getwd()

setwd("F:\\files\\r project")

rs\_train=read.csv("housing\_train.csv ", stringsAsFactors = FALSE)

rs\_test=read.csv("housing\_test.csv ", stringsAsFactors = FALSE)

#Combining train & test

rs\_test$Price=NA

rs\_train$data='train'

rs\_test$data='test'

rs\_all=rbind(rs\_test,rs\_train)

#data preparation

CreateDummies=function(data,var,freq\_cutoff=100){

t=table(data[,var])

t=t[t>freq\_cutoff]

t=sort(t)

categories=names(t)[-1]

for( cat in categories){

name=paste(var,cat,sep="\_")

name=gsub(" ","",name)

name=gsub("-","\_",name)

name=gsub("\\?","Q",name)

name=gsub("<","LT\_",name)

name=gsub("\\+","",name)

name=gsub(">","GT\_",name)

name=gsub("=","EQ\_",name)

name=gsub(",","",name)

name=gsub("/","\_",name)

data[,name]=as.numeric(data[,var]==cat)

}

data[,var]=NULL

return(data)

}

char\_logical=sapply(rs\_all,is.character)

cat\_cols=names(rs\_all)[char\_logical]

cat\_cols

cat\_cols=cat\_cols[!(cat\_cols %in% c('data','Price'))]

cat\_cols

for(col in cat\_cols){

rs\_all=CreateDummies(rs\_all,col,50)

}

library(tidyverse)

str(rs\_all)

rs\_all=rs\_all[!((is.na(rs\_all$Price)) & rs\_all$data=='train'), ]

for(col in names(rs\_all)){

if(sum(is.na(rs\_all[,col]))>0 & !(col %in% c("data","price"))){

rs\_all[is.na(rs\_all[,col]),col]=mean(rs\_all[rs\_all$data=='train',col],na.rm=T)

}

}

library(dplyr)

rs\_train=rs\_all %>% filter(data=='train') %>% select(-data)

rs\_test=rs\_all %>% filter(data=='test') %>% select(-data,-Price)

set.seed(2)

s=sample(1:nrow(rs\_train),0.8\*nrow(rs\_train))

train=rs\_train[s,]

val=rs\_train[-s,]

library(car)

fit=lm(Price~.,data=train)

summary(fit)

vif(fit)

sort(vif(fit),decreasing = T)

fit=lm(Price~.-CouncilArea\_ ,data=train)

sort(vif(fit),decreasing = T)

fit=lm(Price~.-CouncilArea\_-Postcode ,data=train)

sort(vif(fit),decreasing = T)

fit=lm(Price~.-CouncilArea\_-Postcode-Distance ,data=train)

sort(vif(fit),decreasing = T)

fit=step(fit)

summary(fit)

val.pred=predict(fit,newdata = val)

errors=val$Price-val.pred

rmse=errors\*\*2 %>% mean() %>% sqrt()

rmse

Score = 212467/rmse

Score

fit.final=lm(Price~.,data=rs\_train)

fit.final=step(fit.final)

test.predictions=predict(fit.final,newdata=rs\_test)

write.csv(test.predictions,'Dharma\_dash\_P1\_part2.csv',row.names = F)