# Assignment 1: Getting to know your Data

*Sri Seshadri*

*6/22/2017*

## 1. Introduction

This report discusses the Ames, Iowa housing data set, its quality and observations from Exploratory Data Analysis (EDA). EDA was used to identify potential predictor variables that can be used for predicting value of a "typical" home. It is shown below that a.Living area b.Total basement area c.Garage area d.Total rooms above grade e.Neighborhood f.Number of bathrooms are good selection of predictor variables for modelling the sale price of typical homes in Ames, Iowa.

## 2. Sample Definition

It is assumed that typical home buyers are those that move from apartments to single family or town homes. Also apartments are less likely to be sold to individuals as they remain holdings of owners for rental income. Single family and town homes belong to "Residential Low density" (RL) zoning classification in the city of Ames. Data belonging to only to the RL zone is considered for this analysis. There are a total of 2930 data points, out of which the rows belonging to the following zoning classifications are not sampled. The sampled data however has few two-family conversion and duplex houses, which are not removed from the sample at this time.

Table 1: Drop waterfall

| Zoning | counts |
|---|---|
| A (agr) | 2 |
| C (all) | 25 |
| FV | 139 |
| I (all) | 2 |
| RH | 27 |
| RM | 462 |

## 3. Data cleaning and Sanity check

From the Ames, Iowa housing dataset's dictionary and based on some preliminary EDA, the variables listed in Table 2 and Table 3 were chosen to be important for furhter analysis. The dataset is fairy clean and there are 2 points that are missing. Table 2 shows the summary statisics of the numeric variables and it is noted that statistics are within reasonable bounds and appear to be in the units of measure as described in the data dictionary. Table 3 shows the chosen categorical (Nominal) variables and their unique number of values, sample sixe (n) and count of missing data. 4 of the 26 neighborhood are not in the sample frame chosen.

Table 2: Data sanity check for numeric variables

| | min | Q1 | median | Q3 | max | mean | sd | n | missing |
|---|---|---|---|---|---|---|---|---|---|
| LotArea | 1700 | 8339 | 10000 | 12160 | 215245 | 11143.83 | 8428.74 | 2273 | 0 |
| OverallCond | 1 | 5 | 5 | 6 | 9 | 5.52 | 1.03 | 2273 | 0 |
| YearRemodel | 1950 | 1967 | 1993 | 2004 | 2010 | 1985.30 | 19.88 | 2273 | 0 |
| TotalBsmtSF | 0 | 855 | 1056 | 1367 | 6110 | 1111.84 | 448.40 | 2273 | 0 |

| | min | Q1 | median | Q3 | max | mean | sd | n | missing |
|---|---|---|---|---|---|---|---|---|---|
| GrLivArea | 334 | 1149 | 1479 | 1788 | 5642 | 1535.49 | 516.33 | 2273 | 0 |
| BsmtFullBath | 0 | 0 | 0 | 1 | 3 | 0.47 | 0.53 | 2272 | 1 |
| BsmtHalfBath | 0 | 0 | 0 | 0 | 2 | 0.07 | 0.26 | 2272 | 1 |
| FullBath | 0 | 1 | 2 | 2 | 4 | 1.60 | 0.55 | 2273 | 0 |
| HalfBath | 0 | 0 | 0 | 1 | 2 | 0.40 | 0.51 | 2273 | 0 |
| BedroomAbvGr | 0 | 3 | 3 | 3 | 6 | 2.91 | 0.79 | 2273 | 0 |
| TotRmsAbvGrd | 2 | 6 | 6 | 7 | 15 | 6.56 | 1.53 | 2273 | 0 |
| GarageArea | 0 | 370 | 484 | 592 | 1488 | 495.30 | 209.36 | 2273 | 0 |
| MoSold | 1 | 4 | 6 | 8 | 12 | 6.22 | 2.72 | 2273 | 0 |
| YrSold | 2006 | 2007 | 2008 | 2009 | 2010 | 2007.78 | 1.31 | 2273 | 0 |
| SalePrice | 35000 | 137500 | 172000 | 223500 | 755000 | 191283.25 | 81295.74 | 2273 | 0 |

Table 3: Data sanity check for nominal variables

| | # Unique | n | missing |
|---|---|---|---|
| LotConfig | 5 | 2273 | 0 |
| Neighborhood | 22 | 2273 | 0 |
| HouseStyle | 8 | 2273 | 0 |
| KitchenQual | 5 | 2273 | 0 |
| SaleCondition | 6 | 2273 | 0 |

## 4. Initial Exploratory Data Analysis (EDA)

For ease of performing exploratory data analysis using the R software, a unique identification number for each row is added as a variable to the dataset. The year sold (YrSold) and month sold (MoSold) are combined into one variable as yrmonth for example January 2001 is coded as 2001.01. The data is then arranged in chronological order by "Neigborhood". The sales proce of properties are right skewed as seen in Figure 1. It is hypothesized that values in some neighborhoods are higher relative to others.

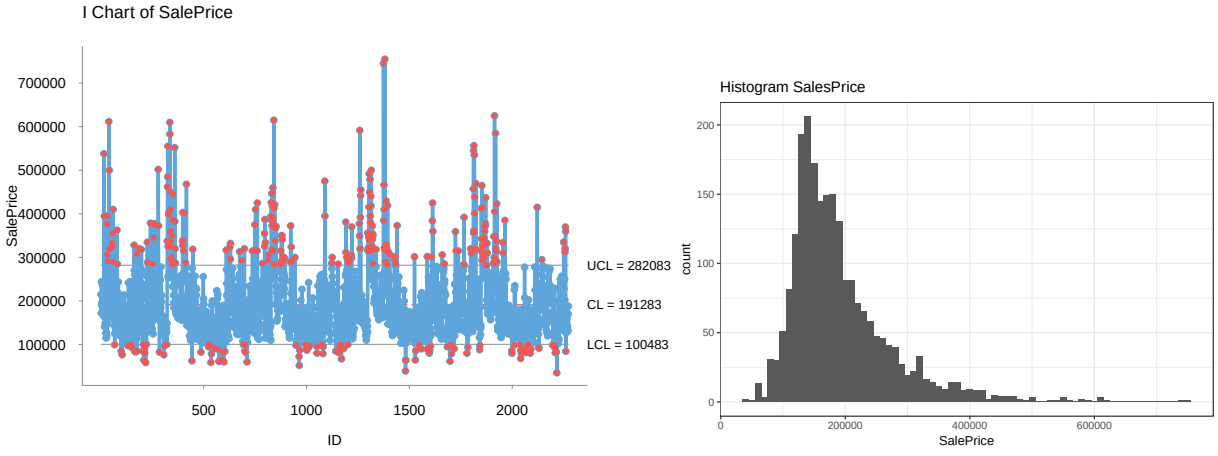


Figure 1: Control Chart and Histogram

### 4.1 Analysis by categorical variables

It is evident from Figure 2 that the sale price that were above the upper control limit on the control chart were more from Stone Brook, Northridge Heights and Northridge than the rest of the neighborhoods. Also it is seen that the some neighborhoods have more sales than others (Figure 3). They may be either new housing developments or likey to have a floating population.

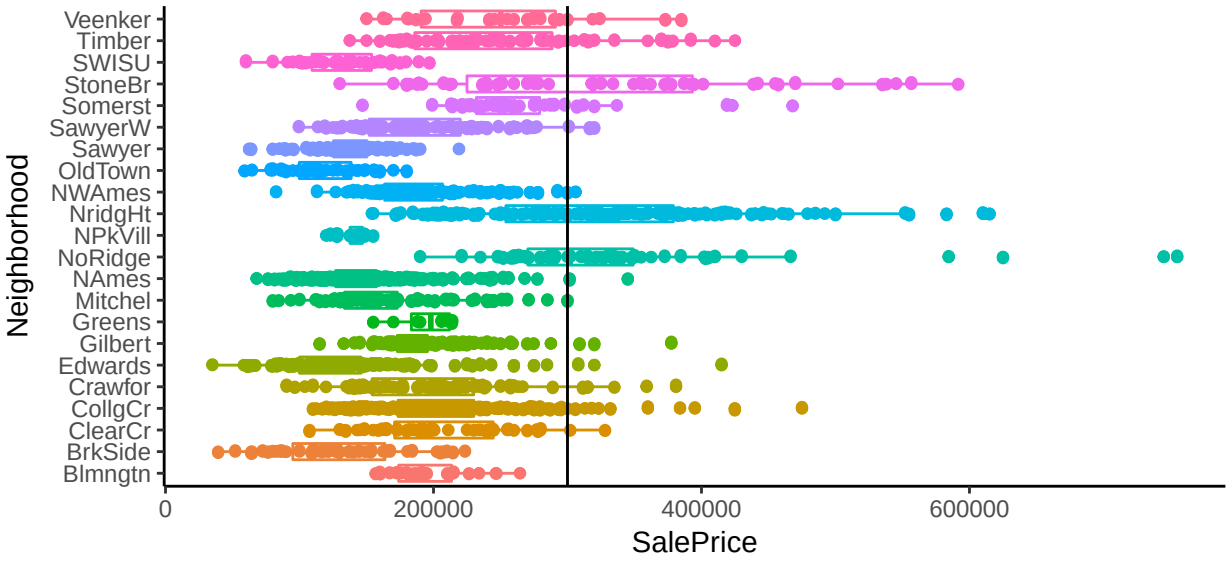


Figure 2: Boxplot of Sales by Neighborhood

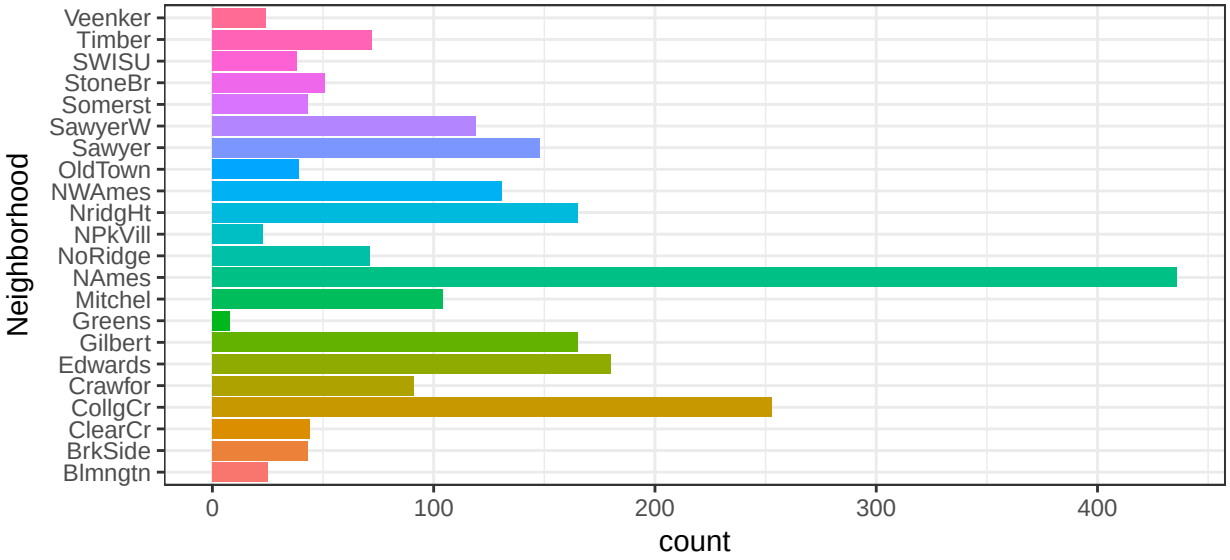


Figure 3: Frequency of Sales by Neighborhood

The sale condition has a pronounced effect on the sale price (see Figure 4). It is seen that homes that were not fully completed when last assessed were expensive. This could mean that they were newly constructed homes that were priced higher than the older ones. This leads us to hypothesize that year built or remodelled would have similar effect on the sales price. Also it would be interesting to see if new constructions are from a specific neighborhood and how do they compare with mature neighborhoods, which we'll explore in the next section.

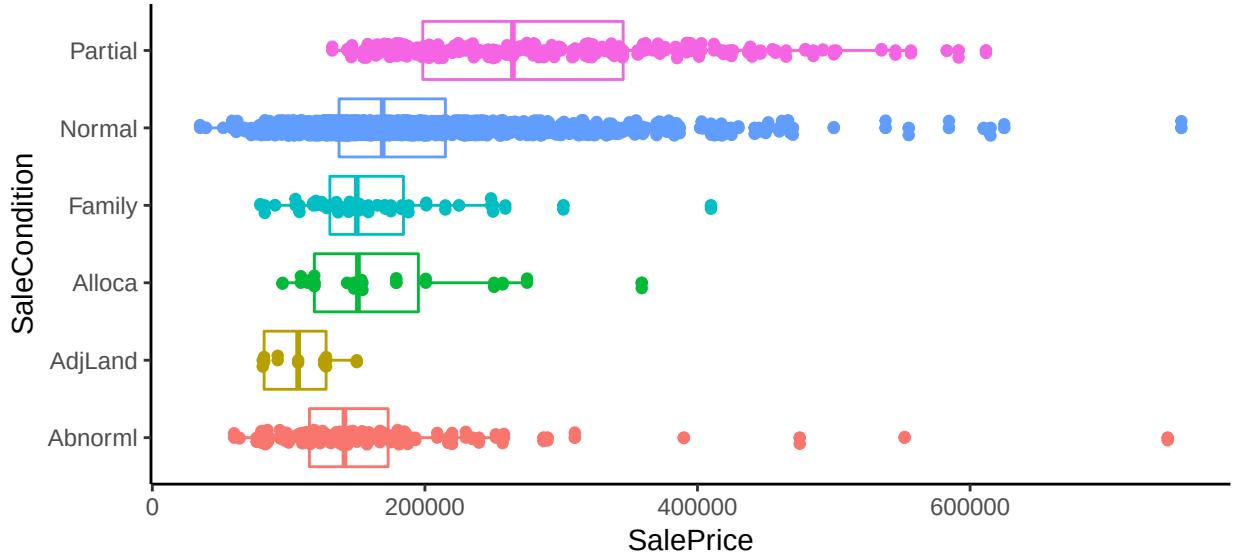


Figure 4: Sales by Sale condition

Homes with excellent kitchen quality have on an average higher sale price, as seen in Figure 5.

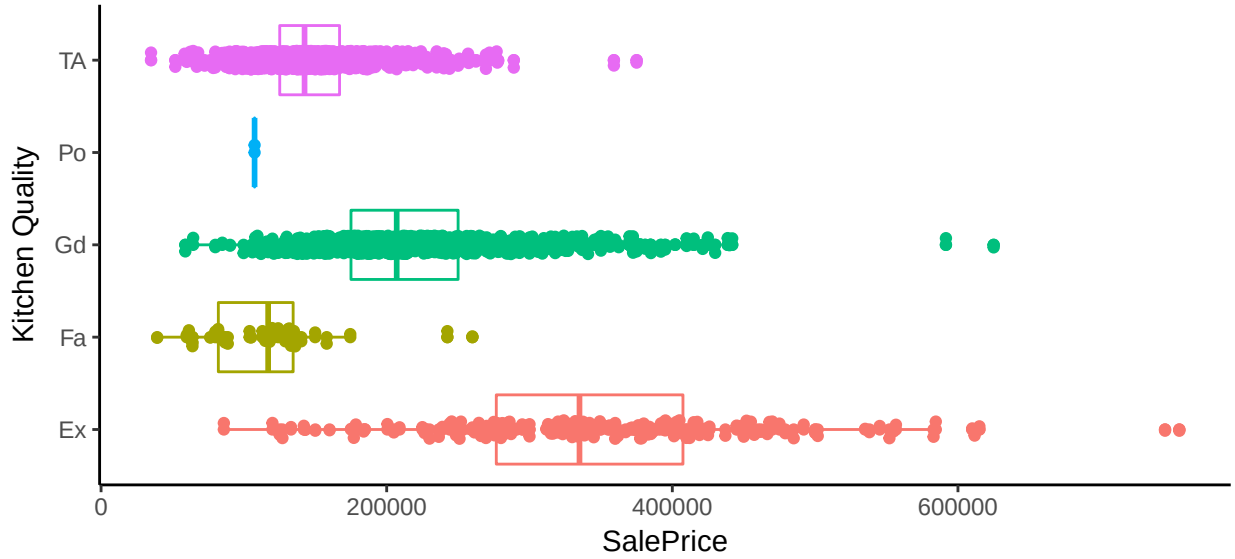


Figure 5: Sale price by kitchen quality assesment

From Figure 6, we can infer that there are more style of houses vailable for sale in lots in the inside of a community than Corners, Cul-de-sac or more frontage lots. Also, Cul-de-sac and 3 frontal properties are valued more compared to the test of the lots (see Figure 7).

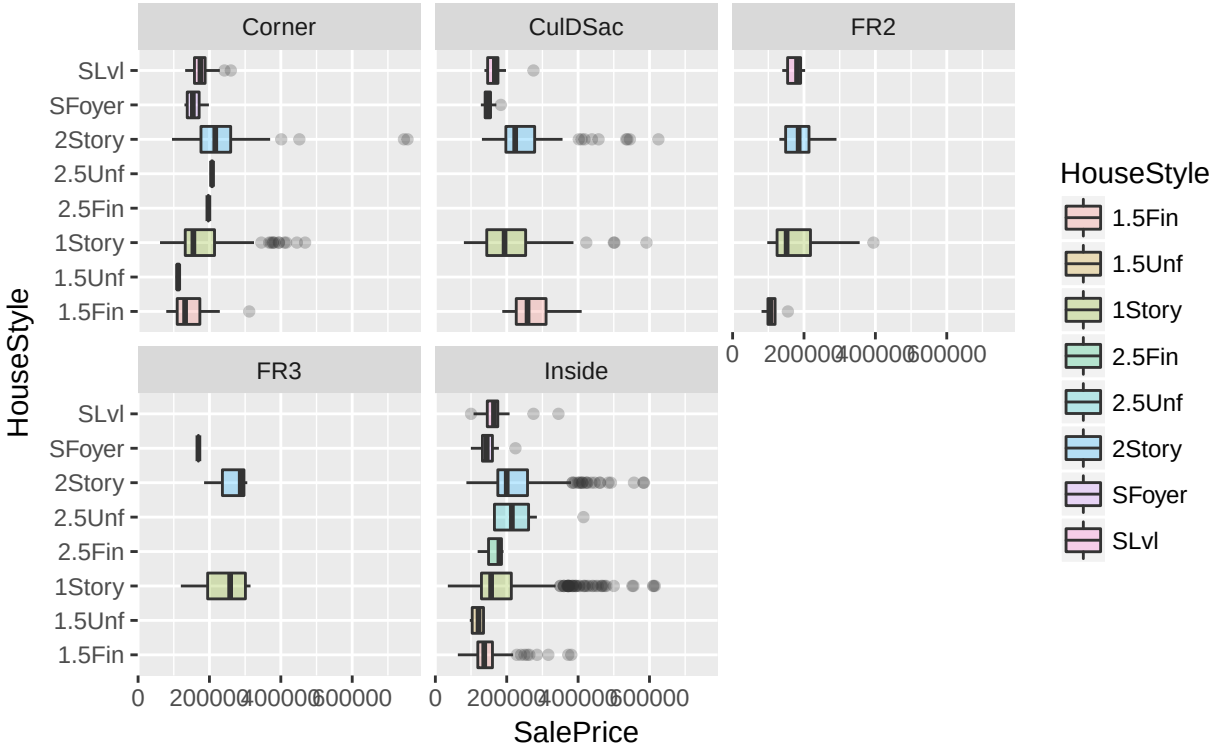


Figure 6: Sale price by lot configuration and style

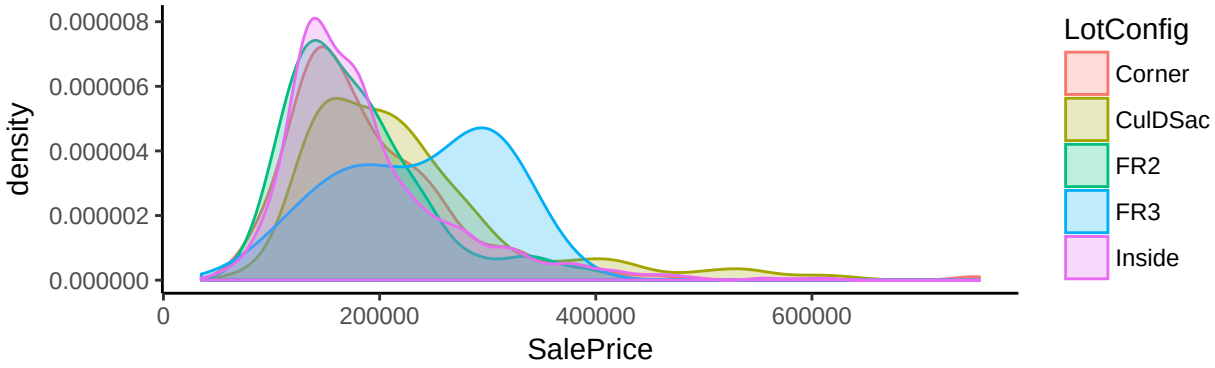


Figure 7: Probability density of Sale price by lot configuration

### 4.2 Analysis by continuous variables

In the previous section we hypothesized that the sale price of homes are likely related to build year or the remodeled year. We'll use the year of remodel year for exploration. We find in figure 8 that newer homes are more expensive than the older homes. There is no strong relationship between the remodel year and the sale prive with a correlation of only 0.53. However, it will be interesting to explore in which neighborhoods there may be much stronger relationship between the two variables. We'll explore this in the next section.

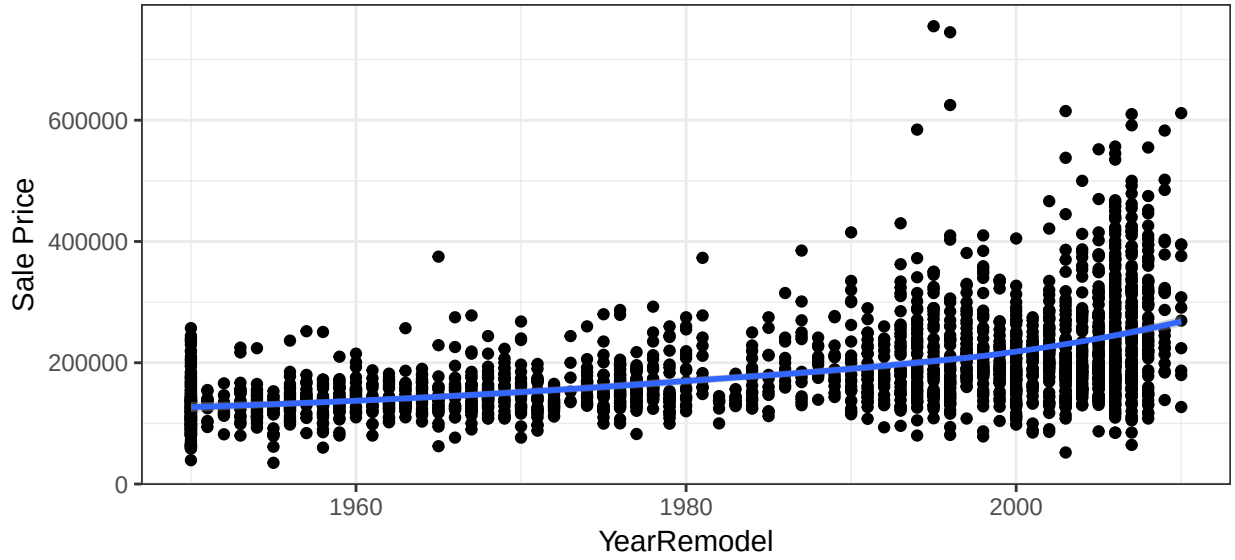


Figure 8: Sale price Vs Year Remodeled

It was hypothesized during variable selection for exploration that the number of rooms in the house is highly correlated with the value of the house. While the correaltion between the two variables may be only 0.53, it may be a good predictor variable among others.

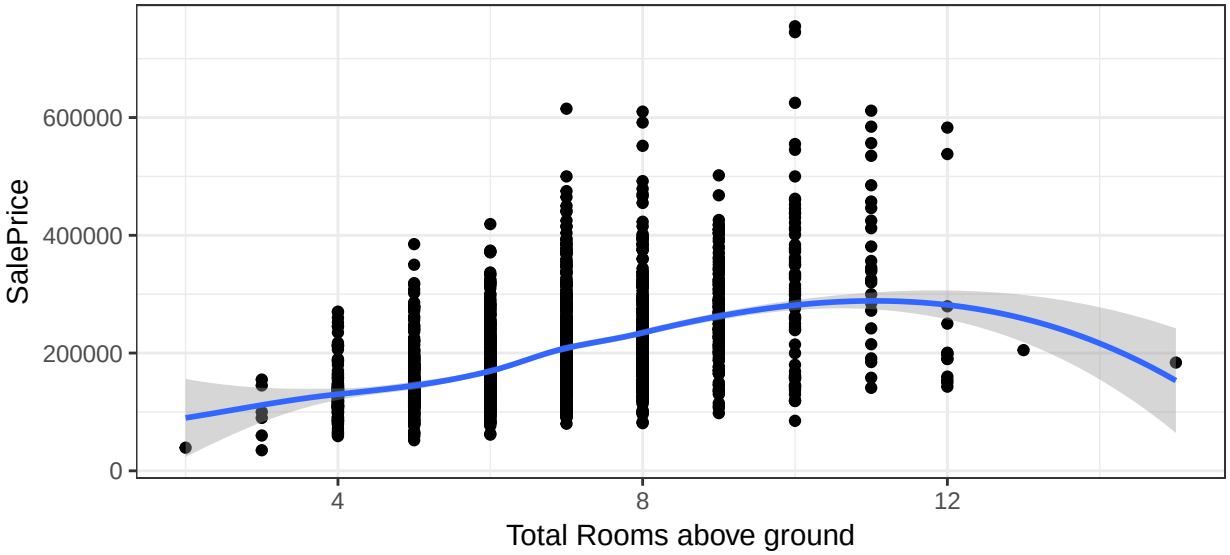


Figure 9: Sale price Vs Number of rooms above ground

Similarly, total basement area was hypothesized to be a good predictor. While basement area by itself may not be a good predictor, it may be a good predictor variable among other variables. It'll be interesting to see which neighborhood have more basement area and if it may be a good predictor in some neighborhoods than others. We will explore the basement area relationship to neighborhood in the next section. Also, the effect of no basement houses need to be removed.

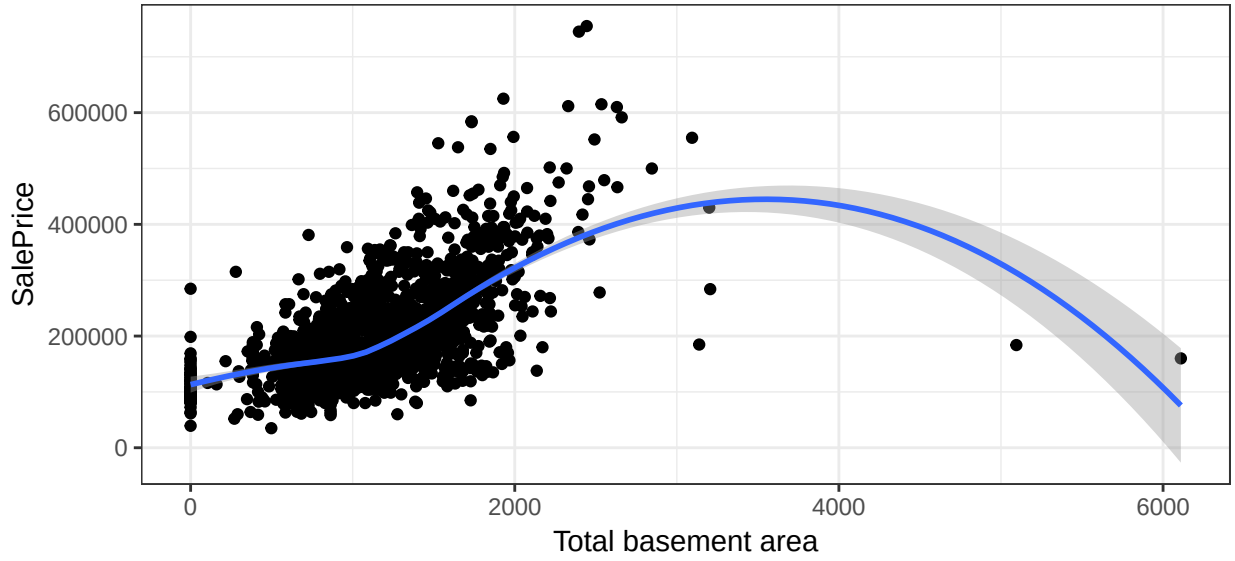


Figure 10: Sale price Vs Total basement area

It is assumed that bigger homes are of high value and are built for high occupancy. High occupancy homes are likely to have more bathrooms and living area (see Figure 12). The total number of bathrooms is a sum of the half baths and the full baths in the house. The total number of baths may be a week predictor of sale price by itself (Figure 11), but may be a good candidate for multivariate regression models among others.

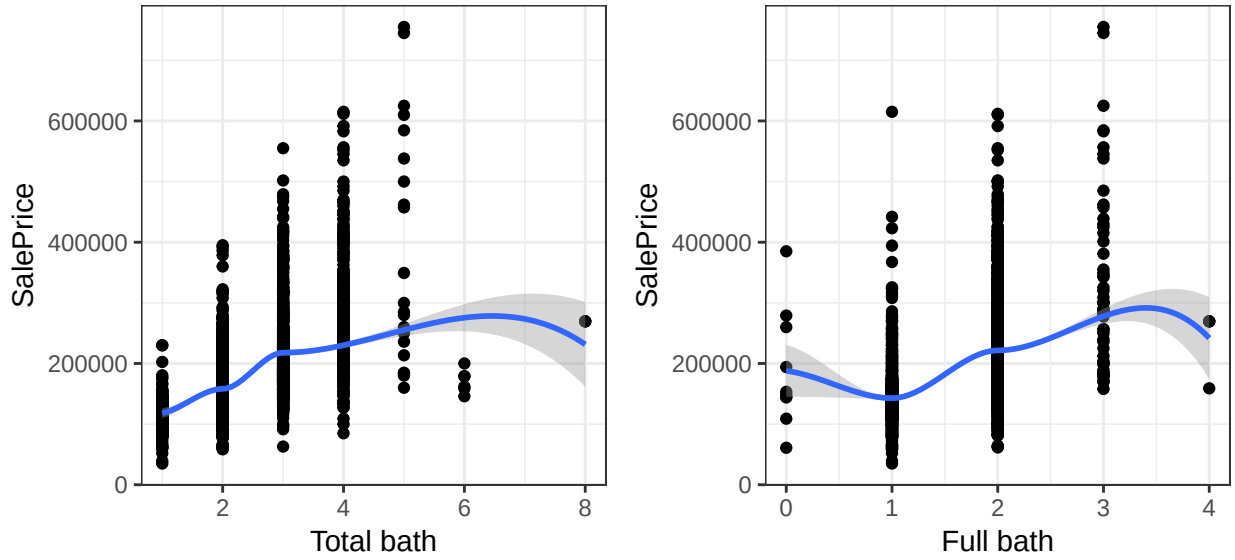


Figure 11: Sale price Vs Number baths

Is is hypothesized that bigger the house, more likely is higher occupancy and sale value. The higher occupancy means a likely living area and garage space for parking. Figure 12 shows there is strong relationship between living area and sale price.

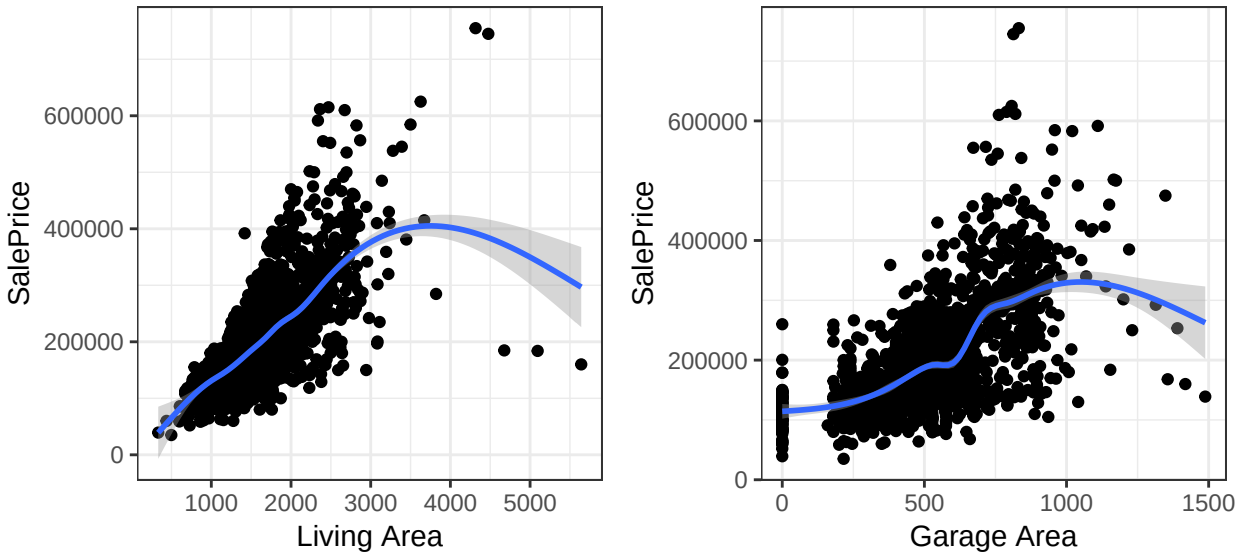


Figure 12: Sale price for high occupancy homes

## 5. Exploratory Data Analysis for Modelling

From the above exploratory analysis, the follwowing parameters come out to be good candidates for further analysis. a) Living Area, b) basement area and c) Sale condition.

Figure 12 hints at a good linear relationship between Living Area and the value of homes. However, the bullhorn shape of the scatter plot suggests that a transformation may help. The sale price and living area variables are transformed logrithmically. Figure 13 shows a much better relationship when the transformed values are used. The blue line shows the regression line.

As seen in Figure 10, there is approximately a linear linerationship between basement area and sale price, when basment area is greater than 0 and less than 2500 square feet. To further examine the relationship, the data corresponding to the target vasement area was subsetted. Figure 14 shows the relationship between sale price and basement area; log of sales vs basement area and log of both sales and basement area. It can be concluded that basment area should be chosen as one of the predictors for the model.

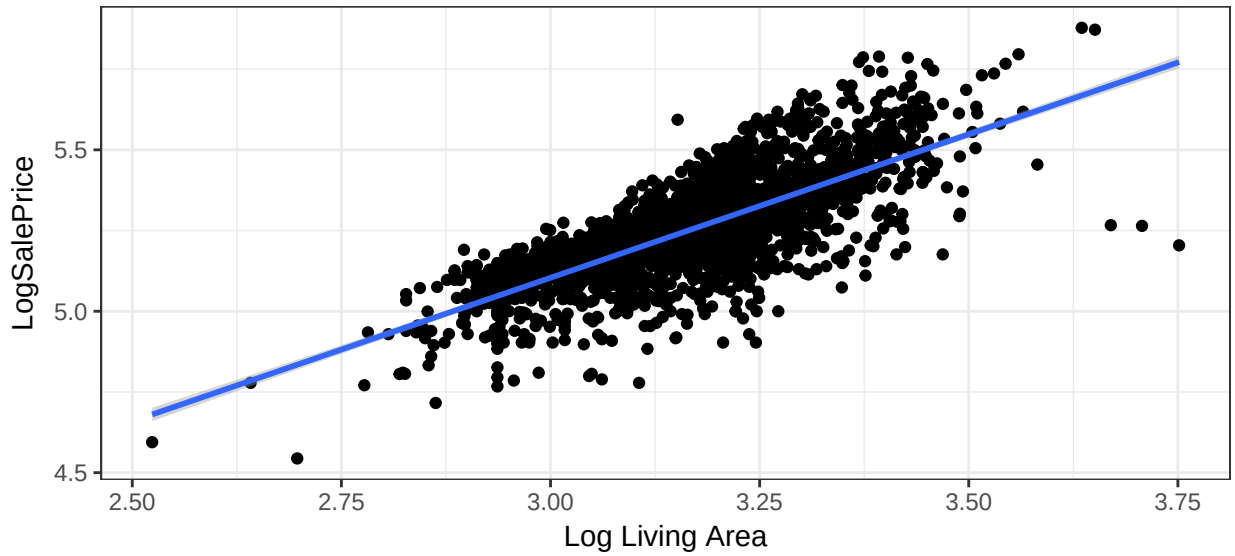


Figure 13: Log Sale price vs Log Living Area

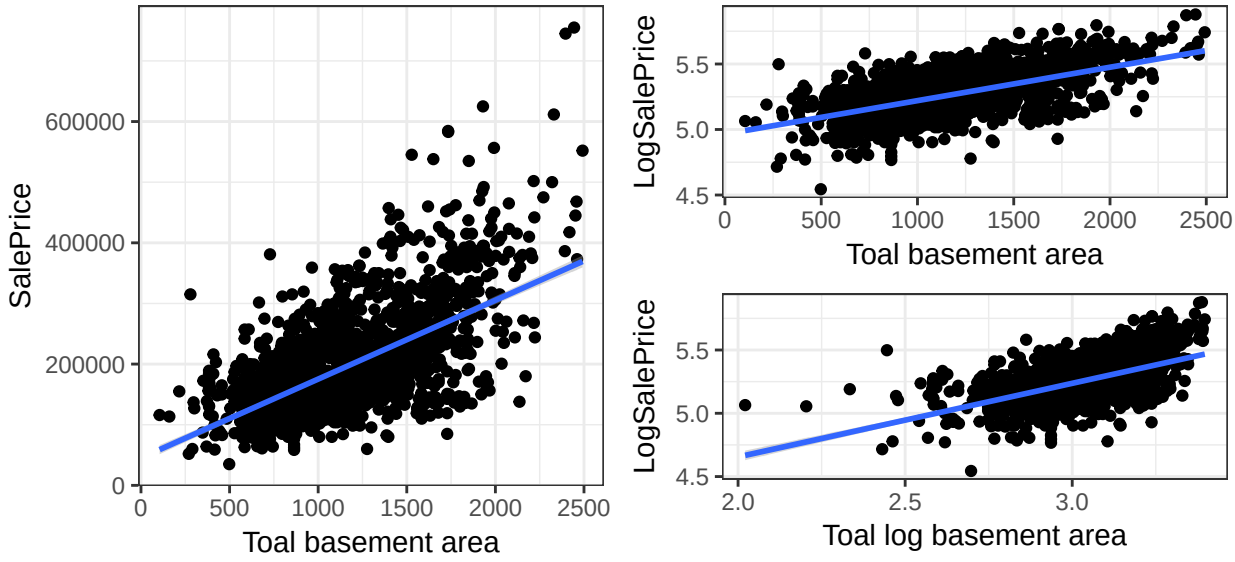


Figure 14: Sale price vs Basement area

We we'll follow up on our hypothesis (Figure 4) from the previous section, if sale condition of partial assessment before sale can be a good predictor across neighborhoods. The data with sale condition equalling "partial" is grouped together and compared against other sale conditions across neighborhood. For ease of presentation and interpretation, only select neighborhoods with number of sales exceeding 100 is shown in Figure 15. It can be infered that Sale condition of "Partial" may be a weak predictor across neighborhoods.

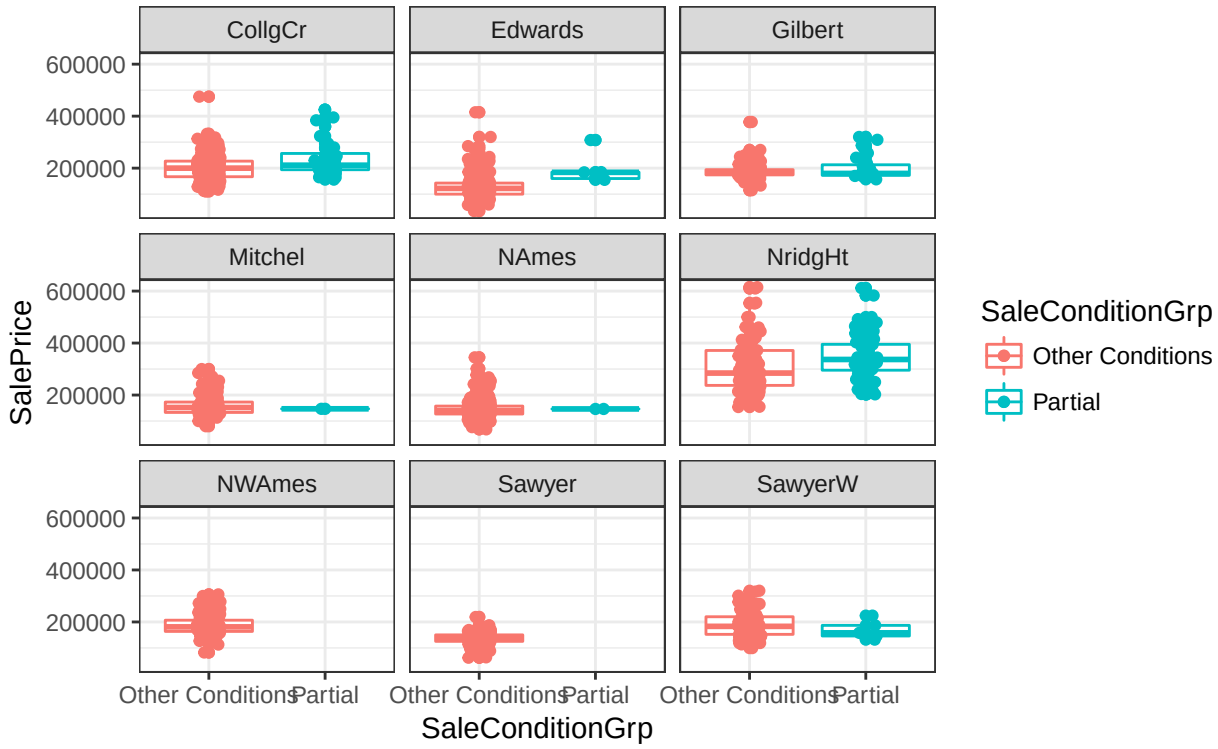


Figure 15: Sale price vs Sale condition group

## Conclusion:

Exploratory Data Analysis (EDA) is very critical to get a deep understanding of the data, which results in proper selection of sample frame and predictor variables for a good predictive model. The analysis serves as a strong foundation for developing and evaluating predictive models.

# APPENDIX

```
knitr::opts_chunk$set(echo = FALSE, tidy.opts = list(width.cutoff = 60),
    tidy = TRUE)
ames <- readr::read_delim(file = "ames_housing_data.csv", delim = ",")
# chamge from scientic notations, to restore to default
# options(scipen = 0)
options(scipen = 999)
library(magrittr)
waterfall <- ames %>% dplyr::filter(Zoning != "RL") %>% dplyr::group_by(Zoning) %>%
    dplyr::summarise(counts = n())

# Print waterfall table
knitr::kable(waterfall, align = c("l", "r"), caption = "Drop waterfall")

# sample population
SampledPop <- dplyr::filter(ames, Zoning == "RL")
# define variables of interest from the data
colsofinterest <- c("LotArea", "LotConfig", "Neighborhood", "HouseStyle",
    "OverallCond", "YearRemodel", "TotalBsmtSF", "GrLivArea",
    "BsmtFullBath", "BsmtHalfBath", "FullBath", "HalfBath", "BedroomAbvGr",
    "KitchenQual", "TotRmsAbvGrd", "GarageArea", "MoSold", "YrSold",
    "SaleCondition", "SalePrice")
library(dplyr)
SampleFrame <- SampledPop %>% select_(.dots = colsofinterest)

library(mosaic)
# populate table with statistics
sanitycheck <- do.call(rbind, dfapply(SampleFrame, favstats,
    select = is.numeric))
# format numbers in the table for mean and sd
sanitycheck$mean <- as.numeric(format(sanitycheck$mean, digits = 1,
    nsmall = 2))
sanitycheck$sd <- as.numeric(format(sanitycheck$sd, digits = 1,
    nsmall = 2))
# subset character columns from sanitycheck
sanitycheckcharacter <- select(SampleFrame, colnames(SampleFrame[1,
    sapply(SampleFrame, class) == "character"]))


library(purrr)
UniqueVals <- sanitycheckcharacter %>% map(unique)
# s <-
# data.frame(names(tst),sapply(tst,function(x){paste(x,collapse
# = ',')}),row.names = NULL)
Counts <- data.frame(sapply(UniqueVals, length), do.call(rbind,
    dfapply(sanitycheckcharacter, length, select = is.character)),
    do.call(rbind, dfapply(sanitycheckcharacter, n_missing, select = is.character)),
    row.names = names(UniqueVals))
colnames(Counts) <- c("# Unique", "n", "missing")
knitr::kable(sanitycheck, caption = "Data sanity check for numeric variables")
knitr::kable(Counts, caption = "Data sanity check for nominal variables",
    align = c("l", "r", "r", "r"))
```

```
# Add attribute columns for exploring purposes
SampleFrame <- SampleFrame %>% mutate(ID = 1:nrow(SampleFrame)) %>%
    mutate(yrmonth = YrSold + MoSold/100) %>% mutate(logSalePrice = log10(SalePrice)) %>%
    arrange(Neighborhood, yrmonth)

# Explore time series source('Multiplot.R')
library(qicharts)
library(ggplot2)

controlchart <- qic(data = SampleFrame, y = SalePrice, x = ID,
    chart = "i", xlab = "ID", ylab = "SalePrice")
ggplot(SampleFrame) + geom_histogram(mapping = aes(x = SalePrice),
    binwidth = 10000) + theme_bw() + ggtitle("Histogram SalesPrice")

ggplot(data = SampleFrame, mapping = aes(x = Neighborhood, y = SalePrice,
    color = Neighborhood)) + geom_boxplot(orientation = "h") +
    geom_point() + geom_jitter(position = position_jitter(width = 0.1,
    height = 0.1)) + coord_flip() + theme_classic() + theme(legend.position = "none") +
    geom_hline(yintercept = 300000)
ggplot(data = SampleFrame) + geom_bar(mapping = aes(x = Neighborhood,
    fill = Neighborhood)) + coord_flip() + theme_bw() + theme(legend.position = "none")
ggplot(data = SampleFrame, mapping = aes(x = SaleCondition, y = SalePrice,
    color = SaleCondition)) + geom_boxplot(orientation = "h") +
    geom_point() + geom_jitter(position = position_jitter(width = 0.1,
    height = 0.1)) + coord_flip() + theme_classic() + theme(legend.position = "none")
ggplot(data = SampleFrame, mapping = aes(x = KitchenQual, y = SalePrice,
    color = KitchenQual)) + geom_boxplot(orientation = "h") +
    geom_point() + geom_jitter(position = position_jitter(width = 0.1,
    height = 0.1)) + coord_flip() + theme_classic() + theme(legend.position = "none") +
    xlab("Kitchen Quality")
ggplot(data = SampleFrame, mapping = aes(y = SalePrice, x = HouseStyle,
    fill = HouseStyle)) + geom_boxplot(alpha = 0.25) + coord_flip() +
    facet_wrap(~LotConfig)
ggplot(data = SampleFrame, mapping = aes(x = SalePrice, color = LotConfig,
    fill = LotConfig)) + geom_density(alpha = 0.25) + theme_classic()
p <- ggplot(data = SampleFrame) + theme_bw()
p + geom_point(mapping = aes(x = YearRemodel, y = SalePrice)) +
    geom_smooth(mapping = aes(x = YearRemodel, y = SalePrice),
        method = "loess", se = T) + ylab("Sale Price")
p + geom_point(mapping = aes(x = TotRmsAbvGrd, y = SalePrice)) +
    geom_smooth(mapping = aes(x = TotRmsAbvGrd, y = SalePrice),
        method = "loess", se = T) + xlab("Total Rooms above ground")
p + geom_point(mapping = aes(x = TotalBsmtSF, y = SalePrice)) +
    geom_smooth(mapping = aes(x = TotalBsmtSF, y = SalePrice),
        method = "loess", se = T) + xlab("Total basement area")
library(gridExtra)
SampleFrame <- SampleFrame %>% mutate(TotalBath = BsmtFullBath +
    BsmtHalfBath + FullBath + HalfBath)
Totalbaths <- ggplot(data = SampleFrame) + geom_point(mapping = aes(x = TotalBath,
    y = SalePrice)) + xlab("Total baths") + geom_smooth(mapping = aes(x = TotalBath,
    y = SalePrice), method = "loess", se = T) + xlab("Total bath") +
    theme_bw()
```

```
Fullbaths <- ggplot(data = SampleFrame) + geom_point(mapping = aes(x = FullBath,
    y = SalePrice)) + xlab("Full baths") + geom_smooth(mapping = aes(x = FullBath,
    y = SalePrice), method = "loess", se = T) + xlab("Full bath") +
    theme_bw()

grid.arrange(Totalbaths, Fullbaths, ncol = 2)

LivingArea <- ggplot(SampleFrame) + geom_point(mapping = aes(x = GrLivArea,
    y = SalePrice)) + xlab("Living Area") + geom_smooth(mapping = aes(x = GrLivArea,
    y = SalePrice), se = T) + theme_bw()

GargeArea <- ggplot(SampleFrame) + geom_point(mapping = aes(x = GarageArea,
    y = SalePrice)) + xlab("Garage Area") + geom_smooth(mapping = aes(x = GarageArea,
    y = SalePrice), se = T) + theme_bw()

grid.arrange(LivingArea, GargeArea, ncol = 2)
SampleFrame <- SampleFrame %>% mutate(LogSalePrice = log10(SalePrice)) %>%
    mutate(LogLivArea = log10(GrLivArea)) %>% mutate(LogBasementArea = log10(TotalBsmtSF))

ggplot(data = SampleFrame, mapping = aes(x = LogLivArea, y = LogSalePrice)) +
    geom_point() + geom_smooth(method = "lm", se = T) + xlab("Log Living Area") +
    theme_bw()
PositiveBsmtSF <- SampleFrame %>% filter(TotalBsmtSF > 0 & TotalBsmtSF <
    2500)

RestrictedBasement <- ggplot(data = PositiveBsmtSF, mapping = aes(x = TotalBsmtSF,
    y = SalePrice)) + geom_point() + geom_smooth(method = "lm",
    se = T) + xlab("Toal basement area") + theme_bw()

LogsalesPrice <- ggplot(data = PositiveBsmtSF, mapping = aes(x = TotalBsmtSF,
    y = LogSalePrice)) + geom_point() + geom_smooth(method = "lm",
    se = T) + xlab("Toal basement area") + theme_bw()

LogsalesBsmt <- ggplot(data = PositiveBsmtSF, mapping = aes(x = LogBasementArea,
    y = LogSalePrice)) + geom_point() + geom_smooth(method = "lm",
    se = T) + xlab("Toal log basement area") + theme_bw()

grid.arrange(RestrictedBasement, arrangeGrob(LogsalesPrice, LogsalesBsmt),
    ncol = 2)
Exceed <- 100

SampleFrame <- SampleFrame %>% mutate(SaleConditionGrp = ifelse(SaleCondition ==
    "Partial", "Partial", "Other Conditions"))

SelectHood <- SampleFrame %>% dplyr::count(Neighborhood) %>%
    filter(n > Exceed)
df <- SampleFrame %>% dplyr::filter(Neighborhood %in% SelectHood$Neighborhood)
ggplot(data = df, mapping = aes(x = SaleConditionGrp, y = SalePrice,
    color = SaleConditionGrp)) + geom_boxplot() + geom_point() +
    geom_jitter(position = position_jitter(width = 0.1, height = 0.1)) +
    facet_wrap(~Neighborhood) + theme_bw()
```