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("boulder-2019-residential_sales.csv")
```

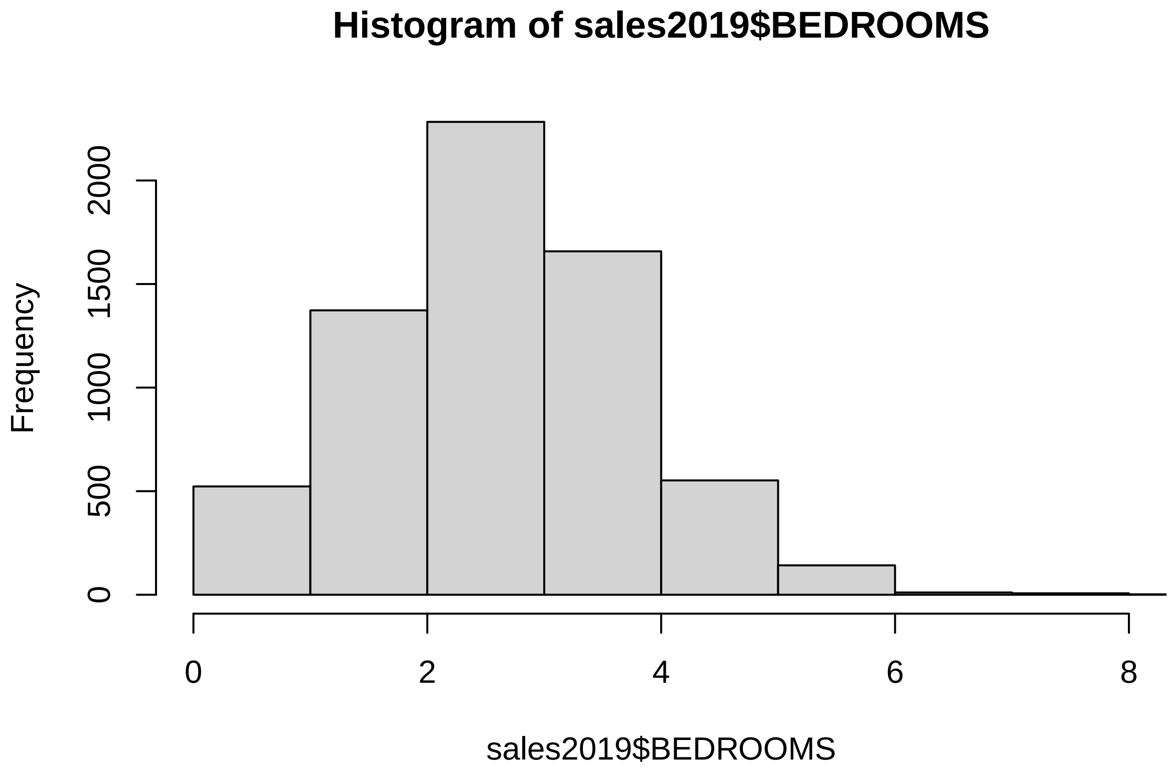
There are 37 variables in this data.

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

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

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

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).

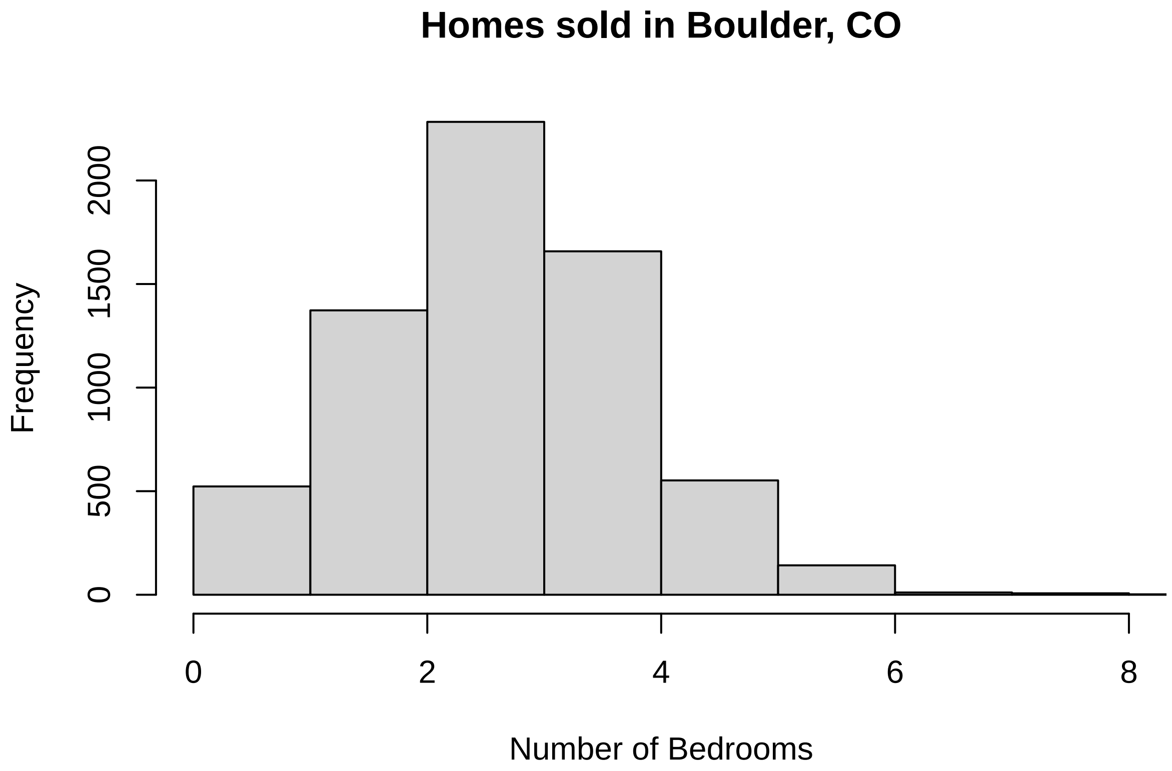
```
hist(sales2019$BEDROOMS, breaks = 27, xlim = c(0,8))
```



The histogram is skewed because a majority of the properties are typically in a range from around 0-6 bedrooms.

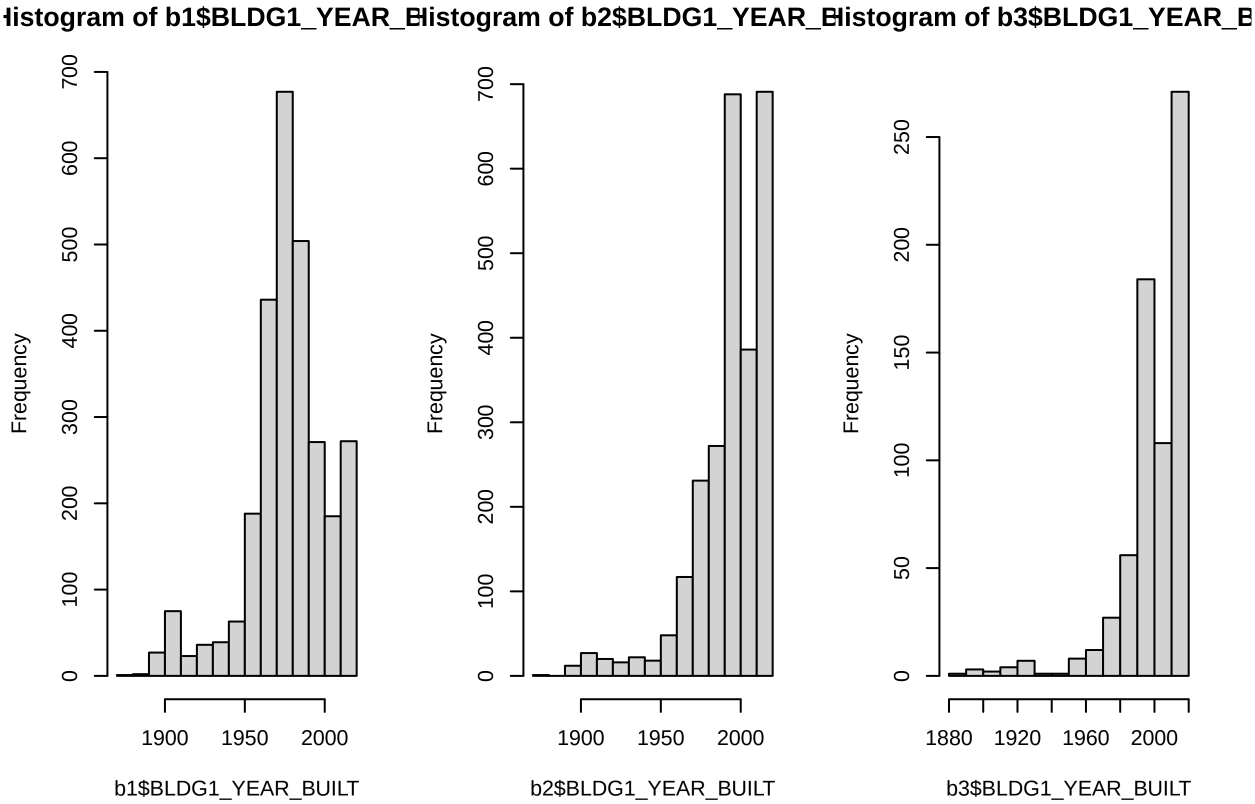
4. Modify the plot from the prior question to improve the title and x-axis label. These should make the plot understandable for a casual observer.

```r
hist(sales2019$BEDROOMS,
     breaks = 27,
     xlim = c(0,8),
     xlab ="Number of Bedrooms" ,
     main="Homes sold in Boulder, CO")
```



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.

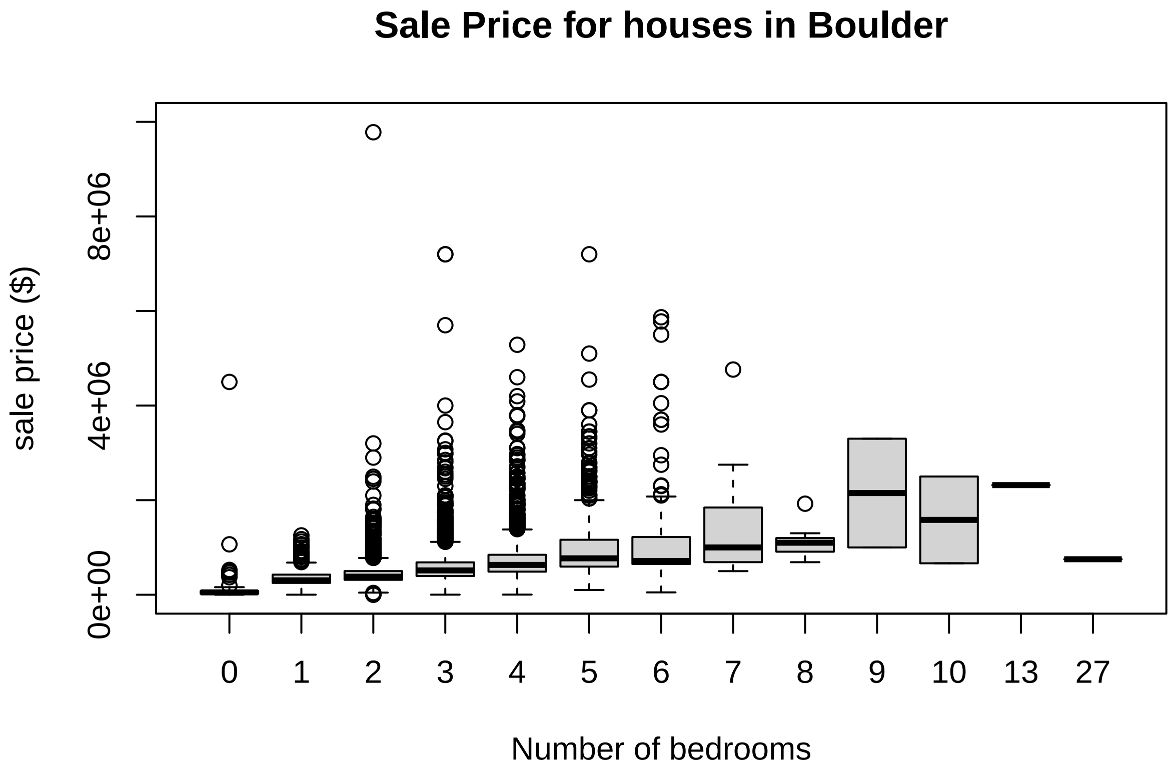
```r
par(mfrow=c(1, 3))
b1 <- sales2019[which(sales2019$FULL_BATHS == 1),]
hist(b1$BLDG1_YEAR_BUILT)

b2 <- sales2019[which(sales2019$FULL_BATHS == 2),]
hist(b2$BLDG1_YEAR_BUILT)

b3 <- sales2019[which(sales2019$FULL_BATHS == 3),]
hist(b3$BLDG1_YEAR_BUILT)
```



One major similarity for these three histograms include that they are all skewed in the same direction (left). One difference is that in b1, the frequency is highest near the middle but the other two are towards the end/last years.

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.

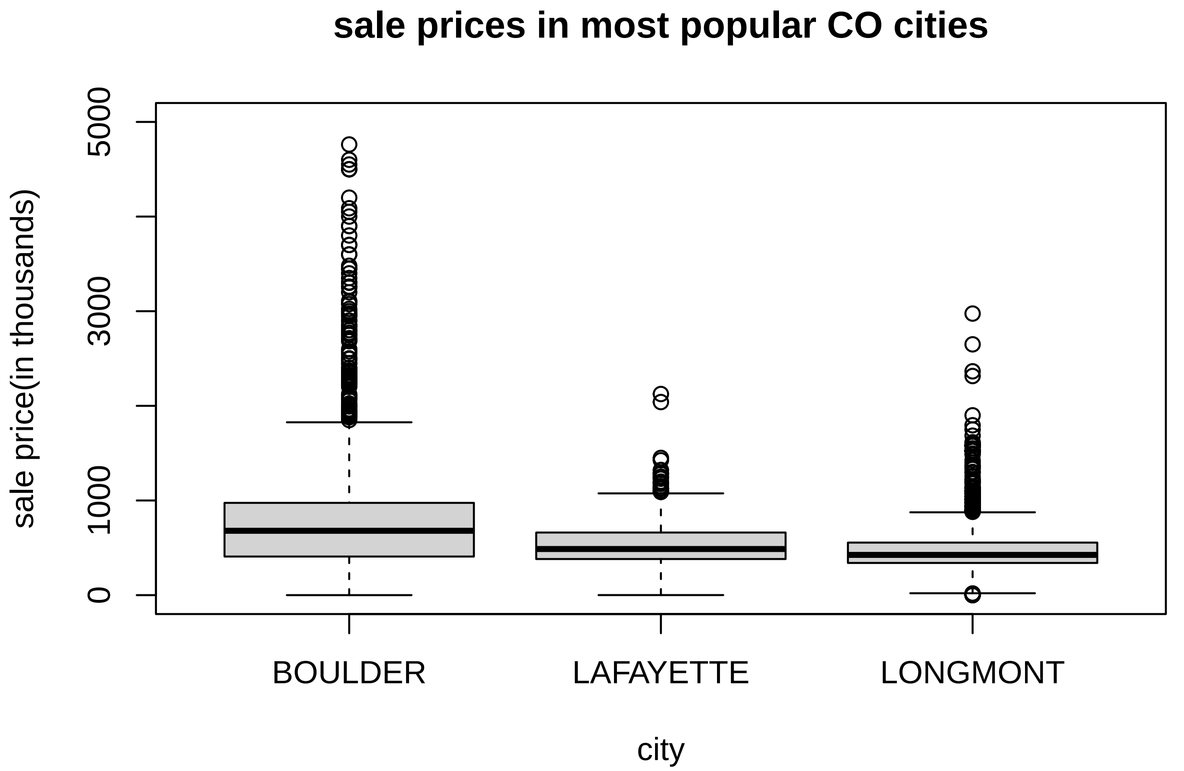
```r
boxplot(sales2019$SALE_PRICE ~ sales2019$BEDROOMS,
        ylim = c(0,1e07),
        main = "Sale Price for houses in Boulder",
        ylab= "sale price ($)",
        xlab = "Number of bedrooms")
```



In this plot, I see that as the number of bedrooms increases, the sale price usually increases as well.

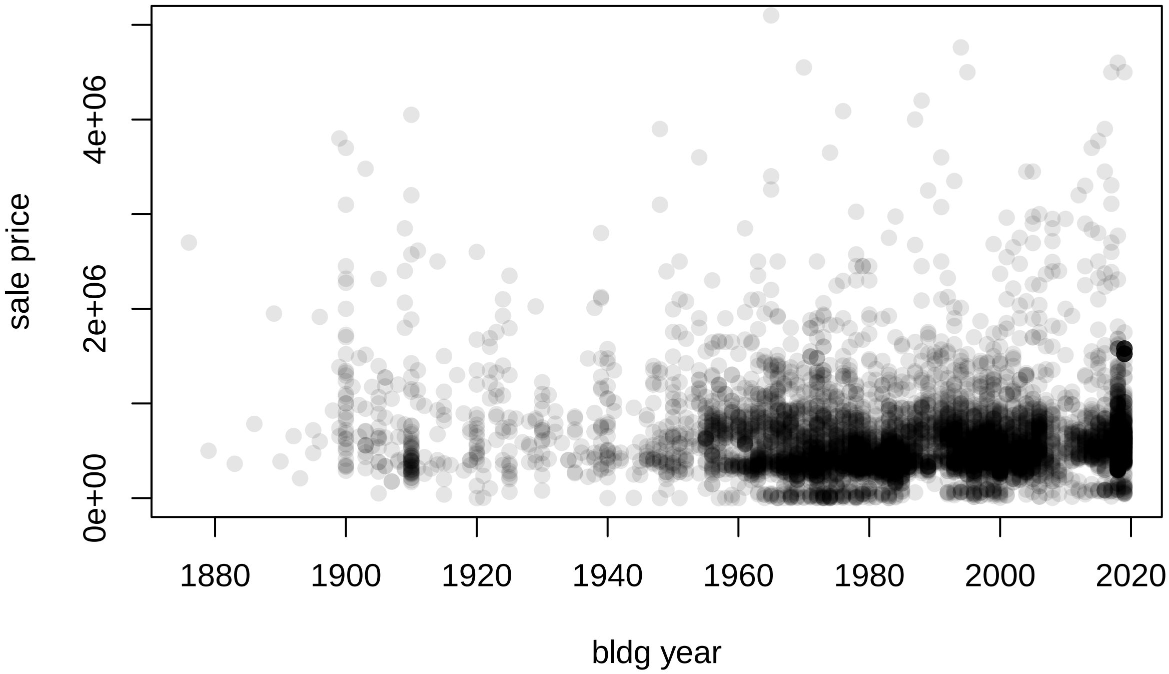
7. Filter the data to the three cities in Boulder County that had the most sales. Compare their housing prices with a side-by-side boxplot. Change the scale of the y-axis to be in thousands of dollars.

```
#unique(sales2019$CITY)
#sort(table(sales2019$CITY))
# LONGMONT BOULDER LAFAYETTE
city <- sales2019[sales2019$CITY == 'LONGMONT'|
                         sales2019$CITY == 'BOULDER'|
                         sales2019$CITY == 'LAFAYETTE',]

city$SALE_PRICE <- city$SALE_PRICE/ 1000

boxplot(city$SALE_PRICE ~ city$CITY,
        main = "sale prices in most popular CO cities",
        xlab = 'city',
        ylab = 'sale price(in thousands)',
        ylim = c(0, 5000))
```



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,
     ylim = c(0,5e06),
     xlab = 'bldg year',
     ylab = 'sale price',
     pch = 19,
     col = rgb(0,0,0,.1))
```



This tells me as the building year is later, the sale price typically increases. This shows me that there is a positive correlation between the year a house was built and its sales price.