Project Proposal

Brady Lamson

Provide the the following information for your project.

1. What research question(s) do you hope to answer?

I want to use linear regression to help understand the most important variables in the price of a home. Or, at the very least, get as much insight into it as I can with the data set I have at hand. The conversation around home ownership has changed a lot over the past couple decades. I think it's a very important topic to understand as it becomes harder and harder for young people to break into home ownership and gentrification forces more people out of their homes.

2. From where did you obtain the data you will use to answer your research question?

I acquired my data from kaggle. The dataset can be found here.

3. How many observations does your data set have?

```
df %>% nrow()
```

```
## [1] 4140
```

4. Are you merging multiple data sets?

No.

5. Provide a table listing each variable you are considering for analysis, briefly describe each variable (e.g., the number of disease cases in each region), and the variable type (e.g., numeric, factor, date, etc.).

```
## # A tibble: 6 x 7
##   date                  price bedrooms bathrooms sqft_living yr_built floors
##   <dttm>                <dbl>    <dbl>     <dbl>       <dbl>    <dbl>  <dbl>
## 1 2014-05-09 00:00:00  376000        3      2           1340     2008      3
## 2 2014-05-09 00:00:00  800000        4      3.25        3540     2007      2
## 3 2014-05-09 00:00:00 2238888        5      6.5         7270     2010      2
## 4 2014-05-09 00:00:00  324000        3      2.25         998     2007      2
## 5 2014-05-10 00:00:00  549900        5      2.75        3060     1979      1
## 6 2014-05-10 00:00:00  320000        3      2.5         2130     2003      2
```

Variable	Description	Variable Type
date	The date the property was sold. Will be used for filtering to a specific year.	Date
price	Response variable. Price in USD of the home.	numeric
bedrooms	Number of bedrooms	numeric
bathrooms	Number of bathrooms	numeric
sqft_living	Square footage of the living room	numeric
yr_built	The year the building was built	numeric
floors	Number of floors	numeric

6. What will your response variable be for answering the research question(s)?

Response variable is price in USD. I feel it is somewhat self explanatory how this response variable answers my research question.

7. Provide a numeric summary of your response variable.

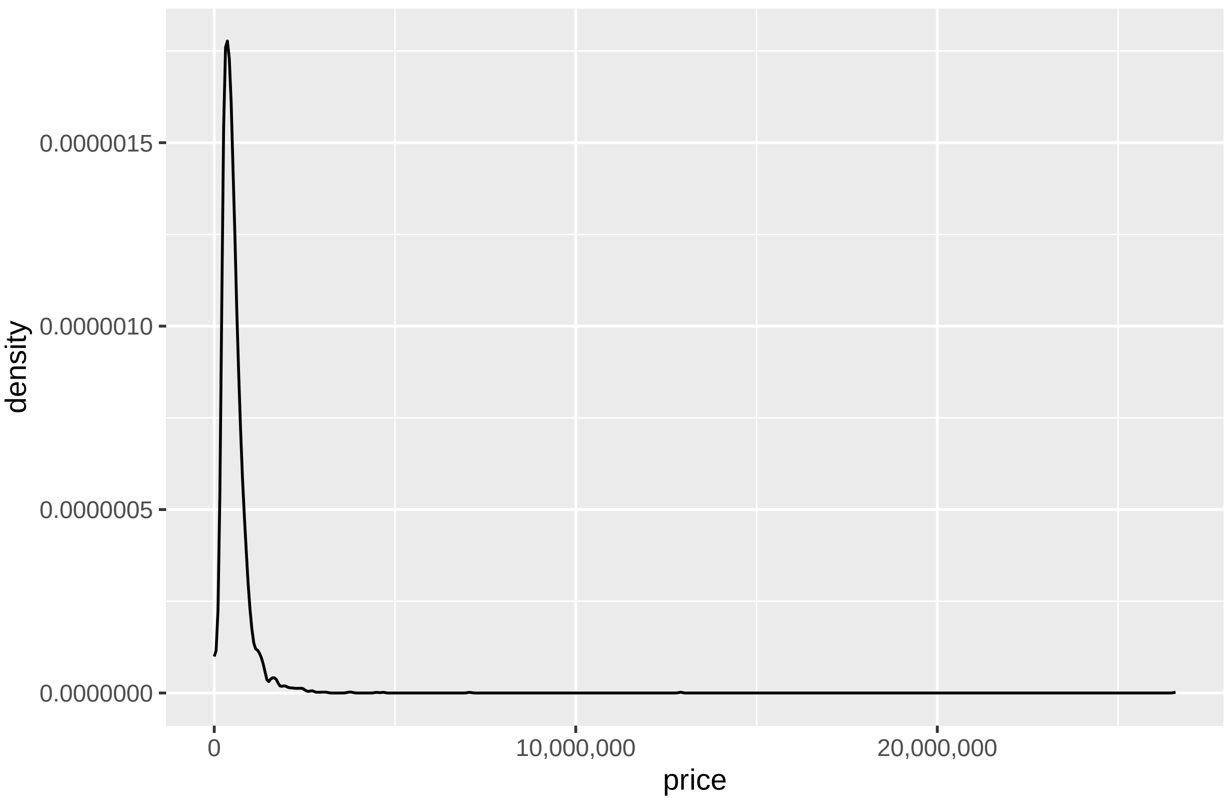
```r
df$price %>% summary()
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##        0   320000   460000   553063   659125 26590000
```

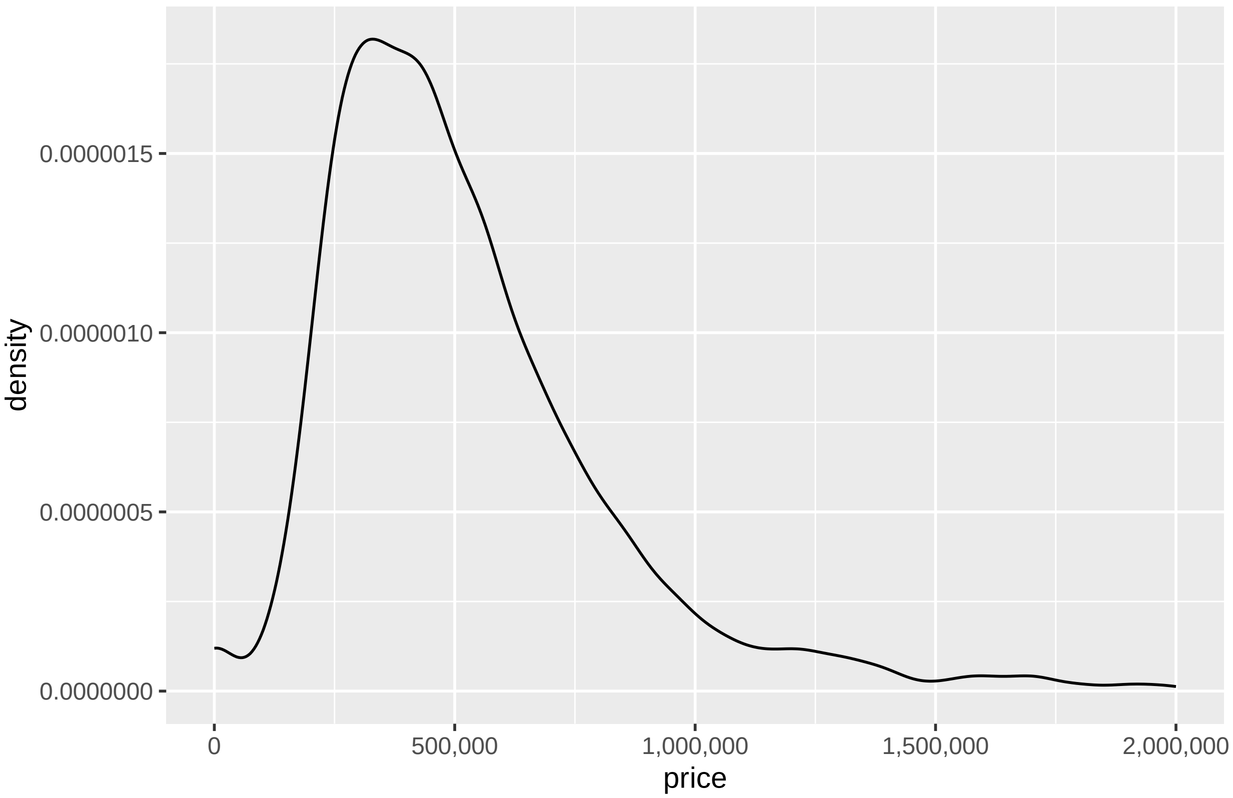
8. Provide a visual summary of your response variable (histogram if discrete, density plot if continuous, bar plot if categorical.)

For this I will include two density plots as there is an extreme positive skew in my data. First will be the unfiltered dataset. Second will filter down to just homes under $2,000,000 in price so we get a better view of the distribution at lower prices.

Distribution of Housing Prices

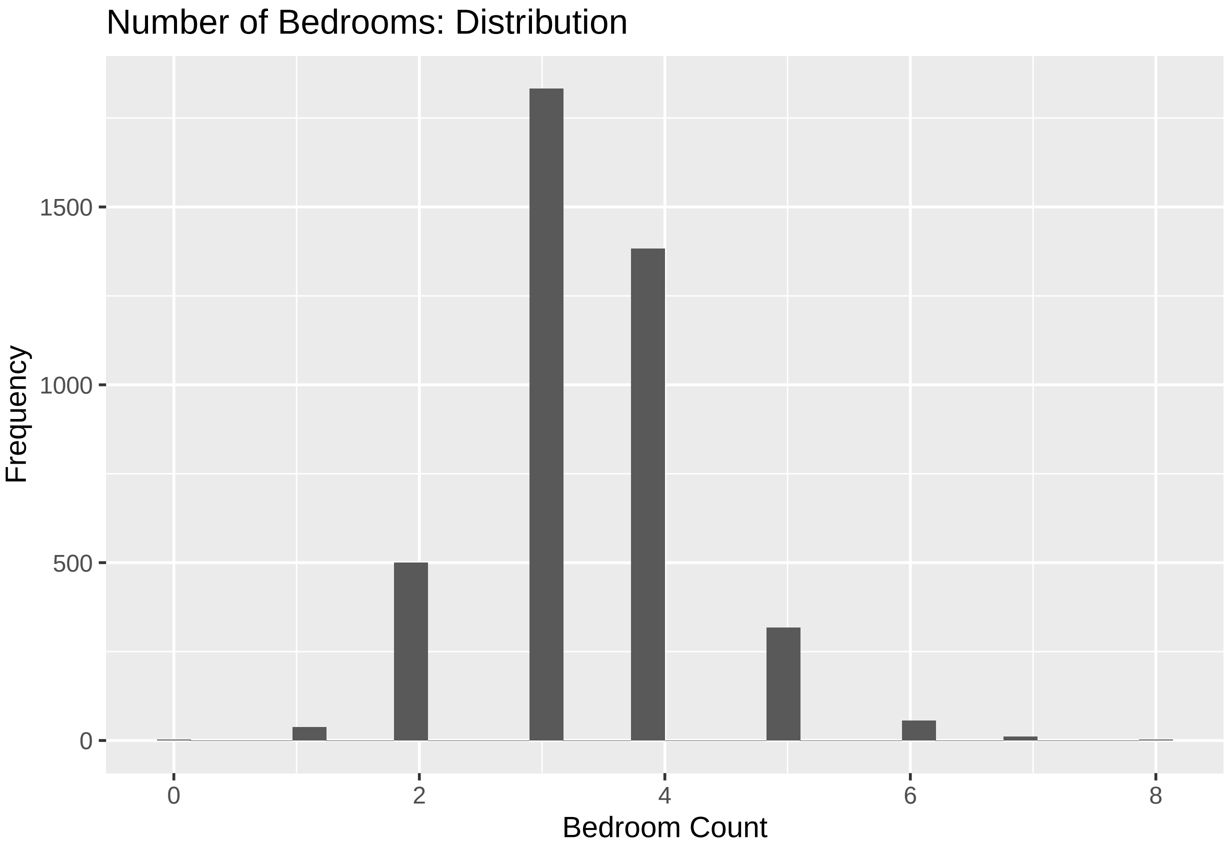


Distribution of Housing Prices Below $2,000,000



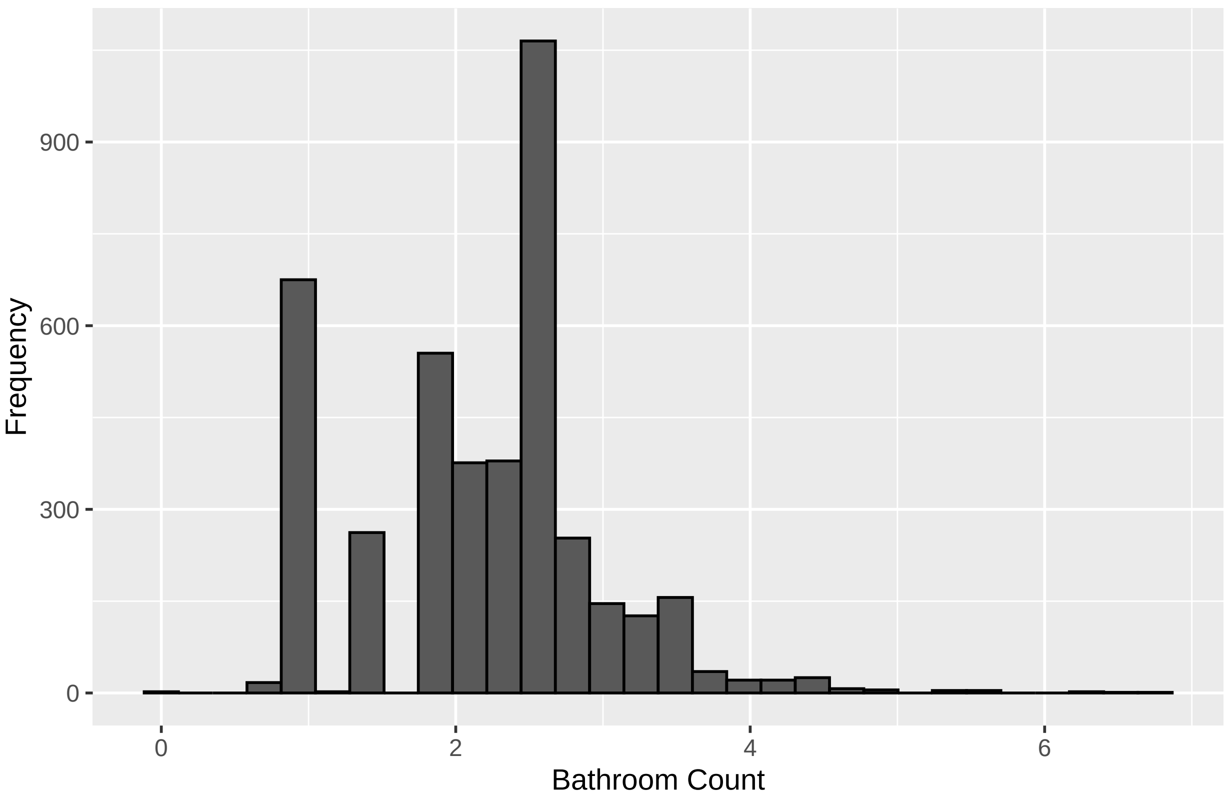
9. Provide an appropriate graphical summary for each predictor variable.

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

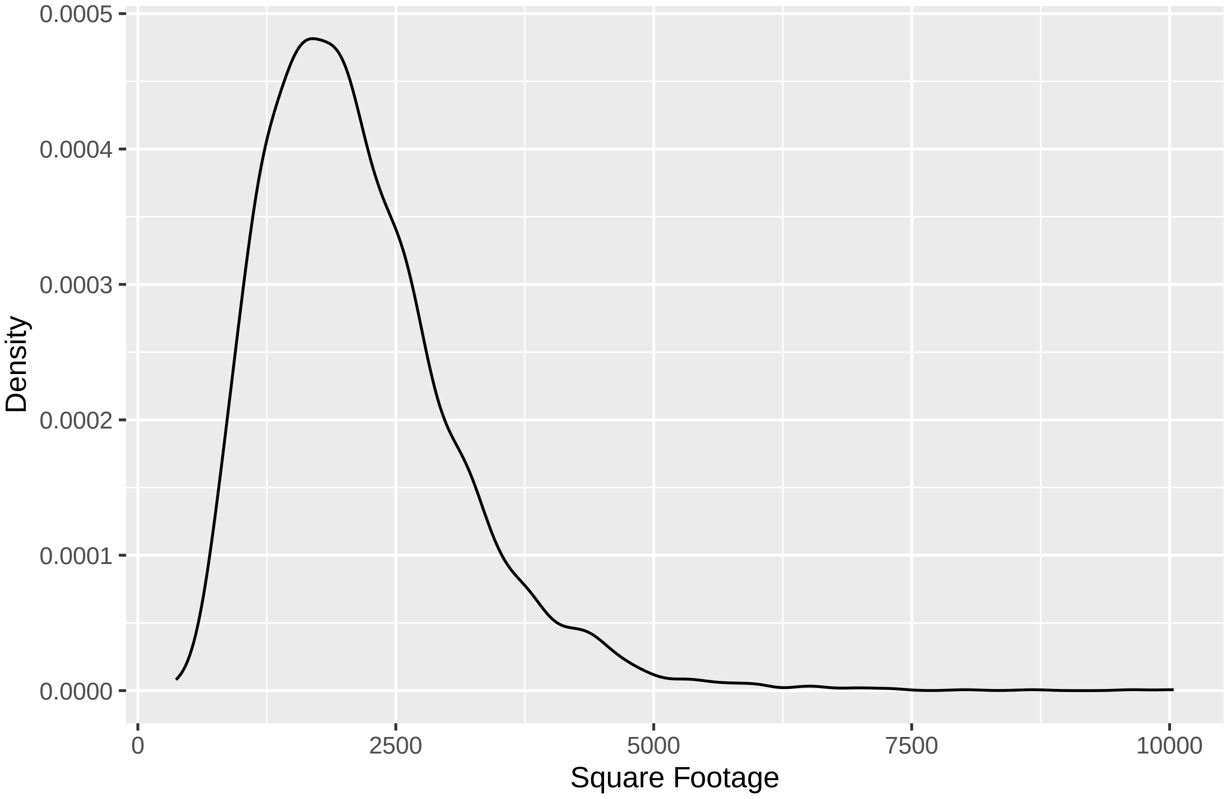


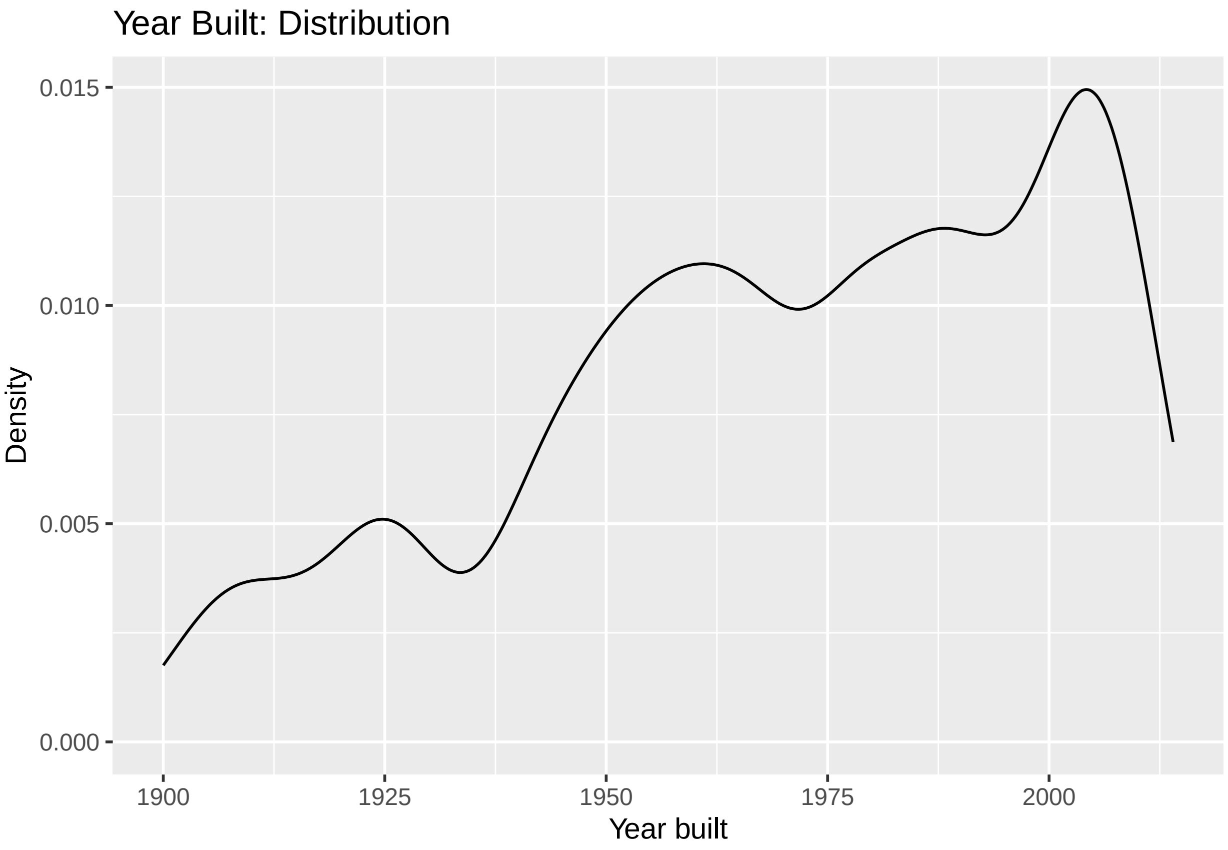
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

Number of Bathrooms: Distribution

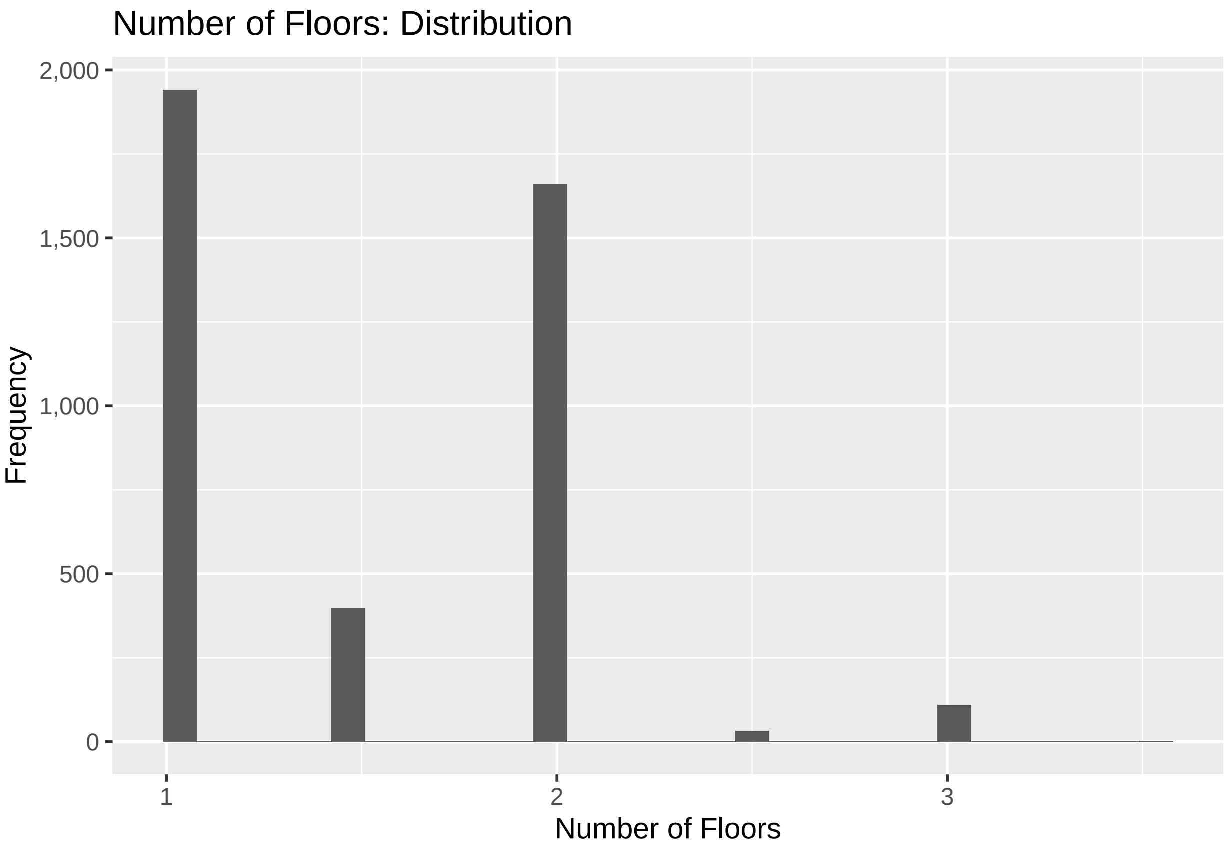


Living Room Square Footage: Distribution





```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Number of Floors: Distribution
Frequency
2,000
1,500
1,000
500
0
1
2
3
Number of Floors