2nd Assignment BDA

Samar Khan

2022-12-01

## Reading the given Data set

housing.dataset\_melbourne <- read.csv("/Users/samarkhan/Downloads/melbourne\_housing\_data.csv")  
str(housing.dataset\_melbourne)

## 'data.frame': 48433 obs. of 14 variables:  
## $ X : int 1 2 3 4 5 6 7 8 10 11 ...  
## $ Suburb : chr "Abbotsford" "Abbotsford" "Abbotsford" "Aberfeldie" ...  
## $ Address : chr "49 Lithgow St" "59A Turner St" "119B Yarra St" "68 Vida St" ...  
## $ Rooms : int 3 3 3 3 2 2 2 3 3 3 ...  
## $ Type : chr "h" "h" "h" "h" ...  
## $ Price : int 1490000 1220000 1420000 1515000 670000 530000 540000 715000 1925000 515000 ...  
## $ Method : chr "S" "S" "S" "S" ...  
## $ SellerG : chr "Jellis" "Marshall" "Nelson" "Barry" ...  
## $ Date : chr "1/04/2017" "1/04/2017" "1/04/2017" "1/04/2017" ...  
## $ Postcode : int 3067 3067 3067 3040 3042 3042 3042 3042 3206 3020 ...  
## $ Regionname : chr "Northern Metropolitan" "Northern Metropolitan" "Northern Metropolitan" "Western Metropolitan" ...  
## $ Propertycount: int 4019 4019 4019 1543 3464 3464 3464 3464 3280 2185 ...  
## $ Distance : num 3 3 3 7.5 10.4 10.4 10.4 10.4 3 10.5 ...  
## $ CouncilArea : chr "Yarra City Council" "Yarra City Council" "Yarra City Council" "Moonee Valley City Council" ...

summary(housing.dataset\_melbourne)

## X Suburb Address Rooms   
## Min. : 1 Length:48433 Length:48433 Min. : 1.000   
## 1st Qu.:15797 Class :character Class :character 1st Qu.: 2.000   
## Median :31587 Mode :character Mode :character Median : 3.000   
## Mean :31562 Mean : 3.072   
## 3rd Qu.:47365 3rd Qu.: 4.000   
## Max. :63021 Max. :31.000   
## Type Price Method SellerG   
## Length:48433 Min. : 85000 Length:48433 Length:48433   
## Class :character 1st Qu.: 620000 Class :character Class :character   
## Mode :character Median : 830000 Mode :character Mode :character   
## Mean : 997898   
## 3rd Qu.: 1220000   
## Max. :11200000   
## Date Postcode Regionname Propertycount   
## Length:48433 Min. :3000 Length:48433 Min. : 39   
## Class :character 1st Qu.:3051 Class :character 1st Qu.: 4280   
## Mode :character Median :3103 Mode :character Median : 6567   
## Mean :3123 Mean : 7566   
## 3rd Qu.:3163 3rd Qu.:10412   
## Max. :3980 Max. :21650   
## Distance CouncilArea   
## Min. : 0.0 Length:48433   
## 1st Qu.: 7.0 Class :character   
## Median :11.7 Mode :character   
## Mean :12.7   
## 3rd Qu.:16.7   
## Max. :55.8

# changing to date format  
housing.dataset\_melbourne$Date<-as.Date(housing.dataset\_melbourne$Date,format="%y/%m/%d")

## Exploring the data-set

* Removing the Outliers

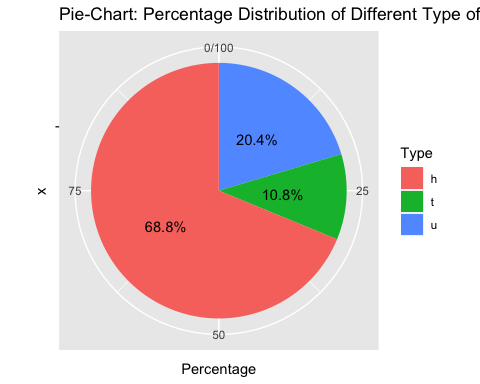
```r  
#Price variable with Outliers present  
boxplot(housing.dataset\_melbourne$Price, col = "blue", main="With Outliers")   
```  
  
![](2-ASSINGMENT-BDA\_files/figure-docx/unnamed-chunk-1-1.png)<!-- -->  
  
```r  
housing.dataset\_melbourne1<-data.frame(housing.dataset\_melbourne)  
Q3\_price = quantile(housing.dataset\_melbourne1$Price,0.65)# gives the 3rd quantile value  
Q1\_price = quantile(housing.dataset\_melbourne1$Price,0.25)# gives the 1rd quantile value  
IQR=IQR(housing.dataset\_melbourne1$Price)  
#Removing outliers in Price attribute###  
housing.dataset\_melbourne1<-subset(housing.dataset\_melbourne1,housing.dataset\_melbourne1$Price>(Q1\_price-IQR)& housing.dataset\_melbourne1$Price<(Q3\_price+1.5\*IQR))  
# Price variable without Outliers present  
boxplot(housing.dataset\_melbourne1$Price, col ="lightblue", main="Without Outliers")   
```  
  
![](2-ASSINGMENT-BDA\_files/figure-docx/unnamed-chunk-1-2.png)<!-- -->  
  
```r  
nrow(housing.dataset\_melbourne1)# printing the no of rows after rmoving outlier  
```  
  
```  
## [1] 45275  
```

* ***Representation of the different types of house- Using pie chart***

##   
## Attaching package: 'dplyr'

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

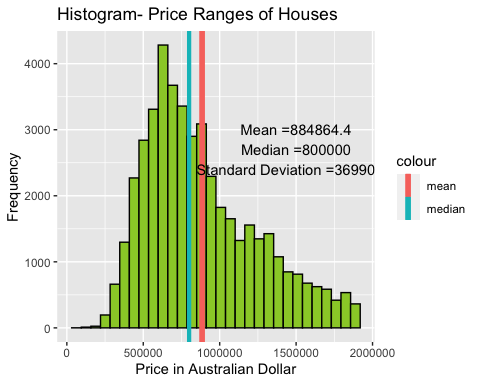
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union



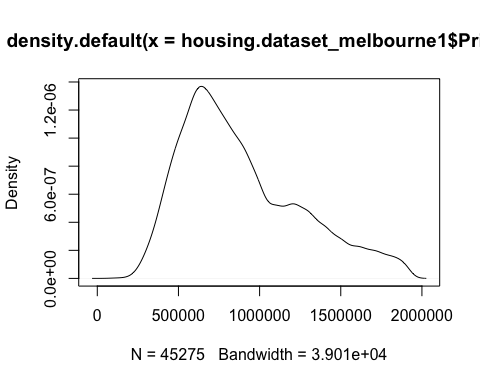
* ***Exploring the Price of the houses in the given data set-using Histogram also showing the mean***
* ***median and standard Deviation***

Mean\_Price<- round(mean(housing.dataset\_melbourne1$Price),1)  
Median\_Price<-median(housing.dataset\_melbourne1$Price)  
StdDev\_Price<-round(sd(housing.dataset\_melbourne1$Price),1)  
#ploting the histogram of price coloumn and adding the mean,median, mode value and x-intercept  
Histogram\_Price<-ggplot(data = housing.dataset\_melbourne1, aes(x=Price)) +   
geom\_histogram(fill="yellowgreen",col="BLACK") +  
geom\_vline(aes(xintercept=Mean\_Price,color='mean'),show.legend = TRUE, size=2) +  
geom\_vline(aes(xintercept=Median\_Price,color='median'),show.legend = TRUE,size=1.5) +  
annotate("text",x=1500000,y=3000,label=paste0("Mean =",Mean\_Price))+  
annotate("text",x=1500000,y=2700,label=paste0("Median =",Median\_Price))+  
annotate("text",x=1500000,y=2400,label=paste0("Standard Deviation =",StdDev\_Price))+  
labs(x="Price in Australian Dollar",y="Frequency",title = "Histogram- Price Ranges of Houses")  
print(Histogram\_Price)

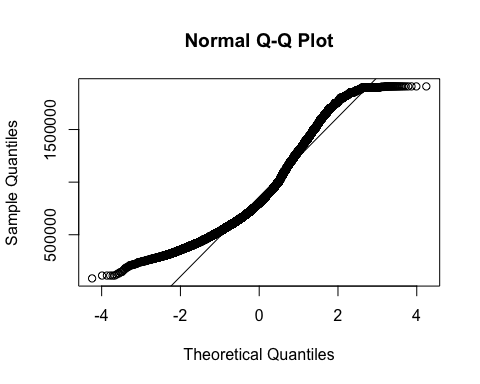
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



plot(density(housing.dataset\_melbourne1$Price)) # gives the desity graph to check if data is normally distributed



qqnorm(housing.dataset\_melbourne1$Price)  
qqline(housing.dataset\_melbourne1$Price)# quantile quantile plot tocheck if data is normally distibuted



summary(housing.dataset\_melbourne1$Price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 85000 605000 800000 884864 1116000 1910000

## Task 1

#### Hypotheses Testing

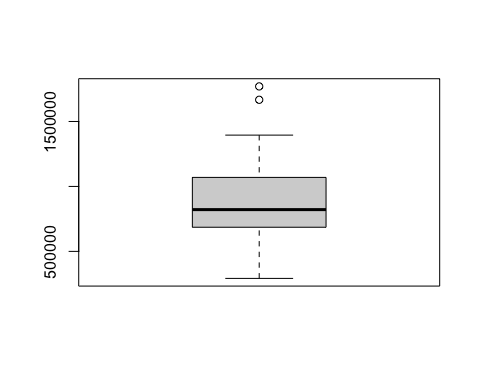
**1.** Testing if the average value of the house price in Melbourne is $ 884864

that means on the sample data set. we will be using one sample test.

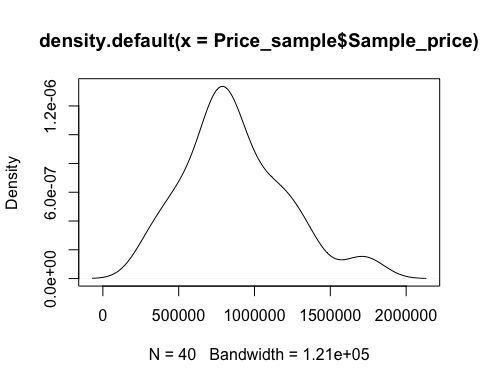
H0 is equals to $884864 (Null hypotheses)

Ha is not equals to $884864 (Alternate Hypotheses)

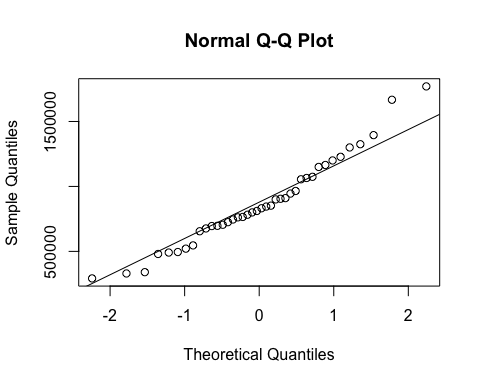
#taking the 40 sample data- from the price collumn.  
set.seed(101) # setting the seed for sample data  
Price\_sample <- data.frame(sample(housing.dataset\_melbourne1$Price, 40))  
#taking renaming the sample price collumn name  
Price\_sample<-Price\_sample %>% rename(Sample\_price=sample.housing.dataset\_melbourne1.Price..40.)  
  
# Boxplot for the sample price  
boxplot(Price\_sample$Sample\_price)



# Draw probabilities plot  
plot(density(Price\_sample$Sample\_price)) # gives the desity graph to check if data is normally



# quantile quantile plot tocheck if data is normally distibuted  
qqnorm(Price\_sample$Sample\_price)  
qqline(Price\_sample$Sample\_price)



As shown below clearly we can significance level is 5% T-statistic or t value is in negative and the P-value is greater than 0.5 we are rejecting the **Alternate Hypotheses** with the standard error of 5%.

t.test(Price\_sample$Sample\_price,mu=884864)

##   
## One Sample t-test  
##   
## data: Price\_sample$Sample\_price  
## t = -0.24894, df = 39, p-value = 0.8047  
## alternative hypothesis: true mean is not equal to 884864  
## 95 percent confidence interval:  
## 763314.3 979773.2  
## sample estimates:  
## mean of x   
## 871543.8

**2.** Test if the different region have same average mean number of rooms in the Melbourne data-set. using the concept of ANOVA.

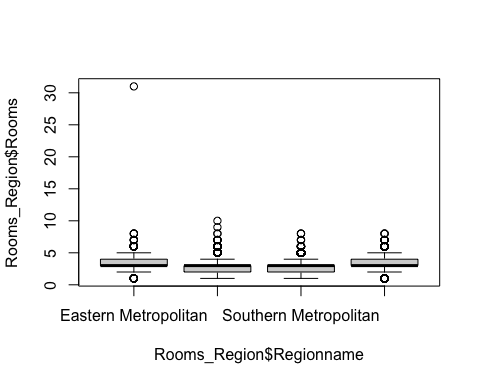
H0 = mean number of room in different region.

H1 != mean number of room in different region.

#creating the new data set or sub seting for different region and no of rooms  
Rooms\_Region<-data.frame(housing.dataset\_melbourne1%>%filter(Regionname %in%c("Northern Metropolitan","Southern Metropolitan","Western Metropolitan","Eastern Metropolitan"))%>%select(Rooms,Regionname))  
head(Rooms\_Region) # printing the result

## Rooms Regionname  
## 1 3 Northern Metropolitan  
## 2 3 Northern Metropolitan  
## 3 3 Northern Metropolitan  
## 4 3 Western Metropolitan  
## 5 2 Western Metropolitan  
## 6 2 Western Metropolitan

#box plot to understand the distribution of rooms in diffrent region  
boxplot(Rooms\_Region$Rooms~Rooms\_Region$Regionname)



We know when we need to compare more than 2 means we use ANOVA concept . As shown below since the p-value is very less hence we can confirm that or we reject the **Null hypothesis**

Rooms\_Region %>%aov(Rooms~Regionname,data = .) %>% summary()

## Df Sum Sq Mean Sq F value Pr(>F)   
## Regionname 3 2275 758.2 966.7 <2e-16 \*\*\*  
## Residuals 40302 31611 0.8   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

As shown below non of them include zero in its lower and upper limt and does not have possibility of have no difference at all and the p value is zero so we reject the **Null hypothesis**

#code for honestly significant difference for diffrent region with each other  
Rooms\_Region %>%aov(Rooms~Regionname,data = .) %>% TukeyHSD()

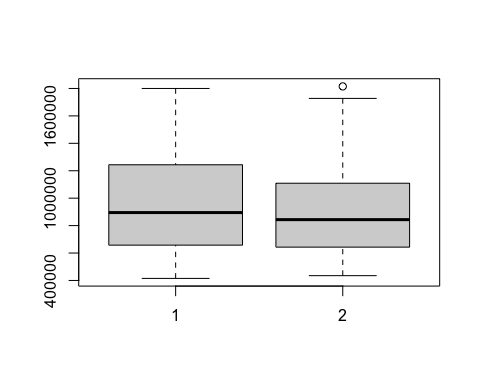
## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Rooms ~ Regionname, data = .)  
##   
## $Regionname  
## diff lwr upr  
## Northern Metropolitan-Eastern Metropolitan -0.4878805 -0.5210352 -0.4547258  
## Southern Metropolitan-Eastern Metropolitan -0.6789121 -0.7138060 -0.6440181  
## Western Metropolitan-Eastern Metropolitan -0.2498148 -0.2852223 -0.2144073  
## Southern Metropolitan-Northern Metropolitan -0.1910316 -0.2209813 -0.1610819  
## Western Metropolitan-Northern Metropolitan 0.2380657 0.2075192 0.2686122  
## Western Metropolitan-Southern Metropolitan 0.4290973 0.3966713 0.4615232  
## p adj  
## Northern Metropolitan-Eastern Metropolitan 0  
## Southern Metropolitan-Eastern Metropolitan 0  
## Western Metropolitan-Eastern Metropolitan 0  
## Southern Metropolitan-Northern Metropolitan 0  
## Western Metropolitan-Northern Metropolitan 0  
## Western Metropolitan-Southern Metropolitan 0

**3** Testing if the mean price value of the house type **h** and **t** are same or not (two sample t-test).

H0=> mean price value of **h** type house=mean price value of **t** type house.

H1=>mean price value of **h** type house !=mean price value of **t** type house.

set.seed(101)  
Price\_T\_Type<-subset(housing.dataset\_melbourne1,Type=='t')# subset on type T  
Price\_T\_Type\_Sample<-sample(Price\_T\_Type$Price,40)# taking the sample  
Price\_h\_Type<-subset(housing.dataset\_melbourne1,Type=='h')# subset on type H  
Price\_H\_Type\_Sample<-sample(Price\_h\_Type$Price,40)#taking the sample  
boxplot(Price\_H\_Type\_Sample,Price\_T\_Type\_Sample) # ceratinthe pox plot for the same



As two groups are involved we are using 2 sample t test. As shown below t value is in negative and P value is ver less so we reject the **Null hypotheses** at 95% confidence

#performing the t-test on price variable  
t.test(Price\_H\_Type\_Sample,Price\_T\_Type\_Sample,var.equal = T)

##   
## Two Sample t-test  
##   
## data: Price\_H\_Type\_Sample and Price\_T\_Type\_Sample  
## t = 0.394, df = 78, p-value = 0.6947  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -136999.7 204604.7  
## sample estimates:  
## mean of x mean of y   
## 979775.0 945972.5