STA 108 Project 1

Tyler Venner, Queena Huang, Yunxiao Li, Anna Yeh, Sherry Zheng
[Justin Santos did not participate]

Instructor: Maxime Pouokam

Residential Area in Seattle, King County, Washington State

Ⅰ. Introduction

The real estate industry is a significant component of the economy, and understanding the factors that influence home prices is of great interest to buyers, sellers, real estate professionals, and policymakers. One important question in the real industry is whether or not there is a relationship between the square footage of a home's living space and its market price if a person wants to move to King County. Understanding this relationship is crucial for stakeholders to make informed decisions about purchasing and selling. This question is pertinent because it can inform potential homebuyers about what to expect in terms of pricing as they search for homes with their desired living space. It also assists sellers and realtors in setting competitive prices for properties on the market. So for this project, we will focus on in King County whether or not square footage does have an effect on market price, then give a series of methods to answer it.

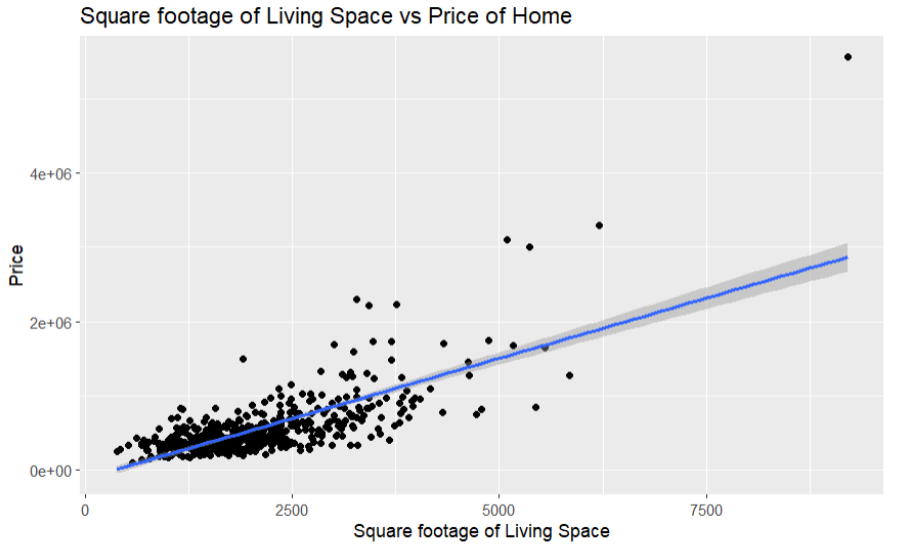
In this project, we begin by summarizing and visualizing the data with Histogram and Boxplot to gain an initial understanding of the data before applying any inferential statistics or predictive models. Then we will apply the QQ plot and Residual plot and two tests, Shapiro-Wiks to determine if the sample is normally distributed and Fligner-Killeen test to examine homoscedasticity, to ensure we satisfy all the requirements to fit a linear regression model. Further, as part of our analysis, we will also explore predictive insights by estimating the potential selling price for homes based on their square footage. To capture the uncertainty inherent in such predictions, we will calculate 90% confidence intervals. These intervals will provide a range within which we expect the true selling price of a home with a given square footage, such as 3200 square feet, to fall. This aspect of our analysis is crucial for stakeholders who require not just point estimates, but also an understanding of the associated prediction uncertainty when making informed decisions in the real estate market.

Ⅱ. Summary of Data

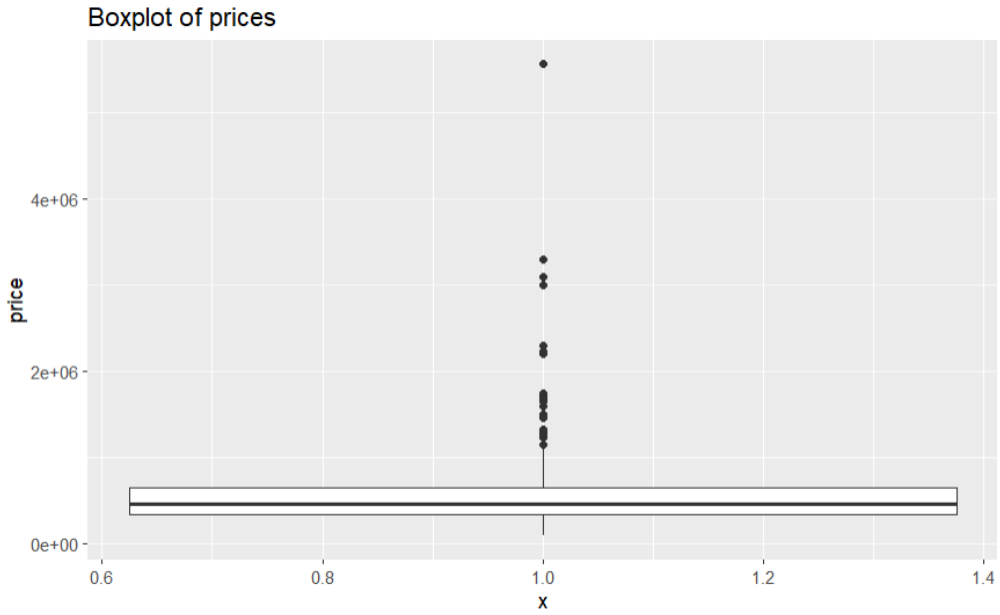
In order to construct our analysis, we need to put our data into graphs for visualization. Our goal is to see whether or not we can predict the sale price based on only one square footage of the home. The graphs would allow us to get a better sense of whether or not there really is a way to predict price based on size. We will be able to see the mean/median, the spread, minimum and maximum, quantile ranges, and the variance and standard deviation to help better understand our analysis.

We computed the sample mean. The average sale price of houses is approximately $566,594. With a standard deviation of $434,558. This shows a wide range of house prices. The prices range from a minimum of $90,000 to a maximum of $5,570,000. The average square footage of the living space is approximately 2,115 sq ft. The standard deviation is 977 sq ft. The living space ranges from 390 soft to 9200 sq ft.

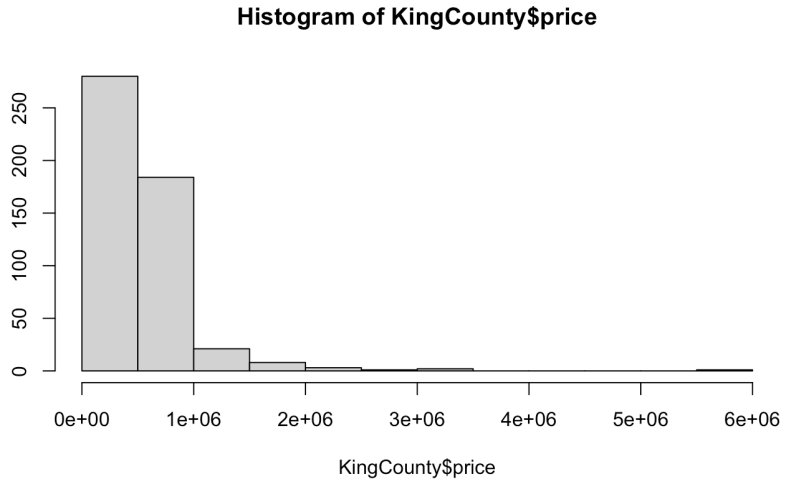
The Mean squared error (MSE) of the data is approximately $81,327,002,574. The estimated slope of the model is approximately 324.7759. This suggests that for every additional square foot of living space. The house price increased by about 324.7759. The model intercept is about -120325.5225 and it doesn't have a practical meaning since the price can't be negative. The total sum of squares(SSTO) is about 94.23 trillion. The SSR is about 50.24 trillion. This number is a large portion of the total variation in house prices which indicates that the square footage of living space is a meaningful predictor of house price in King County.



We can clearly see that the majority of the points are grouped up in one area and are only slightly disbursed out. There are also very clear outliers. The graph suggests that there is a positive linear relationship between the square footage of the houses and the corresponding price. From the plot, it appears that square footage does have an impact on price. However, prices associated with square footage of living space greater than 4000 ft^2 appear to be spread out more relative to values corresponding to square footage of living space less than 4000 ft^2.



According to the box plot, we can see that price has many outliers, giving us the sense that a prediction of price would be unlikely and unreliable. The outliers affect the rest of the plot such as the mean, the range, and especially the variance. The mean and especially the median has a greater impact from the outliers making it unreliable to get a good sense of the center from the boxplot. Though the range in price is very large, we can see that most of the data lies around below the one million price range.



The box plot above shows that the large majority of prices are below one million dollars. It also suggests that if we were to make a prediction, it would be within the range of the first two boxes though we should be aware that there are outliers that may have an effect on the prices.

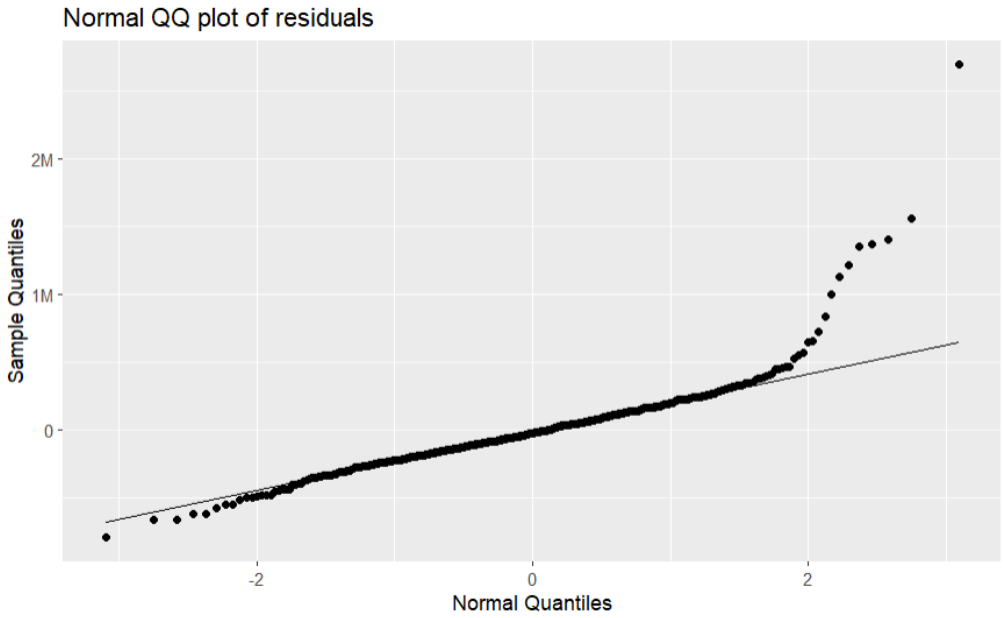
Ⅲ. Diagnostics

Before fitting a simple linear model, we have to examine the requirements and check if the data meets those requirements. A requirement is that the residual terms must follow a normal distribution. This means that for any X_i the residuals about the estimated regression line are allowed to vary in accordance to their distribution. The last requirement is that the residuals vary about the estimated regression line equally for all X_i.

Ⅲ.1 Normal Residuals with QQ plot

The first requirement that we will take a look at for fitting a mean simple linear regression line is that the errors follow a normal distribution. Below is a Normal

Quantile-Quantile plot (Normal QQ plot). The Normal QQ plot graphs the quantile of a distribution on the x axis and quantiles for another distribution of the y axis. If the x and y distributions are the same, then the plot should have a linear relationship. In this case, the normal distribution quantiles are plotted on the x axis and the sample quantiles are plotted on the y axis. If our residuals follow a normal distribution, then the QQ plot should show a linear relationship. It is below



From the graph above, we can see that the sample quantiles roughly follow a linear relationship with the normal quantiles indicating that there may be a normal distribution of our residuals. However, for extreme positive values of the normal distribution, we can see that the values of the sample quantiles tend to tail upwards, providing doubt if our residuals follow the normal distribution.

Ⅲ.2 Normal Residuals with Shapiro-Wiks

To test whether our residuals follow the normal distribution, we employ the Shapiro-Wiks test. The Shapiro-Wiks test is for determining if a sample is normally distributed. In this paper,

we will use the Shapiro-Wiks test for the normality of the residuals of our sample. The hypothesis of this test is as follows. We proceed with controlling for type 1 error at 0.10.

$$H_o: \textit{The errors are normally distributed}$$

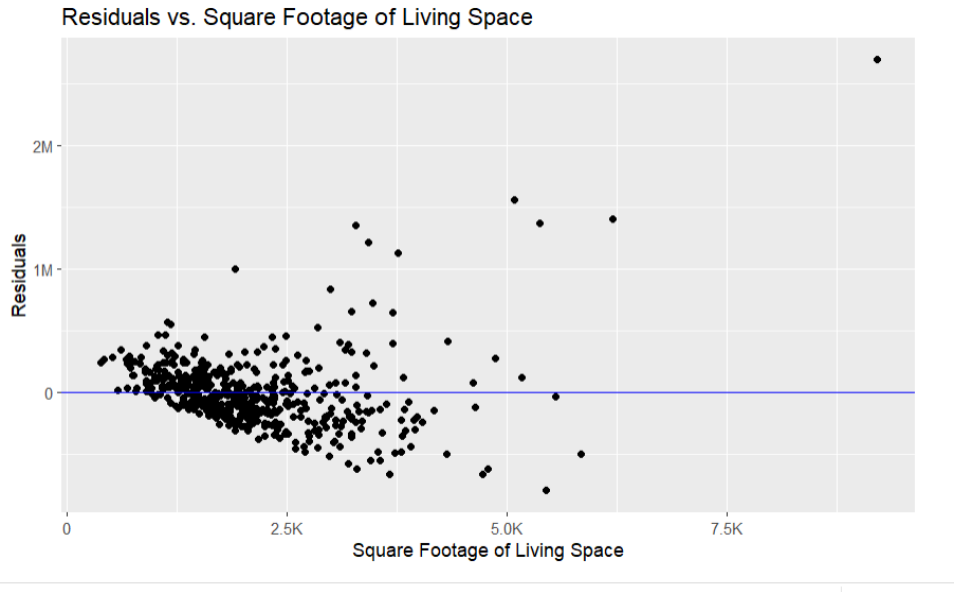
$$H_A: \textit{The errors are not normally distributed}$$

After running the test, we achieved a p-value of $2.2 * 10^{-16}$. Consequently, we conclude that there is strong evidence that the errors are not normally distributed.

Ⅲ.3 Constant variance for residuals with residual plot

We start by showing the variance of the residuals across all X_i. Below is a plot of the residuals vs square footage of living space. The graph below plots the residuals about the fitted line for all X_i. The residuals, e_i is

$$e_i = Y_i - \hat{Y}_i$$



It appears that the residuals have increasing variance as square footage of living space increases. For instance, residuals which are from square footage of living space below 2.5k do not vary much about the estimated regression line. However, living space with above 2.5k appears to have larger residuals. As such, we run a Fligner-Killeen test to determine if the residuals are equal.

Ⅲ.4 Normal Residuals with Fligner-Killeen test

We run the following hypothesis' while controlling for type 1 error at 0.10 as follows

$$H_0: There\ are\ equal\ variances\ for\ the\ upper\ and\ lower\ groups$$
$$H_A: There\ is\ not\ equal\ variances\ for\ the\ upper\ and\ lower\ groups$$

The lower group contains all the values equal or below the median fitted value while the upper group contains all the values above the median fitted value. After running the test, we achieved a p-value of $2.044 * 10^{-6}$. Since the p-value < alpha, we reject the null hypothesis and conclude that there is strong evidence which suggests that there are not equal variances for the upper and lower groups.

From our diagnostic tests, we ran the Shapiro-Wiks test which tests for normality of the estimated errors. From our testing, we found that there was strong evidence to conclude that the estimated errors do not follow the normal distribution with a p-value of $2.2 * 10^{-16}$. We also ran a Fligner-Killeen test which tests for equal variances of residuals. However, we found that there was strong evidence that the residuals do not have constant variance for all values of square footage of living space. As a result, the use of simple linear regression is not justified. Regardless, we will proceed anyway.

Ⅳ. Analysis

In this section we will fit the simple linear model. The simple linear model is below:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Where Y_i represents the true price of the house with X_i square footage of living space. β_1 and β_0 are parameters of the model which represent the amount Y_i changes for every change in one unit of X_i and the value of Y_i when the square footage of living space is zero, respectively. However, in this case, β_0 does not have useful interpretation since a house cannot have zero square footage of living space. ϵ_i represents the random error about the line $\beta_0 + \beta_1 X_i$ that may occur. Additionally, ϵ_i follows a normal distribution with mean 0, and variance σ_ϵ^2. As X_i is treated as a fixed value and β_1 and β_0 are parameters of the model, it follows that Y_i follows the a normal distribution of mean $\beta_0 + \beta_1 X_i$ and variance σ_ϵ^2. Moreover, the estimated regression line is

$$\hat{Y}_i = b_o + b_1 X_i$$

Where b_o and b_1 are estimates of the parameters β_0 and β_1 respectively and the residual, e_i, estimates ϵ_i. Consequently, $\hat{Y}_i$ is an estimate of Y_i. We calculated b_1 and b_o using this equations respectively:

$$b_1 = \frac{\sum X_i Y_i - n\bar{Y}\bar{X}}{\sum X_i^2 - n\bar{X}^2}$$

$$b_o = \bar{Y} - b_1 \bar{X}$$

Getting values of $b_1 = 324.7759$ and $b_o = -120325.5225$. Consequently, we estimate that for every extra square footage of living space would result in a 324.7759 increase in price, on average.

Constructing confidence intervals, we aim to estimate the true population parameters β_0 and β_1 with 90% confidence. The equations for confidence intervals β_0 and β_1 are:

$$(1-\alpha)100\%\ for\ \beta_0 : b_0 \pm t_{\alpha/2,n-2}\sqrt{MSE(1/n + \bar{X}^2/\sum(X_i - \bar{X})^2)}$$

$$(1-\alpha)100\%\ for\ \beta_1 : b_1 \pm t_{\alpha/2,n-2}\sqrt{MSE/\sum(X_i - \bar{X})^2}$$

The 90% confidence interval for the intercept (-172604.43 to -68046.62) and square footage living coefficient (302.33 to 347.22) guides us in understanding the likely range of true values based on our regression analysis. The intervals, not containing zero, imply a negative intercept and a positive slope for square footage living.

Hypothesis tests play a pivotal role in assessing the significance of predictors. The results indicate strong evidence against the null hypothesis that the square footage living coefficient is zero. Both the intercept and square footage living emerge as statistically significant predictors in our regression model, providing valuable insights into their roles in predicting house prices. We also use the hypothesis test result as an information of nulls and alternatives.

We employ t-tests to evaluate the significance of individual coefficients. The t-value, representing the number of standard deviations from the mean, and the p-value, indicating the probability of observing such extreme values, contribute crucial information. The null hypothesis is that the intercept is that beta1 is 0, while the alternative is that beta1 is not 0. The intercept's t-value of -3.793 and corresponding p-value of 0.000167 support the rejection of the null hypothesis, emphasizing a significant difference from zero. Similarly, testing for slope resulted in a high t-value of 23.847 and an extremely low p-value of roughly 2.08e-84 leading to the rejection of the null hypothesis, affirming its significant impact on house prices.

Ⅴ. Interpretation

Though we can definitely make a prediction on the price of the house based on the square footage of the house, the prediction would be deemed quite unreliable based on all the outliers that shift our data which resulted in failing to fulfill the simple linear model assumptions which we used. Based on the plots, mainly the box plot, the outliers are very apparent. Knowing that, we can infer that the mean is heavily affected due to the outliers.

There is strong evidence that the errors are not normally distributed. We achieved this conclusion by using the Shapiro-Wiks test. After performing the Shapiro-Wiks test with a control for type 1 error at 0.10, we achieved a p value of 2.2*10-16. With this, we have clear and strong evidence that the residuals are not normally distributed.

The Fligner-Killeen test tells us whether or not the variances of residuals for the upper and lower groups are equal. After performing this test with the null hypothesis stating they are equal variances and the alternative hypothesis stating that they are not equal, we reject the null hypothesis. We achieved a p-value of 2.044*10-6 which in comparison to our alpha, is less than. Due to this, we reject the fact that the variance of residuals are constant for the upper and lower groups.

Based on the Fligner-Killeen test and the Shapiro-Wiks test, we are given the information that residuals are not normally distributed and the residuals do not have constant variance for all values of square footage of living space. With this, the use of simple linear regression is not justified.

We used hypothesis testing for testing whether or not the slope and intercept are equal to zero. We will have two separate null and alternative hypotheses. The first null hypothesis being that the intercept is equal to zero and the alternative being that the intercept does not equal zero. The second null hypothesis is that the slope is equal to zero and the alternative being the slope does not equal to 0. We used t-tests to evaluate the significance of individual coefficients. We achieved the t-value of the intercept of -3.793 and a p-value of 0.000167 which does not support the null hypothesis for the intercept. This emphasizes that there is a significant difference from zero. We achieved the t-value of the square footage living variable of 23.847 and a p value of

2.08e-84 for the slope which leads to the rejection of the null hypothesis. In summary, both the intercept and the slope are not equal to zero.

VI. Prediction

The equation for the a confidence interval prediction of a single price from square footage of living space is:

$$(1-\alpha)100\%\ for\ Y^* : Y^* \pm t_{\alpha/2,n-2}\sqrt{MSE(1+1/n+(X-\bar{X})^2/\sum(X_i-\bar{X})^2)}$$

Where Y* represents a singular price of a home with X square feet of living space. The confidence interval leaves us with a wide range of possibilities for the prediction. From the prediction confidence interval, the three separate predictions we have range for different scenarios. The predicted price for a home with square footage of 2800 is $789,046.9. The corresponding prediction confidence interval for price if the square footage is: [298518.2, 1279575.6]. In other words, we are 90% confident that the price of the house with a square footage of 2800 sq ft is within the price range of $298,518.2 and $1,279,575.6.

The predicted price for a home with square footage of 3200 is $918,957.3. The corresponding prediction confidence interval for price if the square footage is: [428065.2, 1409849.3]. In other words, we are 90% confident that the price of the house with a square footage of 3200 sq ft is within the price range of $428,065.2 and $1,409,849.3.

The predicted price for a home with square footage of 8000 is $2,477,881. However, this estimate is based on our fitted regression line where most of the data does not have a square footage of 8000. Therefore, this prediction is quite dubious. We caution the reader to use this prediction based on our analysis and data. The corresponding prediction confidence interval for price if the square footage is: [1970115, 2985648]. In other words, we are 90% confident that the price of the house with a square footage of 8000 sq ft is within the price range of $1,970,115 and $2,985,648. Again, we reiterate that this particular prediction is quite dubious with respect to our current data.

Ⅶ. Conclusion

All findings of tests and predictions in this project have provided critical insights into the relationship between home prices and square footage. The Shapiro Wiks test told us that the residuals were not normally distributed. The Fligner-Killeen test told us that the variance of the residuals were not constant for all values of square footage of living space. The hypothesis test told us that both the slope and the intercept are not equal to zero. Through the Shapiro Wiks test, the Fligner-Killeen test, hypothesis testing, and our prediction confidence intervals, we see a decent relationship between price of a house and the square footage of the house. Also, our findings demonstrate that it is possible to predict sale prices based on the square footage within 90% confidence interval. However, the presence of outliers and the results of the Fligner-Killeen and Shapiro-Wilk tests indicate that the assumptions required for a simple linear regression model are not fully satisfied due to the non-normal distribution of residuals and heteroscedasticity. Which means, although the t-tests suggest that square footage is a statistically significant predictor of home prices, the underlying conditions for linear regression do not hold in our dataset. Therefore, our analysis concludes that while there is a potential for prediction within a certain range of square footage, the simple linear regression model may not be the most reliable method for such predictions.

Ⅷ. Appendix

```
library(scales)

library(MASS)

library(car)

kingcounty <- read.csv(file = "C:/Users/tyler/School/Learn R/STA 108/Project 1/KingCounty.csv")
library(tidyverse)

the.model = lm(price ~ sqft_living,data = kingcounty)

ggplot(kingcounty, aes(sample=the.model$residuals)) +
  stat_qq() +
  stat_qq_line() +
  labs(
    title = "Normal QQ plot of residuals",
    x = "Normal Quantiles",
    y = "Sample Quantiles"
  ) +
  scale_y_continuous(labels = scales::label_number_si()) # scaling


# Testing Normality (Shapiro-Wilks)
# Ho: The residuals are normally distributed. Ha: the residuals are not normally distributed.
options(scipen = default_locale()) # get scientific notation back
the.model = lm(price ~ sqft_living,data = kingcounty)
ei = the.model$residuals
the.SWtest = shapiro.test(ei)
the.SWtest


# visual for constant variance for residuals
kingcounty$yhat = the.model$fitted.values

options(scipen = 999) # removes scientific notation

qplot(sqft_living, ei, data = kingcounty) +  ggtitle("Residuals vs. Square Footage of Living Space") +
xlab("Square Footage of Living Space") +
  ylab("Residuals") + geom_hline(yintercept = 0,col = "blue") +
  scale_y_continuous(labels = scales::label_number_si()) +
  scale_x_continuous(labels = scales::label_number_si()) # scaling



options(scipen = default_locale()) # get scientific notation back
```

```
kingcounty <- read.csv(file = "C:/Users/tyler/School/Learn R/STA 108/Project 1/KingCounty.csv")
the.model = lm(price ~ sqft_living,data = kingcounty)
kingcounty$ei = the.model$residuals
kingcounty$yhat = the.model$fitted.values
ei = the.model$residuals
Group = rep("Lower",nrow(kingcounty))
Group[kingcounty$sqft_living > median(kingcounty$sqft_living)] = "Upper"
Group = as.factor(Group)
kingcounty$Group = Group
the.FKtest= fligner.test(kingcounty$ei, kingcounty$Group)
the.FKtest

price_1 = data.frame( sqft_living = 2800)
predict_1 = predict(the.model,price_1)
"prediction price of living square footage 2800:"

price_2 = data.frame( sqft_living = 3200)
predict_2 = predict(the.model,price_2)
"prediction price of living square footage 3200:"

price_3 = data.frame( sqft_living = 8000)
predict_3 = predict(the.model,price_3)
"prediction price of living square footage 8000:"

summary(the.model) #regression

plot(df$sqft_living,df$price, main = "Square footage living and Price",xlab = "square footage living",ylab =
"price",pch = 19)
abline(the.model, col='purple',lwd=2)

ggplot(df,aes(x=sqft_living)) +
  geom_histogram(aes(color=price, fill=price))

##plot a histogram to summary the model

ggplot(df, aes(x=sqft_living, y=price)) +
  geom_boxplot() +
  labs(x="square footage living", y="price") +
  theme_minimal()

#sample means
mean_price=mean(df$price)
mean_sqft=mean(df$sqft_living)
#Standard Deviations
se=summary(the.model)$sigma
#SSE
SSE=round(sum(the.model$residuals^2),4)
SSE

#SSR
```

```
the.table=anova(the.model)
SSR=the.table[1,2]
SSR

#SSTO
SSE=the.table[2,2]
SSTO=SSE+SSR
SSTO

#Analysis
##CI
CIs = confint(the.model,level = 0.9)
"90% CI: "

##hypothesis tests
all.sum = summary(the.model)
all.sum

HT = all.sum$coefficients
"hypothesis tests: "

##ANOVA table
the.table = anova(the.model)
"ANOVA table: "

price_1 = data.frame( sqft_living = 2800)
predict_1 = predict(the.model,price_1,interval = "prediction",se.fit = TRUE,level = 0.90 )
MSE = sum(the.model$residuals^2)/(length(the.model$residuals) -2 )
se.starY_1 = sqrt( MSE + predict_1$se.fit^2)
CI_1 = predict_1$fit[-1]
CI_1

price_2 = data.frame( sqft_living = 3200)
predict_2 = predict(the.model,price_2,interval = "prediction",se.fit = TRUE,level = 0.90 )
se.starY_2 = sqrt( MSE + predict_2$se.fit^2)
CI_2 = predict_2$fit[-1]
CI_2

price_3 = data.frame( sqft_living = 8000)
predict_3 = predict(the.model,price_3,interval = "prediction",se.fit = TRUE,level = 0.90 )
se.starY_3 = sqrt( MSE + predict_3$se.fit^2)
CI_2 = predict_3$fit[-1]
CI_2
```