Visualization of Ames Housing Data

Emily C. Mills

January 2020

Visualization of Ames Housing Data

Using the Ames Housing Data set, (a data set which describes residential property sales in Ames, Iowa from 2006 to 2010, and contains explanatory variables broken down into the following categories: 23 nominal, 23 ordinal, 14 discrete, and 20 continuos. The data set which describes the sale of individual residential properties contains 2930 observations) [1] as an example, we will demonstrate how to easily create data visualizations using portable, convenient and affordable technology. An iPad with basic applications as well as the R Compiling Application was used to produce this project.

Large data sets, such as the Ames Housing Data, can be challenging to understand. The technology typically used to create visual representations of data, such as laptops can be costly and unattainable for many college students. In this study, R Programming was used to create various graphs and charts to gain a clearer understanding of the data using an iPad, a relatively affordable alternative to a laptop or desktop computer. iPad and mobile devices are more compact and convenient. This project was completed for a Statistics 177 course at Middlesex Community College in Bedford, Massachusetts in the fall semester of 2019. An iPad mini 2nd generation was used to perform all calculations, compile and code data, create plots and charts, upload files, as well as edit text. The charts in this project were created using the R Compiler App , and App which allows one to use R Programming to compile coding to create charts and graphs easily. Dropbox was used to upload a file and create a link for the data to use in the R Compiler App. Moreover, the App allows one to visualize and work with data from relatively large data sets and additional packages may be installed and used for compiling code. In this case, the ggplot 2 package was installed and used to create charts and plots. ggplot 2 library allows the implementation of The Grammar of Graphics [2] , which makes ggplot2 effective in describing the components of the graphics, and how visualizations represent data.

The following charts were created on the R Compiler Application using data from the Ames Housing Data set. Although the data used for this project was collected by Ames Assessor's Office for Tax purposes (see [1]), it can be used to show factors that relate to past home sale prices and predict the sale price and values of future home sales. Additionally, data collected from one geographical area or region can be used to reliably predict home sale prices in another region [3]. The original data set can be accessed at https://cran.r-project.org/web/packages/AmesHousing as a R package, we use dropbox just for convenience.

Figure 1

Types of Homes in Study

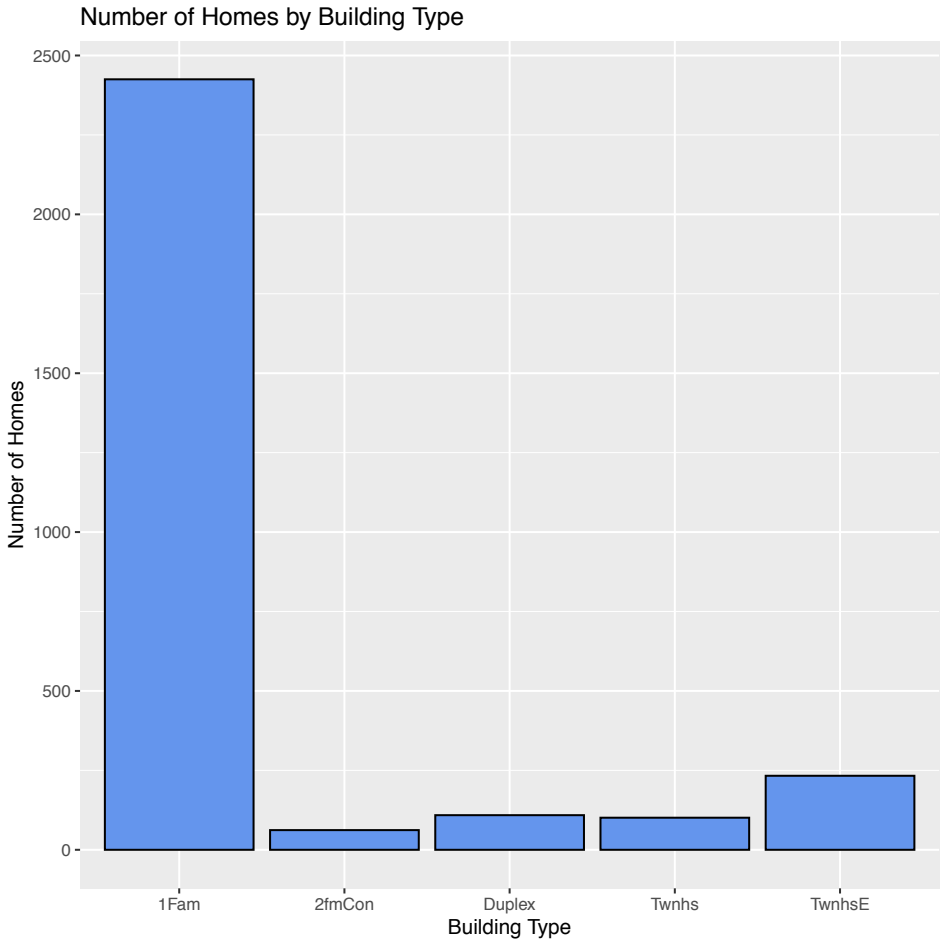


Fig.1 shows the number of homes sorted by building type. The majority are single family detached homes(1Fam). The other building types or type of dwelling that make up a the minority in this study include townhouse end unit (TwnhsE), townhouse inside unit (Twnhs), duplexes(Duplex), and two family conversions; originally built as a one family dwelling(2fmCon).

The following code was used to create this graph:

```
Ameshousing<-read.csv("https://www.dropbox.com/s/d2qgcet7w4a1k28/
AmesHousing.csv?dl=1", header=TRUE, sep=',')
library(ggplot2)
ggplot(data=Ameshousing,aes(x = BldgType)) +
  geom_bar(fill = "cornflowerblue",
           color="black") +
  labs(x = "Building Type",
       y = "Number of Homes",
       title = "Number of Homes by Building Type”
```

Figure 2

Sale Price

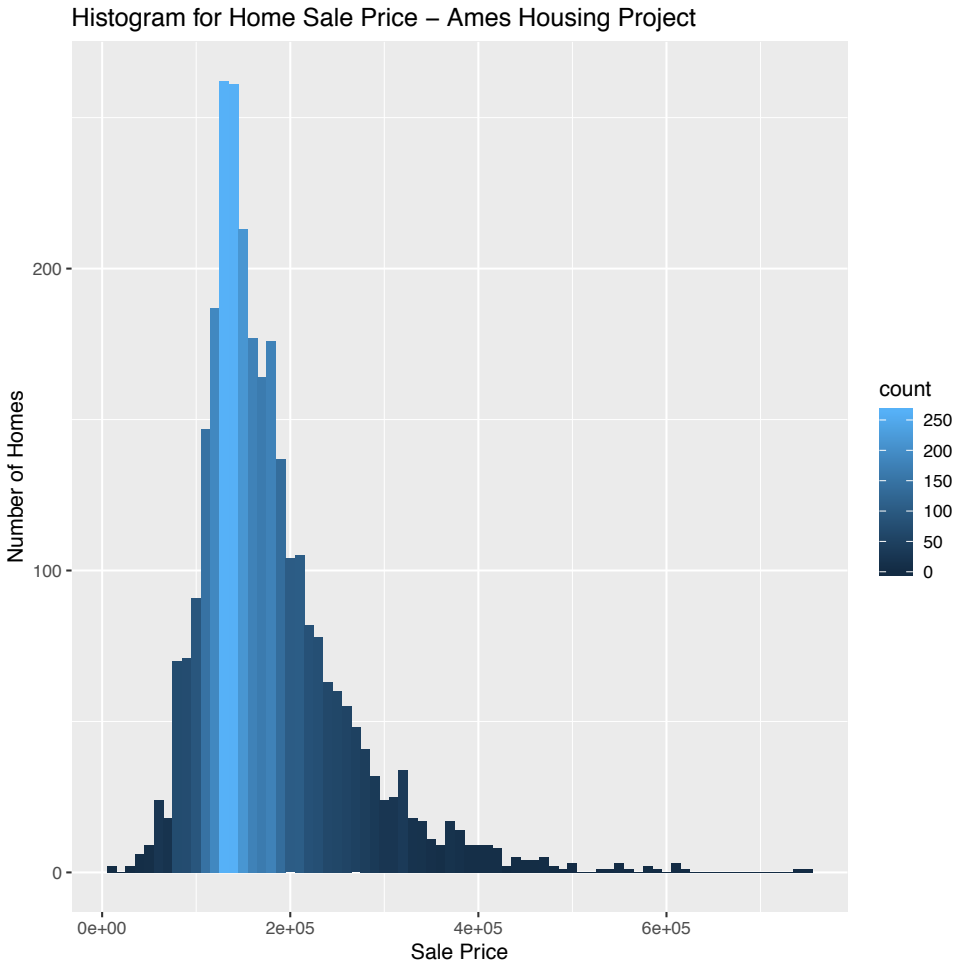
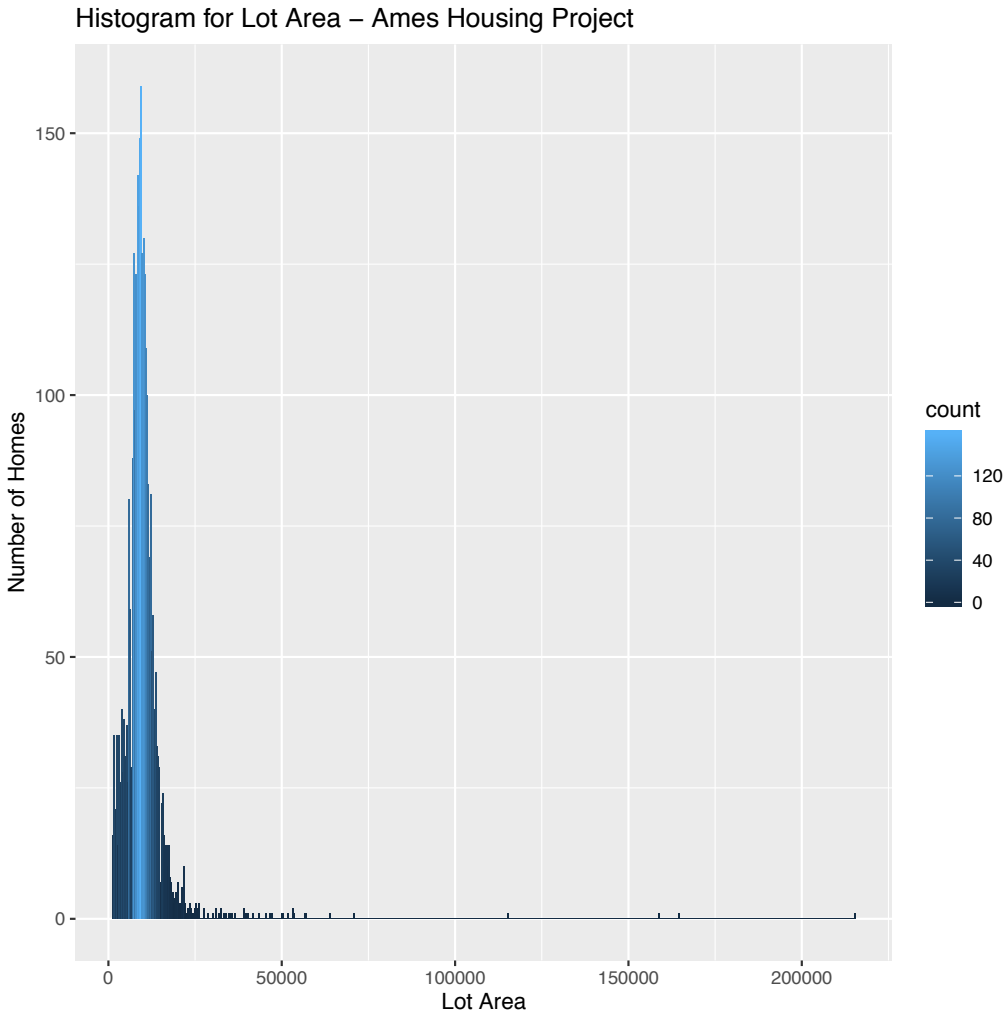


Fig. 2 shows the frequency distribution of the number of homes by sale price. The following code was used to create this graph:

```
Ameshousing<-read.csv("https://www.dropbox.com/s/d2qgcet7w4a1k28/
AmesHousing.csv?dl=1", header=TRUE, sep=',')
library(ggplot2)
ggplot(data=Ameshousing,aes(x=SalePrice))
+geom_histogram(binwidth=10000,aes(y=..density..))+
  labs(x="Sale Price", y="Number of Homes", title = "Histogram for Sale Price - Ames
Housing Project")
```

Figure 3

Lot Area



We again opted for a histogram to present the data for Lot Area. The following code was used to create this graph:

```
Ameshousing<-read.csv("https://www.dropbox.com/s/d2qgcet7w4a1k28/
AmesHousing.csv?dl=1", header=TRUE, sep=',')
library(ggplot2)
ggplot(data=Ameshousing,aes(x=LotArea))
+geom_histogram(binwidth=500,aes(y=..density..))+
  labs(x="Lot Area", y="Number of Homes", title = "Histogram for Lot Area")
```

The histogram clearly presents the data to the reader, which will allow them to easily identify if there are any outliers, as well as see the distribution of Lot Area for properties in this study.

What Raises Sale Price?

Several Scatterplots were created to compare what factors had a greater correlation to the sale price of homes.

Figure 4

Sale Price vs Lot Area

Fig. 4 shows the correlation between Lot Area and Sale Price. The line shows the conditional mean.
The following code was used to create this graph:

```
Ameshousing<-read.csv("https://www.dropbox.com/s/uw0jja0z4usj7t3/
AmesHousingE.csv?dl=1", header=TRUE, sep=',')
library(ggplot2)
ggplot(data=Ameshousing, aes(LotArea, SalePrice)) +
  geom_point(show.legend = FALSE) +
   scale_x_continuous(trans="log2")+
  theme_minimal() +
  geom_smooth(method = lm)+
  labs(x="Lot Area", y="Sale Price", title = "Sale Price vs. Lot Area")
```

Figure 5

Sale Price vs. Overall Quality Rating

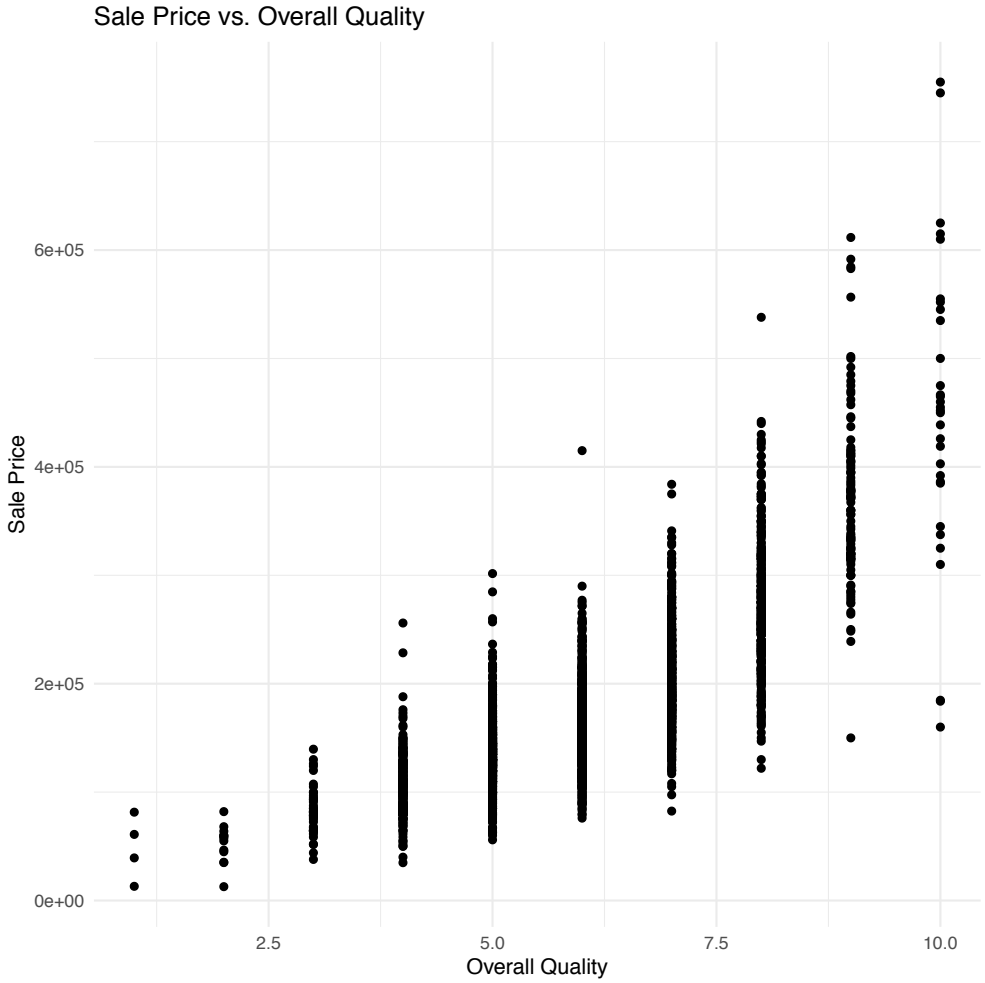


Fig 5. shows the relationship between overall quality rating and sale price. The overall quality in this study was measured on a scale of 1-10 with 1 being the lowest quality and 10 being the highest quality.
The following code was used to create this graph:

```
Ameshousing<-read.csv("https://www.dropbox.com/s/d2qgcet7w4a1k28/
AmesHousing.csv?dl=1", header=TRUE, sep=',')
library(ggplot2)
ggplot(data=Ameshousing, aes(SalePrice, OverallQual)) +
  geom_point(show.legend = FALSE) +
  theme_minimal() +
  labs(x="Sale Price", y="Overall Quality", title = "Sale Price vs. Overall Quality")
```

Figure 6

Sale Price vs. Overall Condition Rating

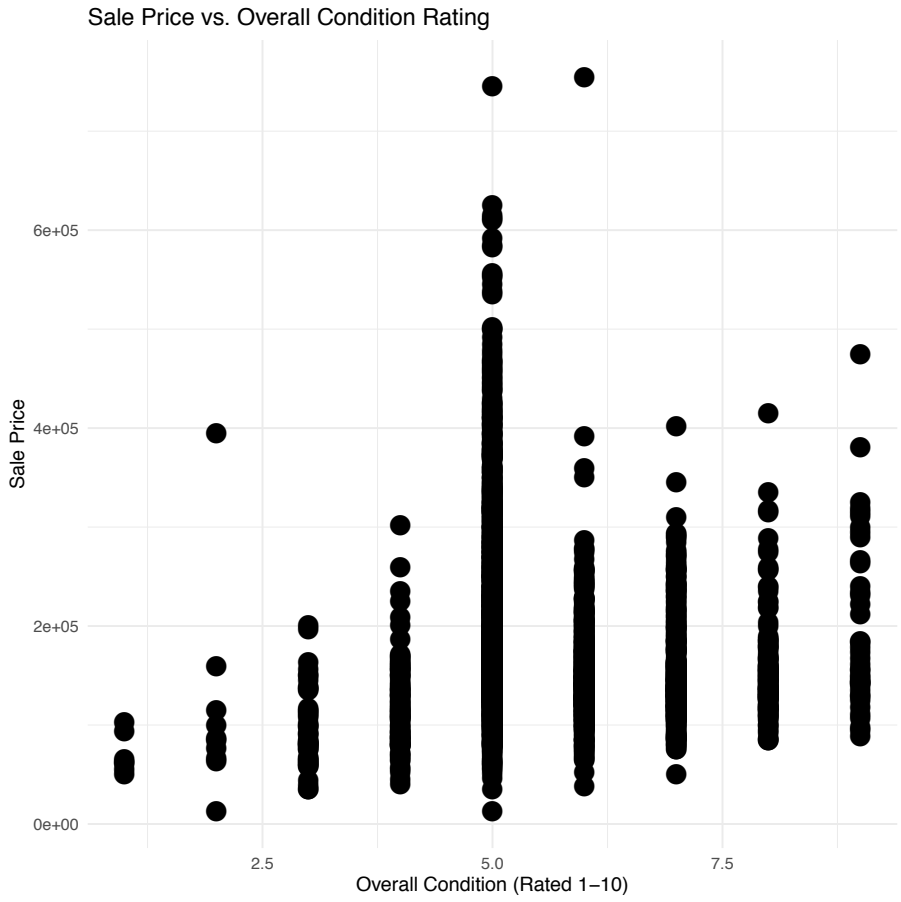


Fig. 6 shows the relationship between overall condition rating of the home and sale price. The overall condition in this study was measured on a scale of 1-10 with 1 being the lowest quality and 10 being the highest.
The following code was used to create this graph:

```
Ameshousing<-read.csv("https://www.dropbox.com/s/d2qgcet7w4a1k28/
AmesHousing.csv?dl=1", header=TRUE, sep=',')
library(ggplot2)
ggplot(data=Ameshousing, aes(OverallQual, SalePrice)) +
  geom_point(show.legend = FALSE) +
  theme_minimal() +
  labs(x="Overall Quality", y="Sale Price", title = "Sale Price vs. Overall Quality")
```

Figure 7

Sale Price vs Year Built

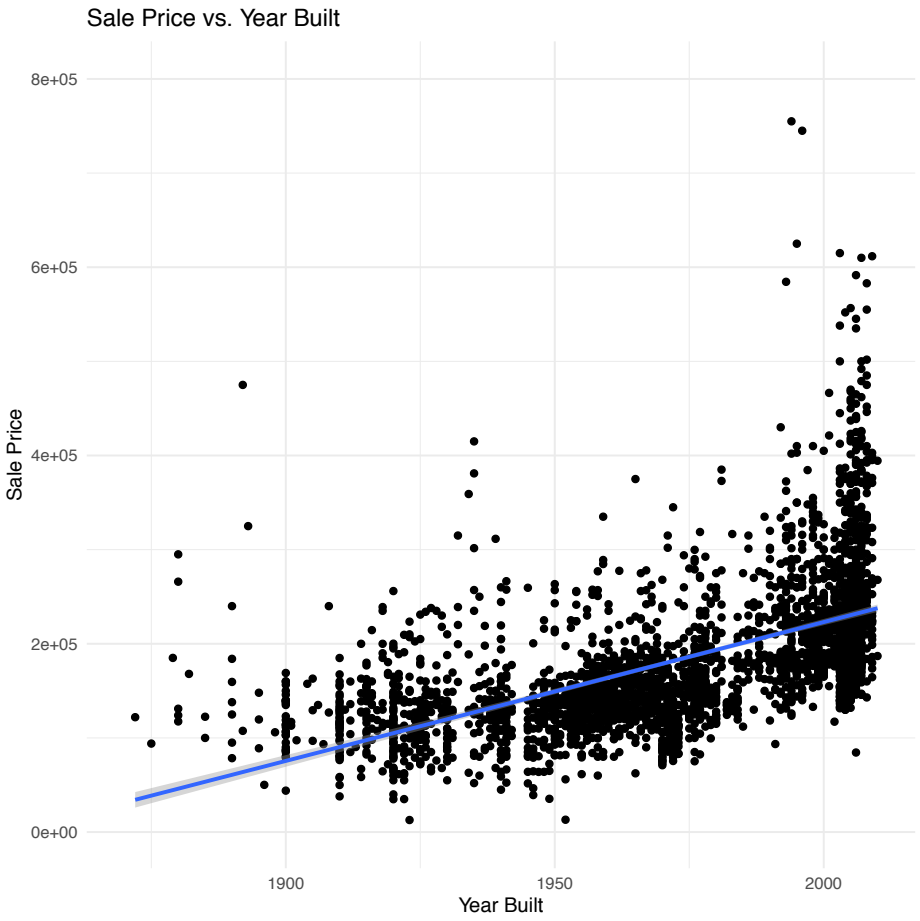


Fig. 7 shows the relationship between the year the home was built and the sale price. The line shows the conditional mean. The following code was used to create this graph:

```
Ameshousing<-read.csv("https://www.dropbox.com/s/d2qgcet7w4a1k28/
AmesHousing.csv?dl=1", header=TRUE, sep=',')
library(ggplot2)
ggplot(data=Ameshousing, aes(YearBuilt, SalePrice)) +
  geom_point(show.legend = FALSE) +
  coord_cartesian(xlim=c(1870, 2010),ylim=c(1500,800000))+
  theme_minimal() +
  geom_smooth(method = lm)+
  labs(x="Year Built", y="Sale Price", title = “Sale Price vs.Year Built")
```

Figure 8

Sale Price vs Total Rooms Above Ground

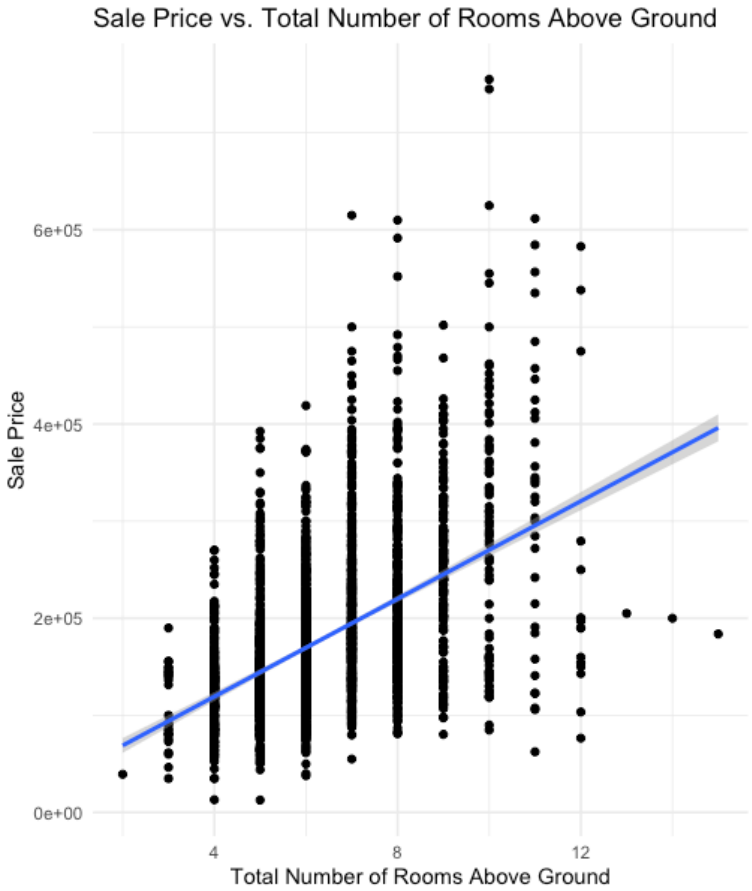


Fig 8. shows the relationship between the total number of rooms above ground in the homes in this study with sale price of the home. The line shows the conditional mean.
The following code was used to create this graph:

```
Ameshousing<-read.csv("https://www.dropbox.com/s/d2qgcet7w4a1k28/
AmesHousing.csv?dl=1", header=TRUE, sep=',')
library(ggplot2)
ggplot(data=Ameshousing, aes(TotRmsAbvGrd, SalePrice)) +
  geom_point(show.legend = FALSE) +
  theme_minimal() +
  geom_smooth(method = lm)+
  labs(x="Total Number of Rooms Above Ground", y="Sale Price", title = "Sale Price vs.
Total Number of Rooms Above Ground")
```

The scatterplots created using R compiler make it easier to understand the data and draw conclusions about various factors. As one can see from the data a larger lot area slightly raised the chances of a property selling for a higher value. The total rooms above ground compared to sale price created a steeper slope and seemed to have a greater impact on home values, with more rooms above ground resulting in a higher sale price. The two factors which showed the greatest impact were the year the home was built and the overall quality rating (1-10) of the homes. The newer home, and the higher overall quality rating both impacting sale price.

When examining the effect overall condition of the home (on a scale of 1-10) compared to sale price, we notice that the homes that sold at the highest price were not the highest rated. In fact, the homes sold at the highest prices were rated in the mid-range (5). One could conclude that a high (8-10) rating of overall condition of the home does not necessarily result in a higher sale price.

iPads are user friendly, easy to use, convenient, affordable, and portable. It allows the user to download additional applications and customize and personalize the device. Essentially, the iPad allows one to perform the same tasks and functions as a laptop or desktop computer. In this case, it was proven that an iPad can be used to conveniently create data visualizations. Using portable technology also allows the user to make edits easily and on go.

References.

1. De Cock, Dean (2011). *Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project*, Journal of Statistics Education", Volume 19, Number 3.
2. Wilkinson, Leland (2005). *The Grammar of Graphics*, 2nd Ed., Springer.
3. Bellotti, A. (2018). *Reliable region predictions for automated valuation models: supplementary material for Ames housing data.*