Final Project

Rafi Rahman (u3191010)

09/05/2021

Abstract

This report analyses the Ames Housing statistics from 2006 to 2010. The data is cleaned up and explored. Along with providing a linear regression model and estimation of sale prices, questions about the relevance of the data variables to the sale prices of properties are asked and addressed. The study ends with suggestions and findings based on the report's material.

Contents

1 Problem Identification

1.1 Background Research

Dean De Cock compiled the Ames Housing dataset for use in data studies. It is an updated and modernised version of the commonly studied Boston Housing dataset. From 2006 to 2010, it details the selling of individual residential properties in Ames, Iowa. The data collection includes 2930 observations as well as a significant range of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) that are used to determine home values [1].

Points of domain expertise are:

1. Season - As this is an American dataset, the seasonal periods are different from us here in Australia. Thus, we can create a new variable for the season sold which relates to the already existing variable for the month sold, i.e., we can find which season occurs in which months. This variable will have to use American terminology, such as "fall" instead of "autumn". The purpose of this new variable is to create more plots and show new relationships and trends between variables, such as the fluctuations in sale price depending on the season.
2. Total Interior Area - This variable can be utilised as a way to show the combined area of the entire house and will be related to the variables 1stFlrSF, 2ndFlrSF, and TotalBsmtSF. We can use this variable to also create more plots and explore the relationship/trend of the total area and the sale price.

As a side note, I have decided to create these variables before EDA, as I want to test these variables to see if they will actually be of use for training the model.

1.2 Questions of Interest

1. How does proximity to main road/railroad affect the sale price? As an extension, is the closeness to main road/railroad considered a benefit or a problem to the general buyer?
2. How much does total interior area affect the sale price? As an extension, would this mean that basement area is also important to customers' buying decisions?
3. Are fireplaces only a feature of the older houses? If so, how does the number of fireplaces affect the sale price?
4. For houses with pools, do they sell better during the summer (season variable)? As an extension, is this an indicator that people think more in the short term than the long term when buying houses?
5. What is the trend in sale price for each neighborhood over the years? Which neighborhood had the biggest growth and which neighborhood had the smallest growth in sale price?

2 Data Preprocessing

```
setwd("C:/Users/rafir/OneDrive/Documents/R Studio")

library(tidyverse)
library(ggplot2)
library(dplyr)
library(reshape2)
library(ggpmisc)
library(modelr)
library(gbm)

# Add CSV files as tibbles
train <- read_csv("train.csv")
test <- read_csv("test.csv")

# Check NA values
sapply(train, function(i) sum(is.na(i)*100/nrow(train)))
```

```
##            Id    MSSubClass      MSZoning   LotFrontage       LotArea
##    0.00000000    0.00000000    0.00000000   17.73972603    0.00000000
##        Street         Alley      LotShape   LandContour     Utilities
##    0.00000000   93.76712329    0.00000000    0.00000000    0.00000000
##     LotConfig     LandSlope  Neighborhood    Condition1    Condition2
##    0.00000000    0.00000000    0.00000000    0.00000000    0.00000000
##      BldgType    HouseStyle   OverallQual   OverallCond     YearBuilt
##    0.00000000    0.00000000    0.00000000    0.00000000    0.00000000
##  YearRemodAdd     RoofStyle      RoofMatl   Exterior1st   Exterior2nd
##    0.00000000    0.00000000    0.00000000    0.00000000    0.00000000
##    MasVnrType    MasVnrArea     ExterQual     ExterCond    Foundation
##    0.54794521    0.54794521    0.00000000    0.00000000    0.00000000
##      BsmtQual      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1
##    2.53424658    2.53424658    2.60273973    2.53424658    0.00000000
##  BsmtFinType2    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating
##    2.60273973    0.00000000    0.00000000    0.00000000    0.00000000
##     HeatingQC    CentralAir    Electrical      1stFlrSF      2ndFlrSF
##    0.00000000    0.00000000    0.06849315    0.00000000    0.00000000
##  LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
##    0.00000000    0.00000000    0.00000000    0.00000000    0.00000000
##      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd
##    0.00000000    0.00000000    0.00000000    0.00000000    0.00000000
##    Functional    Fireplaces   FireplaceQu    GarageType   GarageYrBlt
##    0.00000000    0.00000000   47.26027397    5.54794521    5.54794521
##  GarageFinish    GarageCars    GarageArea    GarageQual    GarageCond
##    5.54794521    0.00000000    0.00000000    5.54794521    5.54794521
##    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch     3SsnPorch
##    0.00000000    0.00000000    0.00000000    0.00000000    0.00000000
##   ScreenPorch      PoolArea        PoolQC         Fence   MiscFeature
##    0.00000000    0.00000000   99.52054795   80.75342466   96.30136986
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##    0.00000000    0.00000000    0.00000000    0.00000000    0.00000000
##     SalePrice
##    0.00000000
```

```r
# Replace NA values of character variables with "None"
for (v in c(1:ncol(train))) {
    if (sapply(train[v], is.character)) {
        train[v][is.na(train[v])] <- "None"
    }
}

# As LotFrontage has a significant number of values, we use the median
for(i in 1:nrow(train)) {
  if(is.na(train[i, 'LotFrontage'])){
    select_area = train[i, 'Neighborhood']
    j = train %>% dplyr::filter(Neighborhood == select_area)
    con_air = median(j$LotFrontage,na.rm = TRUE)
      train[i, 'LotFrontage'] = con_air
  }
}

# Numerical values where the NA can't be replaced and make up a small percentage
# can be omitted without worry.
train <- na.omit(train)

# Add season sold variable
train$SeasonSold <- ifelse(train$MoSold == "12", "Winter",
                           ifelse(train$MoSold == "1", "Winter",
                           ifelse(train$MoSold == "2", "Winter",
```

```
                          ifelse(train$MoSold == "3", "Spring",
                          ifelse(train$MoSold == "4", "Spring",
                          ifelse(train$MoSold == "5", "Spring",
                          ifelse(train$MoSold == "6", "Summer",
                          ifelse(train$MoSold == "7", "Summer",
                          ifelse(train$MoSold == "8", "Summer",
                          ifelse(train$MoSold == "9", "Winter",
                          ifelse(train$MoSold == "10", "Winter",
                          ifelse(train$MoSold == "11", "Winter",
                          NA)))))))))))
train$SeasonSold <- as.factor(train$SeasonSold)

# Add total interior area variable
train$TotalInteriorArea = train$TotalBsmtSF + train$`1stFlrSF` +
  train$`2ndFlrSF`

# Check NA values
sapply(train, function(i) sum(is.na(i)*100/nrow(train)))
```

```
##               Id       MSSubClass         MSZoning      LotFrontage
##                0                0                0                0
##          LotArea           Street            Alley         LotShape
##                0                0                0                0
##      LandContour        Utilities        LotConfig        LandSlope
##                0                0                0                0
##     Neighborhood       Condition1       Condition2         BldgType
##                0                0                0                0
##       HouseStyle      OverallQual      OverallCond        YearBuilt
##                0                0                0                0
##     YearRemodAdd        RoofStyle          RoofMatl      Exterior1st
##                0                0                0                0
##      Exterior2nd       MasVnrType       MasVnrArea        ExterQual
##                0                0                0                0
##        ExterCond       Foundation         BsmtQual         BsmtCond
##                0                0                0                0
##     BsmtExposure     BsmtFinType1       BsmtFinSF1     BsmtFinType2
##                0                0                0                0
##       BsmtFinSF2       BsmtUnfSF      TotalBsmtSF          Heating
##                0                0                0                0
##        HeatingQC       CentralAir       Electrical         1stFlrSF
##                0                0                0                0
##         2ndFlrSF     LowQualFinSF        GrLivArea     BsmtFullBath
##                0                0                0                0
##     BsmtHalfBath         FullBath         HalfBath     BedroomAbvGr
##                0                0                0                0
##     KitchenAbvGr      KitchenQual      TotRmsAbvGrd       Functional
##                0                0                0                0
##       Fireplaces      FireplaceQu       GarageType      GarageYrBlt
##                0                0                0                0
##     GarageFinish       GarageCars       GarageArea       GarageQual
##                0                0                0                0
##       GarageCond       PavedDrive       WoodDeckSF      OpenPorchSF
##                0                0                0                0
##    EnclosedPorch        3SsnPorch      ScreenPorch         PoolArea
##                0                0                0                0
##           PoolQC            Fence      MiscFeature          MiscVal
##                0                0                0                0
##           MoSold           YrSold         SaleType    SaleCondition
##                0                0                0                0
```

```
##         SalePrice        SeasonSold TotalInteriorArea
##                 0                 0                 0
```

```
rm(i, con_air, j, select_area, v)
```

Alleys had a significant number of missing values (around 93%), likely because a lot of houses usually aren't built next to alleys. This was prior to me cleaning it up, resulting in no missing values (0%).

3 EDA

3.1 Significant Variables

```
# Train dataset with only numeric variables
train_numeric <- train %>% dplyr::select_if(is.numeric) %>% select(-Id)

# Creating dataset correlating each variable to SalePrice
SPHeat = cor(train_numeric)
SPHeat_melt = melt(SPHeat)
SPHeat_melt <- SPHeat_melt %>% filter(Var1 != "SalePrice" & Var2 == "SalePrice")

# Heatmap showing correlation strength between each variable and SalePrice
g1 <- ggplot(SPHeat_melt, mapping = aes(x = reorder(Var1, value), y = Var2,
                                        fill = value)) +
  geom_tile() +
    ggtitle("Numerical Variables vs. Sale Price") +
    theme_classic() +
      theme(plot.title = element_text(size = 12, face = "bold", hjust=0.5),
          axis.text.x = element_text(size = 8, hjust = 0.5),
          axis.title.y = element_text(size = 10, hjust = 0.5),
          axis.text.y = element_text(size = 7)) +
    scale_fill_gradient(low = "lightgreen", high = "orange",
                        name = "Strength of \nCorrelation") +
    coord_flip() +
    xlab("Numerical Variables") +
    ylab("")

rm(SPHeat, SPHeat_melt, train_numeric)
```

3.2 Outliers

```
# Find outliers which are either too large or too small
g2 <- ggplot(train, aes(GrLivArea, fill = ..count..)) +
  geom_histogram(bins = 40) +
  ggtitle("Observations of Ground Living Area") +
  scale_fill_gradient(low = "lightgreen", high = "orange",
                      name = "Number of \nObservations") +
  xlab("Ground Living Area (Sq Ft)") +
  ylab("Number of Observations") +
  theme_classic() +
  theme(plot.title = element_text(size = 12, face = "bold", hjust=0.5),
        axis.text.x = element_text(size = 10, hjust = 0.5),
        axis.title.y = element_text(size = 10, hjust = 0.5))

# Remove outliers > 5000 and < 500
```

```
train = train[-which(train$GrLivArea > 5000),]
train = train[-which(train$GrLivArea < 500),]

g3 <- ggplot(train, aes(GrLivArea, fill = ..count..)) +
  geom_histogram(bins = 40) +
  scale_x_log10() +
  ggtitle("Observations of Ground Living Area (Outliers Removed)") +
  scale_fill_gradient(low = "lightgreen", high = "orange",
                      name = "Number of \nObservations") +
  xlab("Ground Living Area (Sq Ft)") +
  ylab("Number of Observations") +
  theme_classic() +
  theme(plot.title = element_text(size = 12, face = "bold", hjust=0.5),
        axis.text.x = element_text(size = 10, hjust = 0.5),
        axis.title.y = element_text(size = 10, hjust = 0.5))
```

3.3 Answering Questions of Interests

3.3.1 Question 1

```
roadsale_mean <- train %>% select(Condition1, SalePrice) %>%
  group_by(Condition1) %>% summarise(AvgSalePrice = mean(SalePrice))

g4 <- ggplot(roadsale_mean, aes(x = Condition1, y = AvgSalePrice, group = 1)) +
  geom_point(colour = "lightgreen") +
  geom_line(colour = "orange") +
  ggtitle("Proximity to Road vs. Sale Price") +
  xlab("Proximity to Road") +
  ylab("SalePrice (USD)") +
  theme_classic() +
  theme(plot.title = element_text(size = 12, face = "bold", hjust=0.5),
        axis.text.x = element_text(size = 10, hjust = 0.5),
        axis.title.y = element_text(size = 10, hjust = 0.5))

rm(roadsale_mean)
```

My intitial hypothesis was that proximity to main roads and proximity to railroads would bring down the sale price. However, only the former was true, as houses near main roads have significantly lower value than normal houses. The latter seemed to have no consistent effect on the sale price, which was very strange to me, as a railroad would likely be louder than a main road. But as I thought of it more, my theory now is that because trains are not as abundant as cars, proximity to a railroad would not make much difference.

3.3.2 Question 2

```
TIAsale_mean <- train %>% select(TotalInteriorArea, SalePrice) %>%
  group_by(TotalInteriorArea) %>% summarise(AvgSalePrice = mean(SalePrice))
TIAsale_mean <- slice_head(TIAsale_mean, n = nrow(TIAsale_mean) - 1)

g5 <- ggplot(TIAsale_mean, aes(x = TotalInteriorArea, y = AvgSalePrice,
                               group = 1)) +
  geom_point(colour = "lightgreen") +
  geom_smooth(colour = "orange") +
  ggtitle("Total Interior Area vs. Sale Price") +
  xlab("Total Interior Area (Sq Ft)") +
  ylab("SalePrice (USD)") +
```

```
  theme_classic() +
  theme(plot.title = element_text(size = 12, face = "bold", hjust=0.5),
        axis.text.x = element_text(size = 10, hjust = 0.5),
        axis.title.y = element_text(size = 10, hjust = 0.5))

rm(TIAsale_mean)
```

As seen in the graph, which has an exponential trendline, there is a strong correlation between total interior area and sale price. In fact, this correlation is actually stronger than the total above ground living area, which we can see in g1. This would indicate that basement area is also important to customers when buying houses, making TotalInteriorArea a crucial variable.

3.3.3 Question 3

```
yearbuiltfire_mean <- train %>% select(YearBuilt, Fireplaces) %>%
  group_by(YearBuilt) %>% summarise(AvgFireplaces = mean(Fireplaces))

g6 <- ggplot(yearbuiltfire_mean, aes(x = YearBuilt, y = AvgFireplaces,
                                     group = 1)) +
  geom_point(colour = "lightgreen") +
  geom_smooth(colour = "orange", method = "lm") +
  ggtitle("Number of Fireplaces vs. Year") +
  xlab("Year") +
  ylab("Number of Fireplaces") +
  theme_classic() +
  theme(plot.title = element_text(size = 12, face = "bold", hjust=0.5),
        axis.text.x = element_text(size = 10, hjust = 0.5),
        axis.title.y = element_text(size = 10, hjust = 0.5))

firesale_mean <- train %>% select(Fireplaces, SalePrice) %>%
  group_by(Fireplaces) %>% summarise(AvgSalePrice = mean(SalePrice))

g7 <- ggplot(firesale_mean, aes(x = Fireplaces, y = AvgSalePrice, group = 1)) +
  geom_point(colour = "lightgreen") +
  geom_smooth(colour = "orange", method = "lm") +
  ggtitle("Number of Fireplaces vs. SalePrice") +
  xlab("Number of Fireplaces") +
  ylab("Sale Price (USD)") +
  theme_classic() +
  theme(plot.title = element_text(size = 12, face = "bold", hjust=0.5),
        axis.text.x = element_text(size = 10, hjust = 0.5),
        axis.title.y = element_text(size = 10, hjust = 0.5))

rm(yearbuiltfire_mean, firesale_mean)
```

I thought that fireplaces were only a feature of older houses, as we have more efficient and modern sources of heat, but in reality, they are present in newer houses as well. In fact, as we can see in g7, more fireplaces actually increases the sale price in a clearly linear trend. This indicates that fireplaces are an important part of American architecture and that American customers really like fireplaces.

3.3.4 Question 4

```
poolsale <- filter(train, PoolArea > 0)

g8 <- ggplot(poolsale, aes(x = SeasonSold, fill = ..count..)) +
```

```
  geom_histogram(stat = "count") +
  ggtitle("Number of Houses Sold With Pools per Season") +
  scale_fill_gradient(low = "lightgreen", high = "orange",
                      name = "Number of \nObservations") +
  xlab("Season") +
  ylab("Houses Sold With Pools") +
  theme_classic() +
  theme(plot.title = element_text(size = 12, face = "bold", hjust=0.5),
        axis.text.x = element_text(size = 10, hjust = 0.5),
        axis.title.y = element_text(size = 10, hjust = 0.5),
        axis.ticks = element_blank())
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
rm(poolsale)
```

Unfortunately, this was a far smaller sample size than I expected, but we can make some decent observations from this. It does indeed seem that houses with pools sell better in the summer, which made my hypothesis correct. This could be an indicator that customers do indeed think in the short term to a certain extent when buying a house, which is interesting, as houses are a long term investment. However, I can't quite make a definitive conclusion due to the small sample size. SeasonSold could be a very useful variable for specific situations, but in terms of determining something general such as sale price, I'm not so sure I can make a lot of use for it.

3.3.5 Question 5

```
salegrowth <- train %>% select(Neighborhood, YrSold, SalePrice) %>%
  group_by(YrSold, Neighborhood) %>% summarise(AvgSalePrice = mean(SalePrice))
salegrowth <- spread(salegrowth, Neighborhood, AvgSalePrice)
salegrowth <- select_if(salegrowth, ~ !any(is.na(.)))
salegrowth <- gather(salegrowth, Neighborhood, AvgSalePrice, Blmngtn:Timber)

g9 <- ggplot(salegrowth, aes(x = YrSold, y = AvgSalePrice, group = Neighborhood,
                             colour = Neighborhood)) +
  geom_point() +
  geom_line() +
  ggtitle("Growth in Neighborhood Sale Prices") +
  xlab("Year") +
  ylab("SalePrice (USD)") +
  theme_classic() +
  theme(plot.title = element_text(size = 12, face = "bold", hjust=0.5),
        axis.text.x = element_text(size = 10, hjust = 0.5),
        axis.title.y = element_text(size = 10, hjust = 0.5))

salegrowth <- spread(salegrowth, Neighborhood, AvgSalePrice)
salegrowth <- as_tibble(salegrowth)
salegrowth <- rbind(slice_head(salegrowth), slice_tail(salegrowth))
salegrowth <- subset(salegrowth, select = -YrSold )
salegrowth <- slice_tail(salegrowth) - slice_head(salegrowth)
salegrowth <- gather(salegrowth, Neighborhood, SaleDiff, Blmngtn:Timber)

g10 <- ggplot(salegrowth, aes(x = Neighborhood, y = SaleDiff)) +
  geom_bar(stat = "identity", fill = "orange") +
  ggtitle("Change in Neighborhood Sale Prices") +
  xlab("Neighborhood") +
  ylab("Change in SalePrice (USD)") +
  theme_classic() +
  theme(plot.title = element_text(size = 12, face = "bold", hjust=0.5),
```

```
        axis.text.x = element_text(size = 10, hjust = 0.5, angle = 90),
        axis.title.y = element_text(size = 10, hjust = 0.5))

rm(salegrowth)
```

Graph g9 shows us the trend in sale price for every neighborhood. I got rid of the neighborhoods which were missing observations for certain years, as these neighborhoods made the comparison look awful and messy, and the lack of sample size, especially relative to the neighborhoods which have observations for all the years, introduces a degree of bias and inaccuracy. As seen in g10, The biggest growth in sale price occurred in Crawford and The smallest (or rather negative) growths occurred in Gilbert and Northridge.

4 Further Preprocessing

```
# Choose significant variables
train <- train %>% dplyr::select(OverallQual, TotalInteriorArea, FullBath,
                                 TotRmsAbvGrd, YearBuilt, Fireplaces,
                                 SalePrice)

# Add TotalInteriorArea to test dataset
TotalBsmtSF_NoNA <- test$TotalBsmtSF
TotalBsmtSF_NoNA <- replace(TotalBsmtSF_NoNA, is.na(TotalBsmtSF_NoNA),
                            0)

test$TotalInteriorArea <- TotalBsmtSF_NoNA + test$`1stFlrSF` +
  test$`2ndFlrSF`

rm(TotalBsmtSF_NoNA)
```

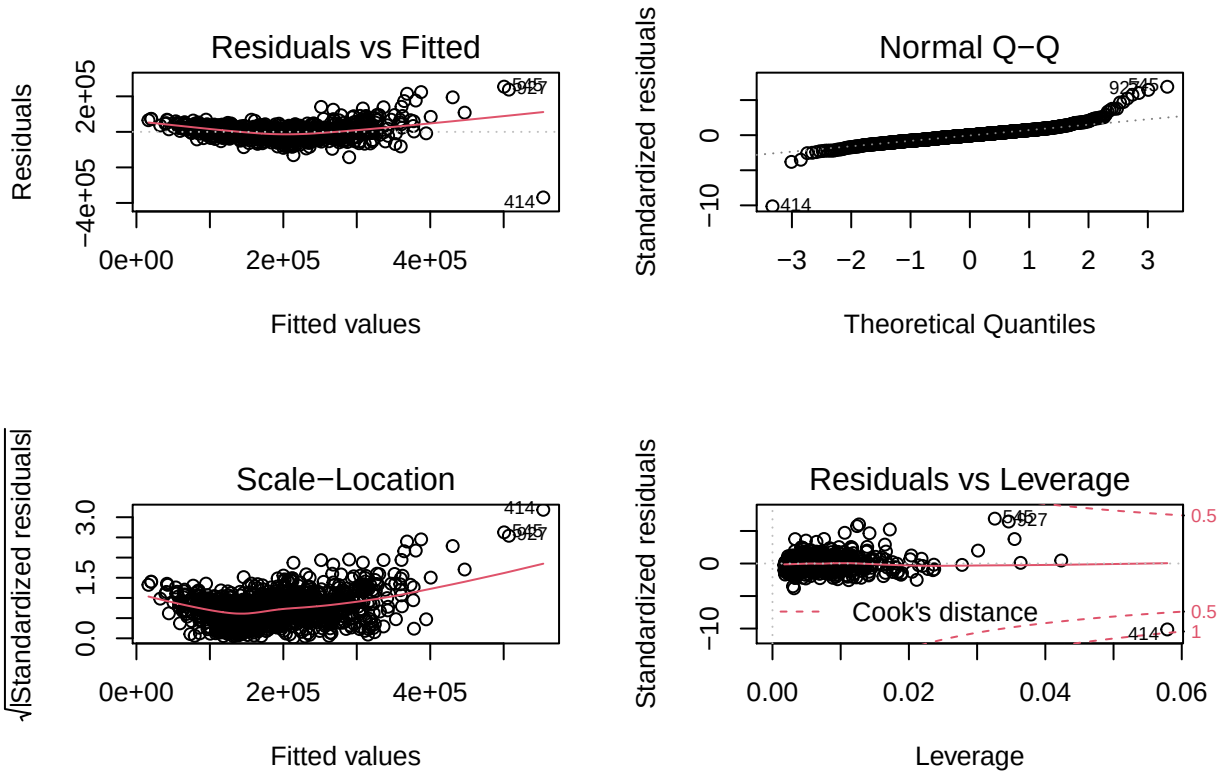
To find the essential variables, I used g1 from the EDA section. Generally, a correlation between 0.25 and 5 is seen as a weak correlation, but this does vary from field to field, as well as situation to situation. Here for instance, I decided to deem anything under 0.45 as a weak correlation, even though initially I was going to do 0.5. This is because in the EDA section, I found fireplaces, which had a score of ~0.47 to be a decent indicator of sale price, as well as important in American households. Thus, any variable with a correlation to SalePrice under 0.45 was excluded. Redundant variables were also excluded. I added my own variables as well, which were SeasonSold and TotalInteriorArea. I had decided to do this earlier, as I wanted to explore and visualise these variables. Thanks to the EDA of these variables, I was able to determine that only total interior area would be useful to me, as season is too hard to quantify and only has use in determining specific variables, such as in my Question 4, rather than something more generalised such as our target SalePrice (people are buying houses all year around, not just at specific seasons).

5 Modelling

5.1 Linear Regression

```
train_lm <- lm(SalePrice ~ OverallQual + TotalInteriorArea + FullBath +
                 TotRmsAbvGrd + YearBuilt + Fireplaces, data = train)

par(mfrow = c(2, 2))
plot(train_lm)
```



5.2 Price Prediction

```
test$SalePrediction <- predict(train_lm, test)

# Actual Sale Price (First 6)
head(test$SalePrice)
```

```
## [1] 105000 172000 189900 195500 191500 175900
```

```
# Predicted Sale Price (First 6)
head(test$SalePrediction)
```

```
##        1        2        3        4        5        6
## 107475.3 175131.1 167333.1 189976.1 217406.9 182234.8
```

6 Evaluation

6.1 Calculate RMSE

```
rmse <- sqrt(mean((test$SalePrice - test$SalePrediction)^2))
```

The RMSE value of my model is approximately 35,982, which isn't too bad all things considered. But perhaps it can be improved using more advanced models. I decided to test an advanced machine learning regression technique called gradient boosting.

```
train_gbm <- gbm(SalePrice ~ ., data = train, cv.folds = 5, distribution = "gaussian", shrinkage = 0.3, n.tree

test$SalePrediction2 <- predict(train_gbm, test)

rmse2 <- sqrt(mean((test$SalePrice - test$SalePrediction2)^2))
```

With a single run of that, I got an RMSE of approximately 29,781, meaning the accuracy was increased by 6,201. As it is a machine learning technique, the results will vary slightly with every run.

6.2 Residuals

```
lm_residuals <- test %>%
  mutate(y = SalePrice) %>%
  mutate(ybar = SalePrediction) %>%
  mutate(diff = abs(y - ybar))

lm_badresiduals <- lm_residuals %>%
  filter(diff > 75000) %>%
  arrange(desc(diff))

g11 <- ggplot(lm_residuals, aes(x = y, y = ybar)) +
  geom_point(aes(x = y, y = ybar)) +
  geom_point(data = lm_badresiduals, colour = "orange") +
  scale_color_gradient(name = "|y - ybar|", limits = c(0, 75000)) +
  geom_abline(slope = 1, intercept = 0) +
  scale_x_continuous(name = "y", labels = scales::comma)+
  scale_y_continuous(name = "ybar", labels = scales::comma)+
  ggtitle("Linear Model Residuals") +
  theme_classic() +
  theme(plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
                    axis.text.x = element_text(size = 10, hjust = 0.5),
                    axis.title.y = element_text(size = 10, hjust = 0.5))

lm_residuals <- test %>%
  mutate(y = SalePrice) %>%
  mutate(ybar = SalePrediction2) %>%
  mutate(diff = abs(y - ybar))

lm_badresiduals <- lm_residuals %>%
  filter(diff > 75000) %>%
  arrange(desc(diff))

g12 <- ggplot(lm_residuals, aes(x = y, y = ybar)) +
  geom_point(aes(x = y, y = ybar)) +
  geom_point(data = lm_badresiduals, colour = "orange") +
  scale_color_gradient(name = "|y - ybar|", limits = c(0, 75000)) +
  geom_abline(slope = 1, intercept = 0) +
  scale_x_continuous(name = "y", labels = scales::comma)+
  scale_y_continuous(name = "ybar", labels = scales::comma)+
  ggtitle("Gradient Boosted Model Residuals") +
  theme_classic() +
  theme(plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
                    axis.text.x = element_text(size = 10, hjust = 0.5),
                    axis.title.y = element_text(size = 10, hjust = 0.5))

rm(lm_residuals, lm_badresiduals)
```

7 Recommendations and Final Conclusions

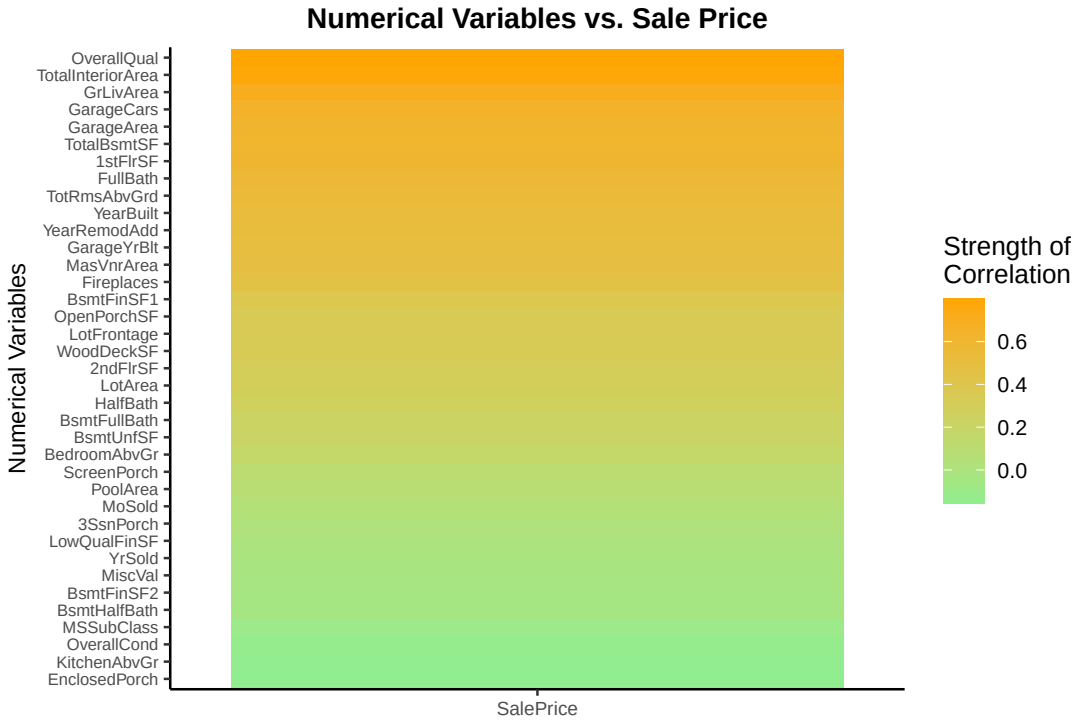
When creating my model, I discovered that numerical variables had a significant effect on the selling prices, which I had found using a heat map I created depicting numerical variables against SalePrice. In question 1, I explored how proximity to certain roads affects sale price. I discovered that proximity to main roads brought down the value of a house, but not proximity to railroads, perhaps due to the abundance of cars compared to trains. In question 2, I wanted to explore a newly created variable for total interior area. Visualising it showed an exponential trendline, which indicated a strong correlation between total interior area and sale price, stronger than with above ground living area, implying that basement area was also a consideration for customers. In question 3, I wanted to explore whether fireplaces were features of older houses. I discovered that this was not the case, and that fireplaces are simply commonplace in American architecture. I also discovered that increasing numbers of fireplaces brought up the sale price, a relationship which I also wanted to explore. In question 4, I wanted to explore another variable I created, which was SeasonSold, by finding out if customers were more inclined to buy houses with pools during the summer. Although the sample size was quite small as there weren't many houses with pools, the data indicated that customers were indeed more likely to buy houses with pools during the summer, as well as possibly indicating that customers to a certain extent think in the short term when buying houses, though it is ultimately hard to conclude on. Finally, in question 5, I explored the trend in sale price for every neighborhood. I found that Crawford had the biggest growth in sale price, while Gilbert and Northridge had declined in sale price. The best RMSE I obtained was around 29,781, using gradient boosting. A lot of the improvements that could've been made revolved around the unclean test data, which contained too many NAs. There was no mention in the assignment sheet either about cleaning it up and when I did try to clean it up, I ended up with RMSE values of 80,000 to a whopping 100,000. This kept me stuck for hours, resulting in a slightly late submission as well. Thus, I had to use the original, unclean dataset. And this was a problem because it probably led to a higher RMSE, as well as the fact I couldn't utilise more regression techniques which required the use of matrices. As for improvements on my side, more advanced variables could've been engineered, and qualitative variables could've been implemented into the modelling, but I was unsure how the RMSE would be affected, especially due to the NA values in the test dataset. All in all, very decent assignment and I learned a lot, so I had a positive experience.

8 References

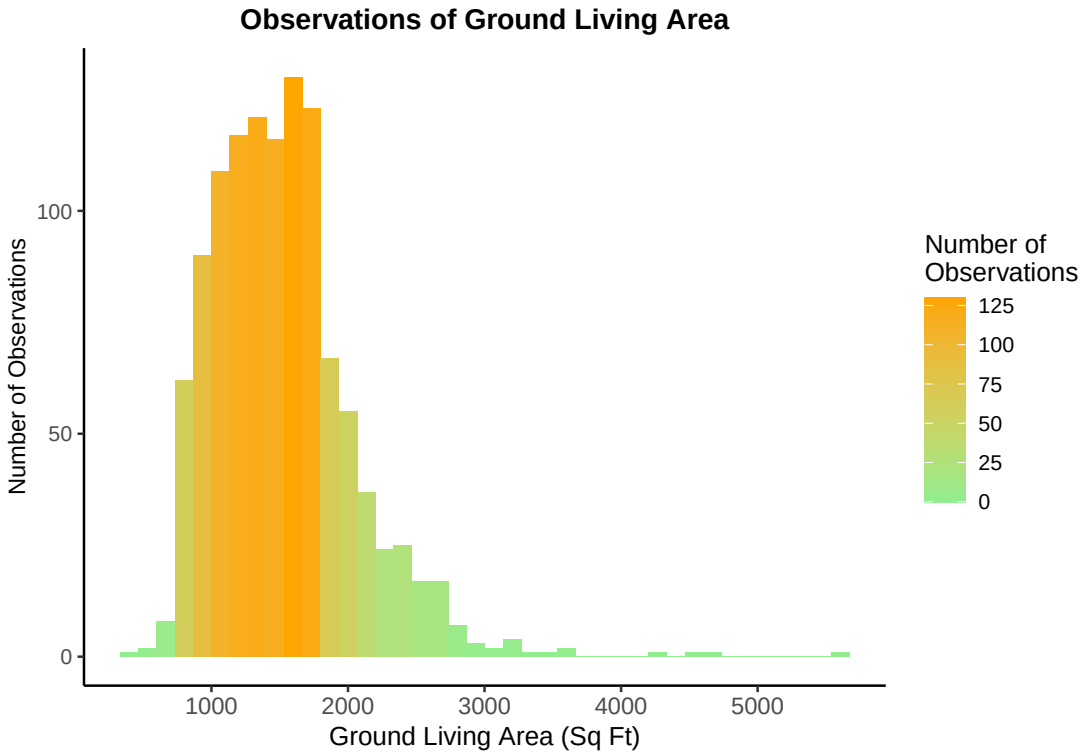
1. De Cock, D 2011, 'Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project', Journal of statistics education, vol. 19, no. 3.

9 Graphs

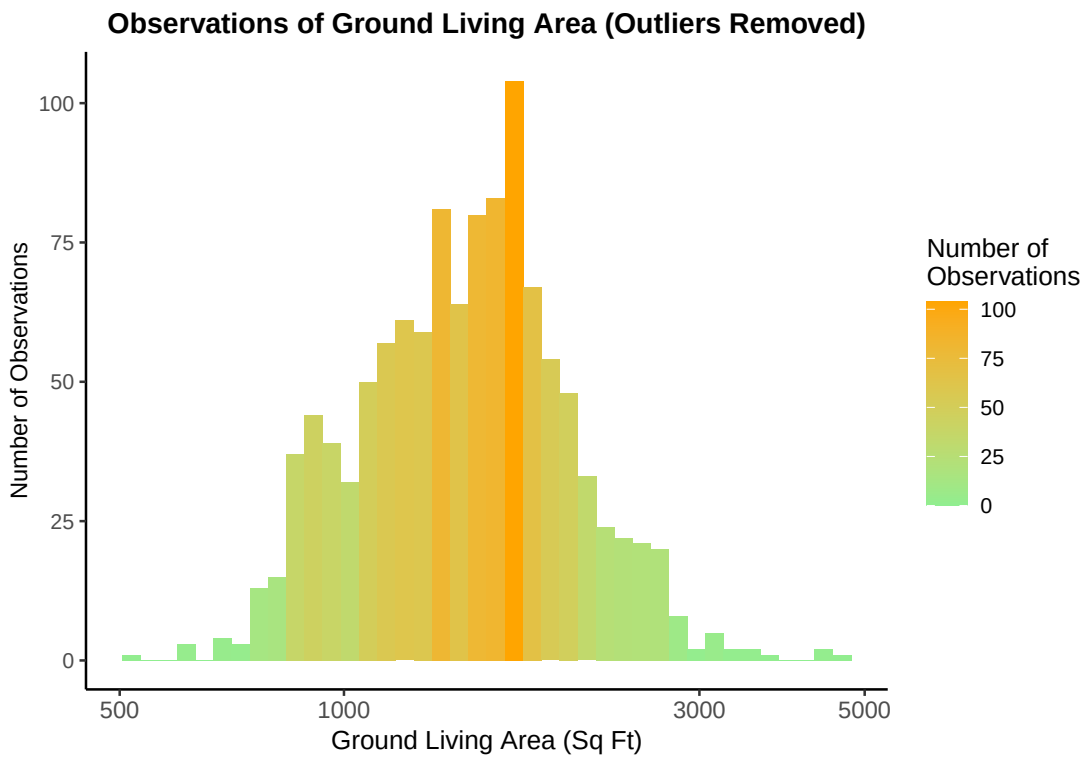
g1



g2



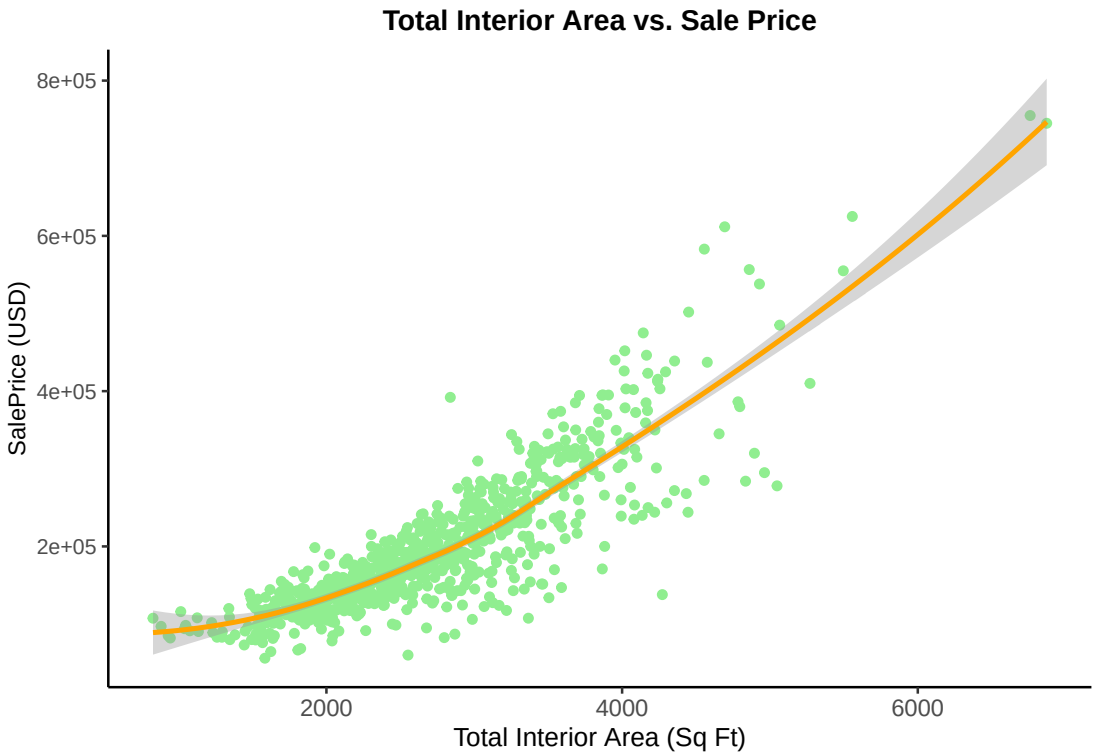
```
g3
```




```
g4
```

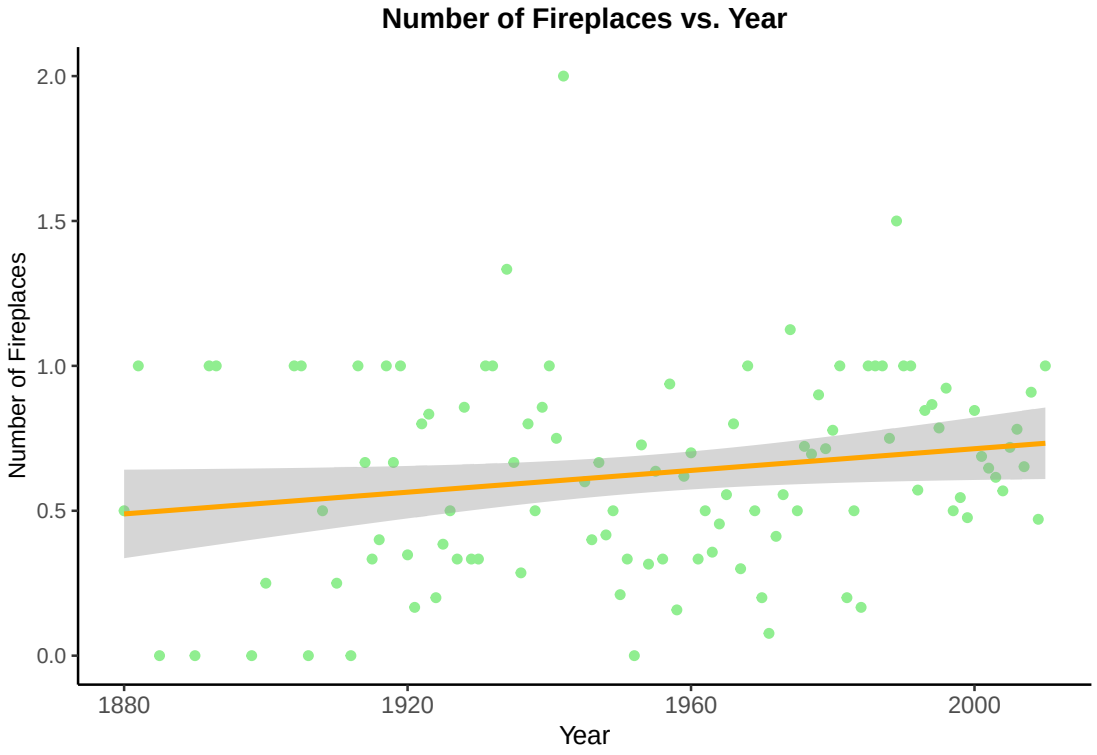

```
g5
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



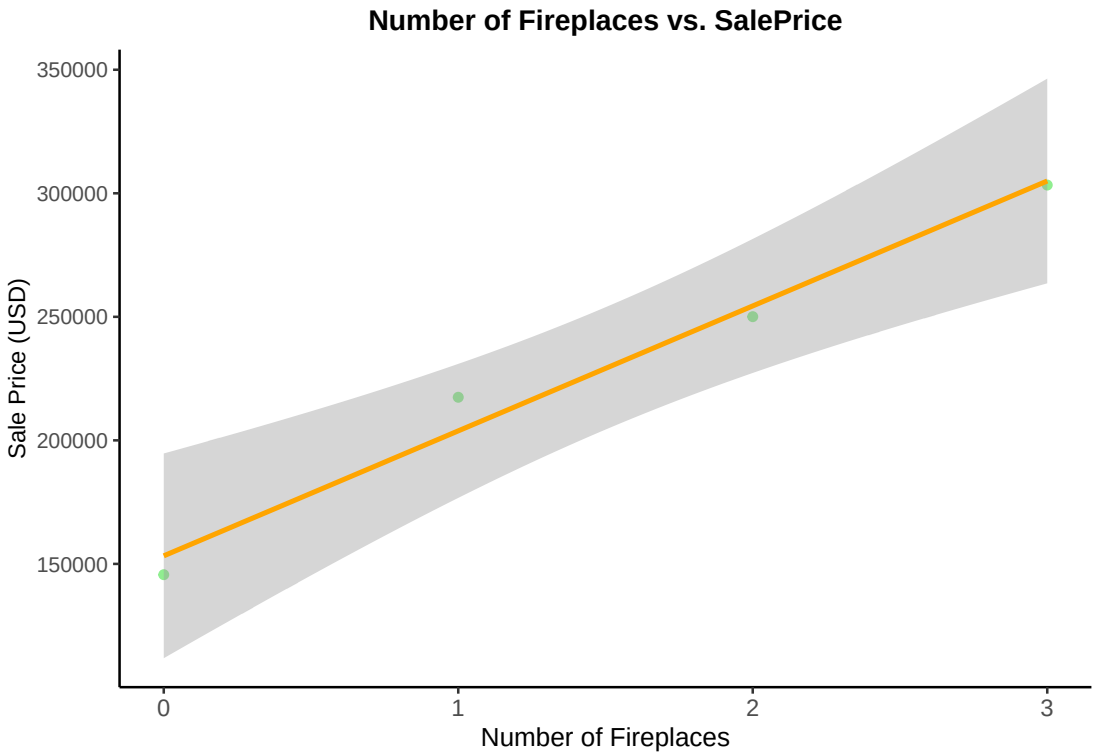

```
g6
```

```
## `geom_smooth()` using formula 'y ~ x'
```

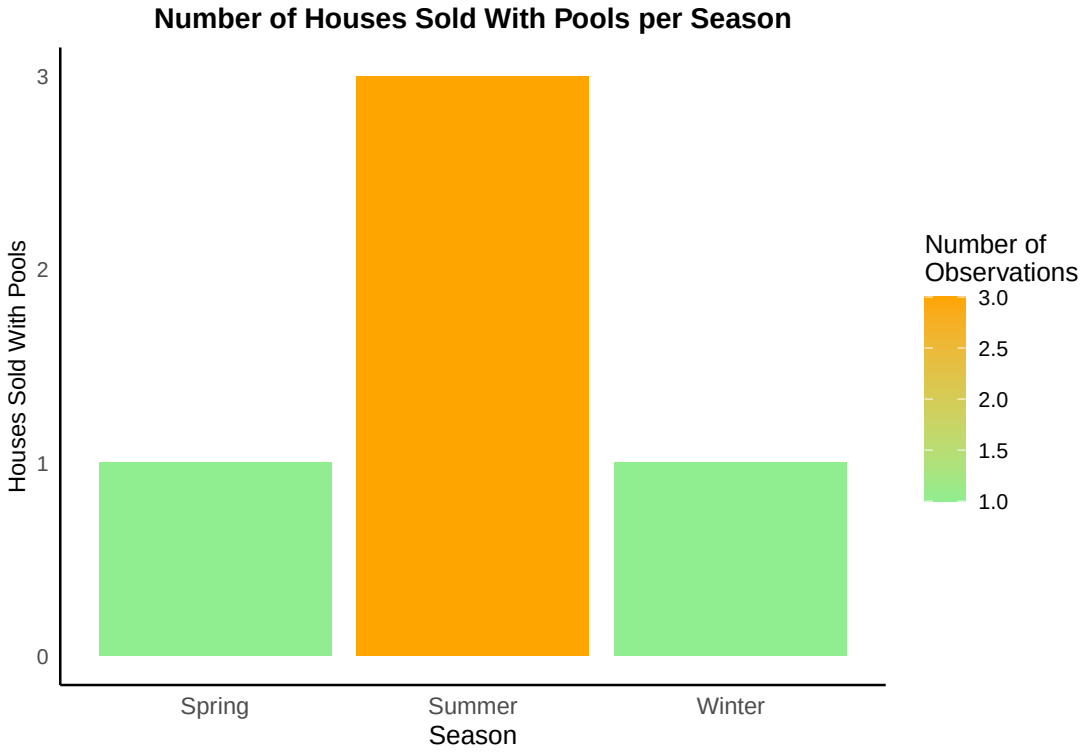



```
g7
```

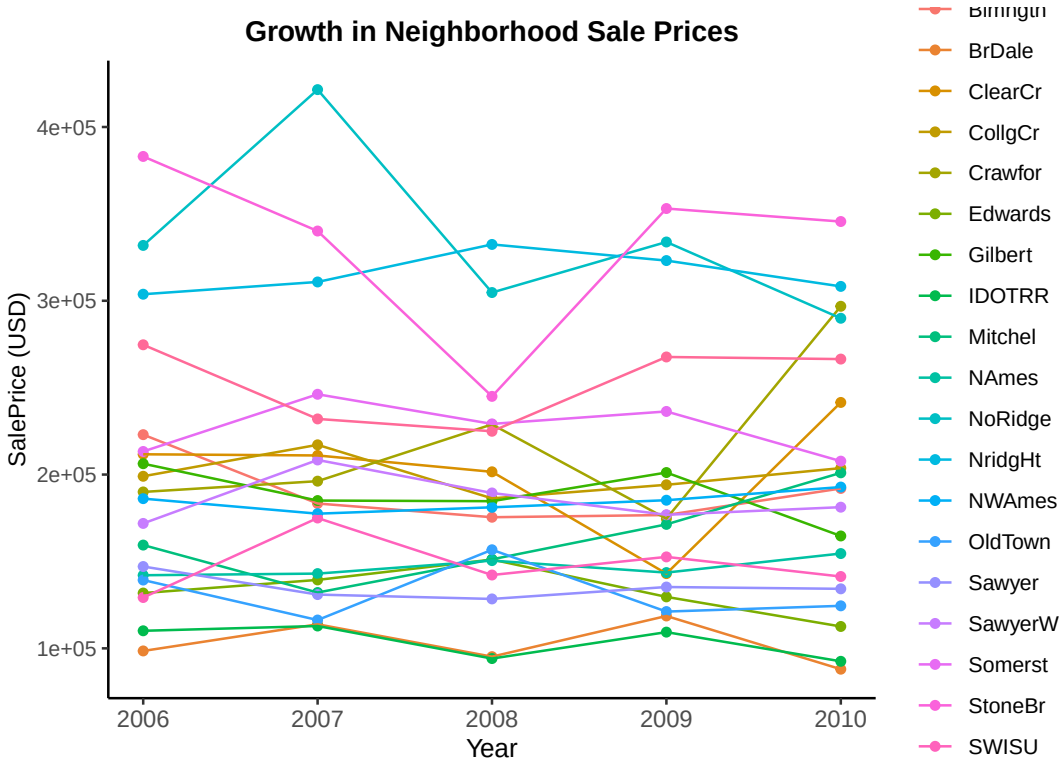
```
## `geom_smooth()` using formula 'y ~ x'
```


Number of Fireplaces vs. SalePrice
350000
300000
250000
200000
150000
Sale Price (USD)
0
1
2
3
Number of Fireplaces

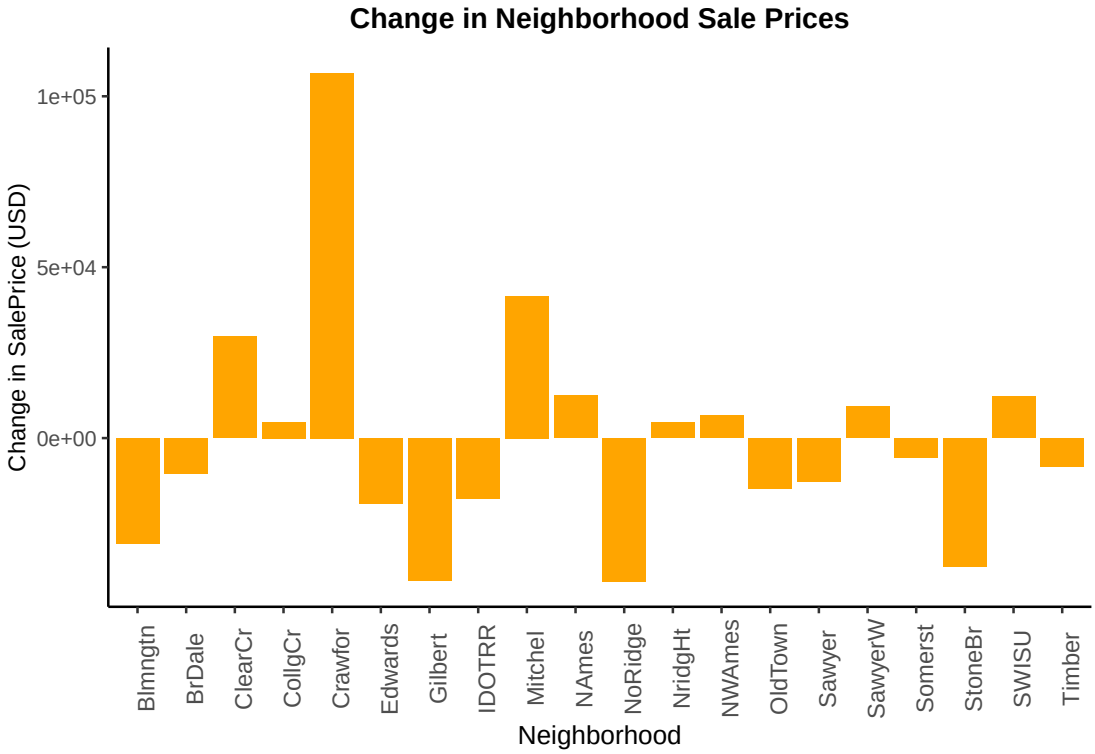
```
g8
```


Number of Houses Sold With Pools per Season
3
2
1
0
Houses Sold With Pools
Spring
Summer
Winter
Season
Number of Observations
3.0
2.5
2.0
1.5
1.0

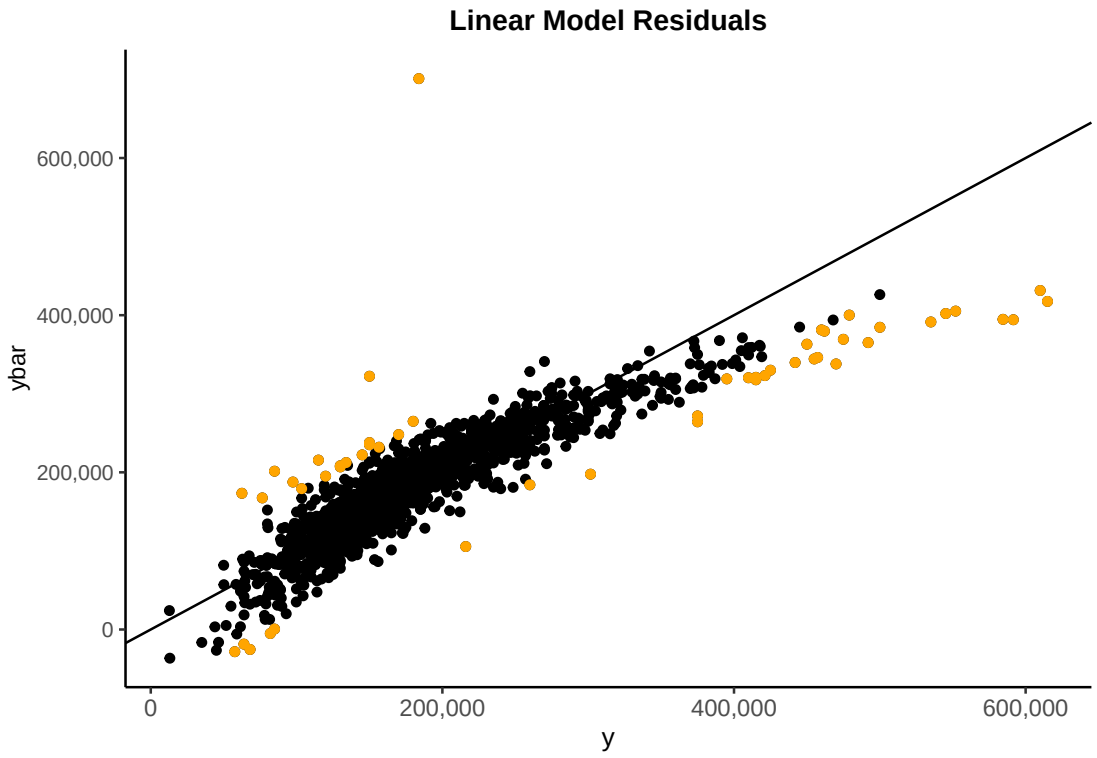
```
g9
```


Growth in Neighborhood Sale Prices
SalePrice (USD)
4e+05
3e+05
2e+05
1e+05
2006
2007
2008
2009
2010
Year
Blmngtn
BrDale
ClearCr
CollgCr
Crawfor
Edwards
Gilbert
IDOTRR
Mitchel
NAmes
NoRidge
NridgHt
NWAmes
OldTown
Sawyer
SawyerW
Somerst
StoneBr
SWISU

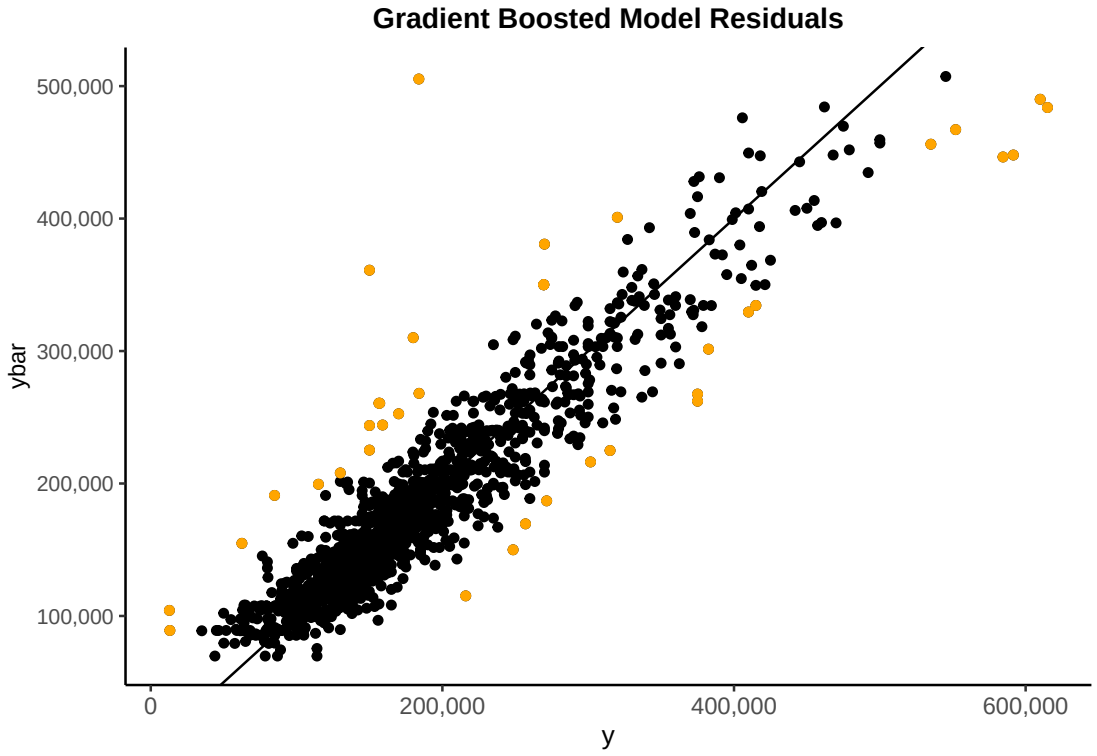
```
g10
```


Change in Neighborhood Sale Prices
Change in SalePrice (USD)
1e+05
5e+04
0e+00
Blmngtn
BrDale
ClearCr
CollgCr
Crawfor
Edwards
Gilbert
IDOTRR
Mitchel
NAmes
NoRidge
NridgHt
NWAmes
OldTown
Sawyer
SawyerW
Somerst
StoneBr
SWISU
Timber
Neighborhood

```
g11
```


Linear Model Residuals
600,000
400,000
200,000
0
ybar
0
200,000
400,000
600,000
y

```
g12
```


Gradient Boosted Model Residuals
500,000
400,000
300,000
200,000
100,000
ybar
0
200,000
400,000
600,000
y