STATS1: First Group Project

. . .

Due Friday, March 15

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(mosaic)
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.
##
## Attaching package: 'mosaic'
##
## The following object is masked from 'package:Matrix':
##
##     mean
##
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
##
## The following object is masked from 'package:purrr':
##
##     cross
##
## The following object is masked from 'package:ggplot2':
##
##     stat
```

```
##
## The following objects are masked from ’package:stats’:
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
##
## The following objects are masked from ’package:base’:
##
##     max, mean, min, prod, range, sample, sum
```

```
#Download cook county housing data
cook_county <- read_csv("~/Stats 172 S24/Class/Project/Project_data/group1_cook_county_data_Project2.cs
```

```
## Rows: 1000 Columns: 27
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (14): pin, construction_quality, garage_attached, basement_type, centra...
## dbl  (12): township_code, num_bedrooms, num_fireplaces, num_full_baths, num_...
## date  (1): sale_date
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

By completing this group project, you will be able to:

- Generate a question that can be answered with provided data.
- Perform EDA for two numerical values.
- Create a linear regression model and interpret it in the context of a problem.
- Write accurate inferential conclusions and interpretations for slope in a simple linear regression.

Useful resources for this homework:

- Book chapters 7 and 24
- In class activities from Unit 3
- Project_data_sets.Rmd (in the Class>Project Folder)

Introduction

For your first group project you will choose one of three datasets provided for you.

- Data Set 1: Cook County Housing Data
- Data Set 2: Canine Assisted Interventions
- Data Set 3: Predicting Classroom Performance

The datasets are available under the folder `Stats 172 S24/Class/Project/Project_data`.

Descriptions of the three datasets are available under the folder `Stats 172 S24/Class/Project/Project_data_sets.Rmd`

Your Job

0. Clearly identify the dataset your group will be working with. Load in that dataset and provide a brief overview of origin of the data.

We will be using Cook County Housing Data dataset for this analysis.

The dataset covers a wide range of properties sold within Cook County, having detailed information about property sales in the Chicago area, including both the square footage, year the properties were built, sale price, and many others. The dataset is publicly available, hosted on GitHub

For this first analysis we will use `year_built2`: Year home was built (explanatory) and `sale_price`: Price of sale in dollars (response)

Question: Is there a relationship between the year a house was built and its sale price?

2. For each of your variables perform a single variable exploratory analysis, including summaries and visualizations. Describe in your own words the results of your summaries and visualizations.

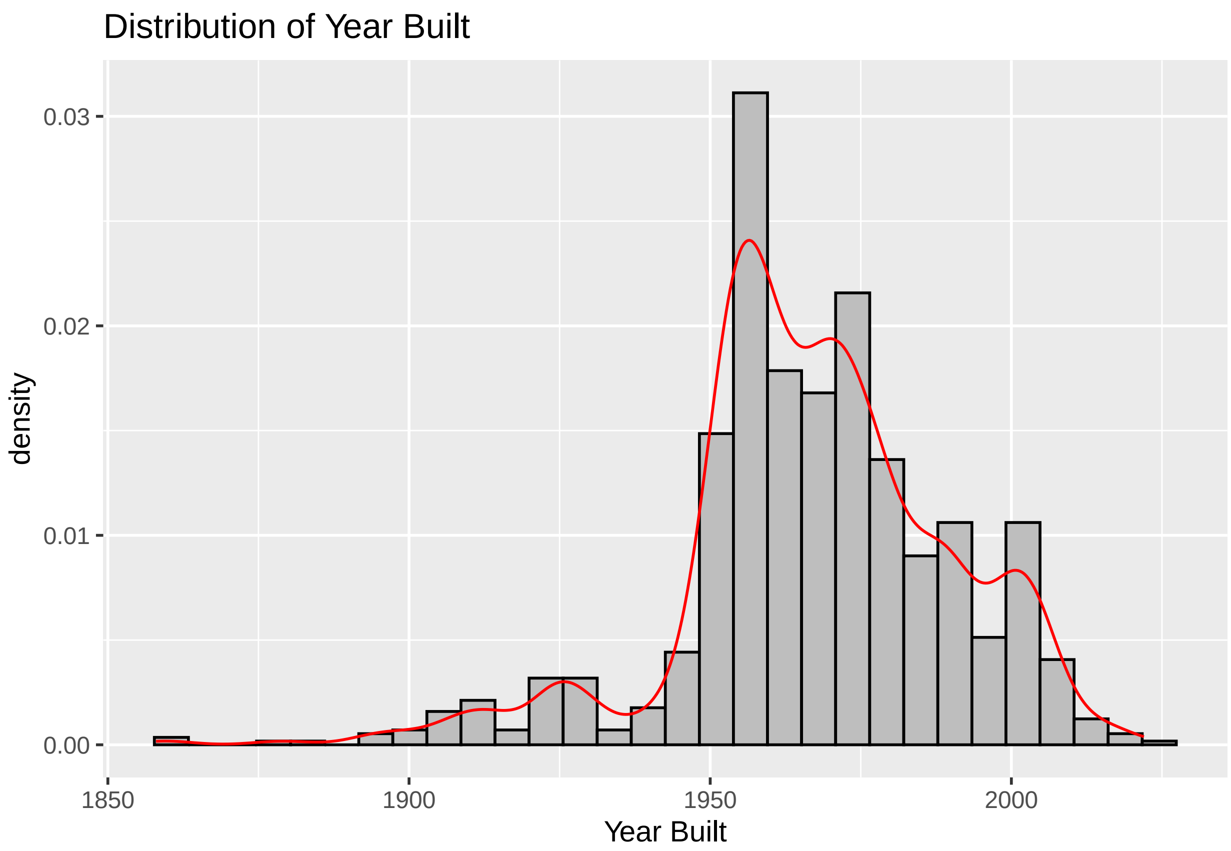
```
summary(cook_county$year_built2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1858    1955    1967    1967    1980    2022
```

```
ggplot(data = cook_county, aes(x = year_built2)) +
  geom_histogram(aes(y = ..density..),fill = "grey", color = "black") +
  geom_density(color = "red") +
  labs(title = "Distribution of Year Built",
       x = "Year Built")
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

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

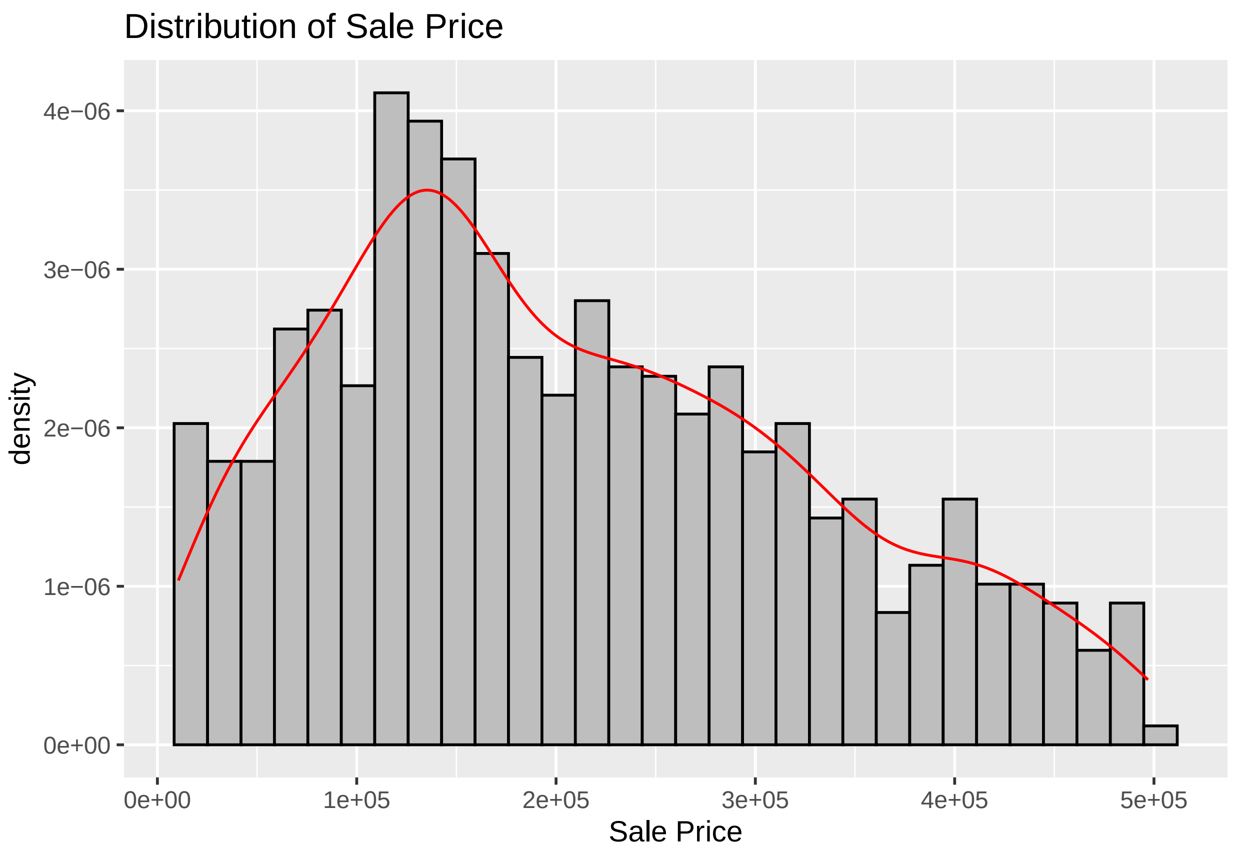


```
summary(cook_county$sale_price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   10500  118000  188500  208699  293000  497000
```

```
ggplot(data = cook_county, aes(x = sale_price)) +
  geom_histogram(aes(y = ..density..),  fill="gray", color = "black") +
  geom_density(color = "red")+
  labs(title = "Distribution of Sale Price",
       x = "Sale Price")
```

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



density+histogram source

In relation to the `year_built2` variable, it is possible to see that the majority of the houses in the dataset were built between the mid-20th century and early 21st century. The range of construction years spans from 1858 to 2022. The median and mean year built are both around 1967, suggesting that the dataset is roughly centered around this year, with half of the houses built before and half after.

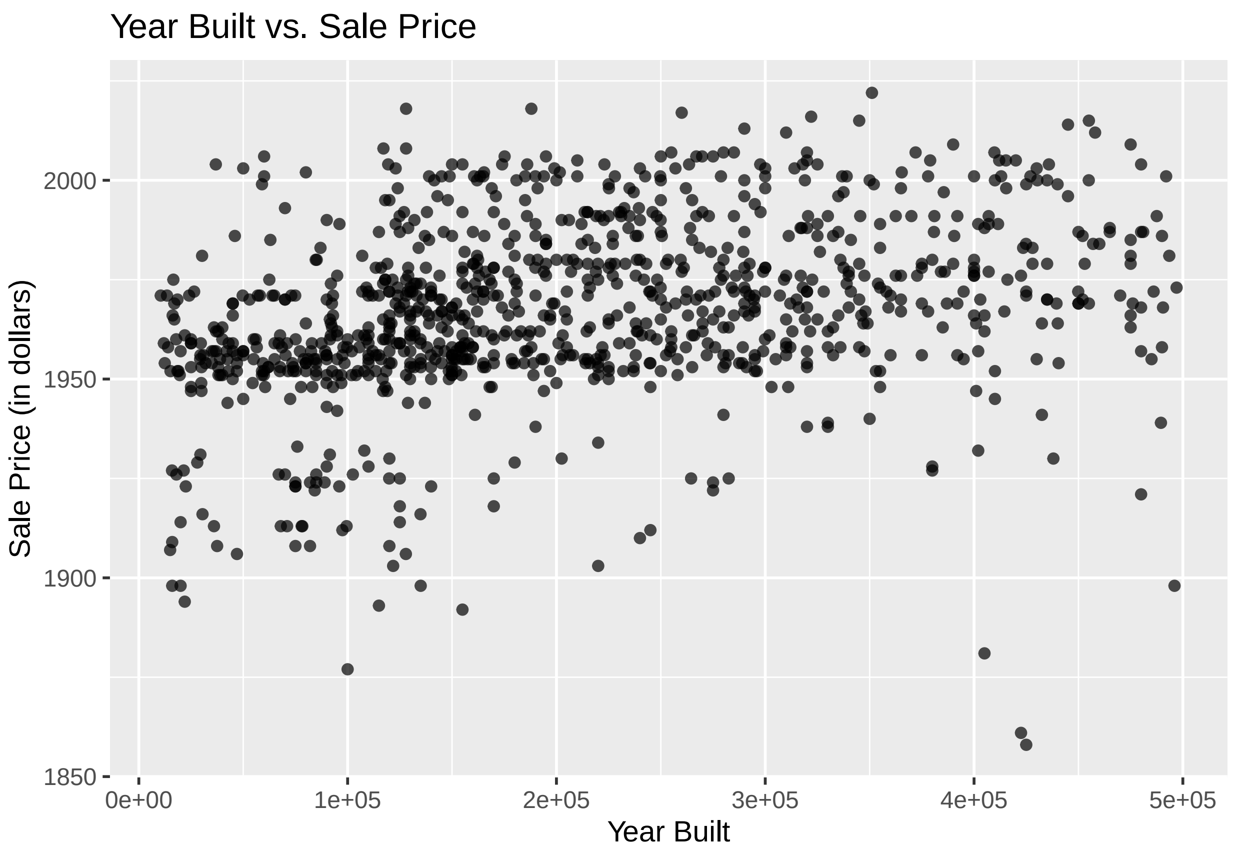
In relation to the `sale_price` variable, we can notice that the median and mean sale prices are 188,500 and 208,699 dollars, respectively. This suggests that the majority of the sale prices fall below the mean, indicating a possible right-skewed distributions.

3. Explore and describe the relationship between the two variables. Make sure to describe the results of your summary and visualization.

```
mosaic::cor(year_built2 ~ sale_price, data=cook_county)
```

```
## [1] 0.323128
```

```
ggplot(data = cook_county, aes(x = sale_price, y = year_built2, )) +
  geom_jitter(width = 0.3, height = 0, alpha = 0.7) +
  labs(title = "Year Built vs. Sale Price",
       x = "Year Built", y = "Sale Price (in dollars)")
```



The correlation coefficient of 0.32 suggests a positive linear relationship between year built and sale price.This suggests that, on average, newer properties tend to have higher sale prices compared to older properties However, the correlation is not very strong. This shows that there is still considerable variability in sale prices that is not explained by the year a property was built alone.

With the plot it is possible to visualize this positive but moderate relation between the variables.

4. Fit a linear regression model and include the fitted line on a plot of your data and write the fitted regression equation. Make sure to report and interpret the following from the regression output

   - Intercept
   - R-squared
   - Slope
   - 95% confidence interval for the slope

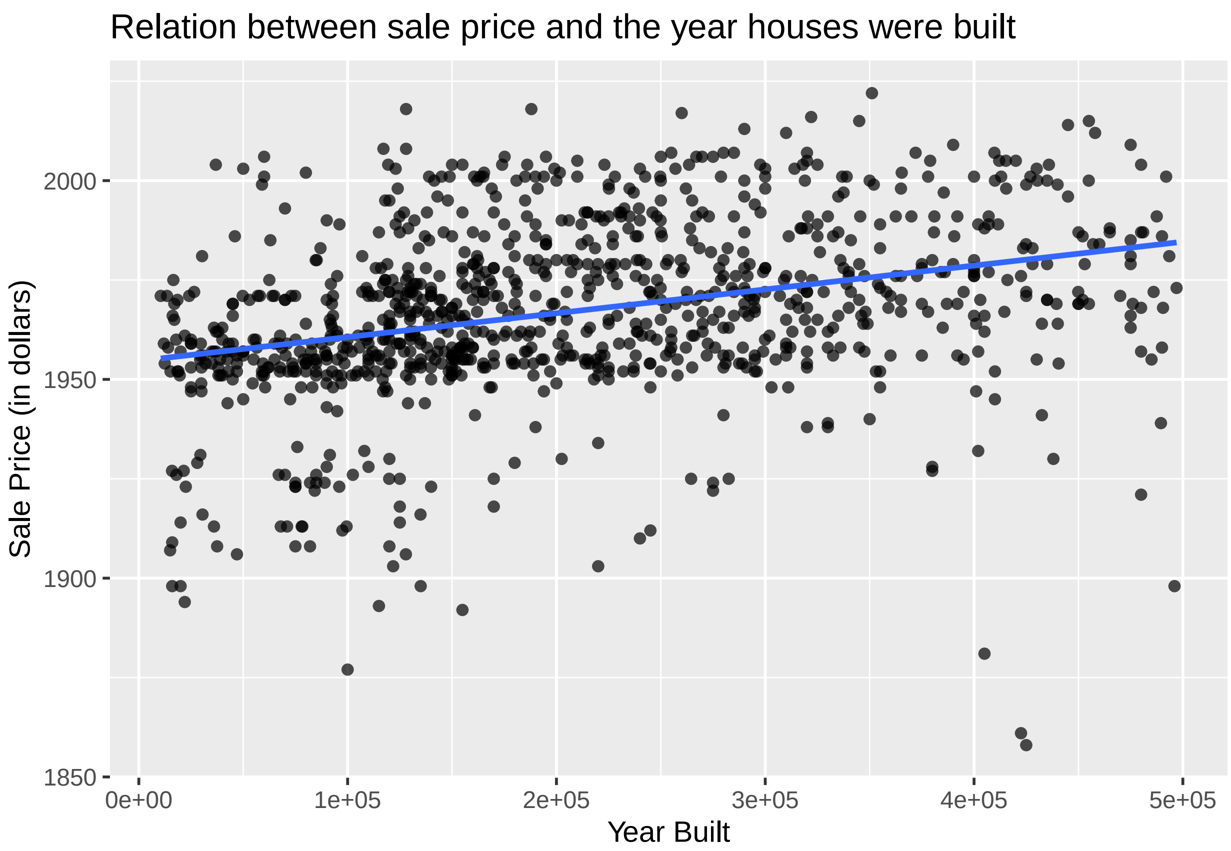
```
fit <- lm(sale_price ~ year_built2, data = cook_county)
summary(fit)
```

```
##
## Call:
## lm(formula = sale_price ~ year_built2, data = cook_county)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -235987  -86928  -25178   77138  407758
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3217490.0   317653.9  -10.13   <2e-16 ***
## year_built2     1741.7      161.5   10.79   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 115000 on 998 degrees of freedom
## Multiple R-squared:  0.1044, Adjusted R-squared:  0.1035
## F-statistic: 116.4 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
ggplot(data = cook_county, aes(x = sale_price, y = year_built2, )) +
  geom_jitter(width = 0.3, height = 0, alpha = 0.7) +
  geom_smooth(method = lm, se = FALSE) +
  labs(title = "Relation between sale price and the year houses were built",
       x = "Year Built", y = "Sale Price (in dollars)",
       color = "Price Range")
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



```
confint(fit)
```

```
##                     2.5 %       97.5 %
## (Intercept) -3840836.140 -2594143.939
## year_built2     1424.837     2058.547
```

$\hat{y} = \beta_0 + \beta_1 x + \varepsilon_i$

$\hat{y} = -3217490 + 1741.7 \cdot yearbuilt2 + \varepsilon_i$

The intercept (-3217490.0): represents the estimated value of the house sale price when the year of construction is zero. However, in this context, it's not meaningful to interpret since the year built cannot be zero.

The slope (1741.7): for each additional year a house is built, the estimated sale price increases by $1741.7.

R-squared (R^2): in this case, the R^2 value is 0.1044, indicating that approximately 10.44% of the variability in sale price can be explained by the linear relationship with the year built.

Confidence intervals: It is possible to be 95% confident that for each year passed in relation to the year the house was built, the sale price increase around $833.71 dollars.

With the plot and the analysis we can see that it suggests a positive trend, though the fit might not be very strong, considering the low R^2 value.

5. State the null and alternative hypothesis in symbols and in words.

$$H_0 : \beta = 0$$

There is no relationship between the year a house was built and its sale price

$$H_A : \beta \neq 0$$

There is a relationship between the year a house was built and its sale price.

6. Report and interpret the P-value and confidence interval for the slope. Based on the p-value for the slope state your statistical conclusion.

```
confint(fit)
```

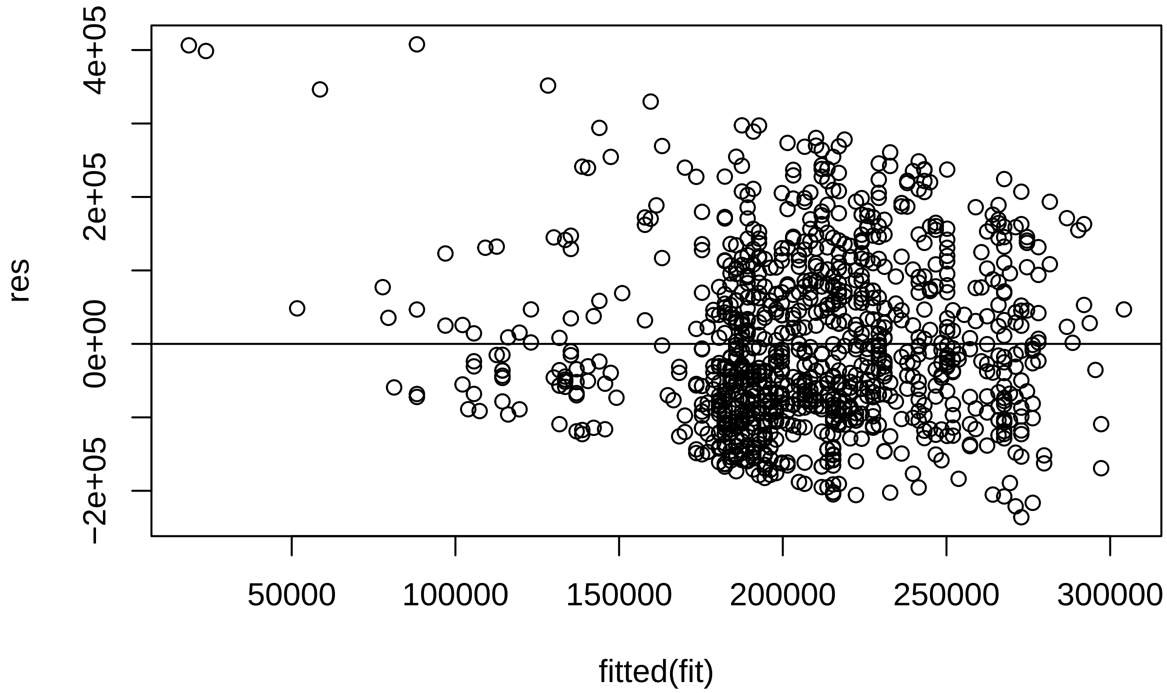
```
##                    2.5 %       97.5 %
## (Intercept) -3840836.140 -2594143.939
## year_built2     1424.837     2058.547
```

Confidence Interval: with this analysis we are 95% confident that for each year passed, the mean price value increases from 1424.8 to 2058.5 dollars.

P-value: a very small p-value, such as 2.2e-16 here, suggests strong evidence against the null hypothesis. This indicates that there is a significant linear relationship between the variables analyzed.
Given the small p-value, we reject the null hypothesis and conclude that there is a statistically significant linear relationship between the explanatory variable and the response variable. Therefore, we can say that as time passes, there is a significant change in the price value, as indicated by the slope of the regression line.

7. Comment on the role of outliers and the four conditions for regression inference in your models.

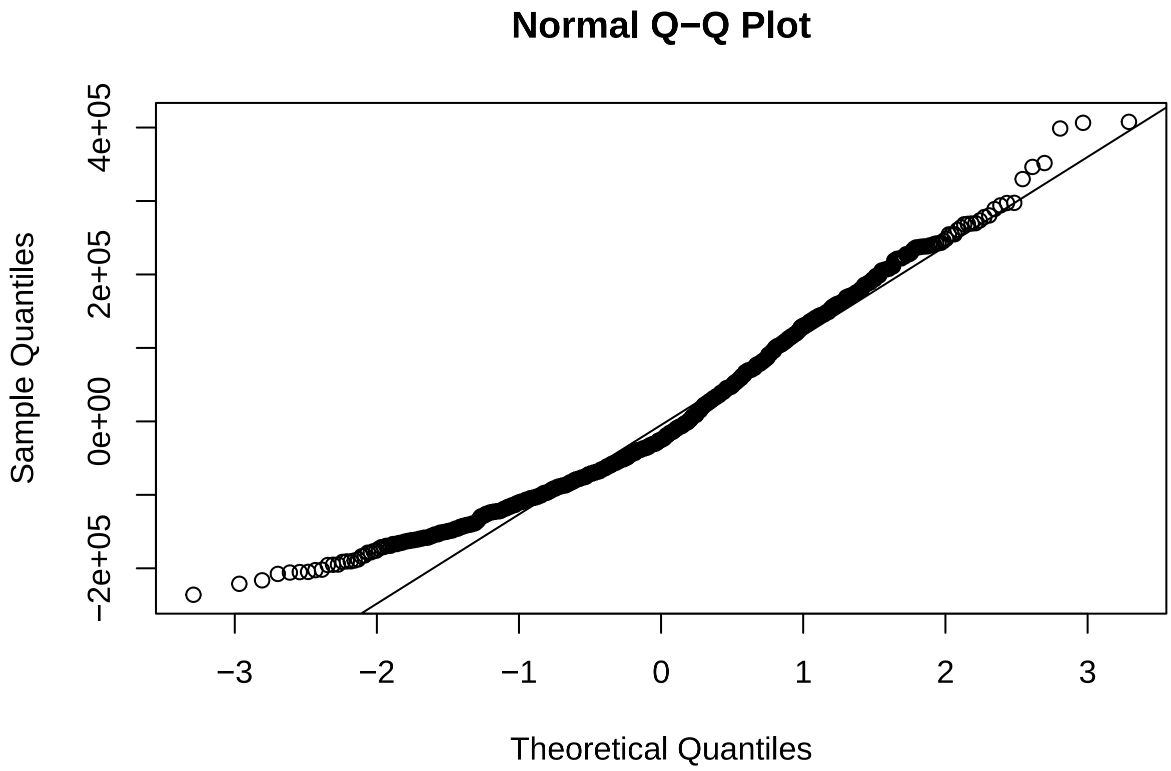
```
#residual vs fitted plot
res <- resid(fit)#These three lines of code will work instead.
plot(fitted(fit), res) + abline(0,0) #add a 0 line to the plot
```



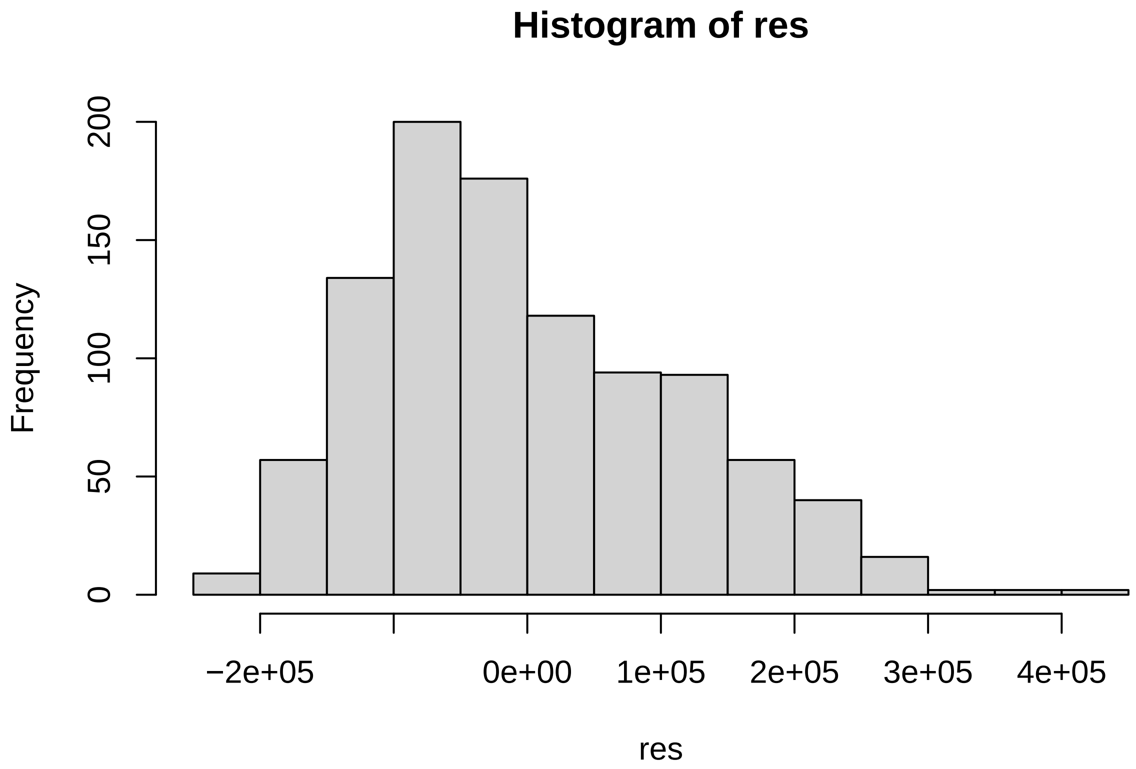
```
## integer(0)
```

```
#qq plot
qqnorm(res) + qqline(res)
```

```
## Error in qqnorm(res) + qqline(res): non-numeric argument to binary operator
```



```
#there is an error in the gg plot but in order to show the graph in the output we used error=TRUE

#histogram
hist(res)
```



Outliers: The outliers here do not have a significant influence in the final analysis.

Linearity: It is possible to identify a linear relation between the two variables analyzed.

Independence: The data exhibits spatial dependence, as all obsrervations are related to the same region at Cook County in Chicago.

Normality: The sale price is not normally distributed at each level of the year the house were built. We can observe it by the failure of the histogram in being normally distributed, presenting a left skewed shape, and also by the presence of a considerable amount of outliers in the q-q plot

Equal Variance: It is possible to observe a considerable equal variance for this analysis even thought there is more data point toward the left of the Residual vs Predicted plot, this could indicate that most of the houses were built in recent years and also that most houses tend to have a higher price rather than a lower.

8. Identify a new explanatory variable (with the same response) or a new response variable (with the same explanatory) and repeat steps 2-7. Your new variable should also be numeric.

We we would be useing: `build_sqft2`: square footage of the home (explanatory) and `sale_price`: Price of sale in dollars (response)

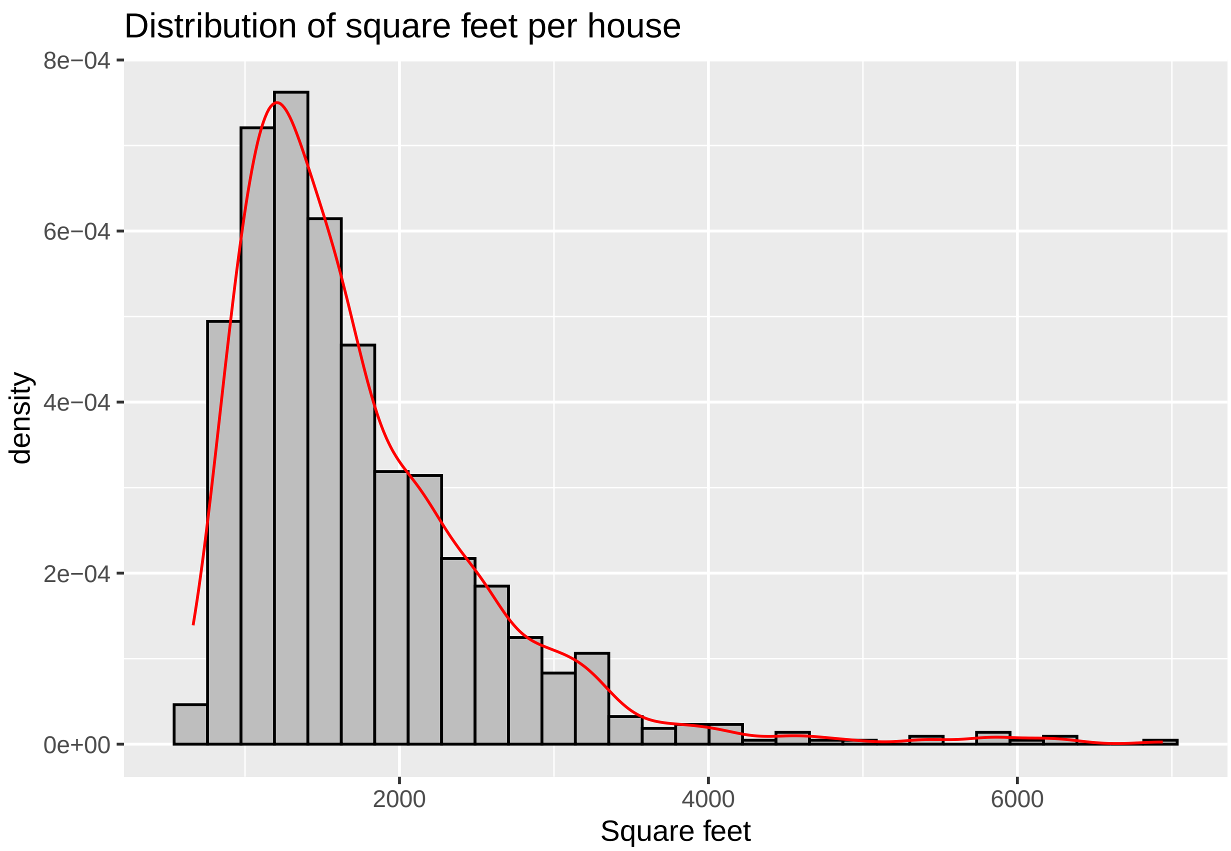
Question: Is there a relationship between the square footage of the house and its sale price?

```
summary(cook_county$build_sqft2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     664    1153    1510    1722    2085    6941
```

```
ggplot(data = cook_county, aes(x = build_sqft2)) +
  geom_histogram(aes(y = ..density..),fill = "grey", color = "black") +
  geom_density(color = "red") +
  labs(title = "Distribution of square feet per house",
       x = "Square feet")
```

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

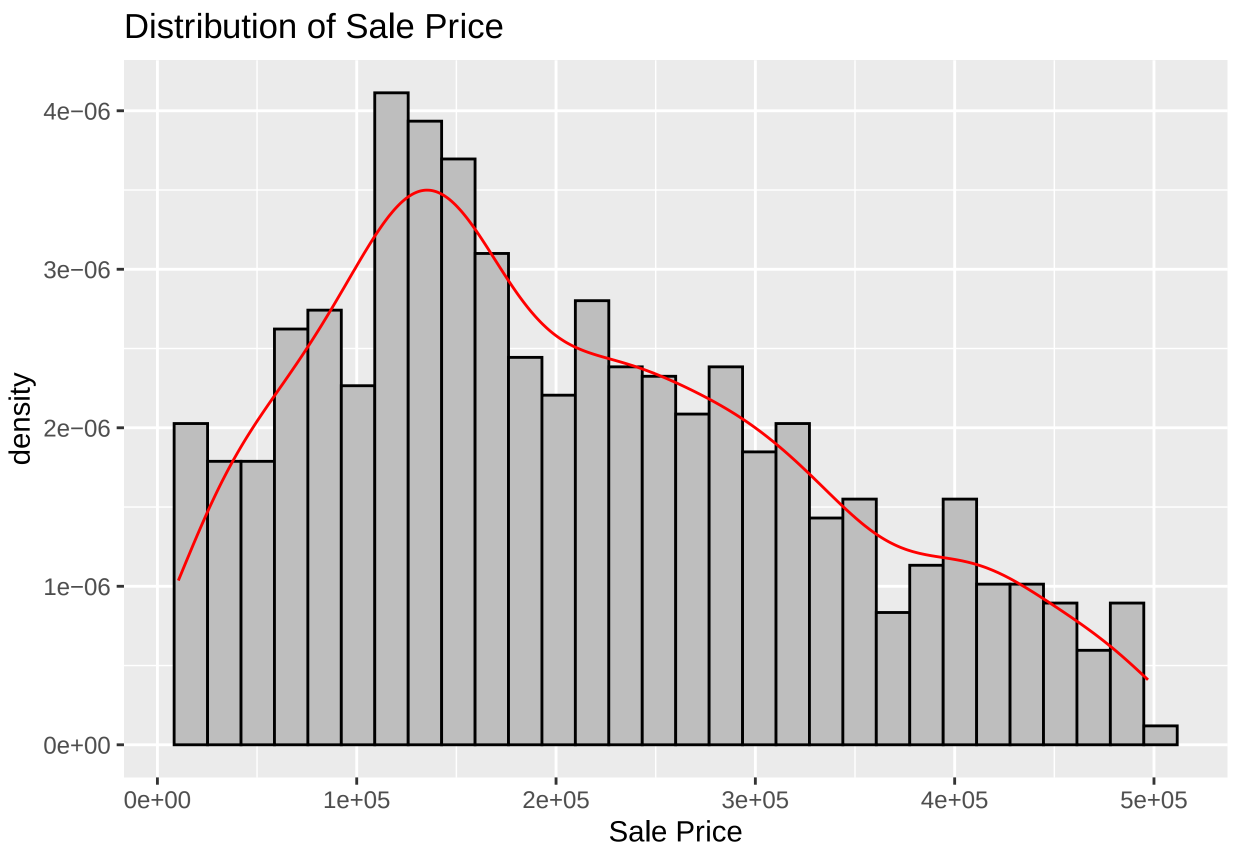


```
summary(cook_county$sale_price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10500  118000  188500  208699  293000  497000
```

```
ggplot(data = cook_county, aes(x = sale_price)) +
  geom_histogram(aes(y = ..density..),  fill="gray", color = "black") +
  geom_density(color = "red")+
  labs(title = "Distribution of Sale Price",
       x = "Sale Price")
```

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



In relation to the `sale_price` variable, we can notice that the median and mean sale prices are 188,500 and 208,699 dollars, respectively. This suggests that the majority of the sale prices fall below the mean, indicating a possible right-skewed distributions.

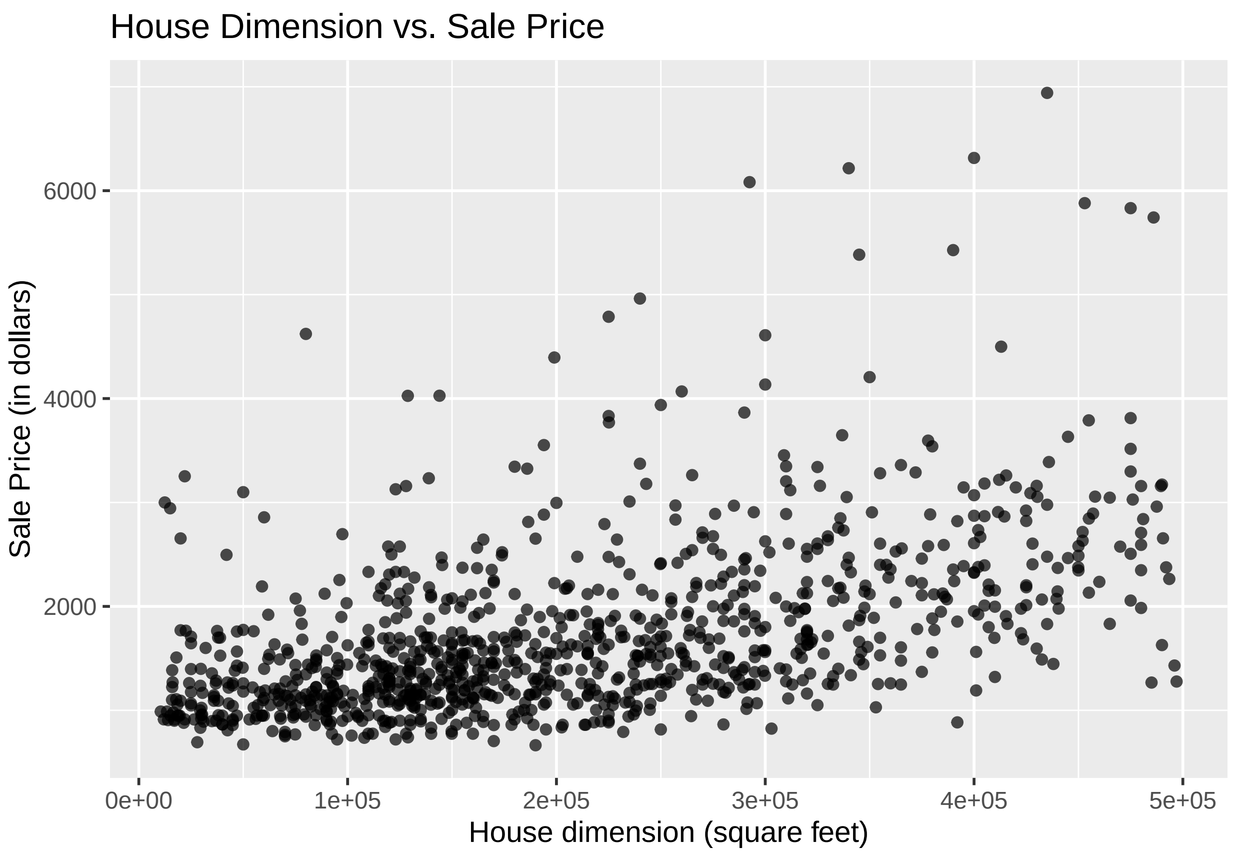
In relation to `build_sqft2` variable, we can see that the media is 1540 square feet while the mean is 1722 square feet, what suggests a right-skewed distribution. The spread is from 664 to 6941 square feet. With the plot it is also possible to see some outliers at around 6000 square feet.

Explore and describe the relationship between the two variables. Make sure to describe the results of your summary and visualization.

```
mosaic::cor(build_sqft2 ~ sale_price, data=cook_county)
```

```
## [1] 0.5136644
```

```
ggplot(data = cook_county, aes(x = sale_price, y = build_sqft2 )) +
  geom_jitter(width = 0.3, height = 0, alpha = 0.7) +
  labs(title = "House Dimension vs. Sale Price",
       x = "House dimension (square feet)", y = "Sale Price (in dollars)")
```



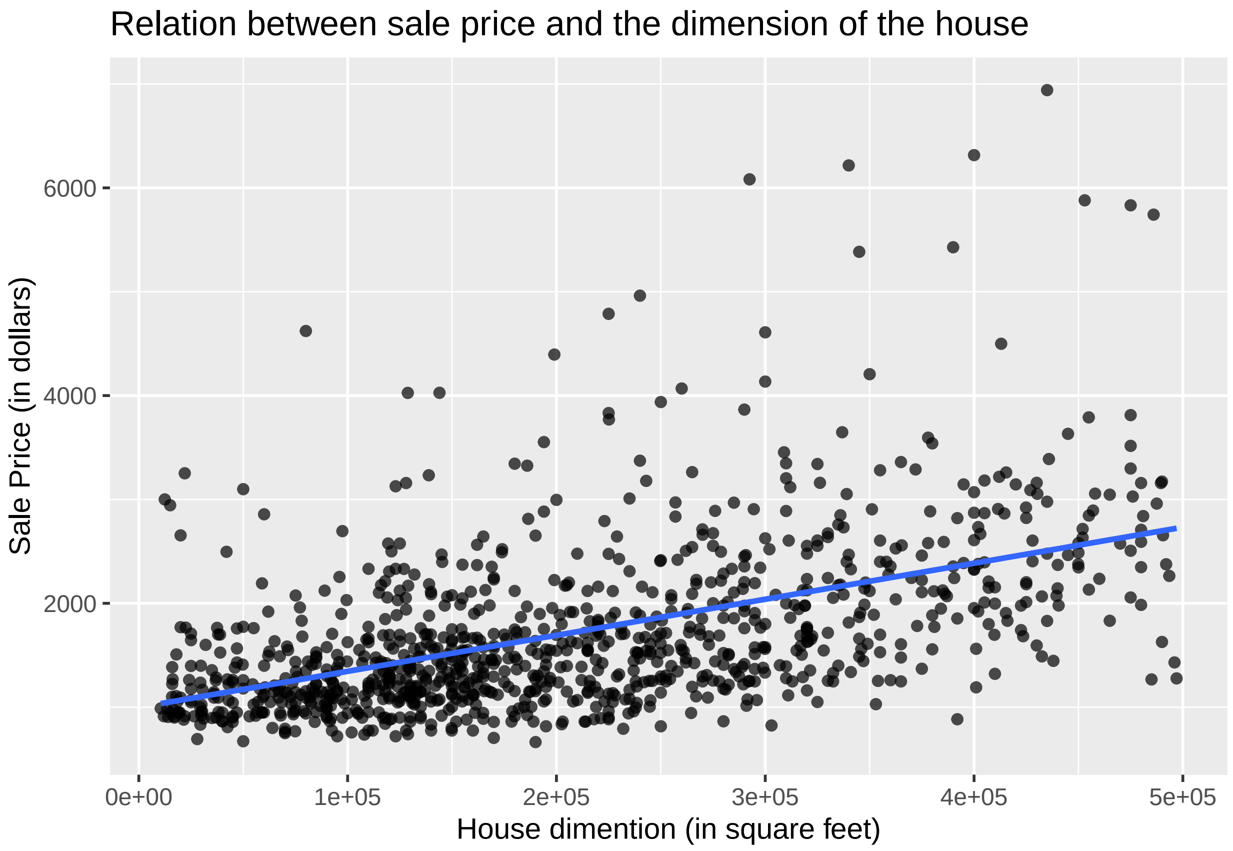
The correlation is 0.51 what indicates a positive relation between the variables. This suggests that, on average, larger houses tend to have higher sale prices. However, the small correlation coefficient indicates that the relationship is not strong. While there is a clear positive trend, there’s still some variability in sale prices.

```
fit2 <- lm(sale_price ~ build_sqft2, data = cook_county)
summary(fit2)
```

```
##
## Call:
## lm(formula = sale_price ~ build_sqft2, data = cook_county)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -349242  -76743  -12213   75533  322174
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 77699.798   7671.071   10.13   <2e-16 ***
## build_sqft2    76.058      4.021   18.91   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 104300 on 998 degrees of freedom
## Multiple R-squared:  0.2639, Adjusted R-squared:  0.2631
## F-statistic: 357.7 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
ggplot(data = cook_county, aes(x = sale_price, y = build_sqft2, )) +
  geom_jitter(width = 0.3, height = 0, alpha = 0.7) +
  geom_smooth(method = lm, se = FALSE) +
  labs(title = "Relation between sale price and the dimension of the house",
       x = "House dimention (in square feet)", y = "Sale Price (in dollars)",
       color = "Price Range")
```

```
## 'geom_smooth()' using formula = 'y ~ x'
```



```
confint(fit2)
```

```
##                   2.5 %      97.5 %
## (Intercept) 62646.51967 92753.07645
## build_sqft2    68.16687    83.94989
```

$\hat{y} = \beta_0 + \beta_1 x + \varepsilon_i$

$\hat{y} = 77699.798 + 76.058 \cdot yearbuilt2 + \varepsilon_i$

The intercept (77699.798): represents the estimated value of the house sale price when the the square foot of the house is zero. However, in this context, it's not meaningful to interpret since the square foot of a house cannot be zero.

The slope (76.058): for each additional square foot a house has, the estimated sale price increases by $76.058.

R-squared (R^2): the R^2 value is 0.2639, indicating that approximately 26.39% of the variability in sale price can be explained by the square feet a house has, so the bigger the house the more expensive it will be to buy it.

Confidence intervals: it is possible to be 95% confident that for each square foot increased in the total dimension of the house, the sale price increase by around 15.8 dollars.
With the plot and the analysis we can see that it suggests a positive trend, though the fit might not be very strong, considering the low R^2 value.

Null and alternative hypothesis:

$$H_0 : \beta = 0$$

There is no relationship between the the square footage of the house and its sale price

$$H_A : \beta \neq 0$$

There is a relationship between the the square footage of the house and its sale price.

Report and interpret the P-value and confidence interval for the slope.

```
confint(fit2)
```

```
##                    2.5 %      97.5 %
## (Intercept) 62646.51967 92753.07645
## build_sqft2    68.16687    83.94989
```

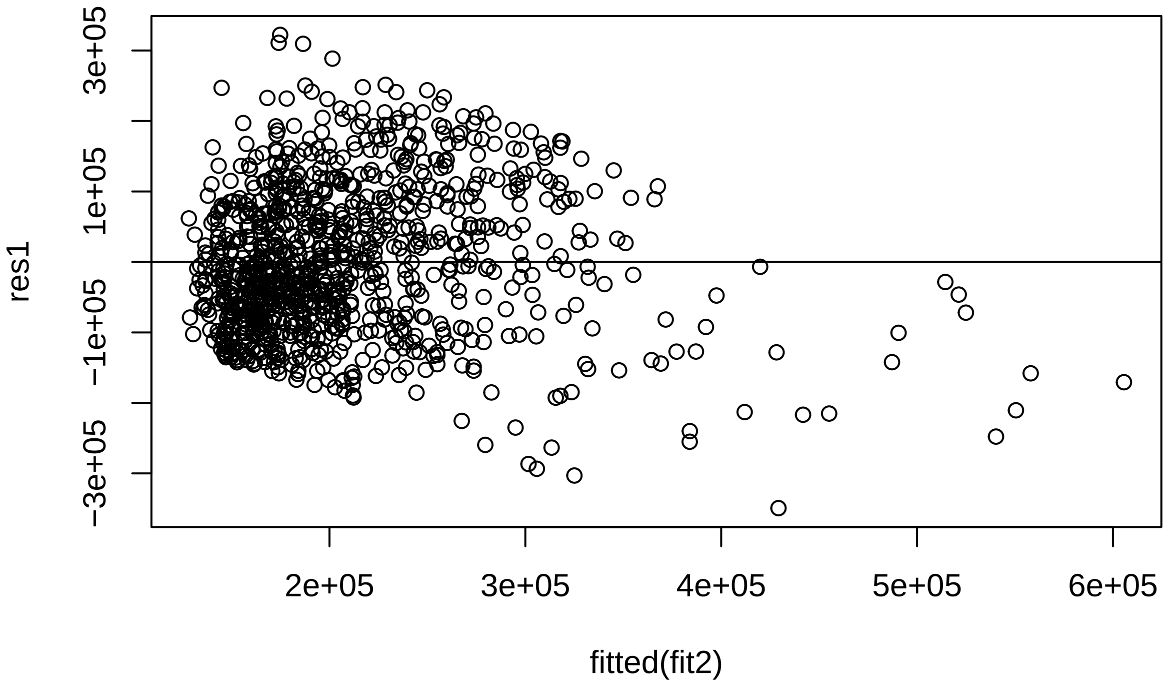
Confidence Interval: with this analysis we are 95% confident that for each square foot increased in the total dimension of the house, the price value increases from 68.17 to 83.95 dollars.

P-value: a very small p-value, such as 2.2e-16 here, suggests strong evidence against the null hypothesis. This indicates that there is a significant linear relationship between the variables analyzed.

Given the small p-value, we reject the null hypothesis and conclude that there is a statistically significant linear relationship between the explanatory variable and the response variable. Therefore, we can say that as the square feet of the house increase, there is a significant change in the price value, as indicated by the slope of the regression line.

Outliers and the conditions for regression inference models.

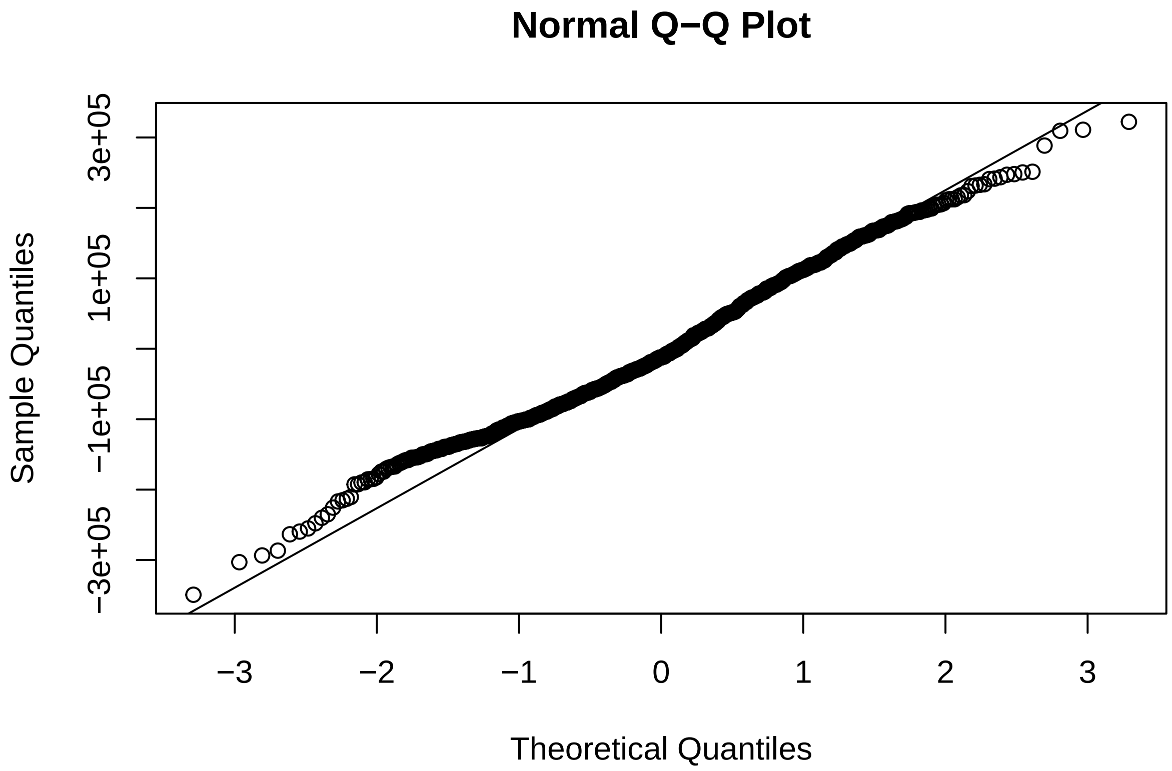
```
#resid_panel(fit)#this is given an error - Laura sent an email saying how to fix it

#residual vs fitted plot
res1 <- resid(fit2)#These three lines of code will work instead.
plot(fitted(fit2), res1) + abline(0,0) #add a 0 line to the plot
```



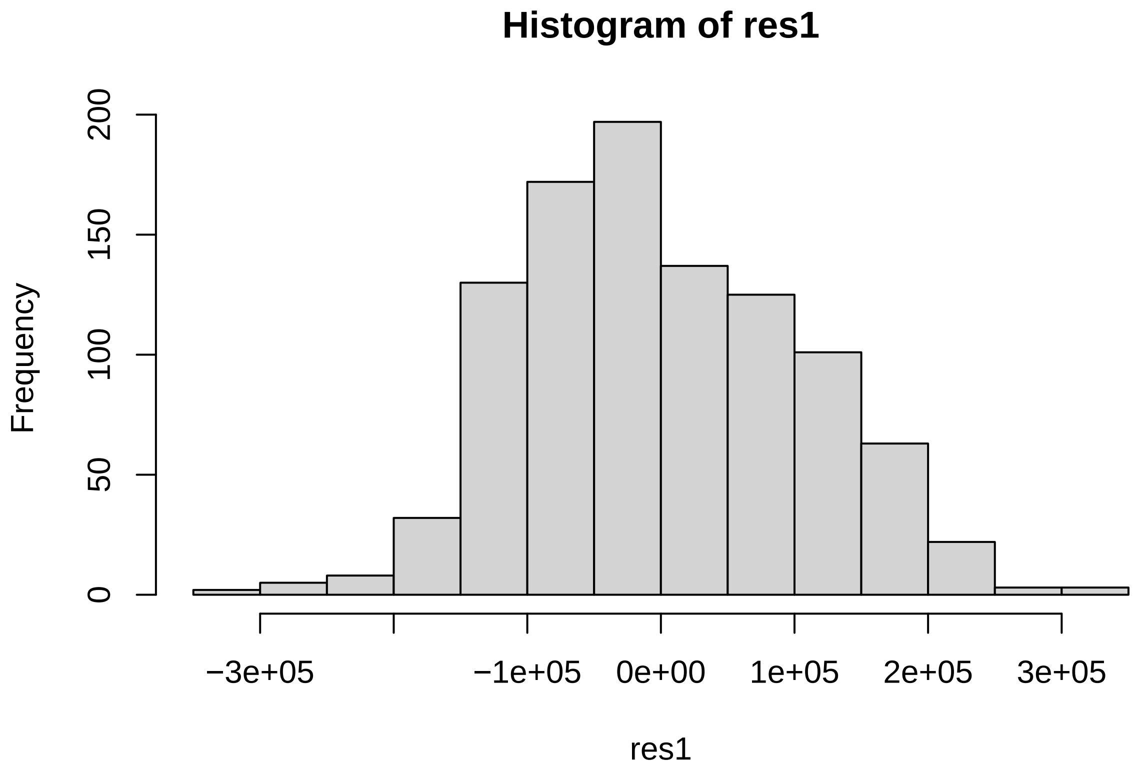
```
## integer(0)
```

```
#qq plot
qqnorm(res1) + qqline(res1)
```

```
## Error in qqnorm(res1) + qqline(res1): non-numeric argument to binary operator
```



```r
#there is an error in the gg plot but in order to show the graph in the output we used error=TRUE

#histogram
hist(res1)
```



Outliers: For this analysis outliers did not have a significant influence on the regression model.

Linearity: The is considerable variance in the normality, however, this variance is not strong enough for the linearity to be disconsidered

Independence: The data shows spatial dependence, given the fact that the all the observations are from the same region at Cook County in Chicago.

Normality: the sale price is normally distributed at each level of the total square feet of the houses. We can observe it by the general normal distribution on the histogram and the Q-Q plot being fairly linear at the normal line

Equal variance: in this analysis it is possible to notice a failure in equal variance since most of the data is cluster together at the far left of the Residual vs Predicted plot and towards the far right of the graph most of the observations fall below the zero line.

9. Comparing your two models, which variable is a better explanatory variable of the response variable you chose? Explain why.

Year Built vs Sale price:

Correlation Coefficient: The correlation coefficient of 0.32 indicates a moderate positive relationship between the year a house was built and its sale price. This means that, on average, newer houses tend to have higher sale prices compared to older houses.

R-squared Value: The R^2 value of 0.1044 suggests that only approximately 10.44% of the variability in sale price can be explained by the year the house was built.

Square Feet vs Sale Price:

Correlation Coefficient: The correlation coefficient of 0.51 indicates a moderately strong positive relationship between the total square footage of a house and its sale price. This suggests that, on average, larger houses tend to have higher sale prices.

R-squared Value: The R^2 value of 0.264 indicates that approximately 26.4% of the variability in sale price can be explained by the total square footage of the house.
With this analysis it is possible to identify that the square footage of the house serves as a better explanatory variable for predicting sale prices compared to the year the house was built. This conclusion is based on the differences in R^2 values between the two analysis. Notably, the square footage vs sale price analysis exhibits a larger R^2 value, indicating that square footage provides greater insight into sale prices than the year the house was built. Additionally, considering the correlation values, we can see that the relationship between square footage and sale price is stronger than that between the year the house was built and sale price.

10. Write a paragraph summarizing your findings to a general audience.

This analysis had as database the Cook County Housing in Chicago that reported a variety of information related to the sale price of houses at that region. We explored two potential explanatory variables: the year a house was built and the total square footage of the house. Our analysis showed that both variables have a positive relationship with sale price, meaning newer houses and larger houses tend to have higher sale prices on average. However, when comparing the models, we found that the total square footage of the house had a slightly stronger explanatory relation with the sale price compared to the year built. This suggests that while the year built provides some insight about sale prices, the size of the house may be a more influential factor. Nonetheless, it's important to note that neither variable alone can fully explain the variability in sale prices, this indicates that other factors, such as geo-location and neighborhood safety, may also play a significant role in determining house sale prices. Overall, our findings highlight the complex relation between diverse factors that influence the real estate markets and the importance of considering multiple variables when analyzing house sale prices.

Knit this file. Submit your pdf to Moodle.