CSI 2300: Intro to Data Science

In-Class Exercise 07: Exploratory Data Analysis

The data for today's exercises are the Colorado Housing data used in the lecture.

1. Download the data for 2019, and then load it into R. How many variables are in each one?

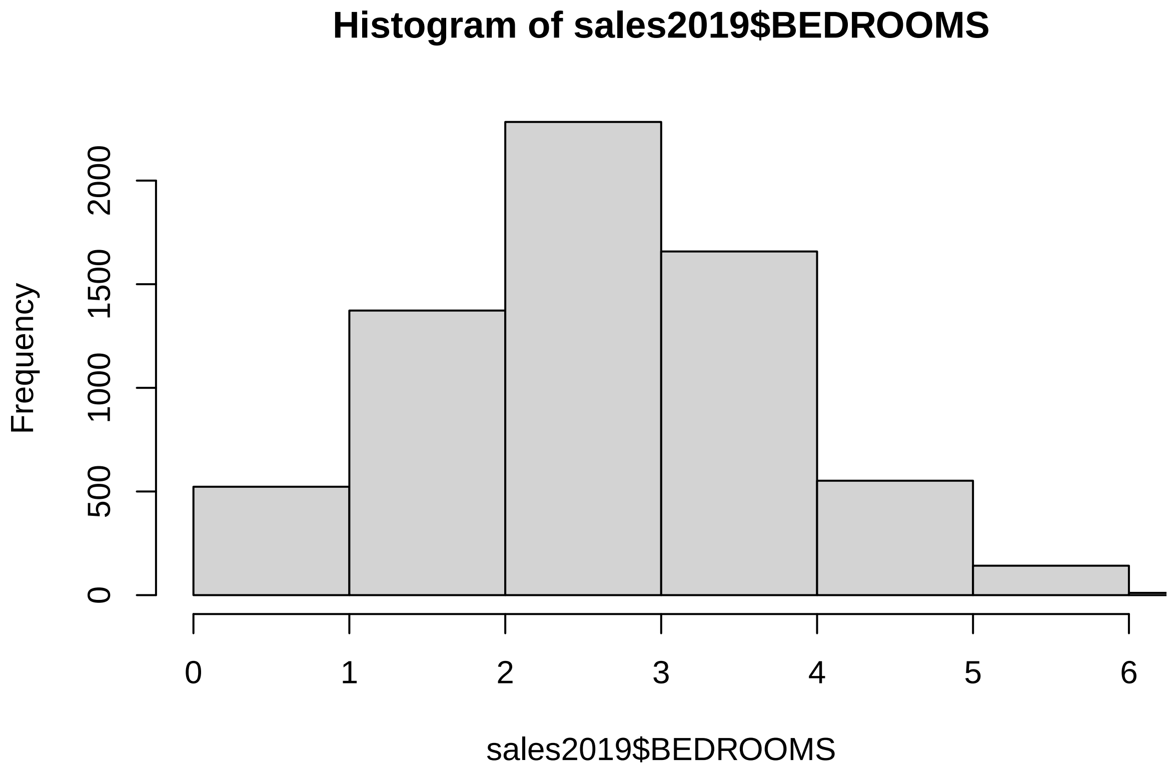
```
sales2019 <- read.csv("dat/boulder-2019-residential_sales.csv")
```

2. As was covered in lecture, we need to strip the dollar signs and commas from the land value, building value, and sale price columns. Show the the complete calls to accomplish this task.

```
sales2019$SALE_PRICE <- gsub(',',"",sales2019$SALE_PRICE)
sales2019$SALE_PRICE <- gsub('\\$',"",sales2019$SALE_PRICE)
sales2019$SALE_PRICE <- as.numeric(sales2019$SALE_PRICE)

sales2019$BLDG_VALUE <- gsub(',',"",sales2019$BLDG_VALUE)
sales2019$BLDG_VALUE <- gsub('\\$',"",sales2019$BLDG_VALUE)
sales2019$BLDG_VALUE <- as.numeric(sales2019$BLDG_VALUE)

sales2019$LAND_VALUE <- gsub(',',"",sales2019$LAND_VALUE)
sales2019$LAND_VALUE <- gsub('\\$',"",sales2019$LAND_VALUE)
sales2019$LAND_VALUE <- as.numeric(sales2019$LAND_VALUE)
```

3. Create a histogram for the number of bedrooms sold for the year 2019. This plot will look right skewed. Why do you think this is? In order to focus on the smaller values, change the number of breaks in the bins, and limit the view of the data by focusing on left-hand range of data. Show your plot and code (only one line of code is needed).

```
hist(sales2019$BEDROOMS, breaks=27, xlim=c(0, 6))
```

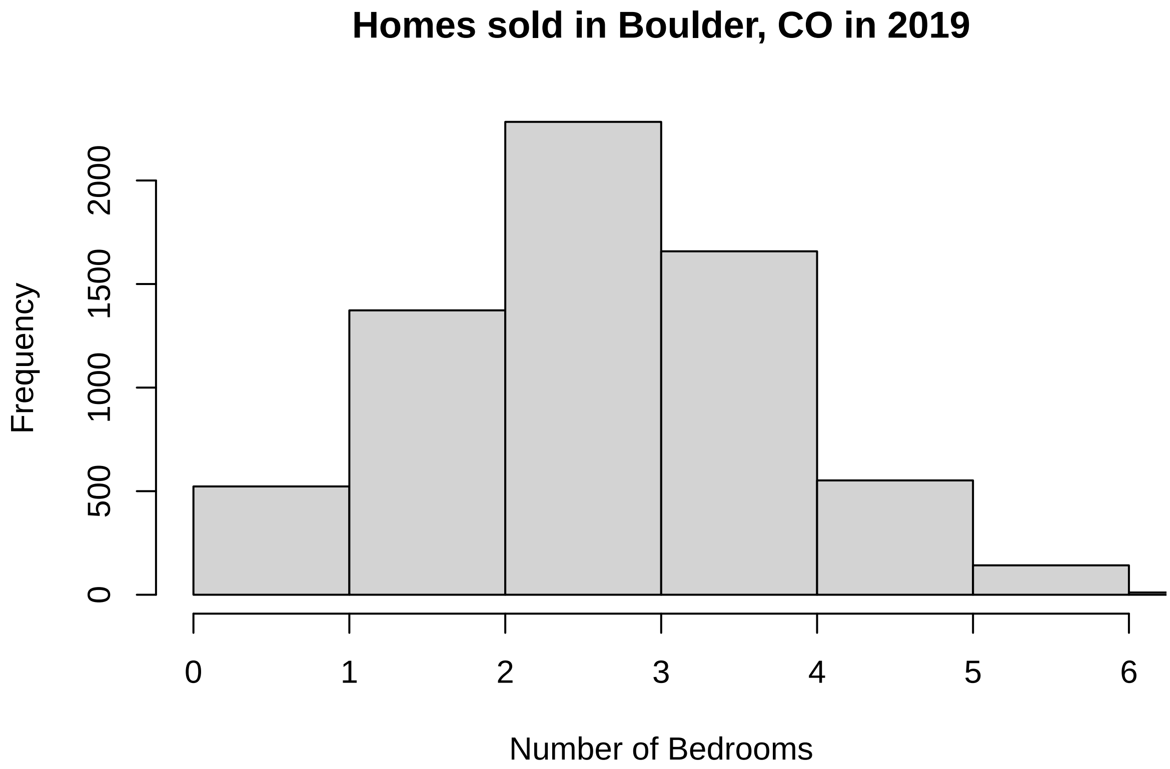


```
#Why is this skewed
#Because it is not possible to have a fractional number of bedrooms. This means that t
```

The histogram is right-skewed because most homes sold in 2019 have 3-4 bedrooms, and fewer homes have 1-2 bedrooms or 5+ bedrooms. Since the number of bedrooms is a discrete variable (no fractional bedrooms), the distribution is naturally skewed toward the lower end (fewer bedrooms).

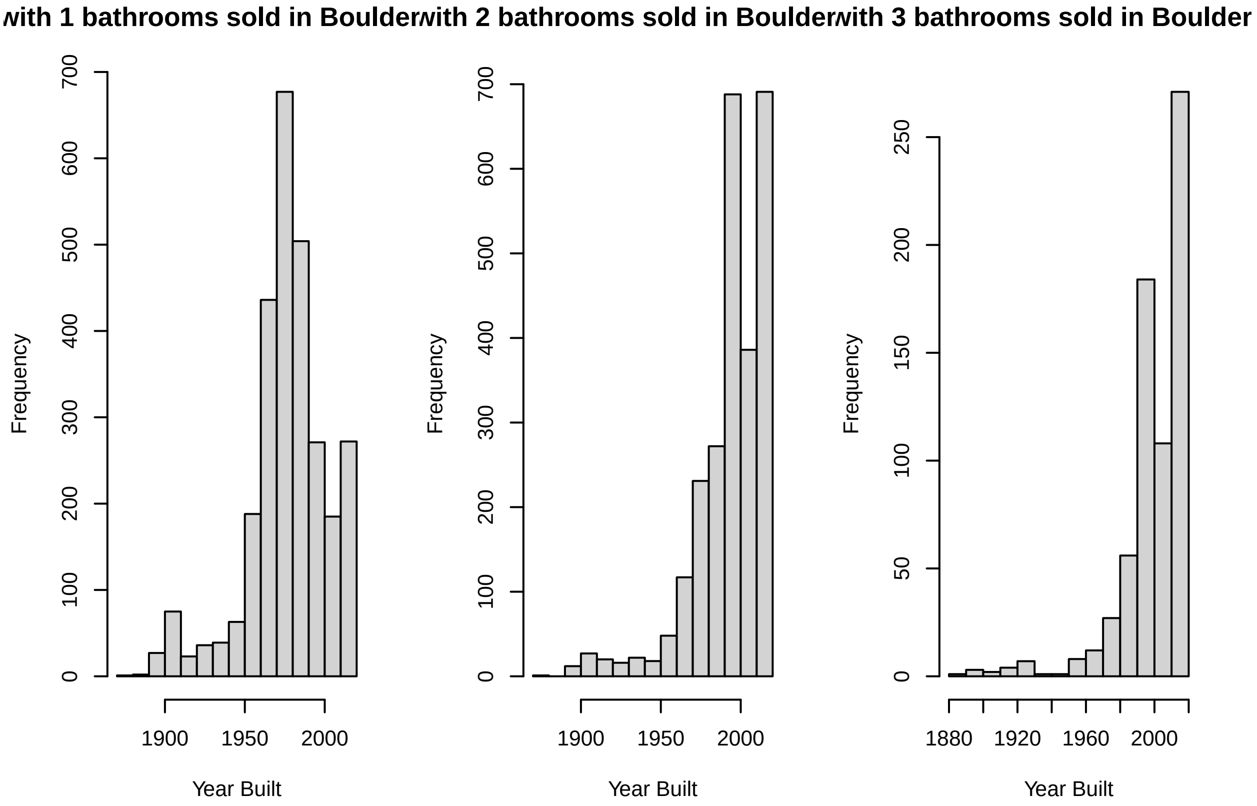
4. Modify the plot from the prior question to improve the title and x-axis label. These should make the plot understandable for a casual observer.

```
hist(sales2019$BEDROOMS,
     breaks = 27,
     xlim = c(0, 6),
     main = "Homes sold in Boulder, CO in 2019",
     xlab = "Number of Bedrooms")
```



5. For the 2019 data, there are houses being sold that were originally built over a wide range of years. We want to investigate how building standards have changed over the years. Create a histogram of the building year for homes with 1 full bathroom. Repeat for homes with 2 full bathrooms and with 3 full bathrooms. Comment on the similarities and differences among these three histograms.

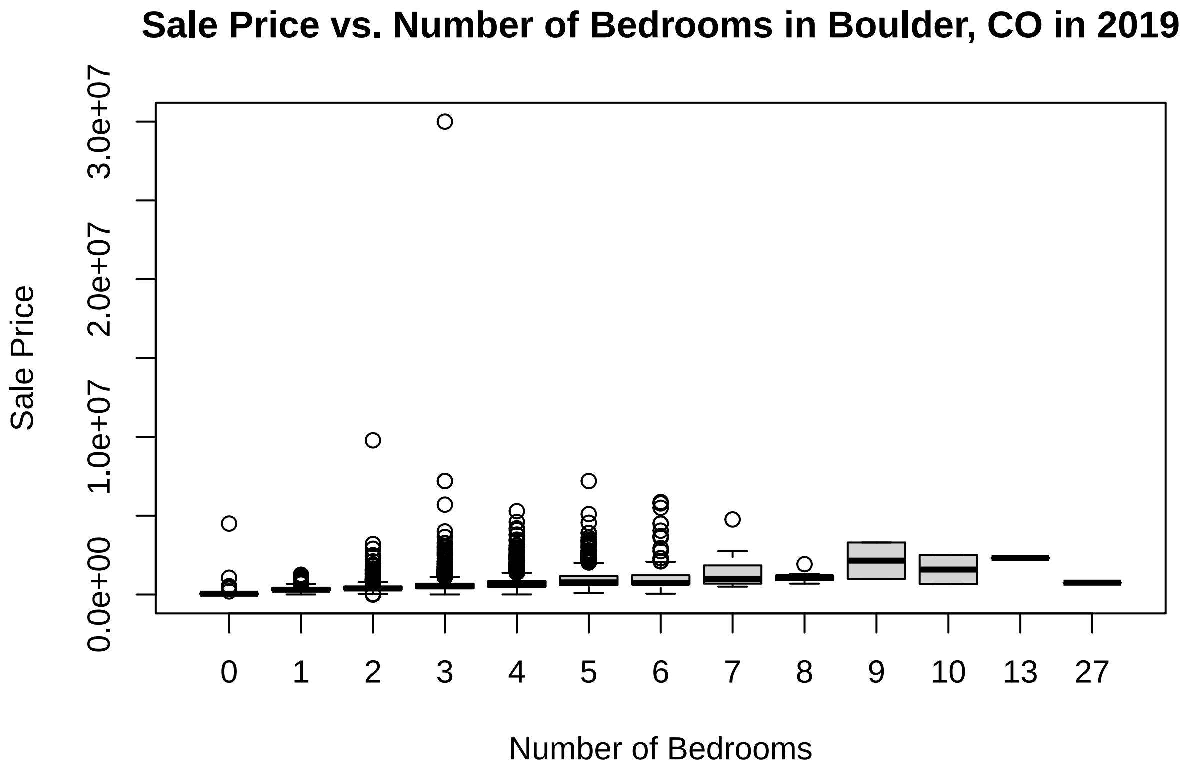
```r
par(mfrow=c(1,3))

for (i in 1:3){
  b <- sales2019[sales2019$FULL_BATHS == i,]
  hist(b$BLDG1_YEAR_BUILT,
       main = paste("Homes with", i, "bathrooms sold in Boulder, CO in 2019"),
       xlab = "Year Built")

}
```



All three histograms show a peak in the number of homes built in the mid-20th century, indicating a period of significant housing development in Boulder.

6. Create a side-by-side boxplot of the sale price against each number of bedrooms in the 2019 sales. Add labels and a title to the plot. Describe what you see in this plot.

```r
boxplot(SALE_PRICE ~ BEDROOMS, data = sales2019,
        main = "Sale Price vs. Number of Bedrooms in Boulder, CO in 2019",
        xlab = "Number of Bedrooms",
        ylab = "Sale Price")
```

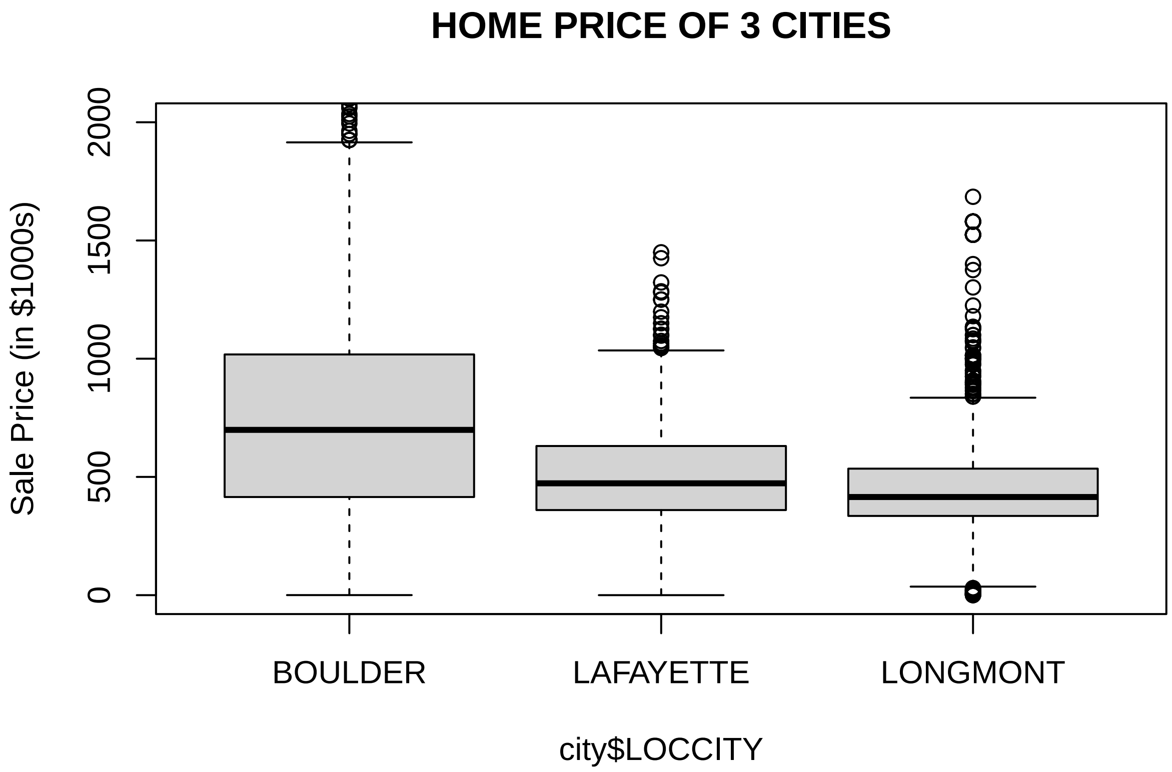


The boxplot shows that as the number of bedrooms increases, the median sale price generally increases. However, there is significant variability in sale prices for homes with more bedrooms (4+ bedrooms), indicating that factors other than the number of bedrooms (e.g., location, lot size) influence the price. Homes with fewer bedrooms (1-2) have a tighter price range, while homes with more bedrooms (5+) show more outliers, suggesting luxury or larger properties.

7. Filter the data to the three cities in Boulder County that had the most sales. Compare their housing prices with a side-by-side boxplot. Change the scale of the y-axis to be in thousands of dollars.

```
unique(sales2019$LOCCITY)
#  [1] "LAFAYETTE"      "BOULDER"        "LONGMONT"       "ERIE"
#  [5] "UNINCORPORATED" "NEDERLAND"      "LOUISVILLE"     "SUPERIOR"
#  [9] "LYONS"          "JAMESTOWN"      "WARD"           "BOULDER "
sort(table(sales2019$LOCCITY))
#
#       BOULDER            WARD      JAMESTOWN      NEDERLAND          LYONS
#              2              8              9             40             45
#       SUPERIOR     LOUISVILLE           ERIE      LAFAYETTE UNINCORPORATED
#            282            404            486            685            871
```

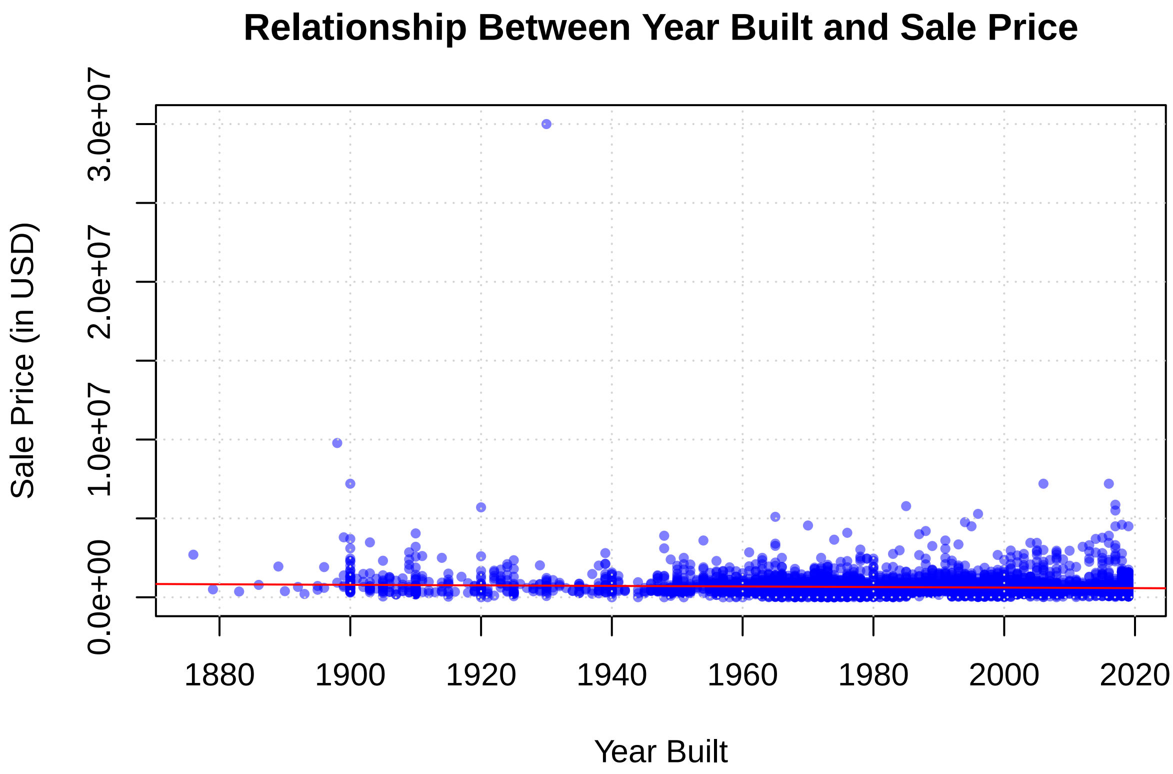
```r
#        BOULDER        LONGMONT
#           1561            2162

city <- sales2019[sales2019$LOCCITY == "BOULDER" | sales2019$LOCCITY == "LONGMONT" | sal

boxplot(city$SALE_PRICE/1000 ~ city$LOCCITY,
        main = "HOME PRICE OF 3 CITIES",
        ylab = "Sale Price (in $1000s)",
        ylim = c(0, 2000))
```



The boxplot shows that Boulder has the highest median sale prices compared to Longmont and Lafayette, reflecting Boulder's more expensive housing market. Longmont and Lafayette have lower median prices, but there are outliers in both cities, indicating some high-priced homes. The scale (in thousands of dollars) makes it clear that Boulder's housing market is significantly more expensive.

8. Investigate the relationship between the year a house was built and its sales price with a scatterplot. What options could be used to improve the plot? What does this plot tell you about the relationship between the year a house was built and its sales price?

```
plot(sales2019$BLDG1_YEAR_BUILT, sales2019$SALE_PRICE,
     main = "Relationship Between Year Built and Sale Price",
     xlab = "Year Built",
     ylab = "Sale Price (in USD)",
     pch = 16,  # Use solid circles for the points
     col = rgb(0, 0, 1, 0.5),  # Semi-transparent blue color
     cex = 0.7)  # Adjust the size of the points

# Add a regression line to show the trend
abline(lm(SALE_PRICE ~ BLDG1_YEAR_BUILT, data = sales2019), col = "red")

# Add grid lines for better readability
grid()
```



The scatterplot shows a weak positive trend, suggesting that newer homes tend to have higher sale prices. However, the relationship is not strong, indicating that other factors (e.g., location, size, condition) play a significant role in determining sale price. The regression line helps visualize the trend, but the wide spread of points shows considerable variability.