Project 1: Prediction Model for house price prediction using Predictive analytics in R.

Problem Statement : Price of a property is one of the most important decision criterion when people buy homes. Real state firms need to be consistent in their pricing in order to attract buyers . Having a predictive model for the same will be great tool to have , which in turn can also be used to tweak development of properties , putting more emphasis on qualities which increase the value of the property.
We have given you two datasets , housing_train.csv and housing_test.csv . You need to use data housing_train to build predictive model for response variable "Price". Housing_test data contains all other factors except "Price", you need to predict that using the model that you developed and submit your predicted values in a csv files.

Solution:

```
setwd("E:/R/real estate project")

library(dplyr)
train=read.csv("housing_train.csv",stringsAsFactors = FALSE,header = T )
#7536 obs,16 variables
test=read.csv("housing_test.csv",stringsAsFactors = FALSE,header = T )
#1885 obs,15 variables
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)
apply(test,2,function(x)sum(is.na(x)))

#Step 2:Data Preparation
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","Oa
kleigh","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)  #9421obs and 16 variables

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)
```

```
#thus train has total obs as 7536 and 70 variables (69+price)

test=all_data %>%
  filter(data=='test') %>%
  select(-data,-Price)#thus test data has original obs 1885 and added new dummy variables totalling to 69 variables

glimpse(train) #7536 obs and 86 variables.
glimpse(test) #1885 obs and 85 variables.

set.seed(123)
s=sample(1:nrow(train),0.75*nrow(train))
train_75=train[s,] #5652
test_25=train[-s,]  #1884

#Step 3: Model Building

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)

#step4: performance measurement of model
PP_test_25=predict(LRf,newdata =test_25)
PP_test_25=round(PP_test_25,1)
class(PP_test_25)

#lets plot the real price vs predicted price for dataset 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

#step5: predict real estate prices for the  final test dataset
PP_test_final=predict(LRf,newdata =test)
PP_test_final=round(PP_test_final,1)
class(PP_test_final)

write.csv(PP_test_final, "price prediction_house_final.csv") #stores the predicted prices in a csv file
on your local repository in pc.

summary(LRf)
```