Log Book Entry 1

Exploration and Visualisation of Ames Housing Data

Pratik Prakash Brahmapurkar

Contents

1.0 Introduction

The Ames housing dataset looks at the characteristics of houses sold in Ames between 2006 and 2010. The data set consists of 81 variables and 2493 observations. The dataset is examined by taking into account nine factors, which are as follows: zoning classification, Building types, Overall Quality, Year Built, Total Basement area, above grade living area, Cars in garage, Garage Area and Sales Price.

The dataset is initially read into a variable before being analysed for outliers and missing values. Before constructing a visualisation, the data is cleaned. Insights are offered with the help of suitable visualisation using ggplot. Different packages are used like readxl, ggplot2 and tidyverse in R.

2.0 Descriptive Statistics

Analysing the missing value in R we can see that only 11 houses have pool area, but there is no data available on pool quality. More than 96% of houses doesn't consists elevator, 2nd garage, shed, tennis court or any other miscellaneous features. Maximum houses do not have alley access and fence.

Variable	Total NA's
Pool.QC	2493
Misc.Feature	2406
Alley	2319
Fence	2010

Table 1. NA's for top 4 variables

<u>Sales Price:</u>
Summarizing the sales price of the house indicates that the mean price is 259,785 USD and the maximum sale price is 10,800,000 USD which is almost 40 times of mean and median.

Summary of Sales Price (in USD)			
Mean	**Median**	**Minimum**	**Maximum**
259785	224000	17905	10800000

Table 2. Shows Summary of Ames Sales Price

<u>Overall quality:</u>
As seen in table 3 below, summarising the over quality reveals that there are extremely few Very Poor, Poor, and Very Excellent overall quality. There is also an 11 rating overall quality with 7 houses, but no data is available in Ames dictionary for 11 number rating.

Overall Quality	Total
11 Unknown	7
10 Very Excellent	26
9 Excellent	94
8 Very Good	296
7 Good	516
6 Above Average	630
5 Average	679
4 Below Average	193
3 Fair	36
2 Poor	13
1 Very Poor	3

Table 3. Shows summary of Overall Quality of house

<u>Above ground living area:</u>

When we sum up the Above grade (ground) living area, we see that there is a significant gap between the third quartile and the highest value, as indicated in the table 4 below. The largest value is about four times the mean.

Above grade (ground) living area square feet				
Minimum	Median	Mean	3rd Quartile	Maximum
334	1450	1500	1743	5642

Table 4. Shows summary of above ground living area

<u>Garage Area:</u>

When we sum up the size of the garage area, we can see that the greatest value is over three times the mean, as indicated in the table 5 below. There are 138 records with no garage as well.

Size of garage in square feet			
Minimum	Mean	Median	Maximum
0	473.3	480	1488

Table 5. Shows summary of garage area

<u>Year Built:</u>

We can observe that there are few year values which are 999. And the latest construction year is 2010.

Original construction date (in Years)			
Minimum	Median	3rd Quartile	Maximum
999	1974	2001	2010

Table 6. Shows summary of year built

<u>Total Basement Area</u>

According to the summary, there is a significant difference between the third quartile and the maximum value, as indicated in the table 7.

Total square feet of basement area			
Mean	Median	3rd Quartile	Maximum
1050.3	989	1302	6110

Table 7. Shows summary of basement area

3.0 Data Visualisations

1. Comparison of Sales Price and year built with different zone types

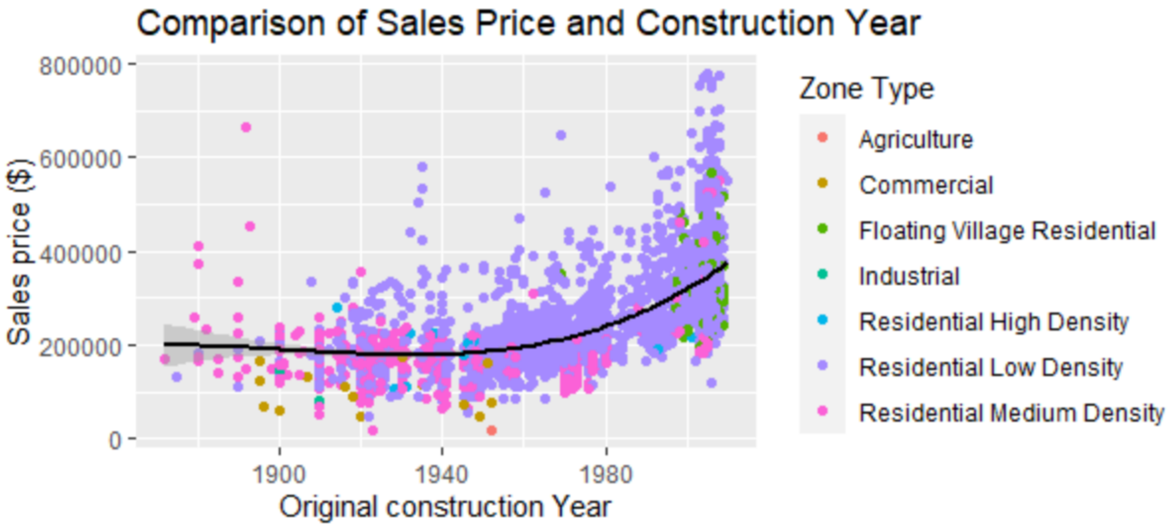

Fig. Visualisation 1.

2. Comparing House Price with respect to overall quality of the house

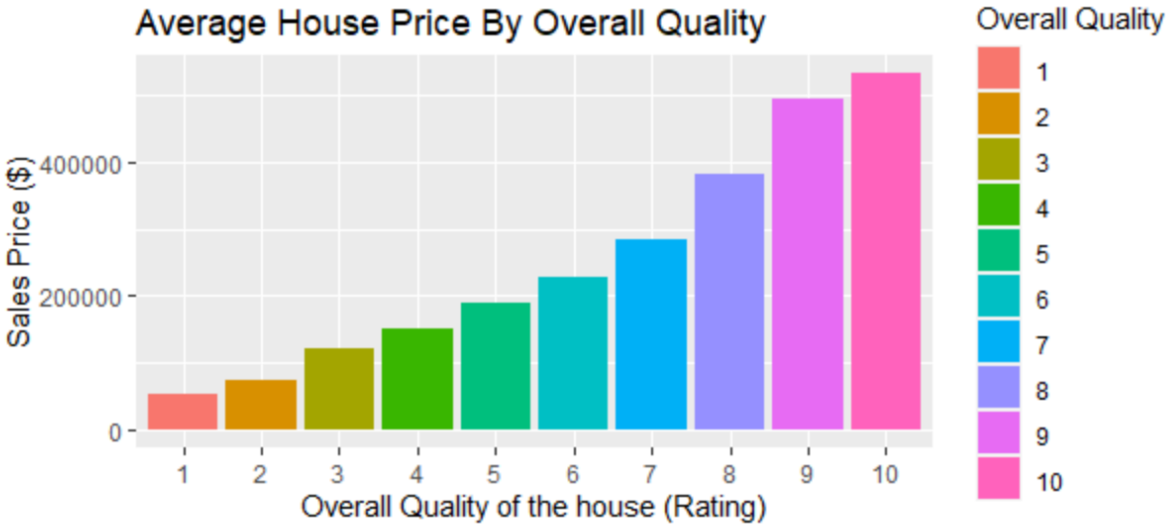

Fig. Visualisation 2.

3. Comparison of sales price and above ground living area with building types

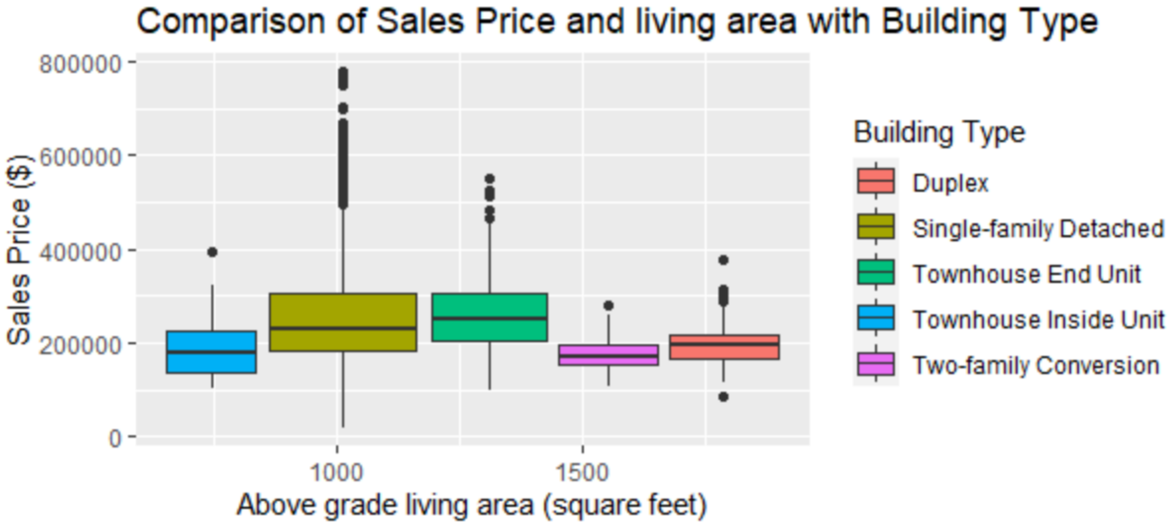

Fig. Visualisation 3.

4. Comparison of Garage area and construction of house year with number of cars

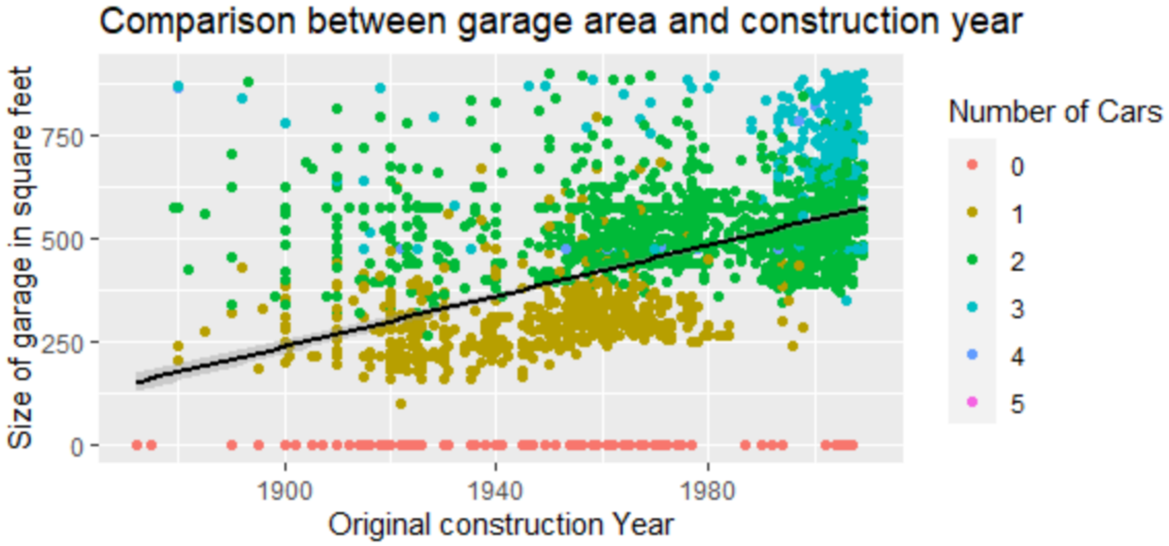

Fig. Visualisation 4.

5. Comparison of Sale Price and Basement area with basement quality

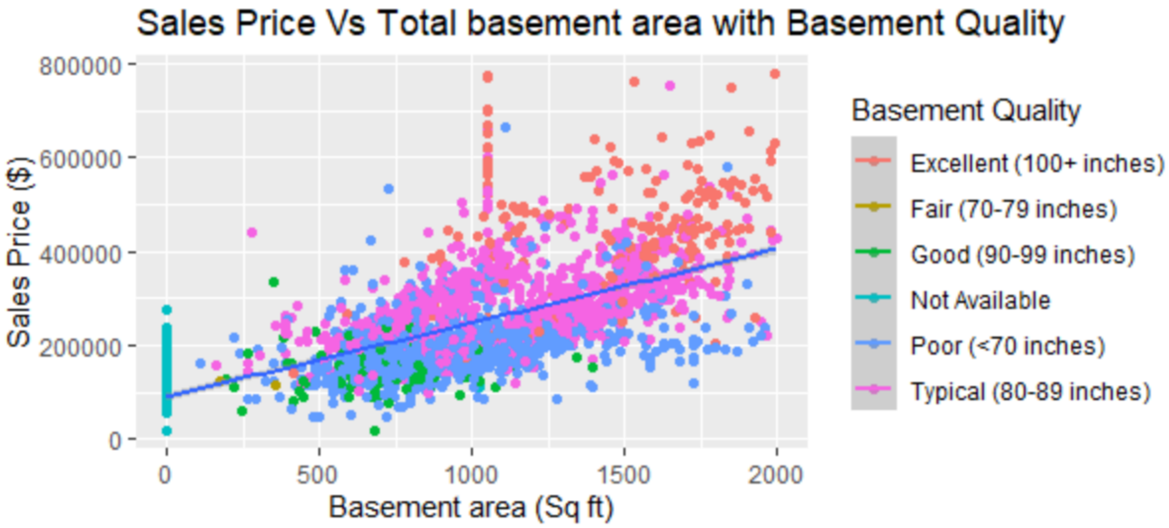

Fig. Visualisation 5.

4.0 Findings

4.1 Data Quality Issues and Actions

Several measures were done to clean the data and remove the outliers. To provide a correct visualisation result, missing data and outliers from specific variables were analysed and processed.

Few variables treated for data quality issue are given below:

1. SalePrice: We can see a huge difference between mean sales price and the maximum amount. We are replacing with mean value for sale price if the value is above 780K USD.
2. Year.Built: There are 5 cells that contain the year 999. It doesn't appear to be a valid year. As a result, we substitute it as round(mean(Year.Built)).
3. Overall.Qual: There are 7 cells, each with an 11-rating, yet the highest rating should be 10. As the data for 7 cells are incorrect, the ratings have been modified to the median value, which is 5. The entire column has also been converted into a factor.
4. Gr.Liv.Area: There is a considerable gap between the mean and maximum values. As a result, the results for areas larger than 2200 square feet are converted to mean values.
5. Garage.Area: There is a great gap between the third and highest quartiles. As a result, the values for garage area in square feet more than 900 are converted to mean value. There is one missing value that must also be considered to be mean.
6. Total.Bsmt.SF: As there is a huge difference between the maximum value which is 6110 and 3rd quartile value, the top values must not be taken in consideration. Values more than 2000 square feet are transformed to a mean value. There is also a blank cell that is transformed to mean.

4.2 Key Insights

By cleaning and analysing the Ames Housing data set, we can see that sales price and built year are the most important elements in identifying maximum insights. The primary conclusion we can draw is that as the area size or quality of the house increases, so does the price.

Visualisation 1 shows that until the early 1940s, there was a tendency of buying houses in residential medium density zones, but after that, people began to favour residential low-density houses. If the house was not too old, the sales price were high. The higher the sales price, the better the quality of the house. In Visualisation 2, we can see a linear graph indicating that the sales price is directly proportional to the overall quality of the property.

Visualisation 3 depicts that majority of consumers choose to purchase single-family detached homes. The highest sales price, as well as the lowest sales price, belongs to this category. Observing Visualisation 4, the average size of the garage area has increased year after year. People used to have only one automobile at first, but by the early 1960s, the majority of home owners had two cars. The trend of three automobiles had also began in the early 2000s.Those individuals who did not have any car didn't have any space for the garage area.

The increasing trend in visualisation 5 states that the larger the basement area, the higher is the sale price of the house. Most basement areas below 1200 square feet had poor basement quality which is less than 70 inches. A typical basement quality which is between 80 to 89 inches is observed if it's above 1200-1700 square feet basement area. The better the basement quality the expensive are the prices of the house. Almost all the expensive houses are of excellent base quality.

Appendix 1: R Code Used

```
#Install the required packages
#Read the Packages
library(readxl)
library(psych)
library(ggplot2)
library(tidyverse)

#Read the excel sheet into variable ames
ames <- read_excel("D:/R_Folder/ames_train.xlsx")

#Check quality of Saleprice variable
hist(ames$SalePrice)

# :: Summarize the Data ::
```

```
#Analyze the Columns with NA
colSums(is.na(ames))

#Summarise the data
summary(ames)


#::::: Data Quality Issues and Action ::::::

#Re-summarize the Saleprice to MEAN above 780K USD
ames$SalePrice[ames$SalePrice > 780000] <- mean(ames$SalePrice, na.rm = TRUE)

#Re-summarize the Year.Built consisting 999 with median year
ames$Year.Built[ames$Year.Built == 999] <- round(mean(ames$Year.Built, na.rm = TRUE))

#Change the rating for Overall.Qual which consists of 11 to median value
ames$Overall.Qual[ames$Overall.Qual == 11] <- median(ames$Overall.Qual, na.rm = TRUE)

#Convert Overall.Qual into factor
ames$Overall.Qual <- as.factor(ames$Overall.Qual)

#For Gr.Liv.Area the value above 2200 is changed to mean value
ames$Gr.Liv.Area[ames$Gr.Liv.Area > 2200] <- mean(ames$Gr.Liv.Area , na.rm = TRUE)

#Garage area 900 are changed to mean value
ames$Garage.Area[ames$Garage.Area > 900] <- mean(ames$Garage.Area , na.rm = TRUE)

#Convert the missing values into Mean for Garage area
ames$Garage.Area[is.na(ames$Garage.Area)] <- mean(ames$Garage.Area , na.rm = TRUE)

#Convert the NA in Garage.Cars to mean value
ames$Garage.Cars[is.na(ames$Garage.Cars)] <- round(mean(ames$Garage.Cars, na.rm = TRUE))

#For Total.Bsmt.SF values above 2000 are converted into mean value
ames$Total.Bsmt.SF[ames$Total.Bsmt.SF > 2000] <- mean(ames$Total.Bsmt.SF, na.rm = TRUE)

#Convert the missing values into Mean for Total.Bsmt.SF
ames$Total.Bsmt.SF[is.na(ames$Total.Bsmt.SF)] <- mean(ames$Total.Bsmt.SF, na.rm = TRUE)

#Blank Basement Quality are converted into 'Not Available'
ames$Bsmt.Qual[is.na(ames$Bsmt.Qual)] <- 'Not Available'

#:: GGPLOT VISUALISATIONS::

#1. First visualisation comparing Sales Price and Built.Year with respect to MS.Zoning
#and also changing the names for MS.Zoning
#Change name for all the zones into fullform
ames$MS.Zoning[ames$MS.Zoning =='A (agr)'] <- 'Agriculture'
ames$MS.Zoning[ames$MS.Zoning =='C (all)'] <- 'Commercial'
ames$MS.Zoning[ames$MS.Zoning =='FV'] <- 'Floating Village Residential'
ames$MS.Zoning[ames$MS.Zoning =='I (all)'] <- 'Industrial'
ames$MS.Zoning[ames$MS.Zoning =='RH'] <- 'Residential High Density'
ames$MS.Zoning[ames$MS.Zoning =='RL'] <- 'Residential Low Density'
ames$MS.Zoning[ames$MS.Zoning =='RM'] <- 'Residential Medium Density'
#YearBuilt Vs Sales Price - Geom Points with respect to Zoning
```

```
ames %>% ggplot(aes(x=(Year.Built), y=(SalePrice), colour = MS.Zoning))+
  geom_point()+
  geom_smooth(method = lm, formula = y ~ splines::bs(x, 1), colour = 'black')+
  labs(title = 'Comparison of Sales Price and Construction Year',
       x="Original construction Year", y= "Sales price ($)", colour = "Zone Type")+
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))

#2. Overall Quality with respect to Sales Price (GeomBar)
ggplot(ames)+
  geom_bar(mapping = aes(x = Overall.Qual, y=(SalePrice), fill = Overall.Qual),
           stat = "Summary", fun.y = "mean")+
  labs(title = "Average House Price By Overall Quality", x="Overall Quality of the house (Rating)", y=
"Sales Price ($)", fill = "Overall Quality")+
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
  scale_fill_brewer(palette = "Dark2")

#3 Living area Vs Sales Price and rename the Building Types
#Renaming the Building types
ames$Bldg.Type[ames$Bldg.Type == "1Fam"] <- "Single-family Detached"
ames$Bldg.Type[ames$Bldg.Type == "2fmCon"] <- "Two-family Conversion"
ames$Bldg.Type[ames$Bldg.Type == "Duplex"] <- "Duplex"
ames$Bldg.Type[ames$Bldg.Type == "Twnhs"] <- "Townhouse Inside Unit"
ames$Bldg.Type[ames$Bldg.Type == "TwnhsE"] <- "Townhouse End Unit"
#Comparison of Sales Price and living area with Building Type (Boxplot)
ames %>% ggplot(aes(x=Gr.Liv.Area,y=SalePrice, fill = Bldg.Type))+
  geom_boxplot()+
  labs(title = "Comparison of Sales Price and living area with Building Type", x="Above grade living
area (square feet)", y= "Sales Price ($)", fill = "Building Type")+
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
  scale_fill_brewer(palette = "Dark2")

#4. Year date VS Garage Area - Geom Points with respect to Garage Cars
ames %>% ggplot(aes(x=(Year.Built), y=(Garage.Area),color = as.factor(na.exclude(Garage.Cars))))+
    geom_point()+
    geom_smooth(method = lm, formula = y ~ x, colour = 'black')+
    labs(title = 'Comparison between garage area and construction year',x="Original construction
Year", y= "Size of garage in square feet", color = 'Number of Cars')

#5. Total Basement VS Sales Price with Year Sold and also changing the name
#Change the name for Basement Qualities
ames$Bsmt.Qual[ames$Bsmt.Qual == 'Ex'] <- 'Excellent (100+ inches)'
ames$Bsmt.Qual[ames$Bsmt.Qual == 'Fa'] <- 'Good (90-99 inches)'
ames$Bsmt.Qual[ames$Bsmt.Qual == 'Gd'] <- 'Typical (80-89 inches)'
ames$Bsmt.Qual[ames$Bsmt.Qual == 'Po'] <- 'Fair (70-79 inches)'
ames$Bsmt.Qual[ames$Bsmt.Qual == 'TA'] <- 'Poor (<70 inches)'
ames$Bsmt.Qual[ames$Bsmt.Qual == 'NA'] <- 'No Basement'
#Total Basement VS Sales Price - Creation of Geom Points with respect to Base Quality
ames %>% ggplot(aes(y=SalePrice,x=Total.Bsmt.SF,color = Bsmt.Qual))+
  geom_point()+
  stat_smooth(aes(group = 1), method = "lm", formula = y ~ x)+
labs(title = 'Sales Price Vs Total basement area with Basement Quality',x="Basement area (Sq ft)", y=
"Sales Price ($)", color = 'Basement Quality')+
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```