

**Log Book Entry 1**

**Exploration and Visualisation of Ames Housing Data**

Joseph Deery

40078793

Contents

[1.0 Introduction 2](#_Toc54708277)

[2.0 Descriptive Statistics 2](#_Toc54708279)

[3.0 Data Visualisations 6](#_Toc54708280)

[4.0 Findings 8](#_Toc54708281)

[4.1 Data Quality Issues and Actions 8](#_Toc54708282)

[4.2 Key Insights 9](#_Toc54708283)

[5.0 Reflective Commentary 10](#_Toc54708284)

[Appendix 1: R Code Used 11](#_Toc54708285)

# Introduction

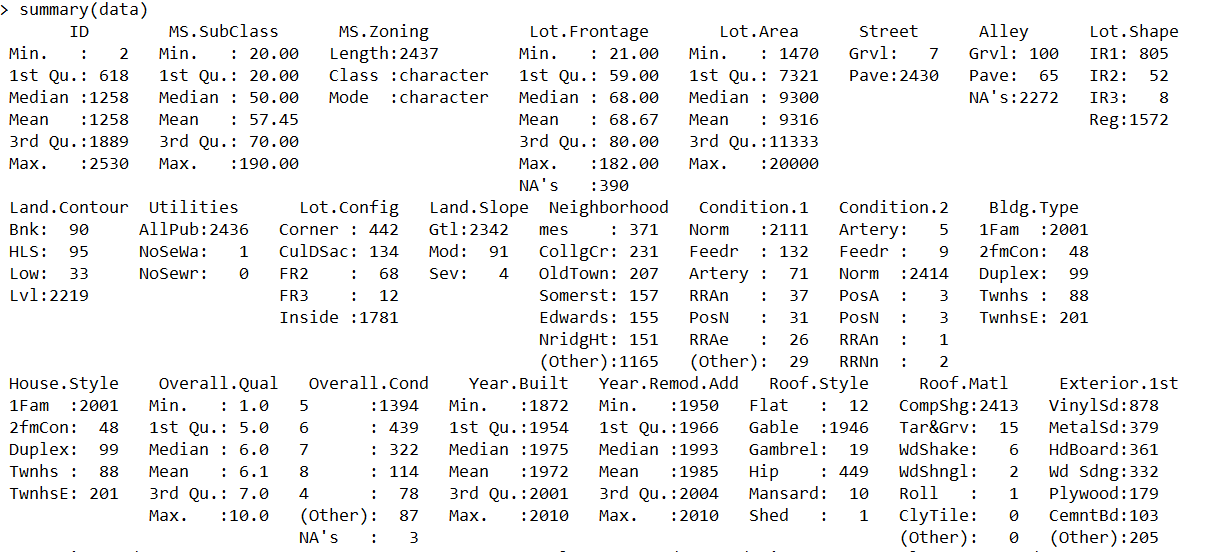
# In this Logbook Entry, I will perform an early-stage analysis of the Ames Housing Data provided, with the purpose of providing insights into what factors influenced sale prices in the area between 2006-2010. This analysis will take the format of any typical analytics project using RStudio; from reading in the data, basic exploring of the data to identify quality issues, ‘cleaning’ /wrangling where necessary and then providing basic descriptive statistics and visual insights using the ‘ggplot2’ package. I will also provide written documentation of findings and key insights that will make the data more valuable/communicable to the typical user.

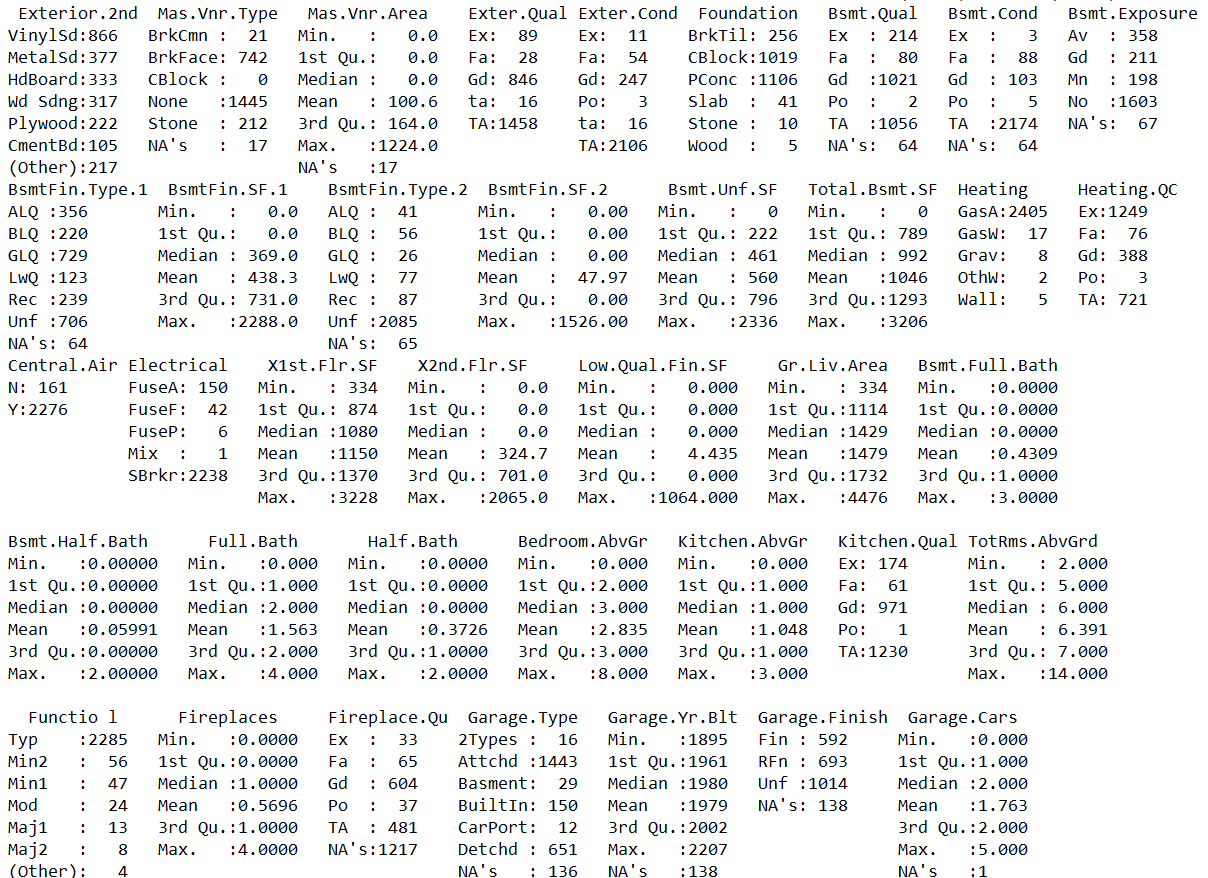
# Descriptive Statistics

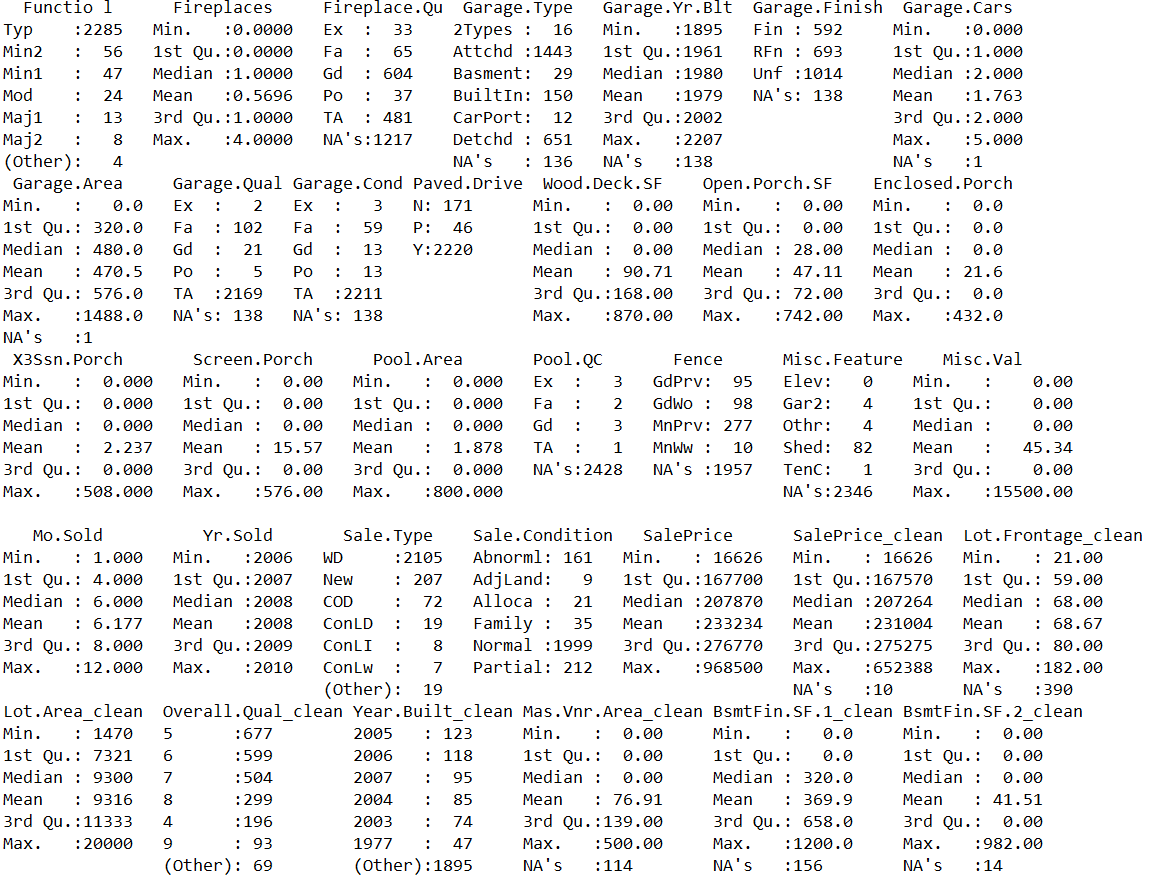
*Summary statistics*:

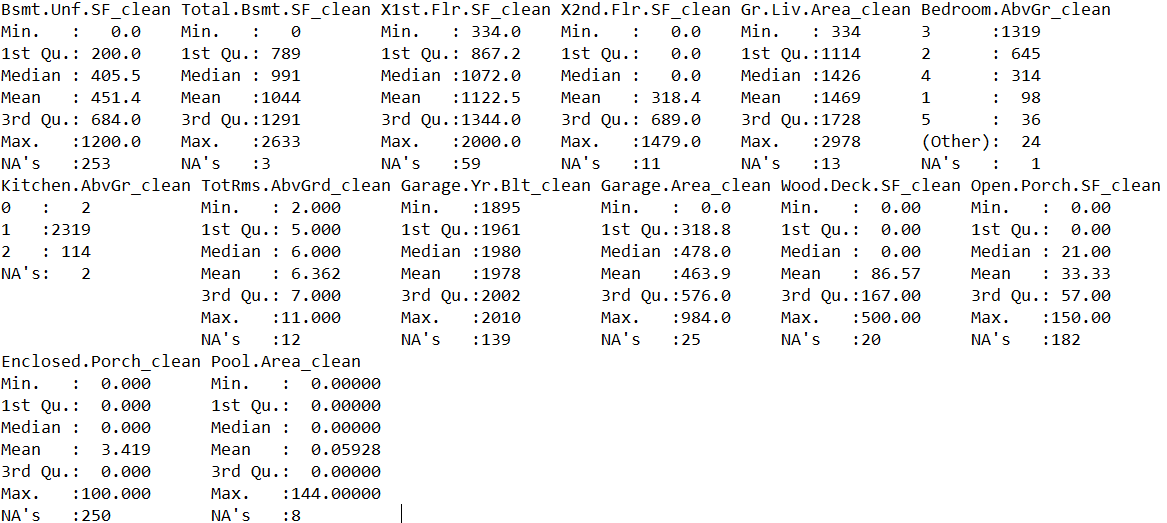
Output in RStudio:

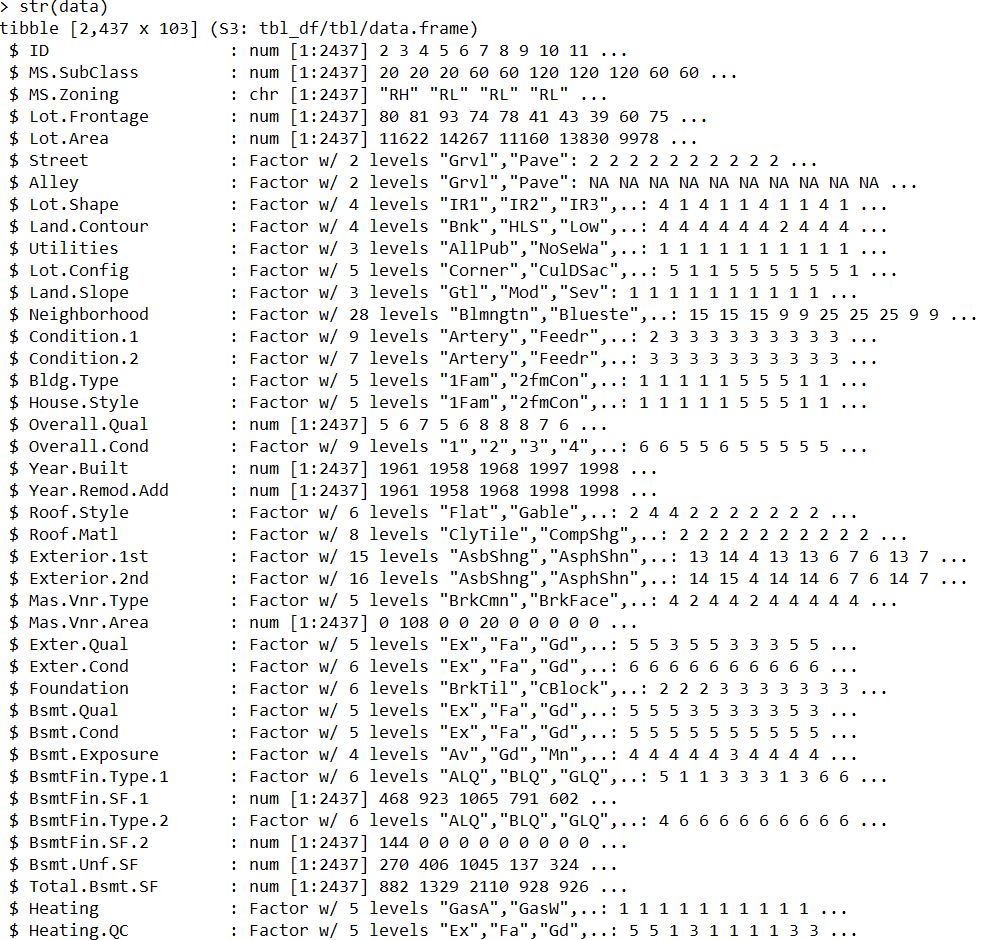
N.B new name given to data that was cleaned e.g “SalePrice\_clean”

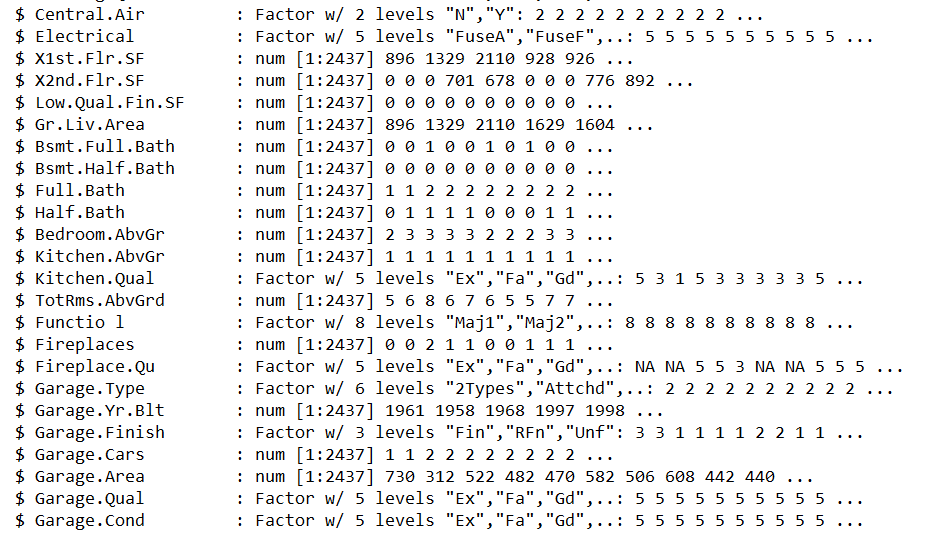


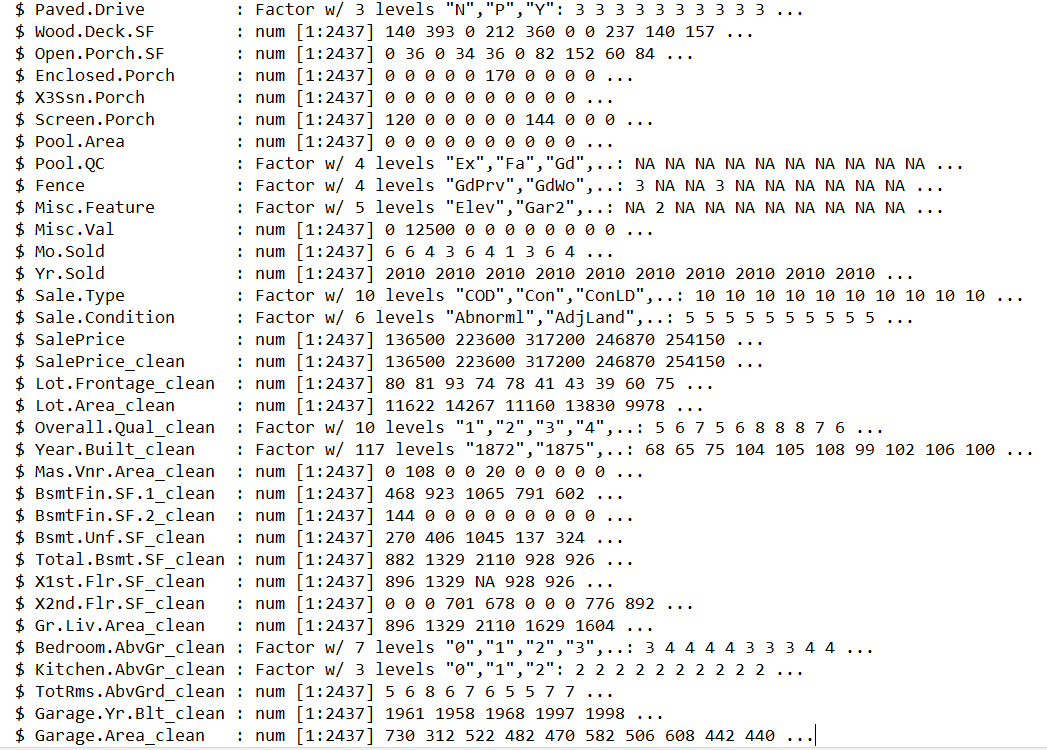






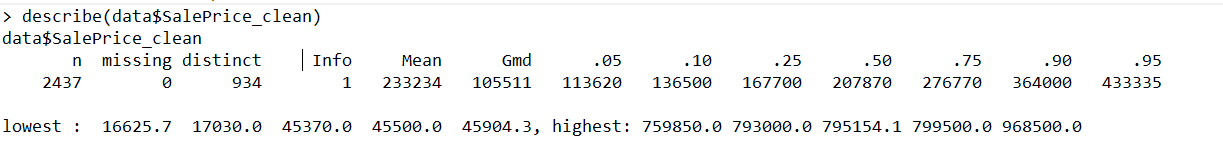


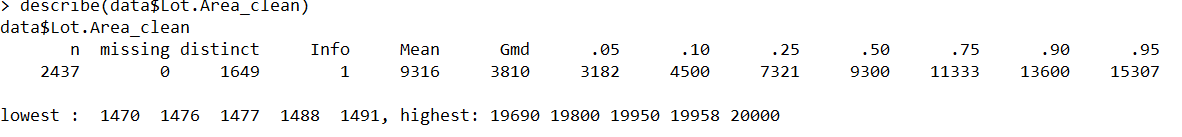




*Descriptives using “Hmisc” Package*

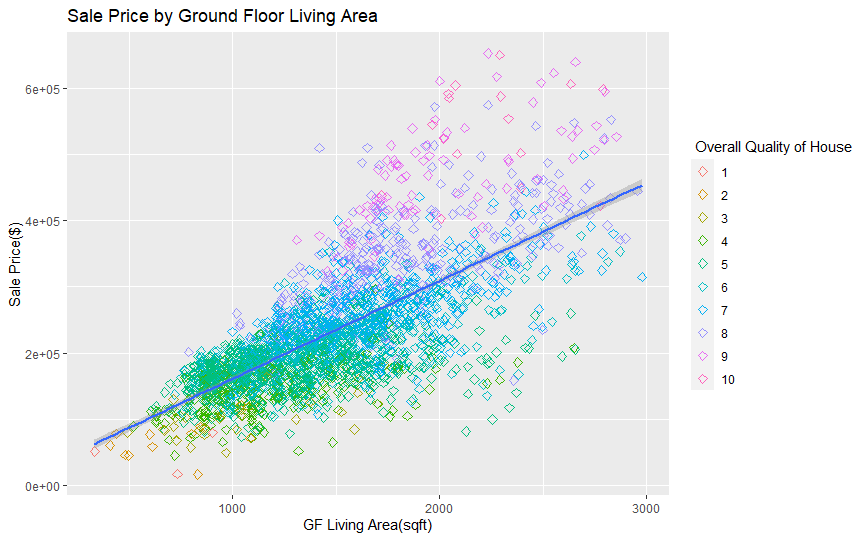
*5 lowest and 5 highest variable scores as well as percentiles, showing effective removal of outliers for important variables.*

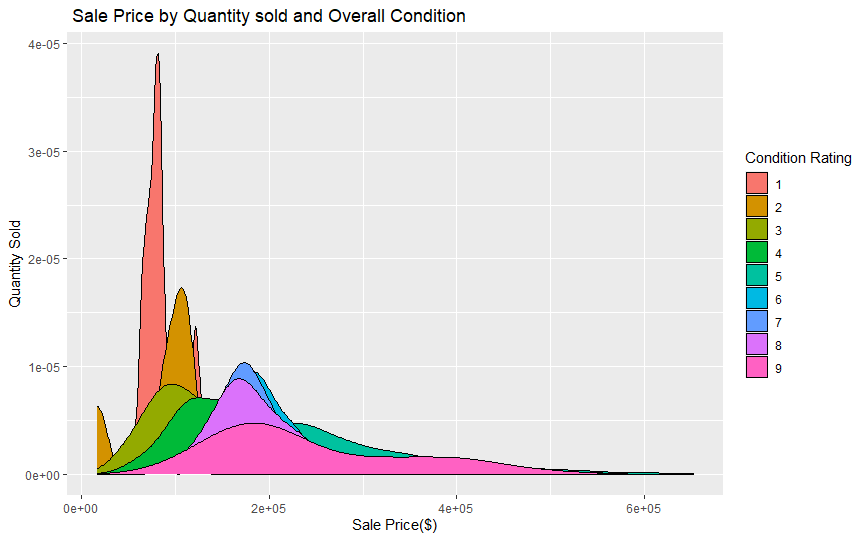


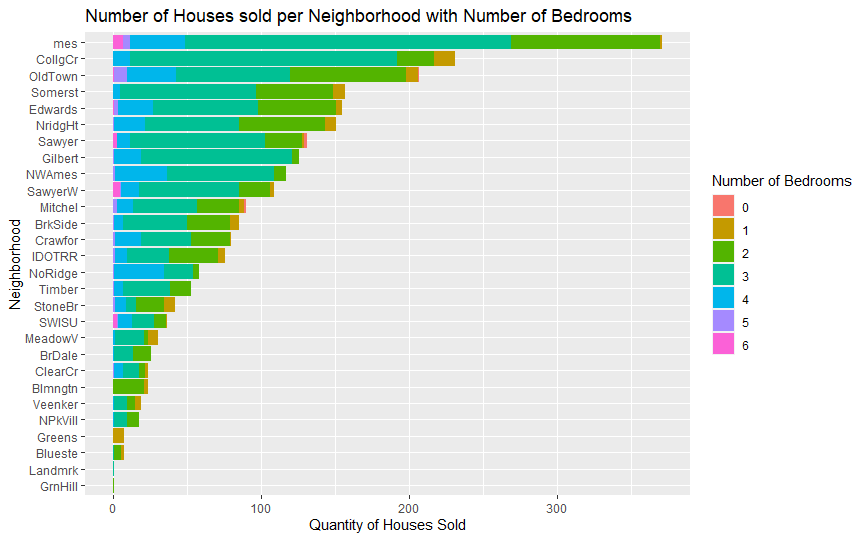


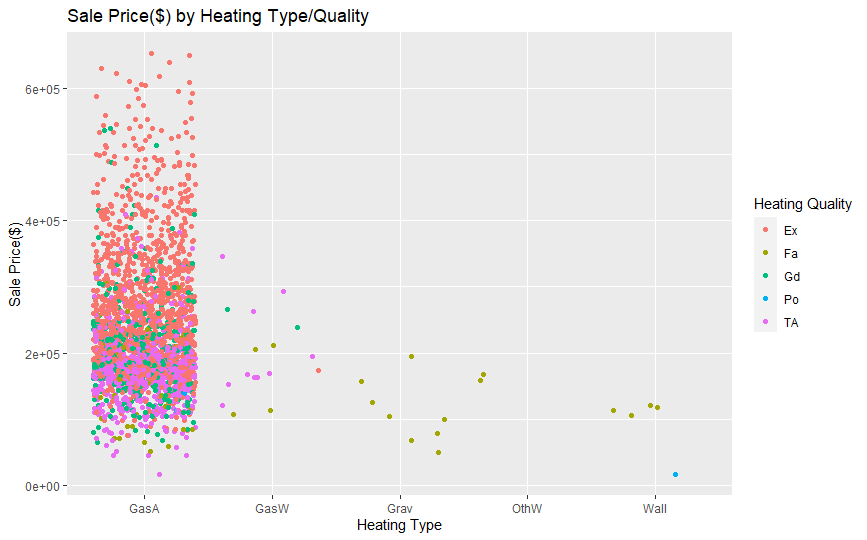
# Data Visualisations

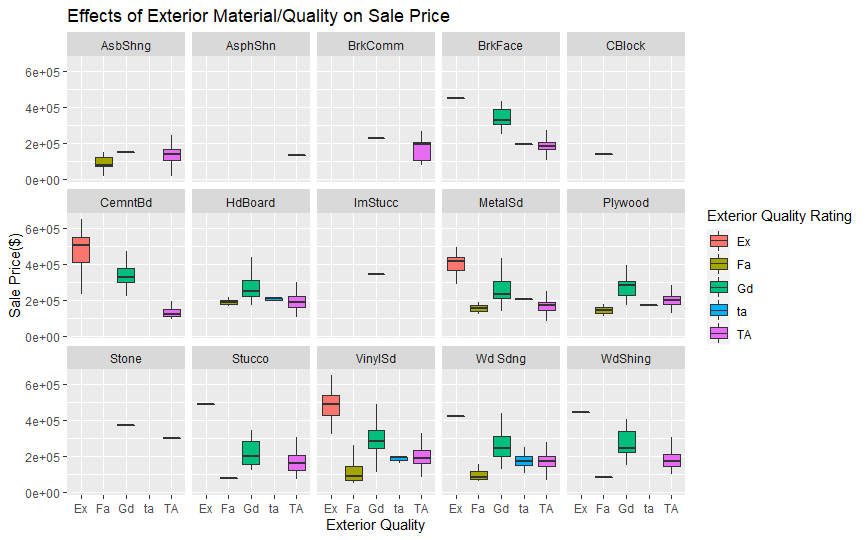
The following visualisations were created using “ggplot2”:

**Figure 1:**

**Figure 2**:

**Figure 3:**

**Figure 4**:

**Figure 5**:

# Findings

## Data Quality Issues and Actions

There were some quality issues in the dataset. Numerical outliers in key variables were removed and new names given as measure of good practice and to maintain the integrity of original data.

Some vectors in dataset were incorrectly formatted. Some were stored as character vectors for example and had to be converted to factors to be of any use for analysis/visualisation.

## Key Insights

Some interesting insights emerged from the analysis (represented visually in Section 3.0).

The scatterplot in *Figure 1* shows a strong positive correlation between the Ground Floor Living Space and eventual sale price of the house. What is particularly interesting is the strength of the correlation, which, on further analysis, proved to be stronger than even the correlation between total lot size and price. This information may be useful for a homeowner/property developer seeking to increase house value by extending the living space. This plot also shows the quality rating of the house has a strong positive correlation, with very few 9/10 rated houses falling below the average sale price. However, more information on what factors determined this quality rating would be of use, particularly for someone seeking to produce a machine-learning model from the data.

*Figure 2* similarly shows the positive correlation between the condition of the house and eventual sale price, as well as the quantity of houses sold within each rating. Houses with a low condition rating sold for a lower price; ratings of 1 or 2 for example sold below the average price, while higher ratings such as 7/8 /9, despite covering a wider range, commanded prices towards the higher end of the scale. This information may be useful for users seeking to increase value, however, more information is needed on what factors/conditions determined these condition ratings.

*Figure 3* gives an insight into the number of houses sold per neighbourhood and the proportion that had a certain number of bedrooms. This gives us a potential insight into what areas were most desirable. Most houses were sold in the North Ames (“mes”) area, with a good distribution of different sized houses in terms of bedroom numbers. This is clearly a popular area in comparison to others, however we do not know from the data how many houses were sold in proportion to the size of the area, crucial information for producing any model using neighbourhood popularity as a variable. We can also see from the data the potentially more affluent areas, by looking at the proportion of 5/6 bedroom houses sold in comparison to the total amount sold, for example Southwest of ISU(SWISU) had a high proportion of houses with 6 bedrooms sold- however this could be student accommodation for example, and more information on these neighbourhoods would be useful.

*Figure 4* produced an interesting insight into the importance of central heating type and quality on house value. The plot shows the importance of gas heating in the area in determining overall house value compared to more traditional heating systems. Houses with warm air gas furnaces commanded prices exclusively over the $400,000 mark- with older heating systems such as wall furnaces barely reaching sale prices of $200,000. Quality appears to be equally important- with gas heating systems demonstrating a higher quality rating. However, it is worth noting that these systems are likely to be newer, particularly in 2006-2010, and may have degraded since then. Nevertheless, heating type/quality seems to be a key factor in determining house sale price.

Lastly, *Figure 5* is a boxplot representing exterior building material/quality and its effect on sale price. The importance of quality is immediately noticeable, regardless of the material, with those with “excellent” quality ratings determining the highest prices, with a median price over $400,000. Houses with a brick face exterior had a higher minimum selling point than any other material. Interestingly, excellent quality cement board seemed to be of value in the area while an asbestos shingle exterior was of the least value.

# Reflective Commentary

There were many stages in my personal development throughout my time completing this Logbook Entry, and of course, a few challenges!

This was my first time sinking my teeth into any sort of programming language. I knew there would be a learning curve, but I was looking forward to the challenge as I need this skill to fulfil my ambition of working in the field of Data Science. I approached the assignment by implementing what I’d learned after each class, giving me time to deal with any struggles, as well as time to digest each step and fully understand it. I found the process to be very satisfying and rewarding. I think somewhat algorithmically, and I find programming in R to be like that. The process of preparing and wrangling the data, for example, was like preparing to cook a good meal!

Of course, I ran into some difficulties along the way. There were some obvious outliers in the data, but others that were not so obvious. I’m looking forward to the feedback as I possibly removed too many of those and got carried away! I also thought I had completed the formatting stage correctly by changing character vectors, however when I went to complete a visualisation I found one of my variables to be numeric when it too should have been a factor. The penny dropped and I formatted those correctly with no further issues. I also had an issue where I had NA’s in 1 or 2 visualisations, but I found the piece of code to remove that on stack overflow, within the plotting function itself (I had left NA’s in many variables as they didn’t appear to be errors, rather lacking features in a house e.g basement).

In conclusion, I thoroughly enjoyed completing this logbook, as I know how crucial these steps are in the early part of any analytics project. I will be able to put into practice/repeat what I have learned in my future career.

# Appendix 1: R Code Used

**Copy and paste below into R to run code(adjust working directory accordingly)**

#this is what the code does....

#get working directory(correct)

getwd()

#install useful packages

install.packages("readxl")

install.packages("tidyverse")

install.packages("ggplot2")

install.packages("psych")

#read in the data

library(readxl)

data<- read\_excel("ames\_train.xlsx")

#summarise the data

summary(data)

#overview of data structure

str(data)

#exploration to identify quality issues

boxplot(data$SalePrice)

hist(data$SalePrice)

boxplot(data$SalePrice, data$Lot.Area)

#deal with numerical outliers/errors

data$SalePrice\_clean <-data$SalePrice

data$SalePrice\_clean[data$SalePrice\_clean>5000000]<- NA

summary(data$SalePrice\_clean)

data$Lot.Frontage\_clean<-data$Lot.Frontage

data$Lot.Frontage\_clean[data$Lot.Frontage\_clean>200]<-NA

data$Lot.Area\_clean <-data$Lot.Area

data$Lot.Area\_clean[data$Lot.Area\_clean>20000] <-NA

data$Overall.Qual\_clean <- data$Overall.Qual

data$Overall.Qual\_clean[data$Overall.Qual\_clean<0] <-NA

data$Year.Built\_clean <- data$Year.Built

data$Year.Built\_clean[data$Year.Built\_clean>2010]<-NA

data$Mas.Vnr.Area\_clean<- data$Mas.Vnr.Area

data$Mas.Vnr.Area\_clean[data$Mas.Vnr.Area\_clean>500] <-NA

data$BsmtFin.SF.1\_clean<- data$BsmtFin.SF.1

data$BsmtFin.SF.1\_clean[data$BsmtFin.SF.1\_clean>1200] <-NA

data$BsmtFin.SF.2\_clean<- data$BsmtFin.SF.2

data$BsmtFin.SF.2\_clean[data$BsmtFin.SF.2\_clean>1000] <-NA

data$Bsmt.Unf.SF\_clean<-data$Bsmt.Unf.SF

data$Bsmt.Unf.SF\_clean[data$Bsmt.Unf.SF\_clean>1200] <-NA

data$Total.Bsmt.SF\_clean<-data$Total.Bsmt.SF

data$Total.Bsmt.SF\_clean[data$Total.Bsmt.SF\_clean>3000] <-NA

data$X1st.Flr.SF\_clean<- data$X1st.Flr.SF

data$X1st.Flr.SF\_clean[data$X1st.Flr.SF\_clean>2000] <-NA

data$X2nd.Flr.SF\_clean<- data$X2nd.Flr.SF

data$X2nd.Flr.SF\_clean[data$X2nd.Flr.SF\_clean>1500]<-NA

data$Gr.Liv.Area\_clean<- data$Gr.Liv.Area

data$Gr.Liv.Area\_clean[data$Gr.Liv.Area\_clean>3000]<-NA

data$Bedroom.AbvGr\_clean<- data$Bedroom.AbvGr

data$Bedroom.AbvGr\_clean[data$Bedroom.AbvGr\_clean>6]<-NA

data$Kitchen.AbvGr\_clean<-data$Kitchen.AbvGr

data$Kitchen.AbvGr\_clean[data$Kitchen.AbvGr\_clean>2]<-NA

data$TotRms.AbvGrd\_clean <- data$TotRms.AbvGrd

data$TotRms.AbvGrd\_clean[data$TotRms.AbvGrd\_clean>11] <-NA

data$Garage.Yr.Blt\_clean<-data$Garage.Yr.Blt

data$Garage.Yr.Blt\_clean[data$Garage.Yr.Blt\_clean>2010]<-NA

data$Garage.Area\_clean <-data$Garage.Area

data$Garage.Area\_clean[data$Garage.Area\_clean>1000]<-NA

data$Wood.Deck.SF\_clean<-data$Wood.Deck.SF

data$Wood.Deck.SF\_clean[data$Wood.Deck.SF\_clean>500]<-NA

data$Open.Porch.SF\_clean<-data$Open.Porch.SF

data$Open.Porch.SF\_clean[data$Open.Porch.SF\_clean>150]<-NA

data$Enclosed.Porch\_clean <-data$Enclosed.Porch

data$Enclosed.Porch\_clean[data$Enclosed.Porch\_clean>100]<-NA

data$Pool.Area\_clean<-data$Pool.Area

data$Pool.Area\_clean[data$Pool.Area\_clean>200] <-NA

#convert categorical data to factors

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

data$Lot.Shape <- as.factor(data$Lot.Shape)

data$Land.Contour <- as.factor(data$Land.Contour)

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

data$Lot.Config <-as.factor(data$Lot.Config)

data$Land.Slope <-as.factor(data$Land.Slope)

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

data$Condition.1 <-as.factor(data$Condition.1)

data$Condition.2 <-as.factor(data$Condition.2)

data$Bldg.Type <- as.factor(data$Bldg.Type)

data$House.Style <- as.factor(data$Bldg.Type)

data$Roof.Style <-as.factor(data$Roof.Style)

data$Roof.Matl <-as.factor(data$Roof.Matl)

data$Exterior.1st<-as.factor(data$Exterior.1st)

data$Exterior.2nd<-as.factor(data$Exterior.2nd)

data$Mas.Vnr.Type <-as.factor(data$Mas.Vnr.Type)

data$Exter.Qual <-as.factor(data$Exter.Qual)

data$Exter.Cond <-as.factor(data$Exter.Cond)

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

data$Bsmt.Qual <-as.factor(data$Bsmt.Qual)

data$Bsmt.Cond <-as.factor(data$Bsmt.Cond)

data$Bsmt.Exposure <-as.factor(data$Bsmt.Exposure)

data$BsmtFin.Type.1 <-as.factor(data$BsmtFin.Type.1)

data$BsmtFin.Type.2 <-as.factor(data$BsmtFin.Type.2)

data$Heating.QC <-as.factor(data$Heating.QC)

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

data$Central.Air <-as.factor(data$Central.Air)

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

data$Kitchen.Qual <-as.factor(data$Kitchen.Qual)

data$`Functio l` <-as.factor(data$`Functio l`)

data$Garage.Type <-as.factor(data$Garage.Type)

data$Garage.Finish <-as.factor(data$Garage.Finish)

data$Garage.Qual <-as.factor(data$Garage.Qual)

data$Garage.Cond <-as.factor(data$Garage.Cond)

data$Paved.Drive <-as.factor(data$Paved.Drive)

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

data$Pool.QC <-as.factor(data$Pool.QC)

data$Misc.Feature <-as.factor(data$Misc.Feature)

data$Sale.Type <-as.factor(data$Sale.Type)

data$Sale.Condition <-as.factor(data$Sale.Condition)

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

data$Fireplace.Qu <-as.factor(data$Fireplace.Qu)

data$Bedroom.AbvGr\_clean<-as.factor(data$Bedroom.AbvGr\_clean)

data$Kitchen.AbvGr\_clean<- as.factor(data$Kitchen.AbvGr\_clean)

data$Year.Built\_clean<-as.factor(data$Year.Built\_clean)

data$Overall.Qual\_clean<-as.factor(data$Overall.Qual\_clean)

data$Overall.Cond<-as.factor(data$Overall.Cond)

summary(data$Overall.Qual\_clean)

#locate and remove missing values where necessary(not removing when potential feature of house eg no basement)

is.na(data)

sum(is.na(data))

data<- subset(data, !is.na(data$SalePrice\_clean))

data<- subset(data, !is.na(data$Lot.Area\_clean))

data<- subset(data, !is.na(data$Overall.Qual\_clean))

data<- subset(data, !is.na(data$Year.Built\_clean))

#descriptive statistics of clean data

summary(data)

str(data)

install.packages("Hmisc")

library(Hmisc)

describe(data$SalePrice\_clean)

describe(data$Lot.Area\_clean)

#data visualisations using ggplot2

library(ggplot2)

library(tidyverse)

##scatterplot

ggplot(data=data, mapping=aes(x=Gr.Liv.Area\_clean, y=SalePrice\_clean))+

geom\_point(size=2, shape=5, mapping=aes(colour=Overall.Qual\_clean))+

geom\_smooth(method="lm")+

labs(title="Sale Price by Ground Floor Living Area", x= "GF Living Area(sqft)", y="Sale Price($)")+

labs(colour=" Overall Quality of House")

##density estimate

data %>%

filter(!is.na(Overall.Cond)) %>%

ggplot() +

geom\_density(mapping = aes(x=SalePrice\_clean, fill=Overall.Cond))+

labs(x="Sale Price($)", y="Quantity Sold", title=" Sale Price by Quantity sold and Overall Condition")+

labs(fill="Condition Rating")

##barchart

data %>%

filter(!is.na(Bedroom.AbvGr\_clean)) %>%

ggplot()+

geom\_bar(mapping = aes(fill=Bedroom.AbvGr\_clean, x= reorder(Neighborhood,Neighborhood,length)))+

coord\_flip()+

labs(fill="Number of Bedrooms",title="Number of Houses sold per Neighborhood with Number of Bedrooms",x="Neighborhood",y="Quantity of Houses Sold")

##scatterplot(jitter)

ggplot(data=data, mapping=aes(x=Heating, y=SalePrice\_clean))+

geom\_jitter(mapping=aes(colour=Heating.QC))+

labs(title="Sale Price($) by Heating Type/Quality", x= "Heating Type", y="Sale Price($)")+

labs(colour="Heating Quality")

##boxplot

ggplot(data=data, mapping=aes( x=Exter.Qual, y=SalePrice\_clean, fill=Exter.Qual))+

geom\_boxplot(outlier.shape = NA)+

facet\_wrap(facets = ~Exterior.1st, nrow = 3, ncol=5)+

labs(fill="Exterior Quality Rating", title="Effects of Exterior Material/Quality on Sale Price", x= "Exterior Quality", y= "Sale Price($)")