library(dplyr)

setwd("C:\\Users\\ADMIN\\Desktop\\r\_project\_2")

train=read.csv("housing\_train.csv",stringsAsFactors = FALSE,header = T )

test=read.csv("housing\_test.csv",stringsAsFactors = FALSE,header = T )

apply(train,2,function(x)sum(is.na(x)))

train$Bedroom2[is.na(train$Bedroom2)]=median(train$Bedroom2,na.rm=T)

apply(train,2,function(x)sum(is.na(x)))

train$Bathroom[is.na(train$Bathroom)]=round(mean(train$Bathroom,na.rm=T),0)

apply(train,2,function(x)sum(is.na(x)))

train$Car[is.na(train$Car)]=round(mean(train$Car,na.rm=T),0)

apply(train,2,function(x)sum(is.na(x)))

train$Landsize[is.na(train$Landsize)]=round(mean(train$Landsize,na.rm=T),0)

apply(train,2,function(x)sum(is.na(x)))

train$BuildingArea[is.na(train$BuildingArea)]=round(mean(train$BuildingArea,na.rm=T),0)

apply(train,2,function(x)sum(is.na(x)))

train$YearBuilt[is.na(train$YearBuilt)]=round(mean(train$YearBuilt,na.rm=T),0)

apply(train,2,function(x)sum(is.na(x)))

apply(test,2,function(x)sum(is.na(x)))

test$Bedroom2[is.na(test$Bedroom2)]=median(test$Bedroom2,na.rm=T)

apply(test,2,function(x)sum(is.na(x)))

test$Bathroom[is.na(test$Bathroom)]=round(mean(test$Bathroom,na.rm=T),0)

apply(test,2,function(x)sum(is.na(x)))

test$Car[is.na(test$Car)]=round(mean(test$Car,na.rm=T),0)

apply(test,2,function(x)sum(is.na(x)))

test$Landsize[is.na(test$Landsize)]=round(mean(test$Landsize,na.rm=T),0)

apply(test,2,function(x)sum(is.na(x)))

test$BuildingArea[is.na(test$BuildingArea)]=round(mean(test$BuildingArea,na.rm=T),0)

apply(test,2,function(x)sum(is.na(x)))

test$YearBuilt[is.na(test$YearBuilt)]=round(median(test$YearBuilt,na.rm=T),0)

test$Price=NA

train$data='train'

test$data='test'

all\_data=rbind(train,test)

apply(all\_data,2,function(x)sum(is.na(x)))

glimpse(all\_data)

t=table(all\_data$Suburb)

View(t)

t1=round(tapply(all\_data$Price,all\_data$Suburb,mean,na.rm=T),0)

View(t1)

t1=sort(t1)

all\_data=all\_data %>%

mutate(

sub\_1=as.numeric(Suburb%in%c("Campbellfield","Jacana")),

sub\_2=as.numeric(Suburb%in%c("Kealba","Brooklyn","Albion","Sunshine West","Ripponlea","Fawkner")),

sub\_3=as.numeric(Suburb%in%c("Glenroy","Southbank","Sunshine North","Keilor Park","Heidelberg West","Reservoir","Braybrook","Kingsbury","Gowanbrae","Hadfield","Watsonia","Footscray","South Kingsville","Balaclava","Melbourne","Maidstone","Sunshine")),

sub\_4=as.numeric(Suburb%in%c("Airport West","Heidelberg Heights","Pascoe Vale","West Footscray","Altona North","Williamstown North","Brunswick West","Keilor East","Oak Park","Maribyrnong","Altona","Flemington","Coburg North","Yallambie","Avondale Heights","Bellfield")),

sub\_5=as.numeric(Suburb%in%c("Strathmore Heights","Glen Huntly","Kensington","Essendon North","St Kilda","Preston","North Melbourne","Coburg","Kingsville","Collingwood","Brunswick East","Gardenvale","Thornbury","Niddrie","West Melbourne","Viewbank")),

sub\_6=as.numeric(Suburb%in%c("Spotswood","Carnegie","Elwood","Heidelberg","Moorabbin","Oakleigh","Rosanna","Docklands","Yarraville","Cremorne","Seddon","Brunswick","Oakleigh South","Ascot Vale","Windsor","Caulfield","Essendon West","Newport")),

sub\_7=as.numeric(Suburb%in%c("Chadstone","South Yarra","Essendon","Bentleigh East","Murrumbeena","Hughesdale","Fairfield","Ashwood","Clifton Hill","Caulfield North","Abbotsford","Carlton","Prahran","Fitzroy","Ivanhoe","Hampton East","Caulfield East")),

sub\_8=as.numeric(Suburb%in%c("Richmond","Travancore","Templestowe Lower","Ormond","Caulfield South","Moonee Ponds","Hawthorn","Box Hill","Bulleen","Burnley","Burwood","Strathmore","Port Melbourne","Fitzroy North","Alphington")),

sub\_9=as.numeric(Suburb%in%c("Doncaster","South Melbourne","Northcote","Aberfeldie","Elsternwick","Bentleigh","Kooyong","Parkville")),

sub\_10=as.numeric(Suburb%in%c("Williamstown","East Melbourne","Seaholme")),

sub\_11=as.numeric(Suburb%in%c("Malvern East","Carlton North","Hawthorn East","Surrey Hills")),

sub\_12=as.numeric(Suburb%in%c("Princes Hill","Mont Albert","Armadale","Kew East","Glen Iris","Ashburton")),

sub\_13=as.numeric(Suburb%in%c("Brighton East","Eaglemont","Hampton")),

sub\_14=as.numeric(Suburb%in%c("Toorak","Ivanhoe East","Camberwell","Balwyn North","Kew")),

sub\_15=as.numeric(Suburb%in%c("Brighton","Middle Park")),

sub\_16=as.numeric(Suburb%in%c("Albert Park","Balwyn","Malvern"))

) %>%

select(-Suburb)

glimpse(all\_data)

all\_data=all\_data %>%

select(-Address)

glimpse(all\_data)

table(all\_data$Type)

all\_data=all\_data %>%

mutate(Type\_t=as.numeric(Type=="t"),

type\_u=as.numeric(Type=="u"))

all\_data=all\_data %>%

select(-Type)

glimpse(all\_data)

table(all\_data$Method)

all\_data=all\_data %>%

mutate(Method\_PI=as.numeric(Method=="PI"),

Method\_SA=as.numeric(Method=="SA"),

Method\_SP=as.numeric(Method=="SP"),

Method\_VB=as.numeric(Method=="VB")) %>%

select(-Method)

glimpse(all\_data)

t=table(all\_data$SellerG)

sort(t)

all\_data=all\_data %>%

mutate(Gnelson=as.numeric(SellerG=="Nelson"),

GJellis=as.numeric(SellerG=="Jellis"),

Ghstuart=as.numeric(SellerG=="hockingstuart"),

Gbarry=as.numeric(SellerG=="Barry"),

GMarshall=as.numeric(SellerG=="Marshall"),

GWoodards=as.numeric(SellerG=="Woodards"),

GBrad=as.numeric(SellerG=="Brad"),

GBiggin=as.numeric(SellerG=="Biggin"),

GRay=as.numeric(SellerG=="Ray"),

GFletchers=as.numeric(SellerG=="Fletchers"),

GRT=as.numeric(SellerG=="RT"),

GSweeney=as.numeric(SellerG=="Sweeney"),

GGreg=as.numeric(SellerG=="Greg"),

GNoel=as.numeric(SellerG=="Noel"),

GGary=as.numeric(SellerG=="Gary"),

GJas=as.numeric(SellerG=="Jas"),

GMiles=as.numeric(SellerG=="Miles"),

GMcGrath=as.numeric(SellerG=="McGrath"),

GHodges=as.numeric(SellerG=="Hodges"),

GKay=as.numeric(SellerG=="Kay"),

GStockdale=as.numeric(SellerG=="Stockdale"),

GLove=as.numeric(SellerG=="Love"),

GDouglas=as.numeric(SellerG=="Douglas"),

GWilliams=as.numeric(SellerG=="Williams"),

GVillage=as.numeric(SellerG=="Village"),

GRaine=as.numeric(SellerG=="Raine"),

GRendina=as.numeric(SellerG=="Rendina"),

GChisholm=as.numeric(SellerG=="Chisholm"),

GCollins=as.numeric(SellerG=="Collins"),

GLITTLE=as.numeric(SellerG=="LITTLE"),

GNick=as.numeric(SellerG=="Nick"),

GHarcourts=as.numeric(SellerG=="Harcourts"),

GCayzer=as.numeric(SellerG=="Cayzer"),

GMoonee=as.numeric(SellerG=="Moonee"),

GYPA=as.numeric(SellerG=="YPA")

) %>%

select(-SellerG)

glimpse(all\_data)

table(all\_data$CouncilArea)

all\_data=all\_data %>%

mutate(CA\_Banyule=as.numeric(CouncilArea=="Banyule"),

CA\_Bayside=as.numeric(CouncilArea=="Bayside"),

CA\_Boroondara=as.numeric(CouncilArea=="Boroondara"),

CA\_Brimbank=as.numeric(CouncilArea=="Brimbank"),

CA\_Darebin=as.numeric(CouncilArea=="Darebin"),

CA\_Glen\_Eira=as.numeric(CouncilArea=="Glen Eira"),

CA\_Monash=as.numeric(CouncilArea=="Monash"),

CA\_Melbourne=as.numeric(CouncilArea=="Melbourne"),

CA\_Maribyrnong=as.numeric(CouncilArea=="Maribyrnong"),

CA\_Manningham=as.numeric(CouncilArea=="Manningham"),

CA\_Kingston=as.numeric(CouncilArea=="Kingston"),

CA\_Hume=as.numeric(CouncilArea=="Hume"),

CA\_HobsonsB=as.numeric(CouncilArea=="Hobsons Bay"),

CA\_MoonValley=as.numeric(CouncilArea=="Moonee Valley"),

CA\_Moreland=as.numeric(CouncilArea=="Moreland"),

CA\_PortP=as.numeric(CouncilArea=="Port Phillip"),

CA\_Stonnington=as.numeric(CouncilArea=="Stonnington"),

CA\_Whitehorse=as.numeric(CouncilArea=="Whitehorse"),

CA\_Yarra=as.numeric(CouncilArea=="Yarra")) %>%

select(-CouncilArea)

glimpse(all\_data)

train=all\_data %>%

filter(data=='train') %>%

select(-data)

test=all\_data %>%

filter(data=='test') %>%

select(-data,-Price)

glimpse(train)

glimpse(test)

set.seed(123)

s=sample(1:nrow(train),0.75\*nrow(train))

train\_75=train[s,] #5652

test\_25=train[-s,] #1884

library(car)

LRf=lm(Price ~ .,data=train\_75)

summary(LRf)

a=vif(LRf)

sort(a,decreasing = T)[1:3]

LRf=lm(Price ~ .-Postcode-sub\_3,data=train\_75)

summary(LRf)

a=vif(LRf)

sort(a,decreasing = T)[1:3]

summary(LRf)

LRf=lm(Price ~ .-Landsize-GRaine-GMoonee-CA\_Bayside-GLITTLE-Gnelson-GSweeney-Ghstuart-CA\_Kingston-Gbarry-GRay-GStockdale-GNoel-GJas-GBiggin-GYPA-CA\_PortP-CA\_Whitehorse-GRendina-GFletchers-GBrad-GHodges-GVillage-GLove-sub\_4-GGary-CA\_Hume-CA\_Boroondara-Method\_SA-GWilliams-GHarcourts-GNick-GGreg-CA\_Monash-GWoodards-CA\_Stonnington-GCayzer-Postcode-sub\_3,data=train\_75)

summary(LRf)

PP\_test\_25=predict(LRf,newdata =test\_25)

PP\_test\_25=round(PP\_test\_25,1)

class(PP\_test\_25)

plot(test\_25$Price,PP\_test\_25)

res=test\_25$Price-PP\_test\_25 #(real value-predicted value)

#root mean square error is as follows

RMSE\_test\_25=sqrt(mean(res^2))

RMSE\_test\_25

212467/RMSE\_test\_25

library(ggplot2)

d=data.frame(real=test\_25$Price,predicted=PP\_test\_25)

ggplot(d,aes(x=real,y=predicted))+geom\_point()

plot(LRf,which = 1) #gives residual vz fitted plot

plot(LRf,which = 2) #gives q-q-plot

plot(LRf,which = 3) #gives scale-location plot

plot(LRf,which = 4) #gives cooks distance

Ganesh\_Patil=predict(LRf,newdata =test)

Ganesh\_Patil=round(Ganesh\_Patil)

class(Ganesh\_Patil)

write.csv(Ganesh\_Patil, "Ganesh\_Patil\_P1\_Part2.csv")

summary(LRf)