BDA Assigment

Lalith Venkat Perugu

22/11/2020

Loading the Melbourne data set.

```
housing.data <- read.csv("C:/Program Files/R/melbourne_data.csv")
```

I have included the below code globally to show the packages that are used in this assignment.

```
knitr::opts_chunk$set(echo = FALSE)
if(!require(tinytex)) install.packages("tinytex", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: tinytex
```

```
if(!require("plotrix")) install.packages("plotrix", repos = "http://cran.us.r-project.org");
```

```
## Loading required package: plotrix
```

```
library("plotrix")
if(!require('ggplot2')) install.packages("ggplot2", repos ="http://cran.us.r-project.org")
```

```
## Loading required package: ggplot2
```

```
library('ggplot2')
if(!require('dplyr')) install.packages("dplyr",repos="http://cran.us.r-project.org")
```

```
## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library("dplyr")
```

Lets see the structure of the dataset to know the no of objects and variables present

```
str(housing.data)
```

```
## 'data.frame':    34857 obs. of  13 variables:
##  $ X            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Date         : chr  "3/09/2016" "3/12/2016" "4/02/2016" "4/02/2016" ...
##  $ Type         : chr  "h" "h" "h" "u" ...
##  $ Price        : int  NA 1480000 1035000 NA 1465000 850000 1600000 NA NA NA ...
##  $ Landsize     : int  126 202 156 0 134 94 120 400 201 202 ...
##  $ BuildingArea : num  NA NA 79 NA 150 NA 142 220 NA NA ...
##  $ Rooms        : int  2 2 2 3 3 3 4 4 2 2 ...
##  $ Bathroom     : int  1 1 1 2 2 2 1 2 1 2 ...
##  $ Car          : int  1 1 0 1 0 1 2 2 2 1 ...
##  $ YearBuilt    : int  NA NA 1900 NA 1900 NA 2014 2006 1900 1900 ...
##  $ Distance     : chr  "2.5" "2.5" "2.5" "2.5" ...
##  $ Regionname   : chr  "Northern Metropolitan" "Northern Metropolitan" "Northern Metropolitan" "Nortl
##  $ Propertycount: chr  "4019" "4019" "4019" "4019" ...
```

Firstly lets clean the dataset by removing/replacing the NA values,incorrect values and make it ready for analysis

Removing all the NA values

```
correctdata<-na.omit(housing.data)
```

we can observe that there many incorrect values and outliers present in many varaiables

For better visualization we have replaced those values by below method

Removing the incorrect values and replacing them.__

Variable Landsize

```
correctdata[correctdata$Landsize<80, "Landsize"] <-250
correctdata[correctdata$Date>1550, "Landsize"]<-600
```

Variable Price

```
correctdata$Price[correctdata$Price > 2000000]<- mean(correctdata$Price,na.rm = T)
correctdata$Price[correctdata$Price < 300000]<- mean(correctdata$Price,na.rm = T)
```

Variable Distance

```
correctdata$Distance<-as.numeric(as.character(correctdata$Distance))
typeof(correctdata$Distance)
```

```
## [1] "double"
```

Variable Propertycount

```
correctdata$Propertycount<-as.factor(as.character(correctdata$Propertycount))
typeof(correctdata$Propertycount)
```

```
## [1] "integer"
```

Variable Car

Variable Regionname

```
correctdata$Regionname<-as.factor(correctdata$Regionname)
typeof(correctdata$Regionname)
```

```
## [1] "integer"
```

```
levels(correctdata$Regionname)
```

```
## [1] "Eastern Metropolitan"       "Eastern Victoria"
## [3] "Northern Metropolitan"      "Northern Victoria"
## [5] "South-Eastern Metropolitan" "Southern Metropolitan"
## [7] "Western Metropolitan"       "Western Victoria"
```

2 Lets find the summary of all the variables of the dataset, The summary gives you the mean,median,1st Quartile,3rd Quartile, minimum and maximum values for the varaiables.

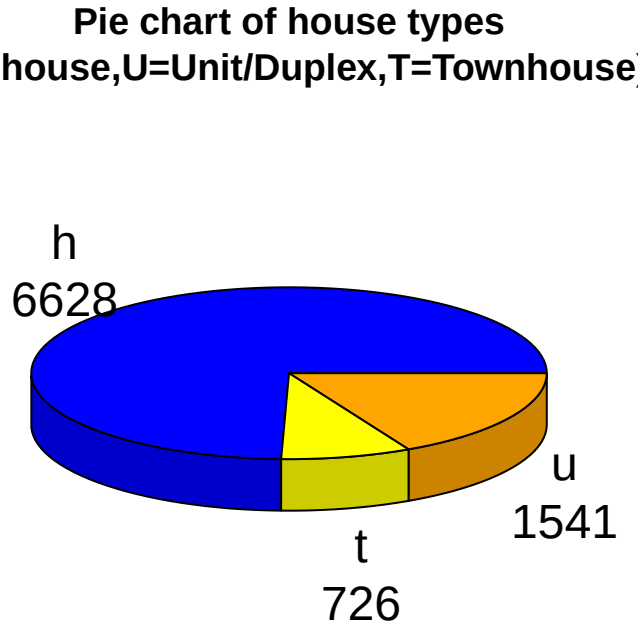
```
summary(correctdata)
```

```
##        X             Date               Type               Price
##  Min.   :    3   Length:8895        Length:8895        Min.   : 300000
##  1st Qu.: 6816   Class :character   Class :character   1st Qu.: 650000
##  Median :13853   Mode  :character   Mode  :character   Median : 905000
##  Mean   :15113                                         Mean   : 958771
##  3rd Qu.:22624                                         3rd Qu.:1185000
##  Max.   :34857                                         Max.   :2000000
```

```
##
##     Landsize        BuildingArea        Rooms           Bathroom
##  Min.   :   80.0   Min.   :   0.0   Min.   : 1.000   Min.   :1.000
##  1st Qu.:  600.0   1st Qu.: 100.0   1st Qu.: 2.000   1st Qu.:1.000
##  Median :  600.0   Median : 132.0   Median : 3.000   Median :2.000
##  Mean   :  587.8   Mean   : 149.3   Mean   : 3.099   Mean   :1.647
##  3rd Qu.:  600.0   3rd Qu.: 180.0   3rd Qu.: 4.000   3rd Qu.:2.000
##  Max.   :21600.0   Max.   :3112.0   Max.   :12.000   Max.   :9.000
##
##       Car          YearBuilt       Distance
##  Min.   :1.000   Min.   :1196   Min.   : 0.0
##  1st Qu.:1.000   1st Qu.:1945   1st Qu.: 6.4
##  Median :2.000   Median :1970   Median :10.2
##  Mean   :1.714   Mean   :1966   Mean   :11.2
##  3rd Qu.:2.000   3rd Qu.:2000   3rd Qu.:13.9
##  Max.   :5.000   Max.   :2019   Max.   :47.4
##
##                        Regionname   Propertycount
##  Southern Metropolitan     :2707   Min.   :  1.0
##  Northern Metropolitan     :2618   1st Qu.: 56.0
##  Western Metropolitan      :2060   Median :186.0
##  Eastern Metropolitan      : 982   Mean   :162.5
##  South-Eastern Metropolitan: 372   3rd Qu.:251.0
##  Northern Victoria         :  62   Max.   :312.0
##  (Other)                   :  94
```

Pie chart

```
numbers<-table(correctdata$Type)
lbls<-paste(names(numbers),"\n", numbers, sep = "")
pie3D(numbers, labels=lbls,radius=1,main= "Pie chart of house types\n (H=house,U=Unit/Duplex,T=Townhouse
```

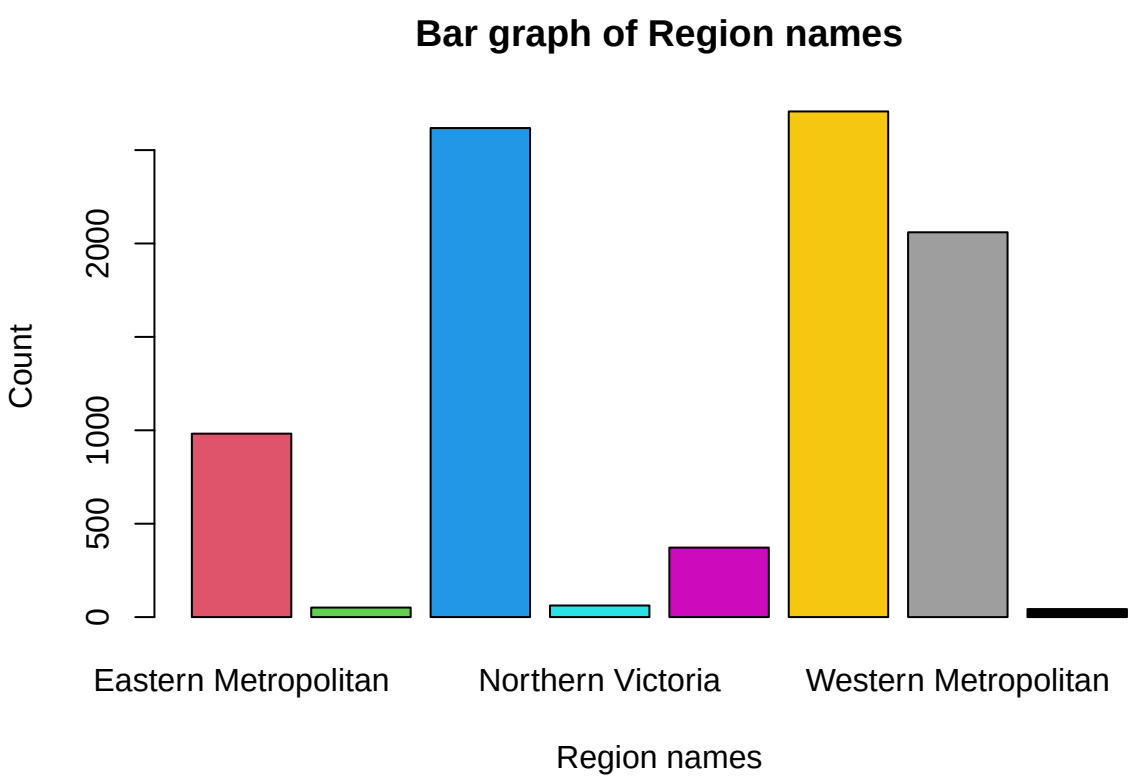


The above pie chart displays the house types with their respective count.

Plotrix package has been used for displying the 3D pie chart.

Bar Garph

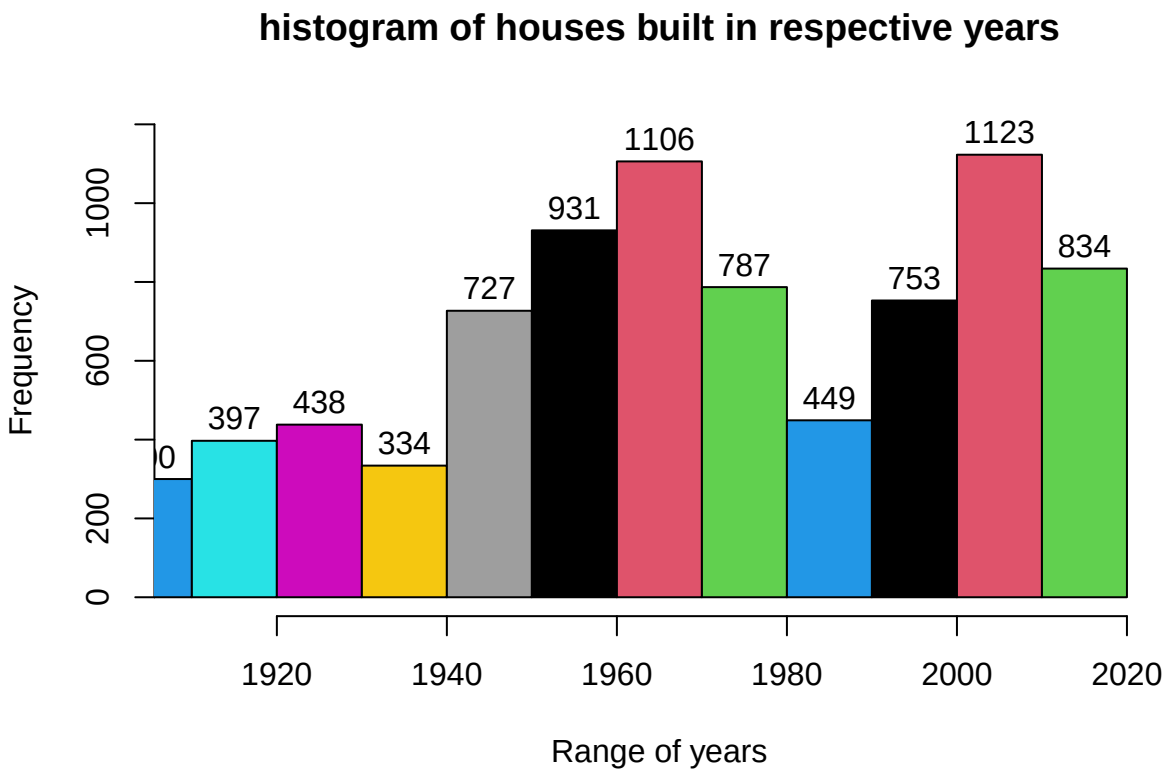
```
barplot(table(correctdata$Regionname),main="Bar graph of Region names", xlab = "Region names",ylab = "C
```



The above bar graph is for Regionname

Histogram

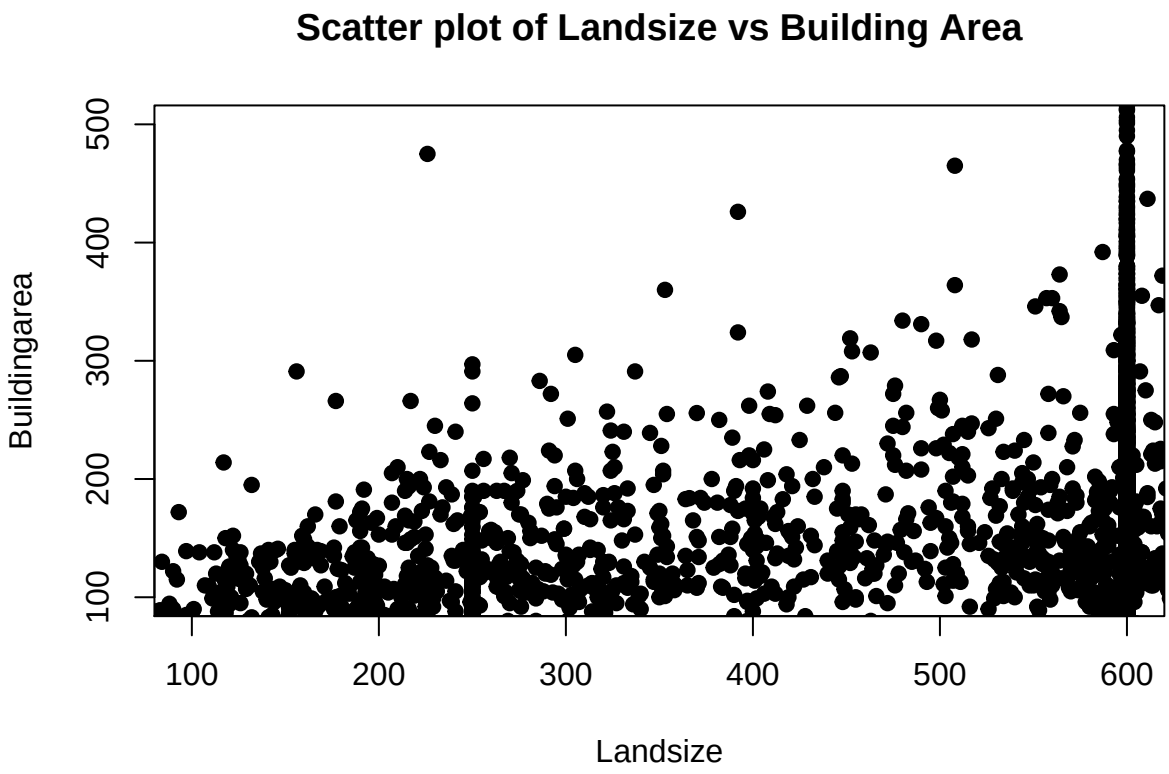
```
histo<-hist(correctdata$YearBuilt,main= "histogram of houses built in respective years", xlab = "Range
text(histo$mids,histo$counts,labels = histo$counts,adj = c(0.5,-0.5))
```



The above histogram shows the number of houses built in those respective range of years.

Scatterplot

```
landsize<-correctdata$Landsize
buildingsize<-correctdata$BuildingArea
plot(landsize,buildingsize,main="Scatter plot of Landsize vs Building Area",xlab="Landsize",ylab="Build
abline(lm(landsize~buildingsize),col="Yellow")
```

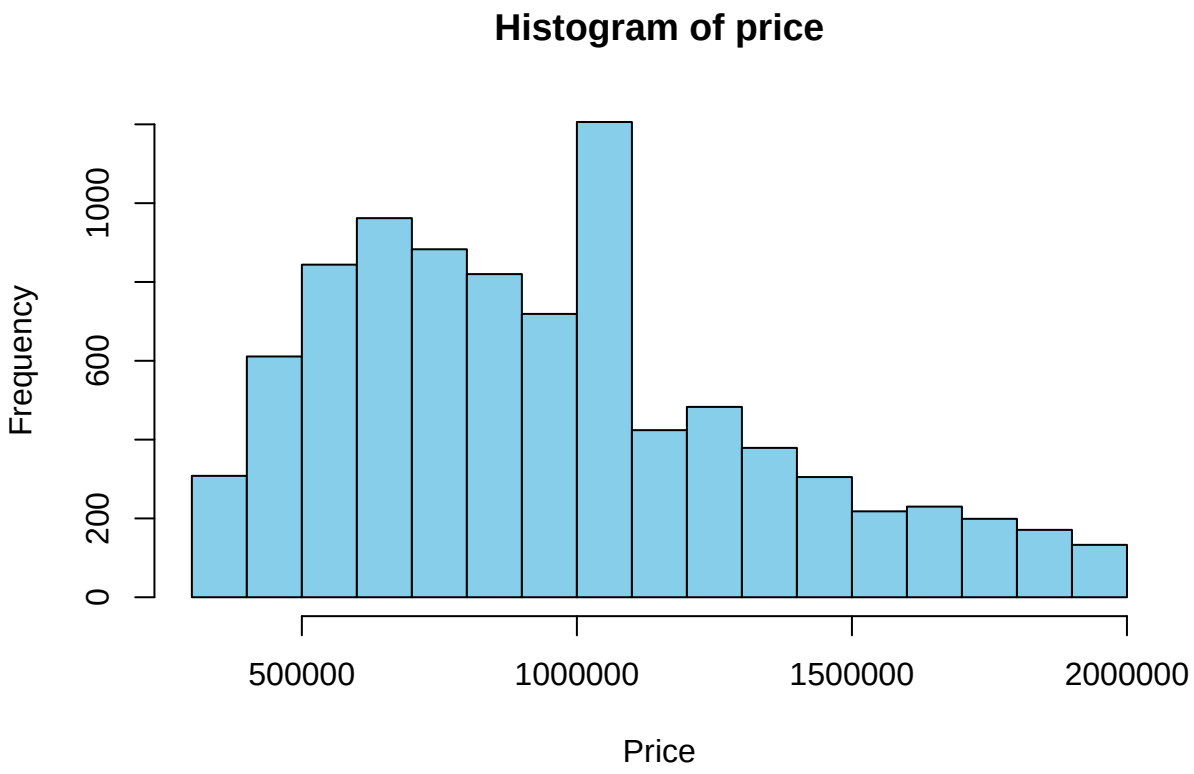


The above graph shows the scatterplot for the landsize and buildingsize.

3 Analysis on Price variable

3a Histogram of price

```
hist(correctdata$Price,xlim = c(300000,2000000),main = "Histogram of price",xlab="Price", ylim=c(0,1200
```



From the above histrogram we can observe that price of houses are much higher in left side of the graph then the right side.

Highest frequencies of prices which is nearly 1200 lies with the price range of 1000000 to 2000000.

```
summary(correctdata$Price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  300000  650000  905000  958771 1185000 2000000
```

The summary of price shows that the minimum price is 300000 and maximum at 2000000 and it also shows the values for mean, median, 1st and 3rd Quartile.

```
var(correctdata$Price)
```

```
## [1] 151049567233
```

var() gives the varience for the price variable.

3b) Grouping of Houses by price ranges

```
grouping<-correctdata %>% group_by(Type) %>% mutate(state= cut(Price,breaks=3,labels=c("Low","Medium","
summary(grouping$state)
```

```
##    Low Medium   high
##   4171   3567   1157
```

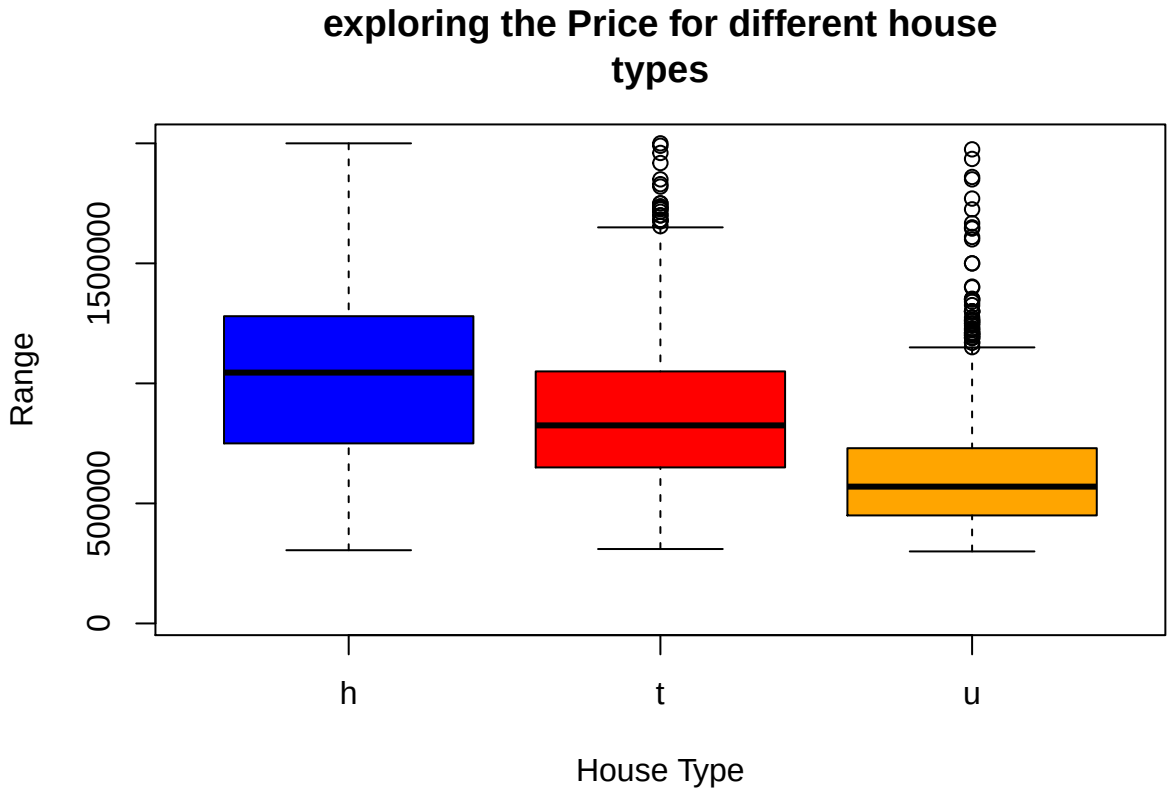
dployr package has been installed to use the group_by() method.

Mutate() function has been used to create a new variable called state.

summary of those are obtained as below.

3c) Exploring prices for different house types

```
housetype<-c("h","t","u")
boxplot(correctdata$Price[correctdata$Type=="h"],
correctdata$Price[correctdata$Type=="t"],correctdata$Price[correctdata$Type=="u"],main= "exploring the
types",ylim=c(30000,2000000),col = c("blue","red","orange"),xlab="House Type"
,ylab="Range",names = housetype)
```



3d) The Variables that are most corelated with price

```
cor(correctdata$Price,correctdata$Rooms, method = "pearson")
```

```
## [1] 0.4052314
```

```
cor(correctdata$Price,correctdata$Distance, method = "pearson")
```

```
## [1] -0.2393597
```

```
cor(correctdata$Price,correctdata$Propertycount, method = "pearson")
```

```
## [1] 0.05752506
```

```
cor(correctdata$Price,correctdata$Bathroom, method = "pearson")
```

```
## [1] 0.3210033
```

```
cor(correctdata$Price,correctdata$Car, method = "pearson")
```

```
## [1] 0.1753096
```

```
cor(correctdata$Price,correctdata$Landsize, method = "pearson")
```

```
## [1] 0.004492762
```

```
cor(correctdata$Price,correctdata$BuildingArea, method = "pearson")
```

```
## [1] 0.3579549
```

```
cor(correctdata$Price,correctdata$YearBuilt, method = "pearson")
```

```
## [1] -0.3586022
```

By checking the corelation of price with all other variables, We can observe that the variables rooms,bathroom,buildingarea are most correlated to the variable price.

4 Listing the Frequencies of various house types

```
table(correctdata$Type)
```

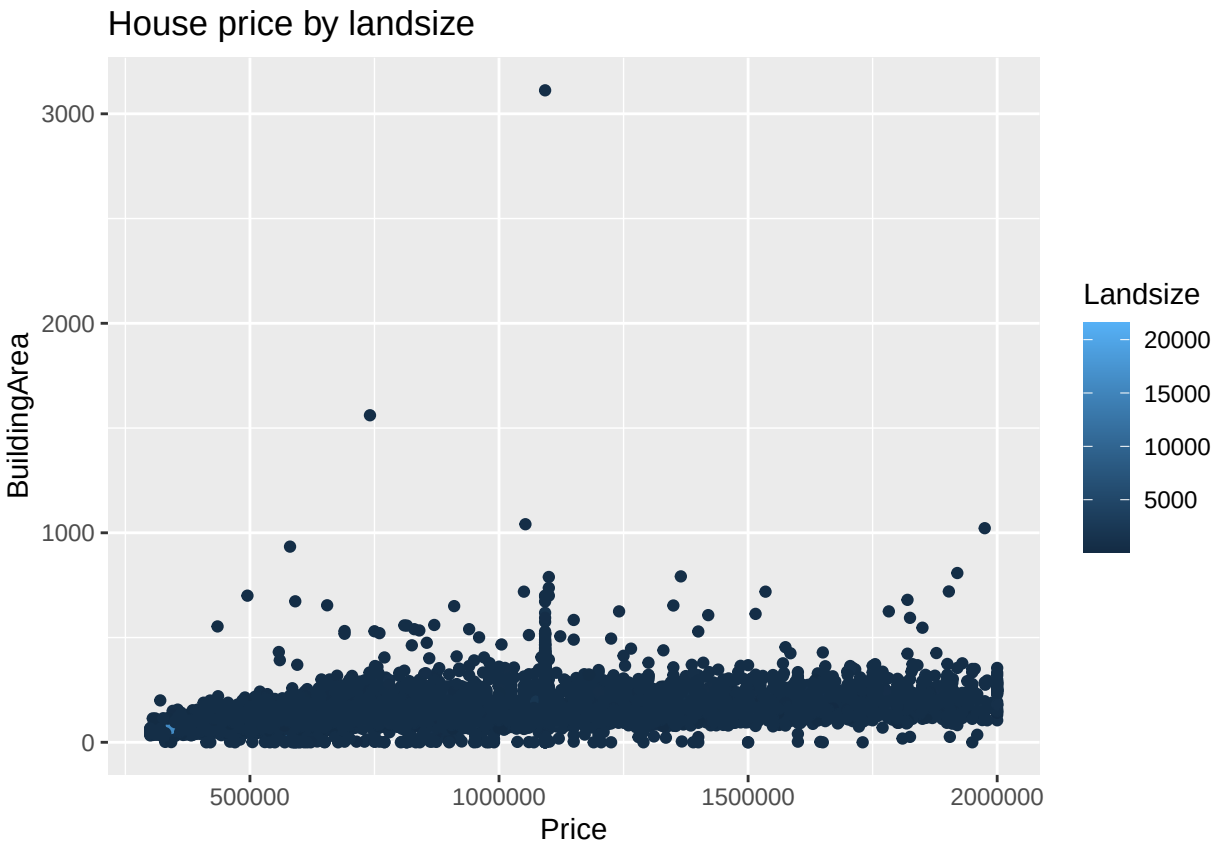
```
##
##    h    t    u
## 6628  726 1541
```

From the above we observe the frequency of each house type.

Scatter plots

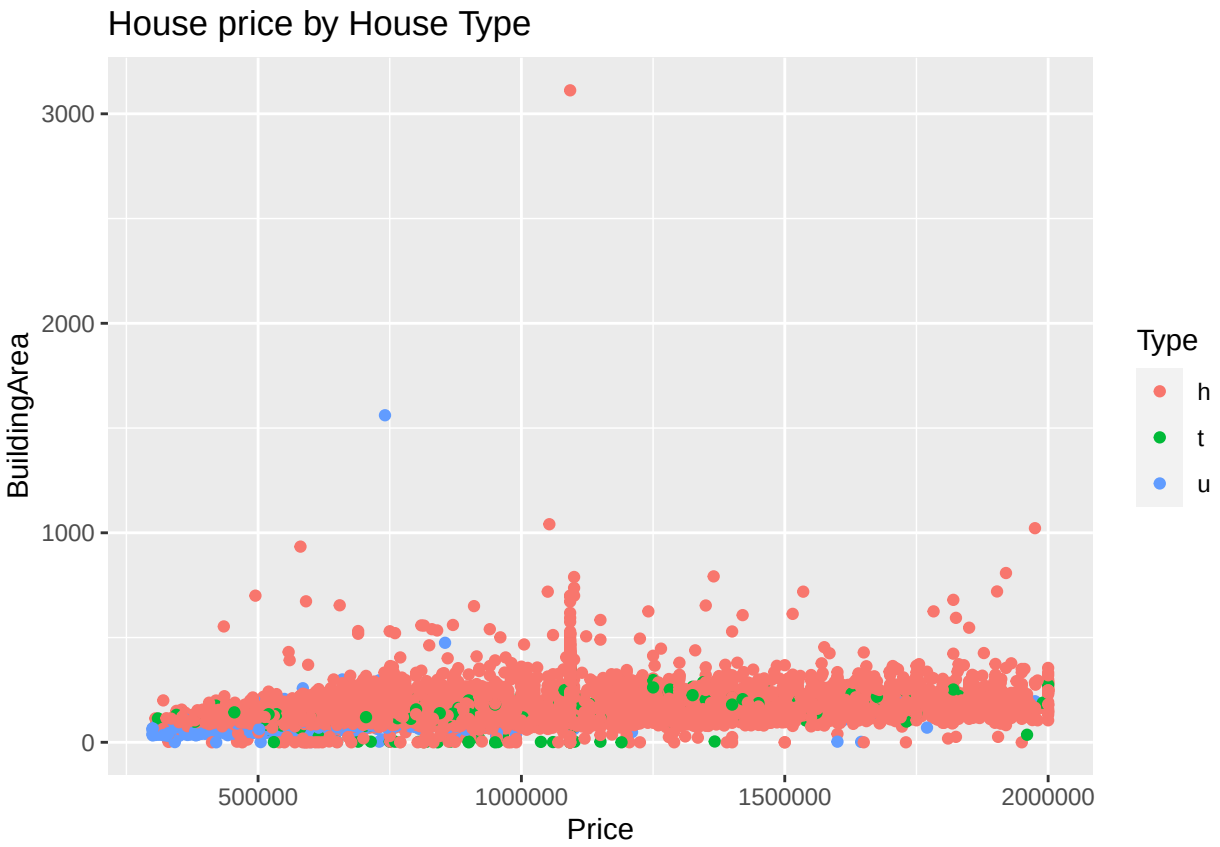
Scatter plot of houseprice with Landsize

```
ggplot(correctdata,aes(Price,BuildingArea,
col=Landsize),xlim(300000,2000000),ylim(0,500))+ geom_point()+ggtitle("House price by landsize")
```



Scatter plot of housing price with House Type

```
ggplot(correctdata,aes(Price,BuildingArea, col = Type))+geom_point()+ggtitle("House price by House Type
```


House price by House Type
BuildingArea
3000
2000
1000
0
500000
1000000
1500000
2000000
Price
Type
h
t
u