a new Ford Exploer 2016 and Nissan Pathfinder 2015 are listed below:

	Ford Explorer 2016	*Nissan Pathfinder 2015*
San Francisco	$32709	$34040
NYC	$30302	$32314
Chicago	$30720	$31975

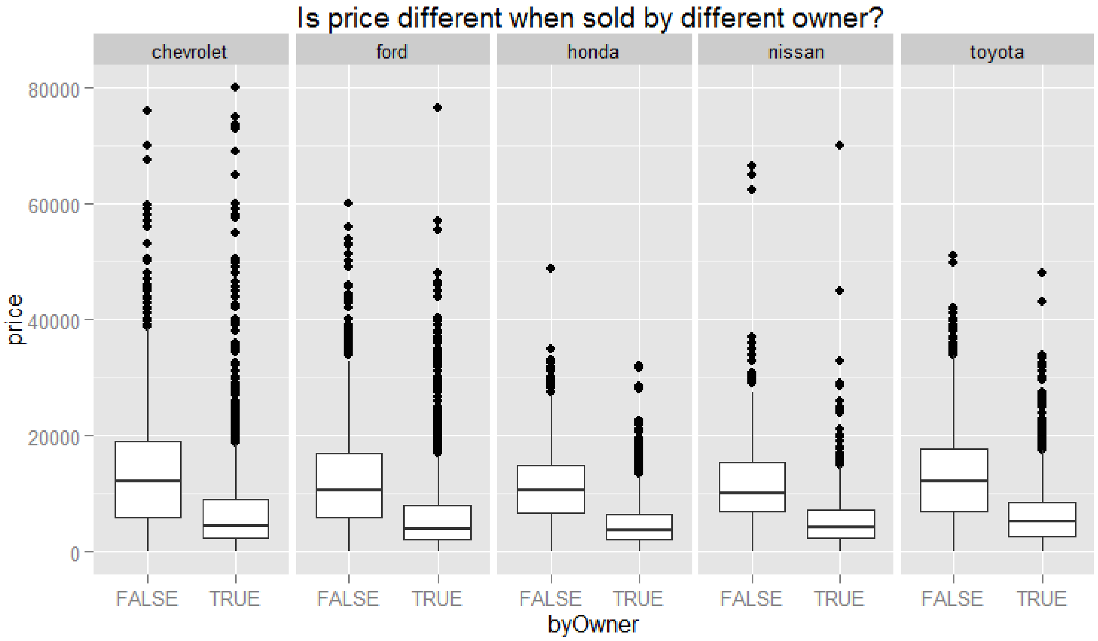
So, the finding generalized from data is common in some degree. In the future, maybe we can buy a car or other vehicles at Chicago.

Reference: https://www.truecar.com/

```
#Price and byOwner
ggplot(vposts,aes(x=byOwner,y=price))+geom_boxplot()+ggtitle("Is price
different when sold by different owner? ")
```



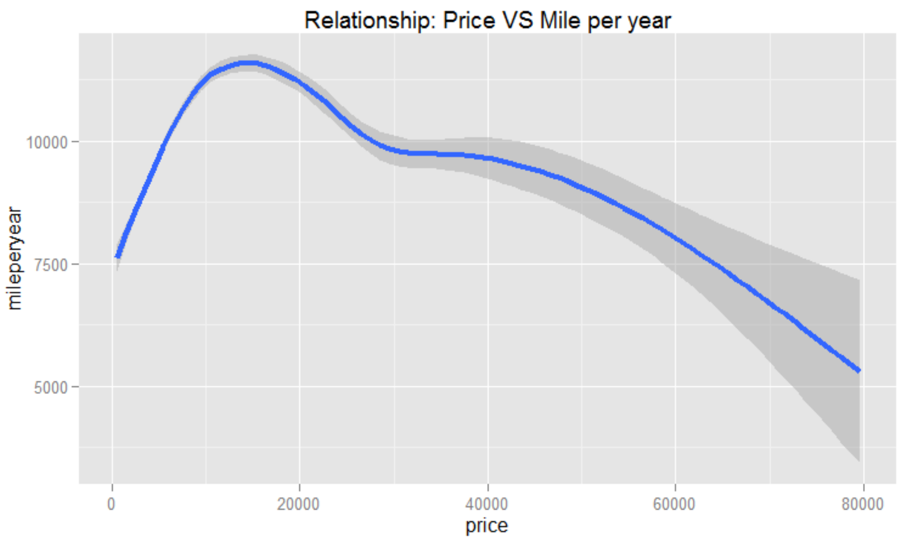
```
#Does it depends on brand?
vposts %>%
  filter(maker=='chevrolet'|maker=='ford'|maker=='honda'|maker=='toyota
'|maker=='nissan') %>%
ggplot(aes(x=byOwner,y=price))+geom_boxplot()+facet_grid(~maker)+ggtitl
e("Is price different when sold by different owner? ")
```



The vehicle sold by saler is always more expensive than by owner. It is comprehensive. Saler need to get the intermediate fee, then the price is higher.

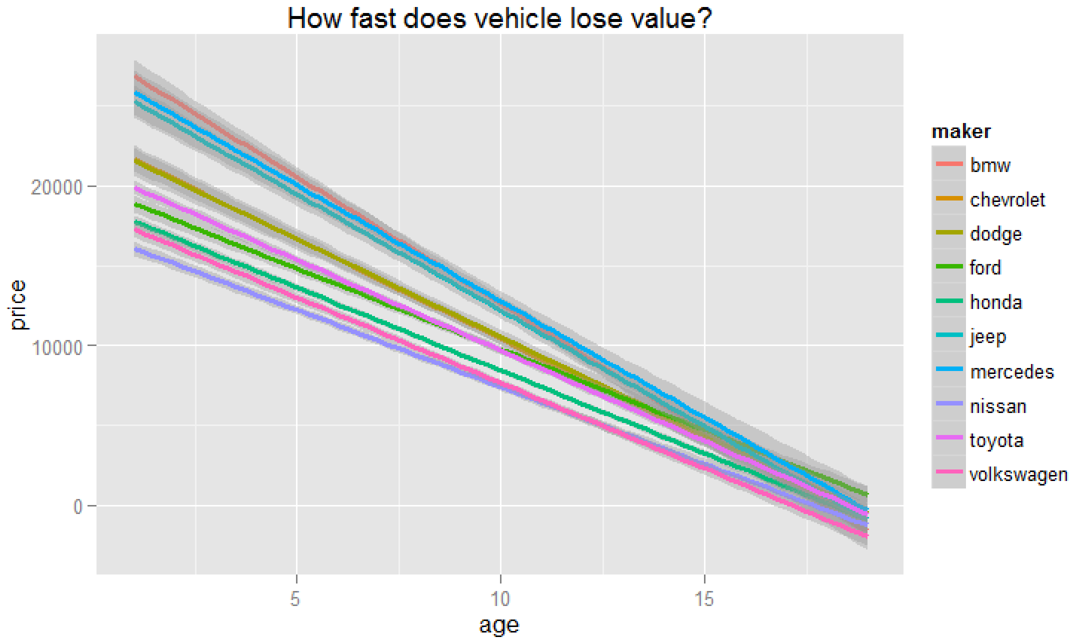
```
#Mile per year and price

#vposts2=add two new variables,age and mileperyear.
#As to calculate mile per year and there are a lot of cars made in 2015,
 I defined age as (2015-year+1)
vposts2=
  vposts %>%
  mutate(age=2015-year+1) %>%
  mutate(mileperyear=odometer/age)

#The 0.75 quantile of price is 13991, so the peak is expensive car.
#Select recent 30 years vehicles, to get clear trend.
vposts2 %>%
  filter(mileperyear<5e4) %>%
  filter(year>1985) %>%
  filter(price>500) %>%
  ggplot(aes(x=price,y=mileperyear))+geom_smooth(size=1.5)+ggtitle("Rel
ationship: Price VS Mile per year")
```

```
## 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.
```

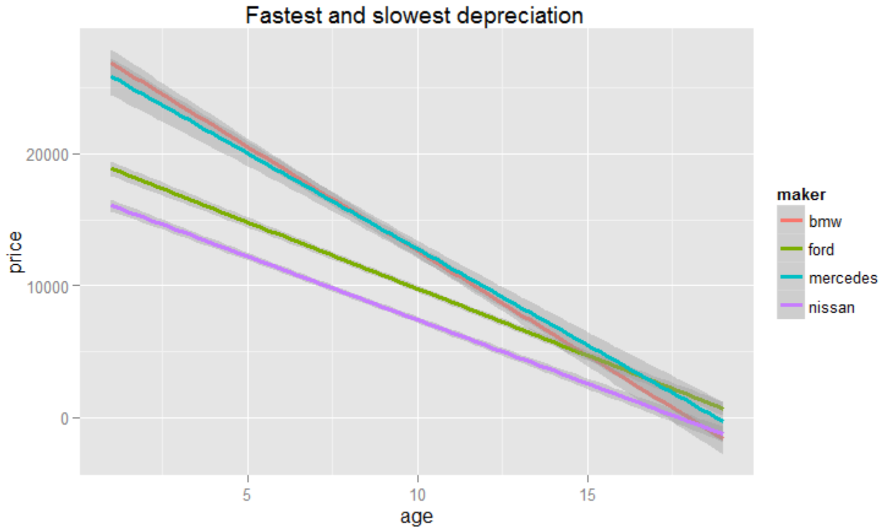


First, I add two variables, age and mile per year which is calculated by odometer reading/age. And I find that there is a peak as price roughly equals to 17500. Since the 0.75 quantile of price is 13700, 17500 is quiet high. Baesd on the graph, before the peak, there is a positive relationship between price and mile per year. I guess the reason is that the higher price, the better quality and the vehicle can run more per year. After the peak, these vehicles are really expensive and may be defined as luxury. And they might be the represent of power or wealth and do not be drived frequently. Thus, as the price becomes higher, mile per year goes down.

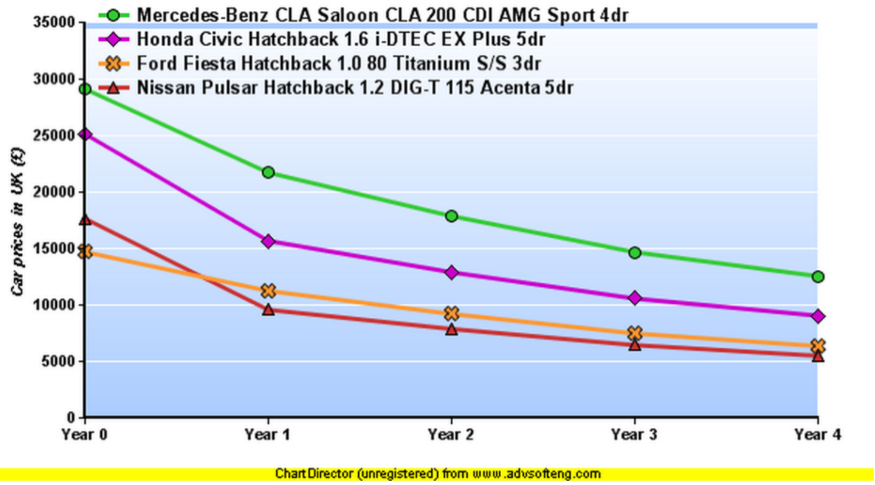
```
#How fast does vehicle lose value?
#Age vs Price among popular makers

vposts2 %>%
  filter(age<20) %>%
  filter(maker=='chevrolet'|maker=='ford'|maker=='honda'|maker=='toyota
'|maker=='nissan'|maker=='dodge'|maker=='bmw'|maker=='jeep'|maker=='mer
cedes'|maker=='volkswagen') %>%
  ggplot(aes(x=age,y=price,col=maker))+geom_smooth(method='lm',size=1.3)
+ggtitle("How fast does vehicle lose value?")
```



```
#BMW, Mercedes and Jeep drop most quickly, then Chevrolet, Dodge, then
Volkswagen, Toyota, then Honda, then Nissan, ford.

#Fastest and slowest depreciation
vposts2 %>%
  filter(age<20) %>%
  filter(maker=='bmw'|maker=='mercedes'|maker=='nissan'|maker=='ford')
 %>%
  ggplot(aes(x=age,y=price,col=maker))+geom_smooth(method='lm',size=1.3)
+ggtitle("Fastest and slowest depreciation ")
```



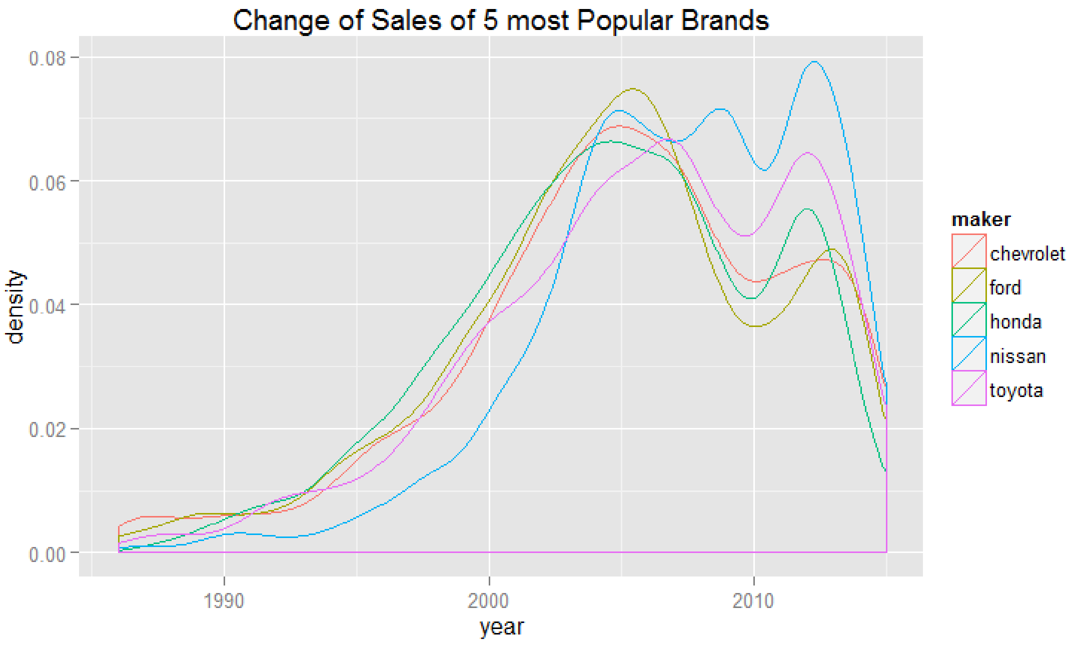
I'm interested in which brand's cars lose their value fast? And I generalized from the graph that the BMW, Mercedes and Jeep depreciate most quickly. At the same time, Ford and Nissan have better capability to keep their value. As far as I see, this finding could provide suggestion for new car buyer. And they could choose vehicles losing value slowly and get a good price if they plan to sale them in the future. And I collect some car depreciation data from Whatcar (website):



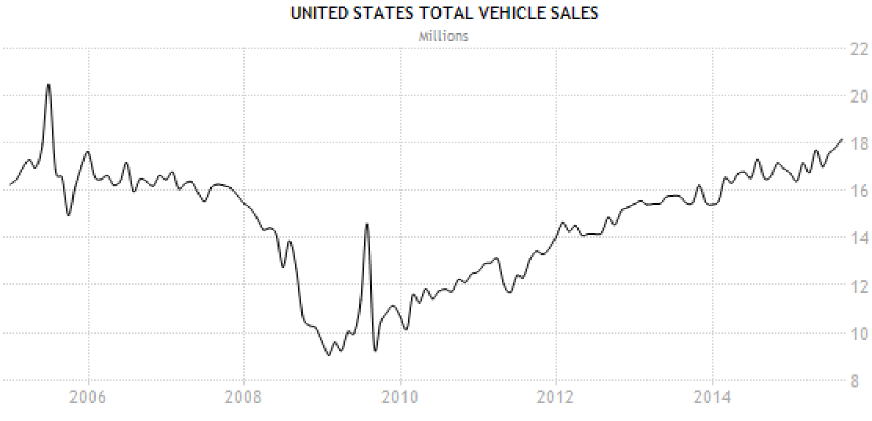
The graph shows similar trend as my finding. Although the speed of depreciation depends on a lot of factors and maker is only one of them, the insight is able to help us a little when we buy a new car.

Reference: http://www.whatcar.com/car-depreciation-calculator/results?edition1=44708&edition2=42343&edition3=45150&makeId=13298&modelVersionId=42268&editionId=44054

```
#Quantity of cars made in different year

#Maker VS Year

#Select 5 most popular brands
head(sort(table(vposts$maker),decreasing = TRUE),5)

##
##      ford chevrolet    toyota     honda    nissan
##      4266      3394      3332      2650      2473

#Focus on recent 30 years data to get a clear graph
vposts %>%
  filter(year>1985)%>%
  filter(maker=='chevrolet'|maker=='ford'|maker=='honda'|maker=='toyota
'|maker=='nissan') %>%
  ggplot(aes(x=year,col=maker))+geom_density()+ggtitle('Change of Sales
 of 5 most Popular Brands')
```



In my opinion, this graph can reflect how many vehicles are made by each maker in each year in some extent. Partly, it tells you the development of every maker and can provide a whole scene of market trend. Because people always sale used vehicles after buying them several years, it is not sound to say the market had been increasing from 1985 to 2005(although it may be true). However, we can figure out some economy recession and expension in recent years, such as the boom around 2005, the depression around 2010 and boom around 2012. To test this feature, I search for the total vehicle sales in recent 10 years. It is the result:



SOURCE: WWW.TRADINGECONOMICS.COM | AUTODATA CORPORATION

This graph also shows the boom around 2005 and depression around 2010, yet no boom in 2012. The reason behind may be that a lot of people sale their cars after 3 years buying. That is why there is a number of used car manufactoried in 2012 to be sold.

Furthermore, we can focus on the behavior among makers during the economy fluctuation cycle. There is a sharp drop of Ford vehicles between 2005 and 2010. It may be caused by the decreasing new car sales from 2005 to 2010. Instead, the amounts of Nissan made in 2005 and made in 2010 differ more little. It reflects that Nissan sales better in the economy recession period.

Reference: http://www.tradingeconomics.com/united-states/total-vehicle-sales

```
#Who make old car?

#This time, we define cars made more than 50 years ago as old cars.
vposts2 %>%
  filter(age>50) %>%
  filter(!is.na(maker)) %>%
  count(maker) %>%
  arrange(desc(n))

## Source: local data frame [33 x 2]
##
##         maker     n
##         (chr) (int)
## 1   chevrolet   139
## 2        ford   127
## 3       dodge    23
## 4  volkswagen    12
## 5      willys    12
## 6    cadillac    11
## 7       buick    10
## 8     mercury    10
## 9     lincoln     9
## 10   plymouth     8
## ..        ...   ...
```

Biggest old car makers were Chevrolet and Ford. It means that these two makers have really long history. This insight might be interesting for people loving vehicle history.