Appendix

code for 1

```
ta1 <- matrix(c( 0.1910393,0.1767595,0.15385,15,13,12),ncol=3,byrow=TRUE)
colnames(ta1) <- c("smoothing","random forest","boosting")
rownames(ta1) <- c("RMSE","Rank")
print("RMSE of Prediction On kaggle With Rank")
as.table(ta1)

ta2 <- matrix(c( 0.1910393,0.1767595,0.15385),ncol=3,byrow=TRUE)
colnames(ta2) <- c("smoothing","random forest","boosting")
rownames(ta2) <- c("RMSE")
print("Prediction For This report")
as.table(ta2)
```

code for 3

```
# input of data here


data = read.csv('housing_price.csv')
library('VIM') # Missing value
library(gbm)
library(mgcv)
```

code for 4.1

```
miss <- aggr(data,prop = FALSE, combined = TRUE, sortVars=TRUE)
```

c

```
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

data$KITCHENS[is.na(data$KITCHENS)] <- getmode(data$KITCHENS[!is.na(data$KITCHENS)])



med_diff_built_remodel <-floor(median(data$YR_RMDL[!is.na(data$YR_RMDL)&!is.na(data$AYB)]-
                                data$AYB[!is.na(data$YR_RMDL)&!is.na(data$AYB)] ))


# Fill AYB and YR_RMDL
data$YR_RMDL[is.na(data$YR_RMDL)&!is.na(data$AYB)] <-data$AYB[is.na(data$YR_RMDL)&!is.na(data$AYB)]


#Missing Both
```

```
missing_both <- is.na(data$YR_RMDL)&is.na(data$AYB)

data$AYB[missing_both]<-floor(median(data$AYB[!is.na(data$AYB)]))

data$YR_RMDL[missing_both]<-data$AYB[missing_both]- med_diff_built_remodel


missing_built_have_remodel <- (!is.na(data$YR_RMDL)&is.na(data$AYB))
data$AYB[missing_built_have_remodel] <- data$YR_RMDL[missing_built_have_remodel]-med_diff_built_remodel

data$STORIES[is.na(data$STORIES)]<-floor(median(data$STORIES[!is.na(data$STORIES)]))}
```

```
data[data$YR_RMDL==20,]
data$YR_RMDL[data$YR_RMDL==20]=data$AYB[data$YR_RMDL==20]+med_diff_built_remodel
data$AC[data$AC ==0] <- getmode(data$AC[data$AC !=0])
data$STORIES[data$STORIES>=14] <- floor(median(data$STORIES[data$STORIES<14]))
```

code for 5.1

```
data$AC <- factor(data$AC,level=c('Y','N'), label=c(1,0))

data$GRADE <- as.numeric(factor(data$GRADE,level=c('Low Quality', 'Fair Quality', 'Average', 'Above Ave

data$CNDTN <- as.numeric(factor(data$CNDTN,level=c('Poor', 'Fair', 'Average', 'Good', 'Very Good','Excel



data$NATIONALGRID <- as.numeric(data$NATIONALGRID)

data$ASSESSMENT_NBHD <- as.factor(data$ASSESSMENT_NBHD)

data$STYLE <- as.numeric(data$STYLE)

#HEAT
#data$STRUCT <- as.numeric(data$STRUCT)

#data$EXTWALL <- as.numeric(data$EXTWALL)
#data$INTWALL <- as.numeric(data$INTWALL)
#data$ROOF <- as.numeric(data$ROOF)
#data$WARD <- as.numeric(data$WARD)
#data$QUADRANT <- as.numeric(data$QUADRANT)



data$HEAT <- as.character(data$HEAT)
data$HEAT[data$HEAT=='Air-oil'|
                data$HEAT=='Electric Rad'|
                 data$HEAT=='Evp Cool'|
                 data$HEAT=='Gravity Furnac'|
                 data$HEAT=='Ind Unit'|
                 data$HEAT=='No Data'|
                 data$HEAT=='Wall Furnace'
```

```
                  ] <- sample(data$HEAT[data$HEAT!='Air-oil'& data$HEAT=='Electric Rad' & data$HEAT!='Ev
                  data$HEAT!='Gravity Furnac'&
                  data$HEAT!='Ind Unit'&
                  data$HEAT!='No Data'&
                  data$HEAT!='Wall Furnace'
                  ],size = 8)
data$HEAT<-as.factor(data$HEAT)


#EXTWALL
#data$EXTWALL[data$EXTWALL == 'Adobe'| data$EXTWALL == 'Default'| data$EXTWALL == 'Plywood' ] <- sample
#data$EXTWALL<-as.factor(data$EXTWALL)



data$QUADRANT[data$QUADRANT == ""] <- sample(data$QUADRANT[data$QUADRANT != ""],size = 65)

#data$INTWALL[data$INTWALL == 'Vinyl Comp'] <- 'Carpet'

library(gbm)
library(mgcv)

# year rebuild
data$SALEYEAR <- as.numeric(substr(data$SALEDATE,0,4))
data$SALEMONTH <- as.numeric(substr(data$SALEDATE,6,7))
data$SALEDATE <- as.numeric(data$SALEDATE)

RMLSE_Score <- function(real,pred, take_log = TRUE){
  if (take_log){
    print(sqrt(1/length(real)* sum( (log(real) -log(pred))^2 ,na.rm=TRUE )))
  }else{
    print(sqrt(1/length(real)* sum( (real -pred)^2 ,na.rm=TRUE )))
  }
}

pairs(~data$HF_BATHRM + PRICE, data = data)
```

code for 5.2

```
t1 <- gam(PRICE~ s(BATHRM) + s(ROOMS) + s(BEDRM)+ s(AYB)  + s(YR_RMDL)+ s(EYB) + s(STORIES)+s(STYLE) + s
summary(t1)


t2 <- gam(PRICE~ s(BATHRM) + s(ROOMS) + s(BEDRM) + HEAT +
EXTWALL+ROOF+INTWALL + AC + STRUCT + KITCHENS + USECODE + ASSESSMENT_NBHD + WARD + QUADRANT , data=data
summary(t2, maxsum = 1)


#########################################################################
xg_at1 <- function(fold_num){
  train<-data[data$fold !=fold_num,]
  train <- subset(train, select= - c(Id,fold, ASSESSMENT_SUBNBHD,FULLADDRESS))
```

```
  train$PRICE = log(train$PRICE)

    gam.object<- gam(PRICE~ s(BATHRM) + s(ROOMS) + s(BEDRM)+ s(AYB)  + s(YR_RMDL)+ s(EYB) + s(STORIES)+s
         (STYLE) + s(FIREPLACES) + s(LANDAREA) + s(ZIPCODE) + s(LATITUDE) + s(LONGITUDE) + s(CENSUS_TRAC


  test<-data[data$fold ==fold_num,]

  test_price <- log(test$PRICE)


  test<-subset(test, select=-c(Id,fold, ASSESSMENT_SUBNBHD,FULLADDRESS))

  predict_price<-predict(gam.object,test)

  return (cbind(data[data$fold==fold_num,]$Id,predict_price,test_price))
}
at1_1 <-xg_at1(1)
at1_2 <-xg_at1(2)
at1_3 <-xg_at1(3)
at1_4 <-xg_at1(4)
at1_5 <-xg_at1(5)
total<-data.frame(rbind(at1_1,at1_2,at1_3,at1_4,at1_5))
RMLSE_Score(total$test_price,total$predict_price, FALSE)

#########################################################################
xg_at2 <- function(fold_num){
  train<-data[data$fold !=fold_num,]
  train <- subset(train, select= - c(Id,fold, ASSESSMENT_SUBNBHD,FULLADDRESS))

  train$PRICE = log(train$PRICE)

    gam.object<- gam(PRICE~ s(BATHRM)+ HF_BATHRM + I(HF_BATHRM^2) + AC + s(ROOMS) + s(BEDRM)+ s(AYB)  +


  test<-data[data$fold ==fold_num,]

  test_price <- log(test$PRICE)


  test<-subset(test, select=-c(Id,fold, ASSESSMENT_SUBNBHD,FULLADDRESS))

  predict_price<-predict(gam.object,test)

  return (cbind(data[data$fold==fold_num,]$Id,predict_price,test_price))
}
at1_1 <-xg_at2(1)
at1_2 <-xg_at2(2)
at1_3 <-xg_at2(3)
at1_4 <-xg_at2(4)
at1_5 <-xg_at2(5)
```

```r
total<-data.frame(rbind(at1_1,at1_2,at1_3,at1_4,at1_5))
RMLSE_Score(total$test_price,total$predict_price, FALSE)
```

code for 5.3

```r
pairs(~ SALEDATE + ZIPCODE + CENSUS_TRACT + LATITUDE + LONGITUDE + LANDAREA + PRICE, data = data)

pairs(~ BATHRM+ HF_BATHRM + I(HF_BATHRM^2) + AC + ROOMS + BEDRM+ AYB  + YR_RMDL+EYB + PRICE, data = data


xg_at1 <- function(fold_num){
  train<-data[data$fold !=fold_num,]
  train <- subset(train, select= - c(Id,fold, ASSESSMENT_SUBNBHD,FULLADDRESS))
  train$PRICE = log(train$PRICE)
    gam.object <- gam(PRICE~ s(BATHRM)+ HF_BATHRM + I(HF_BATHRM^2) + AC + s(ROOMS) + s(BEDRM)+ s(AYB)
                      + STRUCT #0.1949005
                      + GRADE
                      + WARD # 0.1945048
                      + QUADRANT #0.1944545 #[1] 0.1915153
                       , data=train)
  test<-data[data$fold ==fold_num,]

  test_price <- log(test$PRICE)


  test<-subset(test, select=-c(Id,fold, ASSESSMENT_SUBNBHD,FULLADDRESS))

  predict_price<-predict(gam.object,test)
  return (cbind(data[data$fold==fold_num,]$Id,predict_price,test_price))
}
at1_1 <-xg_at1(1)
at1_2 <-xg_at1(2)
at1_3 <-xg_at1(3)
at1_4 <-xg_at1(4)
at1_5 <-xg_at1(5)
total<-data.frame(rbind(at1_1,at1_2,at1_3,at1_4,at1_5))
RMLSE_Score(total$test_price,total$predict_price, FALSE)



######################################################################
foul_num <- 1
xg_at1 <- function(fold_num){
  train<-data[data$fold !=fold_num,]
  train <- subset(train, select= - c(Id,fold, ASSESSMENT_SUBNBHD,FULLADDRESS))
  train$PRICE = log(train$PRICE)
    gam.object <- gam(PRICE~ s(BATHRM)+ HF_BATHRM + I(HF_BATHRM^2) + AC + s(ROOMS) + s(BEDRM)+ s(AYB)
                      + STRUCT #0.1949005
                      + GRADE
                      + WARD # 0.1945048
                      + QUADRANT #0.1944545 #[1] 0.1915153
                      + I(ROOMS^2)
```

```
                      , data=train)
  test<-data[data$fold ==fold_num,]
  test_price <- log(test$PRICE)
  test<-subset(test, select=-c(Id,fold, ASSESSMENT_SUBNBHD,FULLADDRESS))
  predict_price<-predict(gam.object,test)
  return (cbind(data[data$fold==fold_num,]$Id,predict_price,test_price))
}
xg_at2<- function(fold_num){
  train<-data[data$fold !=fold_num,]
  train <- subset(train, select= - c(Id,fold, ASSESSMENT_SUBNBHD,FULLADDRESS))
  train$PRICE = log(train$PRICE)
    gam.object <- gam(PRICE~ s(BATHRM)+ HF_BATHRM + I(HF_BATHRM^2) + AC + s(ROOMS) + s(BEDRM)+ s(AYB)
                      + STRUCT #0.1949005
                      + GRADE
                      + WARD # 0.1945048
                      + QUADRANT #0.1944545 #[1] 0.1915153
                      + I(BEDRM^2)
                       , data=train)
  test<-data[data$fold ==fold_num,]
  test_price <- log(test$PRICE)
  test<-subset(test, select=-c(Id,fold, ASSESSMENT_SUBNBHD,FULLADDRESS))
  predict_price<-predict(gam.object,test)
  return (cbind(data[data$fold==fold_num,]$Id,predict_price,test_price))
}
xg_at3 <- function(fold_num){
  train<-data[data$fold !=fold_num,]
  train <- subset(train, select= - c(Id,fold, ASSESSMENT_SUBNBHD,FULLADDRESS))
  train$PRICE = log(train$PRICE)
    gam.object <- gam(PRICE~ s(BATHRM)+ HF_BATHRM + I(HF_BATHRM^2) + AC + s(ROOMS) + s(BEDRM)+ s(AYB)
                      + STRUCT #0.1949005
                      + GRADE
                      + WARD # 0.1945048
                      + QUADRANT #0.1944545 #[1] 0.1915153
                      + I(LATITUDE^2)
                       , data=train)
  test<-data[data$fold ==fold_num,]
  test_price <- log(test$PRICE)
  test<-subset(test, select=-c(Id,fold, ASSESSMENT_SUBNBHD,FULLADDRESS))
  predict_price<-predict(gam.object,test)
  return (cbind(data[data$fold==fold_num,]$Id,predict_price,test_price))
}
at1_1 <-xg_at1(1)
at1_2 <-xg_at1(2)
at1_3 <-xg_at1(3)
at1_4 <-xg_at1(4)
at1_5 <-xg_at1(5)
total<-data.frame(rbind(at1_1,at1_2,at1_3,at1_4,at1_5))
RMLSE_Score(total$test_price,total$predict_price, FALSE)
at1_1 <-xg_at2(1)
at1_2 <-xg_at2(2)
at1_3 <-xg_at2(3)
at1_4 <-xg_at2(4)
at1_5 <-xg_at2(5)
```

```
total<-data.frame(rbind(at1_1,at1_2,at1_3,at1_4,at1_5))
RMLSE_Score(total$test_price,total$predict_price, FALSE)
at1_1 <-xg_at3(1)
at1_2 <-xg_at3(2)
at1_3 <-xg_at3(3)
at1_4 <-xg_at3(4)
at1_5 <-xg_at3(5)
total<-data.frame(rbind(at1_1,at1_2,at1_3,at1_4,at1_5))
RMLSE_Score(total$test_price,total$predict_price, FALSE)

#########################################################################
#final k
xg_at1 <- function(fold_num){
  train<-data[data$fold !=fold_num,]
  train <- subset(train, select= - c(Id,fold, ASSESSMENT_SUBNBHD,FULLADDRESS))
  train$PRICE = log(train$PRICE)
gam.object <- gam(PRICE~ s(BATHRM)+ HF_BATHRM + I(HF_BATHRM^2) + AC + s(ROOMS) + s(BEDRM)+ s(AYB)  + s(
                + STRUCT
                + GRADE
                + WARD
                + QUADRANT
                + I(BEDRM^2)
                , data=train)

  test<-data[data$fold ==fold_num,]

  test_price <- log(test$PRICE)


  test<-subset(test, select=-c(Id,fold, ASSESSMENT_SUBNBHD,FULLADDRESS))

  predict_price<-predict(gam.object,test)
  return (cbind(data[data$fold==fold_num,]$Id,predict_price,test_price))
}
at1_1 <-xg_at1(1)
at1_2 <-xg_at1(2)
at1_3 <-xg_at1(3)
at1_4 <-xg_at1(4)
at1_5 <-xg_at1(5)

total<-data.frame(rbind(at1_1,at1_2,at1_3,at1_4,at1_5))
RMLSE_Score(total$test_price,total$predict_price, FALSE)


sm <- gam(PRICE~ s(BATHRM)+ HF_BATHRM + I(HF_BATHRM^2) + AC + s(ROOMS) + s(BEDRM)+ s(AYB)  + s(YR_RMDL)+
                    + STRUCT
                    + GRADE
                    + WARD
                    + QUADRANT
                    + I(BEDRM^2)
                    , data=data)

summary(sm)
```

code for 5.3