Assignment Week 4

Karlie Schwartzwald

2022-07-03

First I will import the data from scores.csv

```
scores <- read.csv("scores.csv")
```

What are the observational units in this study?

```
print("The observation units in this study are course grades for these two sections of students")
```

```
## [1] "The observation units in this study are course grades for these two sections of students"
```

Identify the variables mentioned in the narrative paragraph and determine which are categorical and quantitative?

```
print("The variables are Section (Categorical), Score (Quantitative) and Count (Quantitative)")
```

```
## [1] "The variables are Section (Categorical), Score (Quantitative) and Count (Quantitative)"
```

```
str(scores)
```

```
## 'data.frame':    38 obs. of  3 variables:
##  $ Count  : int  10 10 20 10 10 10 10 30 10 10 ...
##  $ Score  : int  200 205 235 240 250 265 275 285 295 300 ...
##  $ Section: chr  "Sports" "Sports" "Sports" "Sports" ...
```

Create one variable to hold a subset of your data set that contains only the Regular Section and one variable for the Sports Section.

```
regular_section <-subset(scores, scores$Section=="Regular")
sports_section <-subset(scores, scores$Section=="Sports")
print(regular_section)
```

```
##    Count Score Section
## 6     10   265 Regular
## 7     10   275 Regular
## 9     10   295 Regular
## 10    10   300 Regular
## 13    10   305 Regular
## 14    10   310 Regular
```

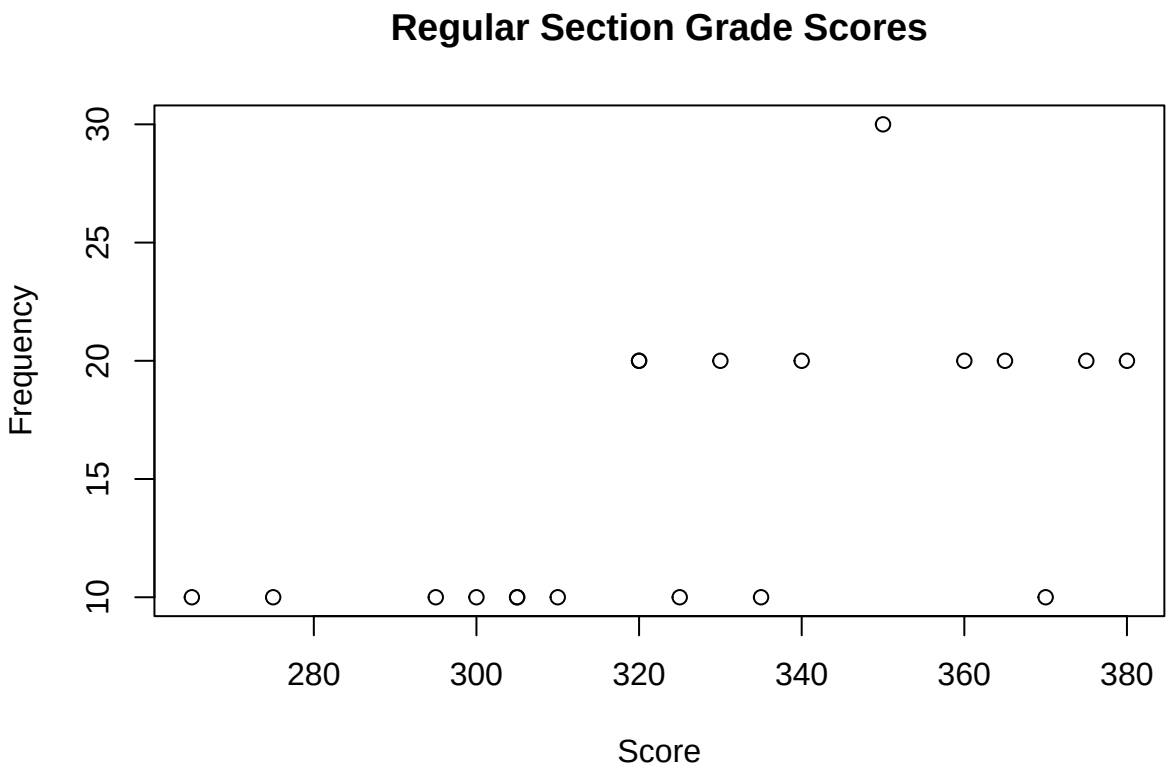
```
## 16    20   320 Regular
## 17    10   305 Regular
## 19    20   320 Regular
## 20    10   325 Regular
## 22    20   330 Regular
## 25    10   335 Regular
## 26    20   340 Regular
## 28    30   350 Regular
## 29    20   360 Regular
## 31    20   365 Regular
## 34    10   370 Regular
## 35    20   375 Regular
## 37    20   380 Regular
```

```
print(sports_section)
```

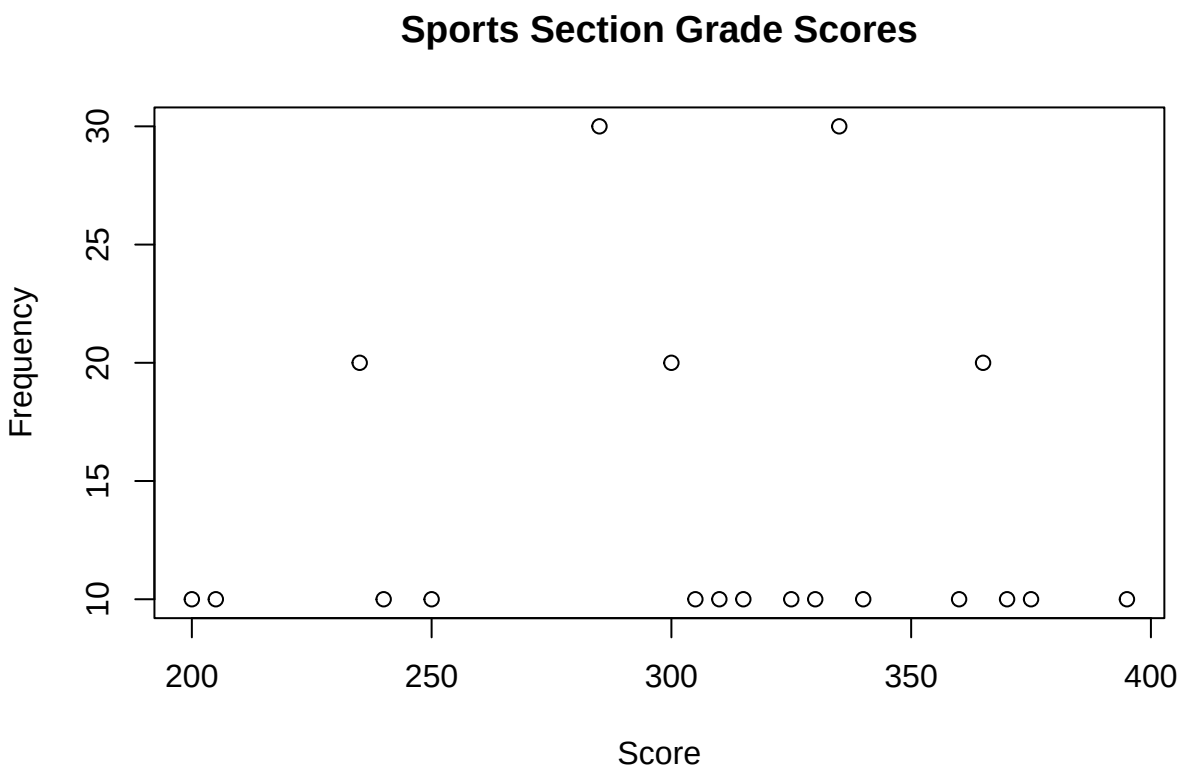
```
##    Count Score Section
## 1     10   200  Sports
## 2     10   205  Sports
## 3     20   235  Sports
## 4     10   240  Sports
## 5     10   250  Sports
## 8     30   285  Sports
## 11    20   300  Sports
## 12    10   305  Sports
## 15    10   310  Sports
## 18    10   315  Sports
## 21    10   325  Sports
## 23    10   330  Sports
## 24    30   335  Sports
## 27    10   340  Sports
## 30    10   360  Sports
## 32    20   365  Sports
## 33    10   370  Sports
## 36    10   375  Sports
## 38    10   395  Sports
```

Use the Plot function to plot each Sections scores and the number of students achieving that score. Use additional Plot Arguments to label the graph and give each axis an appropriate label. Once you have produced your Plots answer the following questions:

```
# plots
plot(regular_section$Score, regular_section$Count, main="Regular Section Grade Scores", ylab = "Frequen
```




```
plot(sports_section$Score, sports_section$Count, main="Sports Section Grade Scores", ylab = "Frequency"
```



1. Comparing and contrasting the point distributions between the two section, looking at both tendency and consistency: Can you say that one section tended to score more points than the other? Justify and explain your answer.

```
regular_mean = sum(regular_section$Score*regular_section$Count)/sum(regular_section$Count)
sports_mean = sum(sports_section$Score*sports_section$Count)/sum(sports_section$Count)
sprintf("The regular section has a mean score of %f.", regular_mean)
```

```
## [1] "The regular section has a mean score of 335.000000."
```

```
sprintf("The sports section has a mean score of %f.",sports_mean)
```

```
## [1] "The sports section has a mean score of 306.923077."
```

```
print("Although the sports section had a higher top grade in the class, it tended to score lower than t
```

```
## [1] "Although the sports section had a higher top grade in the class, it tended to score lower than
```

2. Did every student in one section score more points than every student in the other section? If not, explain what a statistical tendency means in this context.

```
print("Every student in one section did not score better than every student in another section. What we
```

```
## [1] "Every student in one section did not score better than every student in another section. What w
```

3. What could be one additional variable that was not mentioned in the narrative that could be influencing the point distributions between the two sections?

```
print("An additional variable not mentioned in the narrative that could be influencing the point distri
```

```
## [1] "An additional variable not mentioned in the narrative that could be influencing the point distr
```

Now we open the 2014 American Community Survey data

```
library(readxl)
week_6_housing <- read_excel("week-6-housing.xlsx")
head(week_6_housing)
```

```
## # A tibble: 6 x 24
##   `Sale Date`         `Sale Price` sale_reason sale_instrument sale_warning
##   <dttm>                     <dbl>       <dbl>           <dbl> <chr>
## 1 2006-01-03 00:00:00       698000           1               3 <NA>
## 2 2006-01-03 00:00:00       649990           1               3 <NA>
## 3 2006-01-03 00:00:00       572500           1               3 <NA>
## 4 2006-01-03 00:00:00       420000           1               3 <NA>
## 5 2006-01-03 00:00:00       369900           1               3 15
## 6 2006-01-03 00:00:00       184667           1              15 18 51
## # ... with 19 more variables: sitetype <chr>, addr_full <chr>, zip5 <dbl>,
## #   ctyname <chr>, postalctyn <chr>, lon <dbl>, lat <dbl>,
## #   building_grade <dbl>, square_feet_total_living <dbl>, bedrooms <dbl>,
## #   bath_full_count <dbl>, bath_half_count <dbl>, bath_3qtr_count <dbl>,
## #   year_built <dbl>, year_renovated <dbl>, current_zoning <chr>,
## #   sq_ft_lot <dbl>, prop_type <chr>, present_use <dbl>
```

1. Use the apply function on a variable in your dataset.

```
# using the apply function with mean on the sale price variable
sale_price = matrix(week_6_housing$`Sale Price`)
apply(sale_price, 2, mean)
```

```
## [1] 660737.7
```

2. Use the aggregate function on a variable in your dataset

```
# cut the sq feet by number of bedrooms, with mean
aggregate(square_feet_total_living ~ bedrooms, week_6_housing, mean)
```

```
##   bedrooms square_feet_total_living
## 1        0                2576.8421
## 2        1                 882.7273
```

```
## 3          2                1719.1049
## 4          3                2066.3708
## 5          4                2979.4591
## 6          5                3436.1032
## 7          6                3752.7711
## 8          7                7425.4545
## 9          8                6230.0000
## 10         9                4460.0000
## 11        10                6340.0000
## 12        11                7980.0000
```

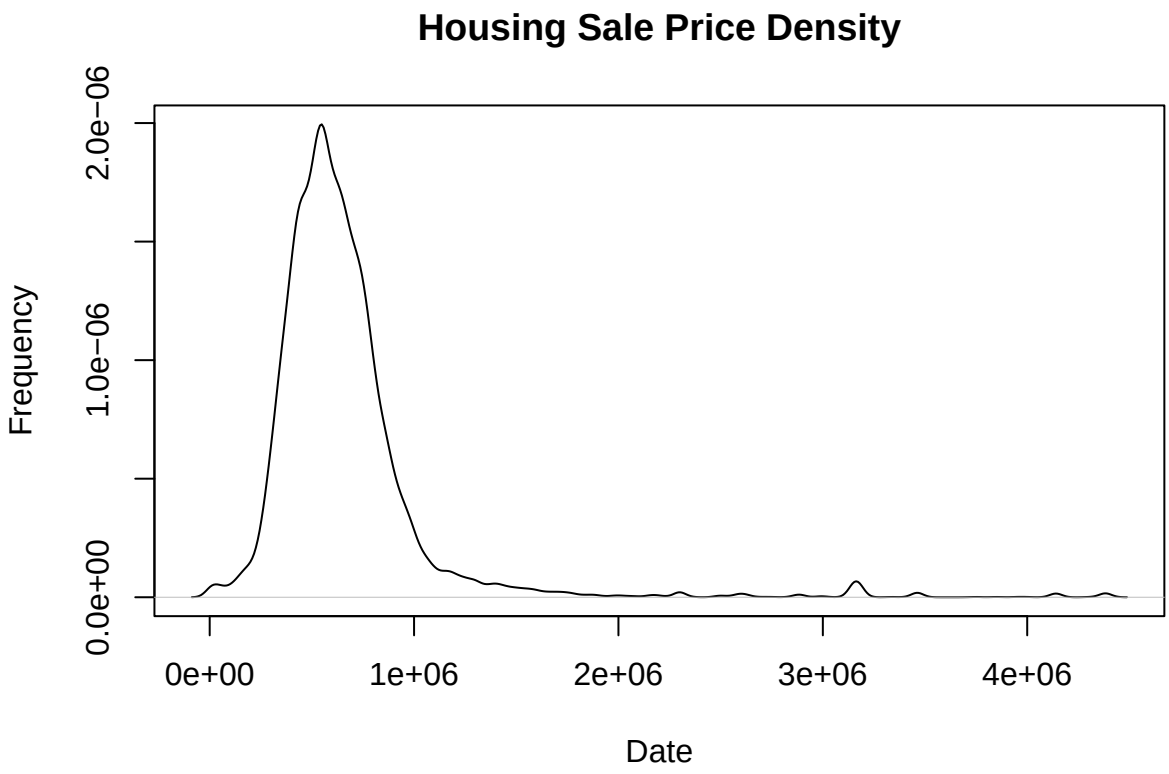
3. Use the plyr function on a variable in your dataset – more specifically, I want to see you split some data, perform a modification to the data, and then bring it back together

```
library(plyr)
price_per_bedroom = ddply(
 .data = week_6_housing,
 .variables = "bedrooms",
 .fun = function(x) mean(x$'Sale Price')
)
price_per_bedroom
```

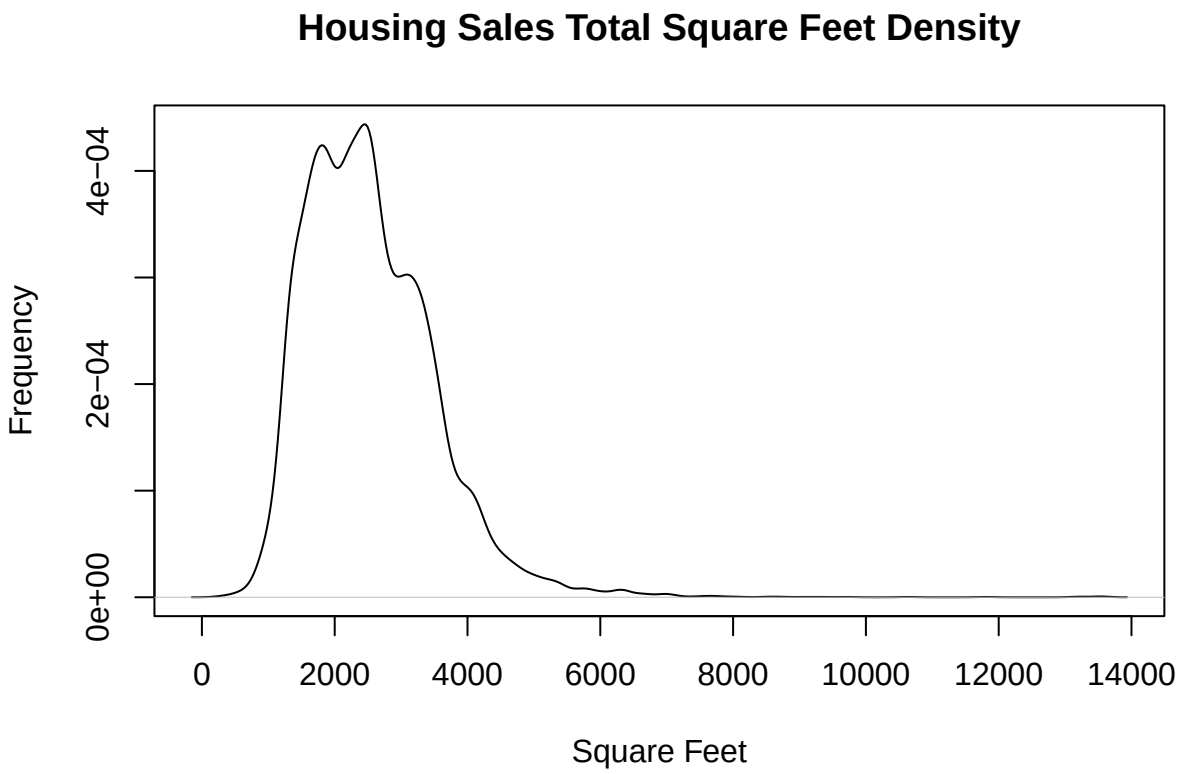
```
##    bedrooms        V1
## 1         0  844059.5
## 2         1  722814.1
## 3         2  544946.4
## 4         3  564958.6
## 5         4  735910.0
## 6         5  836974.0
## 7         6  767494.3
## 8         7 1307281.7
## 9         8 1122500.0
## 10        9  581500.0
## 11       10  450000.0
## 12       11 1825000.0
```

4. Check distributions of the data

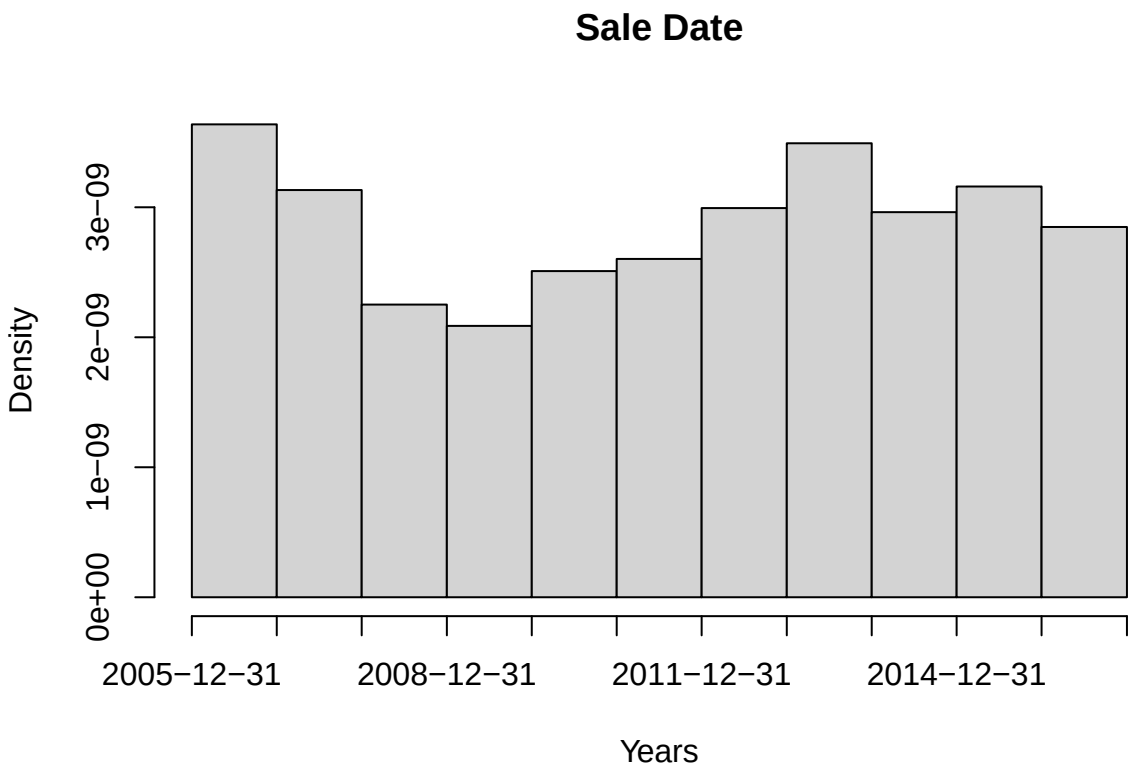
```
# Below is a selection of distributions from some variables in the data
d1 <- density(week_6_housing$`Sale Price`)
plot(d1, main="Housing Sale Price Density", xlab="Date", ylab="Frequency")
```



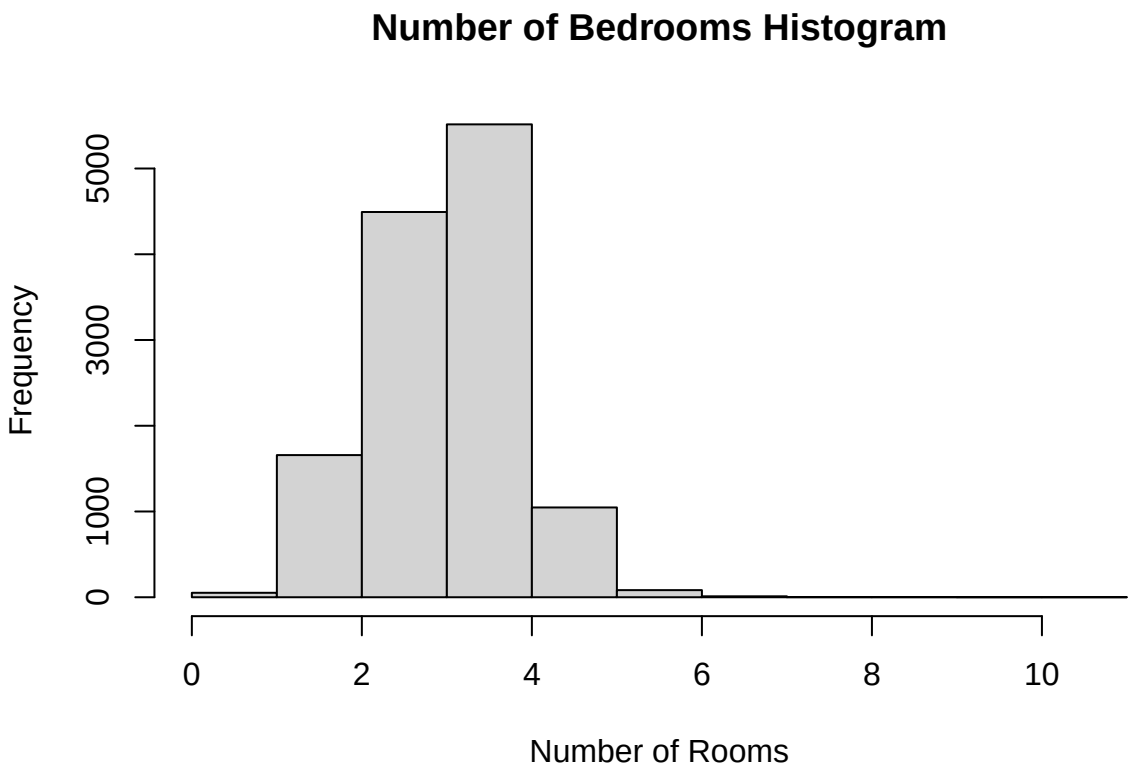

```
d2<- density(week_6_housing$square_feet_total_living)
plot(d2, main="Housing Sales Total Square Feet Density", xlab="Square Feet", ylab="Frequency")
```



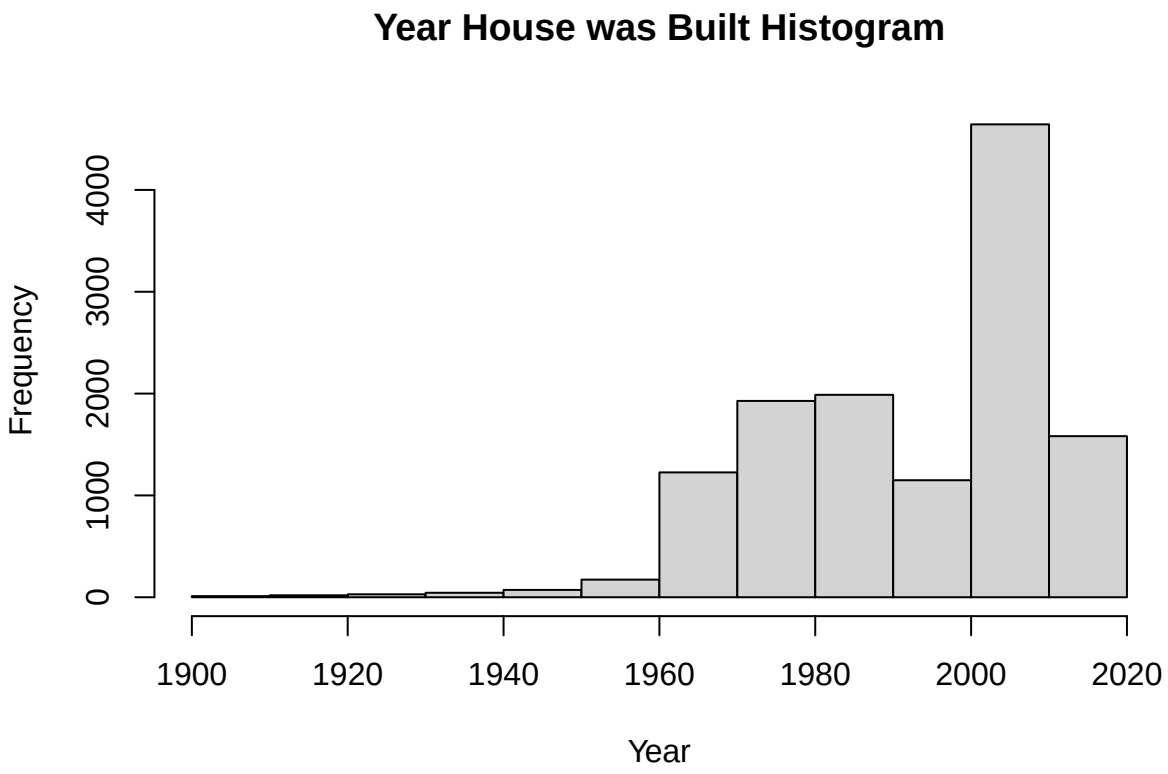

```
hist(week_6_housing$`Sale Date`, main="Sale Date", xlab = "Years", breaks ="years")
```




```
hist(week_6_housing$bedrooms, main="Number of Bedrooms Histogram", xlab="Number of Rooms", ylab="Freque
```




```
hist(week_6_housing$year_built, main="Year House was Built Histogram", xlab="Year", ylab="Frequency")
```



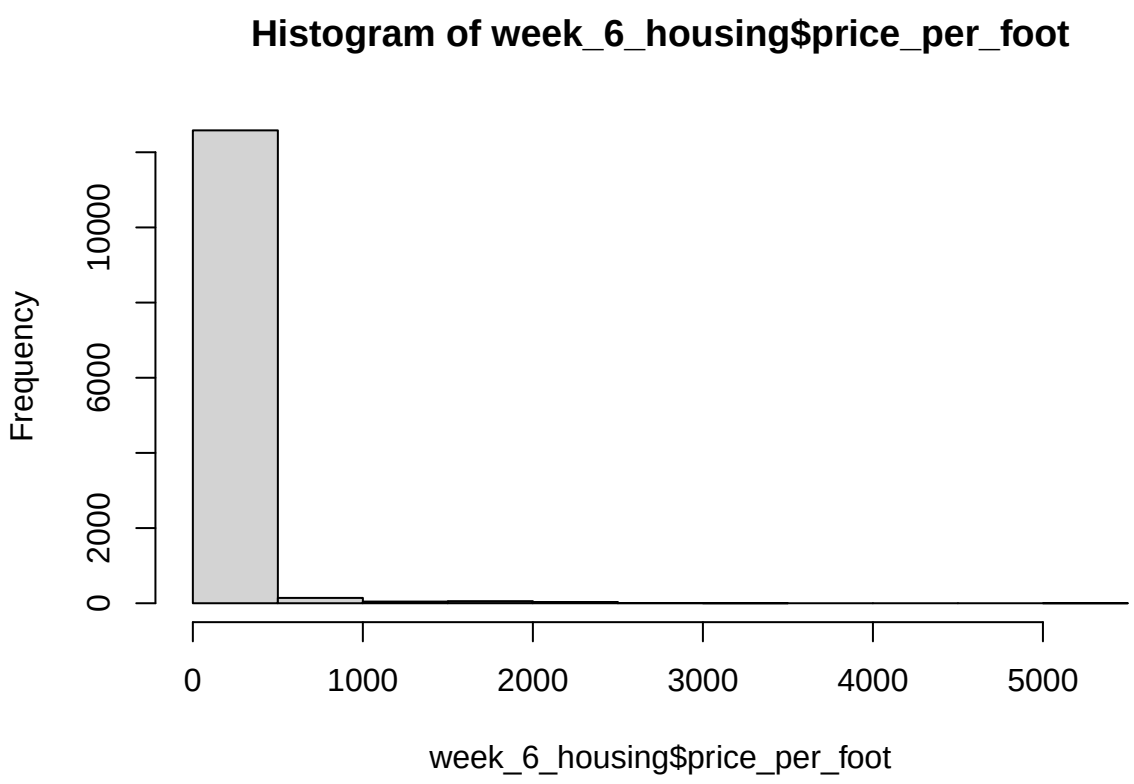
5. Identify if there are any outliers

```
print("There are outliers in the year the ouses were built which were built before 1920. There are outli
```

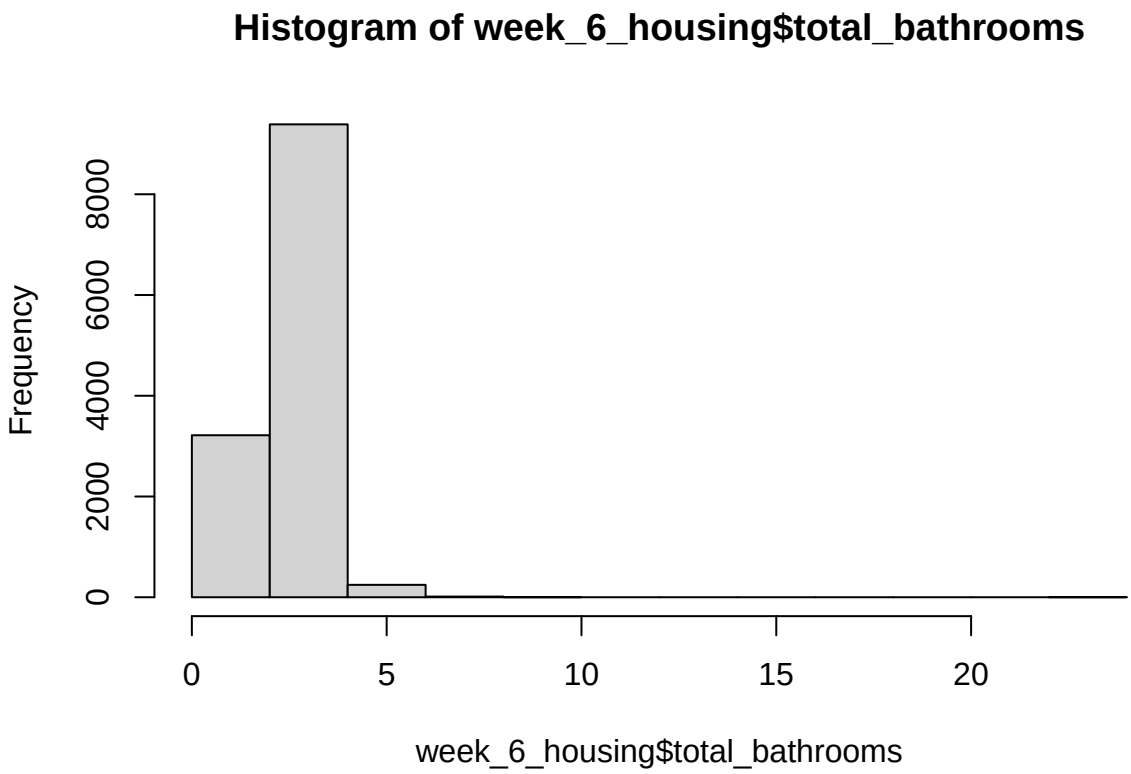
```
## [1] "There are outliers in the year the ouses were built which were built before 1920. There are outl
```

6. Create at least 2 new variables

```
# create price_per_foot variable
week_6_housing$price_per_foot = week_6_housing$`Sale Price`/week_6_housing$square_feet_total_living
# create  variable
week_6_housing$total_bathrooms = week_6_housing$bath_full_count + 0.5*(week_6_housing$bath_half_count) +
# make histograms of the new variables
hist(week_6_housing$price_per_foot)
```




```
hist(week_6_housing$total_bathrooms)
```


Histogram of week_6_housing$total_bathrooms
Frequency
0
2000
4000
6000
8000
0
5
10
15
20
week_6_housing$total_bathrooms