Correlation and regression analysis
CS51 Spring 2020

This report investigates the house sale prices for King County from May 2014 to May 2015 based on the sample data obtained from a survey by Kaggle.

Dataset

I got this data from Kaggle survey that included variables like price, number of bathrooms, number of bedrooms and others like zip code, latitude, longitude and year renovated. The data can be found here.

The variables of interest in this analysis are house sale prices and size of the living-room in square foot, which are the response and predictor variables respectively. Both variables are quantitative, numeric and continuous data types. House sale price is continuous because it can range anywhere (real numbers) from zero to positive infinity, with the inclusion of decimals. Also, the size of a living room is continuous while it includes positive real numbers. Both are numeric because calculations such as mean, standard deviation and variance can be made on it. Furthermore, house sales price is an interval measurement because subtraction can be done and it would make sense. An example is that 34-35 dollars has the same distance as 18-19 dollars. The predictor (independent) variable is the living room size in square foot because it used in regression to predict the other variable and house sale price is the response(dependent) because it relies on factors such as living room size.[1]

Methods

I read the dataset into Python using Pandas package. First, I created a list to store the data. Then I deleted some inputs in the data that were not suitable for analysis [See Appendix A]. I selected the necessary inputs in both house sale prices and size of the living-room.

[1] #variable: I accurately identified, classified, and described the variables and parameters of the model; I also accurately defined and provided a detailed description of the relationship between the independent and dependent variables

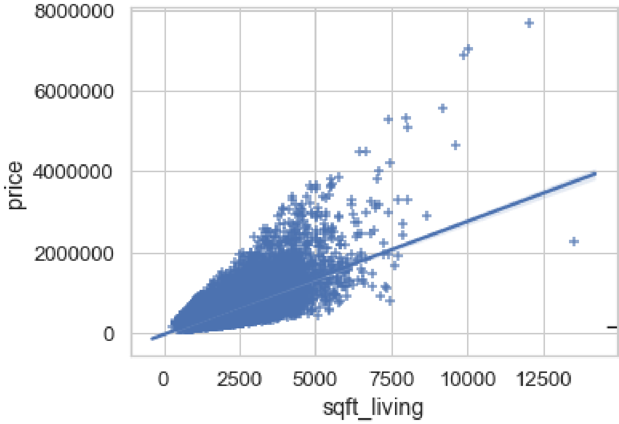


Figure 1. Correlation between price and size of the living room

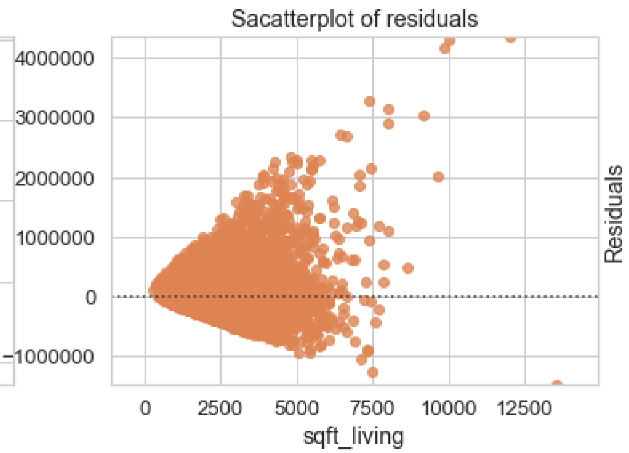
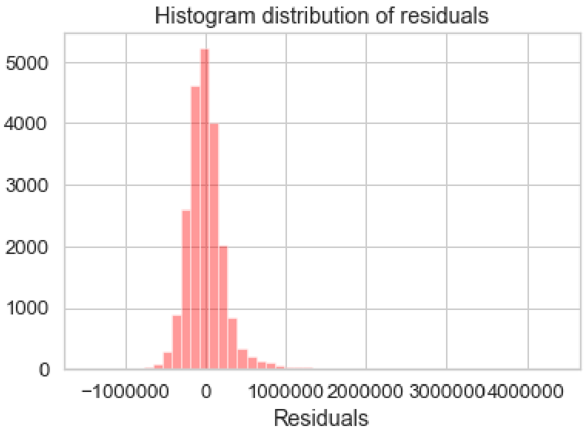
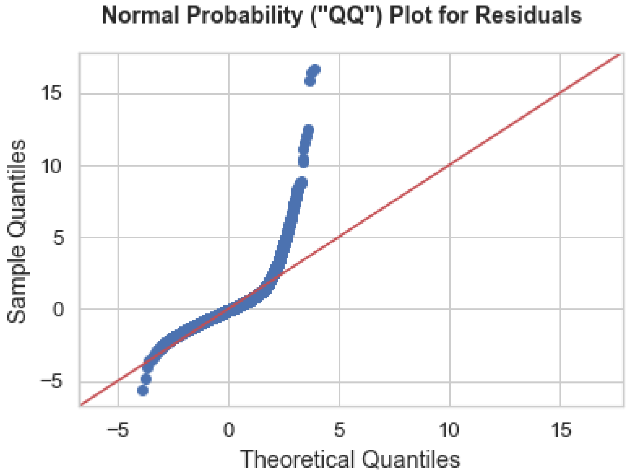


Figure 2. Residual plot between price and size of living room





[2]

Figure 3. Distribution of the residuals

Figure 4. Normal probability plot for residuals

In order to fit a least square line to show the correlation between the two variables, the following assumptions are required [Appendix B]:

[2] #dataviz: I effectively generated a detailed data visualization appropriate for the data; effectively analyzed and interpreted the data visualization with justification; I also critiqued the qq plot and clearly explained why it would be problematic for the model

Linearity: The above figure shows a linear trend. As the size of the living room in sqft increases, the price in dollar increases.

Independent observations: The dataset is not a time series.They are not sequential data obtained from a progress over a period of time such as the growth of a plant with time. Therefore they are independent. Also, they are less than ten percent of the entire population, which is the number of houses in the United states.

Nearly normal residuals: As seen in figure 3, the histogram that shows the distribution of the residuals is close to normal. The residual is the difference between the observed and the expected response of a predicted model.

Equal variance: The variability of points around the least squares line should remain roughly constant. The variability around the line of best fit in figure 1 is non-constant. The variability increases with larger values of x. This could be problematic for the predictive power of our model.

The normal distribution plot, gotten by comparing the normal distribution plot to the residual distribution, is clearly showing outliers. This can reduce the predictivity of the regression model.

The pearson's correlation coefficient is a measure of the linear association between the size of the living room and the house sale price. In the regression model, the pearson's correlation coefficient is 0.7 (computed with formula in Appendix C). This shows a high positive association between the size of the living room and the house sale price. A high association implies that the line of best fit almost perfectly describes the relationship between the predictor and response variables. Furthermore, it shows that as the size of the living room increases, the house sale price increases.[3]

The coefficient of determination is the proportion of the variance in the dependence that is explained by the independent variable. It is also how closely the data cluster around the line of best fit. The coefficient of determination is also called the R-squared value. The general formula is

$$\text{R-squared} = 1\text{- SSE/SSTO}$$

The R-squared can be gotten from squaring the pearson coefficient(r) value because our model is a simple linear least square regression with an intercept term. The R-squared value is 0.493 (computed in Appendix B). This interpretes as 49.3% of the variance in the house sale

[3] #correlation: I accurately computed and interpreted the correlation coefficient while also providing an explanation of the context; I recognized and effectively explained the difference between correlation and causation at the end of the paper, just as I identified extraneous variables

price is explained by the variance in the size of the living room. The adjusted R-squared is 0.493 (also computed in Appendix B). This is the same as the R-squared value because there are not multiple independent variables.

Regression equation: price = 280.624 * sqft_living + -43580.743

The general equation for a line of regression is price = b1 * sq_living + b0

Where b0 is the y intercept estimate and b1 is the slope estimate

The slope describes the estimated difference in the response(y) variable when the predictor(x) variable is one unit larger. A slope of 280.624 signifies that for every squarefoot increase in the size of the living room there is a 280.624 dollar increase in the house sale price. The y-intercept value of -43580.743 shows that when the size of the living room is 0sqft, the house sale price will be -43580.743 dollars. However, this is a problem of extrapolation because of the impossibility of having a negative dollar value. Extrapolation is the application of a model estimate out of the realm of original data.[4]

A statistical significant test will use the dataset to summarize evidence about a hypothesis by comparing sample estimate by the value predicted by the hypothesis. We will find out whether the slope is significantly higher or lower than zero. We significant level is set to $\alpha = 0.05$ and a two-tailed test is carried out. The null and alternative hypotheses respectively are:[5]

- Ho: there is no linear relationship between *living room size* and *price* in the population; slope, $\beta1 = 0$.
- HA: there is some linear relationship between *living room size* and *price* in the population; $\beta1 \neq 0$.

Since the population standard deviation is estimated from the sample standard error, a t-distribution would be used for inference. This gives a better and more reliable estimate than the normal distribution because the population standard deviation is not given. Also, a t-distribution has fatter tails than a normal distribution which will correct the errors from approximating the standard deviation. To validate the use of T-distribution, the observation should be independent by obtaining from less than 10% of the population and nearly normal. This is true for our data

[4] #regression: I accurately interpreted a regression model with a well-justified explanation of the relation between dependent and independent variables; I also computed and interpreted the regression equation with clear and detailed explanation then I interpreted the coefficient of determination with clear and detailed explanation

[5] #significance: I accurately applied and interpreted the statistical significance test with well justified reasoning, while defining the null and alternative; I calculated and interpreted the results of the significance test in a subtle scenario.

because we take the King county as a sample and the United States as a population, which is less than 10%. The populated standard deviation is estimated from the sample standard deviation. The **LINE** acronym (**L**inearity, **I**ndependent observations, **N**ormality of residuals, and **E**qual variance) must also be fulfilled. These have been discussed earlier.

$$\text{T score} = b1-0 \, / \, SE(b1)$$

The T-score of 144.920 resulted in a two-tailed test p-value of 0.000 [Appendix E]. Since 0.000 < alpha value of 0.05, we reject the null hypothesis and consider the alternative. In this regression model, we are in favor of a linear relationship between the living room size and price.

The model can be used to predict a house's sale price based on the size of its living room. The conclusions are inductive because the regression model for the house sale price was generalized from a sample of data from the population. Also, the significance of the test give some evidence to support the hypothesis but they do not give 100% support. However, as inductive reasoning in hypothesis testing is usually plagued by biases such as confirmation bias. Confirmation bias is the tendency to accept rather than deny a current hypothesis based on existing beliefs. This was eliminated by setting the alpha level to 0.05 before getting the p-value and the number of tails was determined before viewing the data, to avoid bias. Also, rather than being surprised by our result, I considered the alternative[6]. This makes the inductive reasoning strong. It is also strong because a sample size in King County of 504 made it possible to generalize to the whole USA. It is sound because there is always a positive correlation between the size of the living room and the price of the house in the United States (Pinsker, 2019).[7]

Correlation does not necessitate a causation. It is fallacious (post hoc ergo procto hoc) to believe that the correlation between two variables equate causation. Therefore, appropriate evidence based on research should confirm before concluding that it is causation.

A possible extraneous variable is the house's location. Some places might be more expensive than others irrespective of its size.

[6] #biasmitigation: I explained the relationship and underlying mechanisms for how implemented strategies avoid psychological(confirmation) bias

[7] #induction: I effectively analyzed and applied inductive reasoning as well as providing a clear, detailed explanation of the generalization; I evaluated the strength of the induction and provided justification.

Reference

[8]Sabo, M. (2016). House sales. *Kaggle*. Retrieved 3 October 2017.

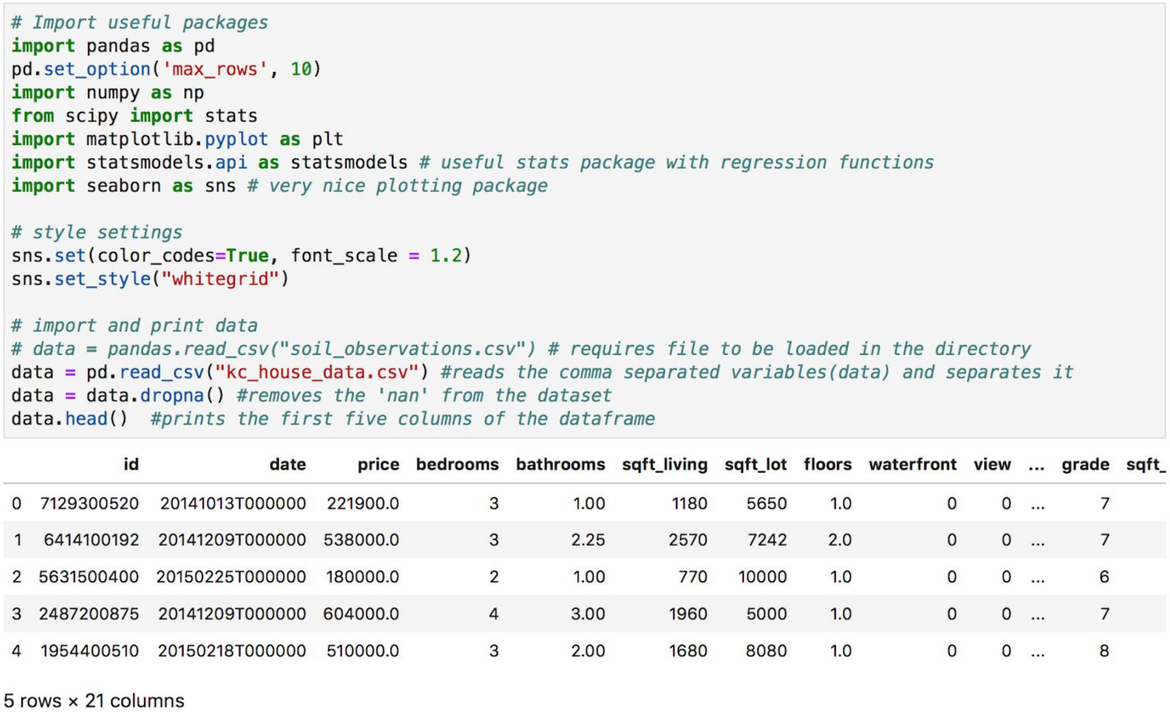
Retrieved from https://www.kaggle.com/harlfoxem/housesalesprediction

Pinsker, J. (2019, September 12). Why Are American Homes So Big? Retrieved from

https://www.theatlantic.com/family/archive/2019/09/american-houses-big/597811/

Appendix

Appendix A: Importing and data analysis



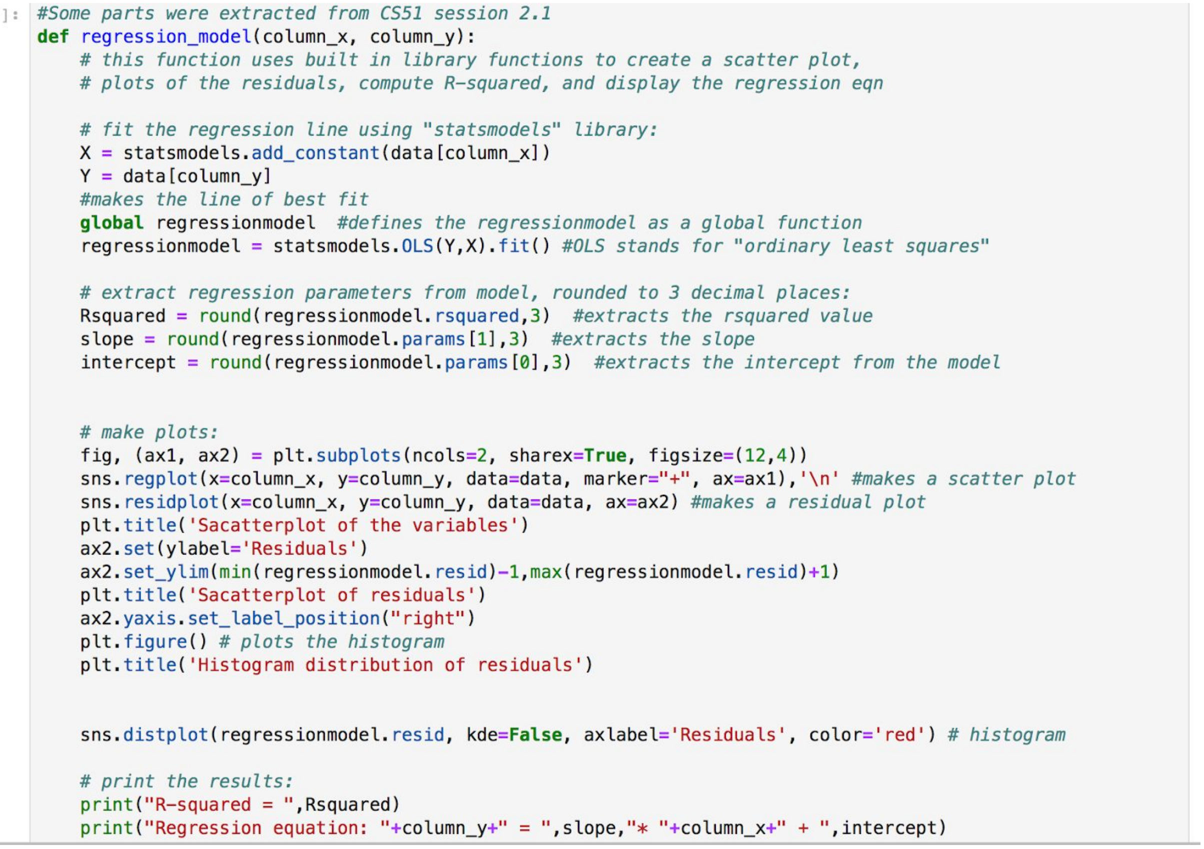
```
# Import useful packages
import pandas as pd
pd.set_option('max_rows', 10)
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt
import statsmodels.api as statsmodels # useful stats package with regression functions
import seaborn as sns # very nice plotting package

# style settings
sns.set(color_codes=True, font_scale = 1.2)
sns.set_style("whitegrid")

# import and print data
# data = pandas.read_csv("soil_observations.csv") # requires file to be loaded in the directory
data = pd.read_csv("kc_house_data.csv") #reads the comma separated variables(data) and separates it
data = data.dropna() #removes the 'nan' from the dataset
data.head()  #prints the first five columns of the dataframe
```

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	grade	sqft_
0	7129300520	20141013T000000	221900.0	3	1.00	1180	5650	1.0	0	0	...	7	
1	6414100192	20141209T000000	538000.0	3	2.25	2570	7242	2.0	0	0	...	7	
2	5631500400	20150225T000000	180000.0	2	1.00	770	10000	1.0	0	0	...	6	
3	2487200875	20141209T000000	604000.0	4	3.00	1960	5000	1.0	0	0	...	7	
4	1954400510	20150218T000000	510000.0	3	2.00	1680	8080	1.0	0	0	...	8	

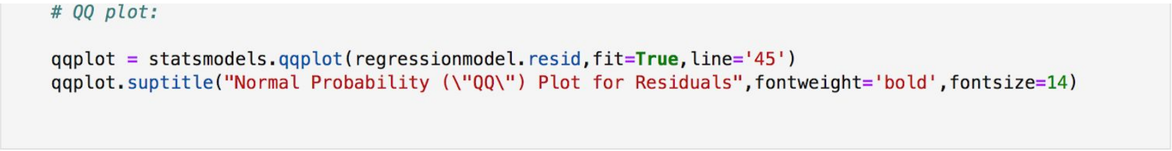
5 rows × 21 columns

[8] #professionalism: I demonstrated a deep grasp of how to present work products in a professional manner by following nuanced conventions for the audience, context, and discipline. I also proofread the assignment with Grammarly and followed the APA guidelines

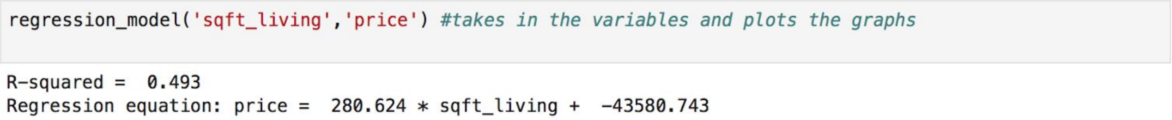
Appendix B: Data visualization



```
#Some parts were extracted from CS51 session 2.1
def regression_model(column_x, column_y):
    # this function uses built in library functions to create a scatter plot,
    # plots of the residuals, compute R-squared, and display the regression eqn

    # fit the regression line using "statsmodels" library:
    X = statsmodels.add_constant(data[column_x])
    Y = data[column_y]
    #makes the line of best fit
    global regressionmodel  #defines the regressionmodel as a global function
    regressionmodel = statsmodels.OLS(Y,X).fit() #OLS stands for "ordinary least squares"

    # extract regression parameters from model, rounded to 3 decimal places:
    Rsquared = round(regressionmodel.rsquared,3)  #extracts the rsquared value
    slope = round(regressionmodel.params[1],3)  #extracts the slope
    intercept = round(regressionmodel.params[0],3)  #extracts the intercept from the model


    # make plots:
    fig, (ax1, ax2) = plt.subplots(ncols=2, sharex=True, figsize=(12,4))
    sns.regplot(x=column_x, y=column_y, data=data, marker="+", ax=ax1),'\n' #makes a scatter plot
    sns.residplot(x=column_x, y=column_y, data=data, ax=ax2) #makes a residual plot
    plt.title('Sacatterplot of the variables')
    ax2.set(ylabel='Residuals')
    ax2.set_ylim(min(regressionmodel.resid)-1,max(regressionmodel.resid)+1)
    plt.title('Sacatterplot of residuals')
    ax2.yaxis.set_label_position("right")
    plt.figure() # plots the histogram
    plt.title('Histogram distribution of residuals')


    sns.distplot(regressionmodel.resid, kde=False, axlabel='Residuals', color='red') # histogram

    # print the results:
    print("R-squared = ",Rsquared)
    print("Regression equation: "+column_y+" = ",slope,"* "+column_x+" + ",intercept)
```

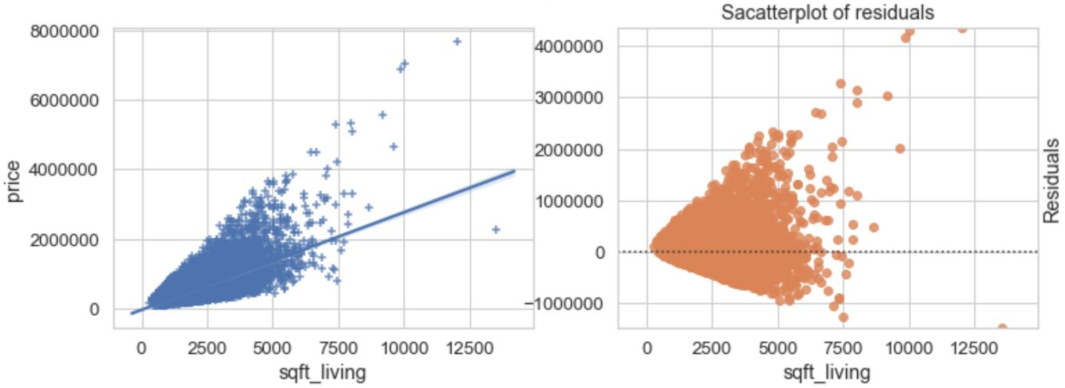


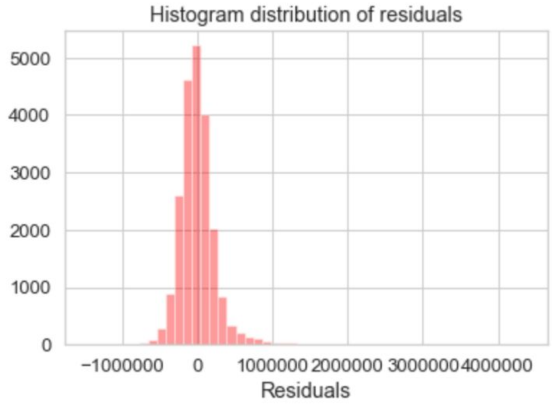
```
    # QQ plot:

    qqplot = statsmodels.qqplot(regressionmodel.resid,fit=True,line='45')
    qqplot.suptitle("Normal Probability (\"QQ\") Plot for Residuals",fontweight='bold',fontsize=14)
```

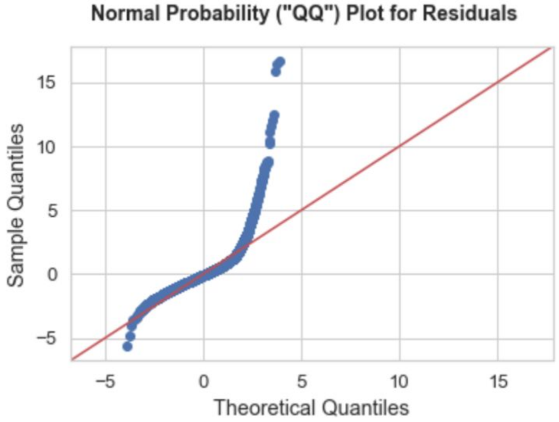


```
regression_model('sqft_living','price') #takes in the variables and plots the graphs
```

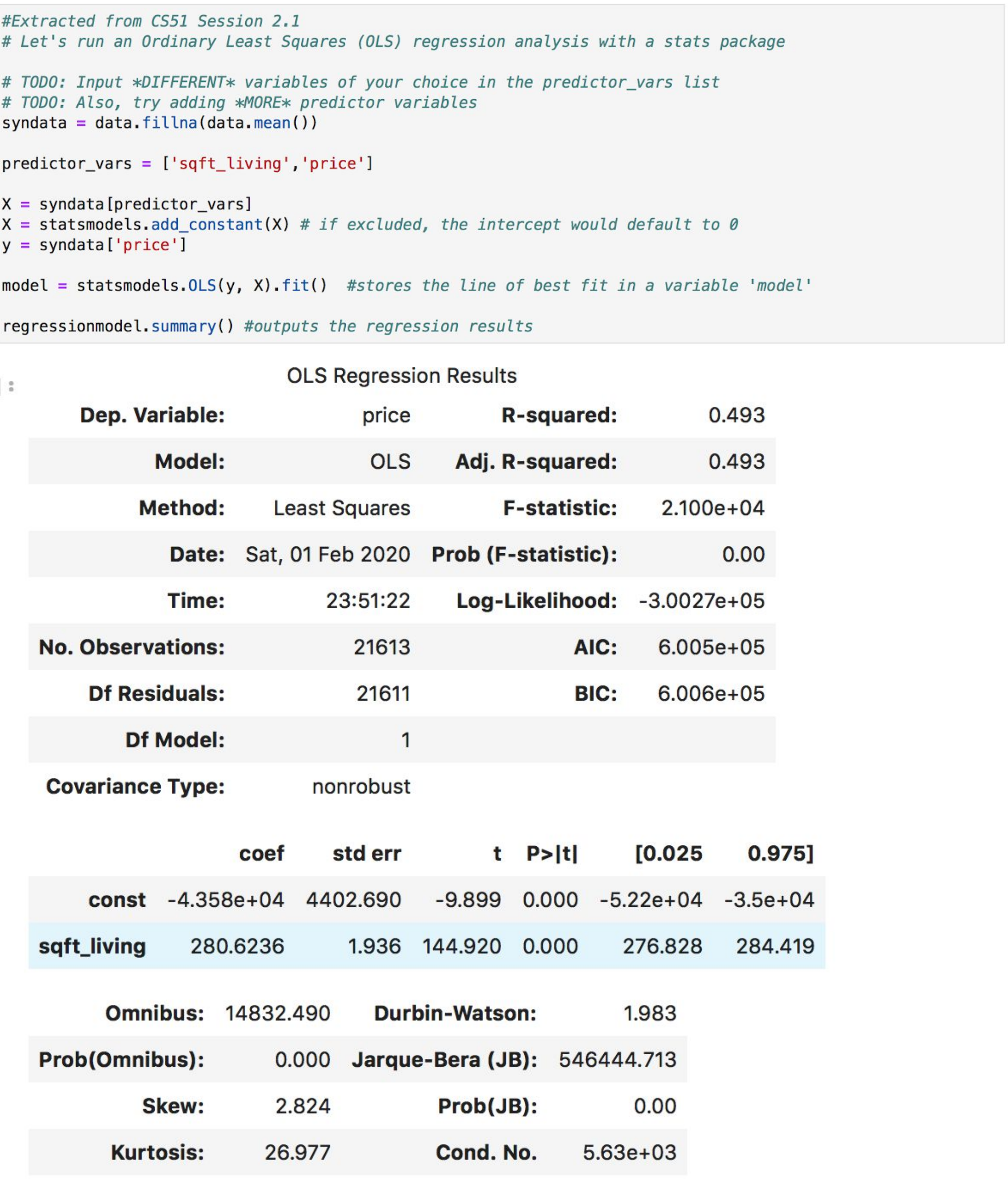
```
R-squared =  0.493
Regression equation: price =  280.624 * sqft_living +  -43580.743
```







Appendix C: Regression model summary



```
#Extracted from CS51 Session 2.1
# Let's run an Ordinary Least Squares (OLS) regression analysis with a stats package

# TODO: Input *DIFFERENT* variables of your choice in the predictor_vars list
# TODO: Also, try adding *MORE* predictor variables
syndata = data.fillna(data.mean())

predictor_vars = ['sqft_living','price']

X = syndata[predictor_vars]
X = statsmodels.add_constant(X) # if excluded, the intercept would default to 0
y = syndata['price']

model = statsmodels.OLS(y, X).fit()  #stores the line of best fit in a variable 'model'

regressionmodel.summary() #outputs the regression results
```

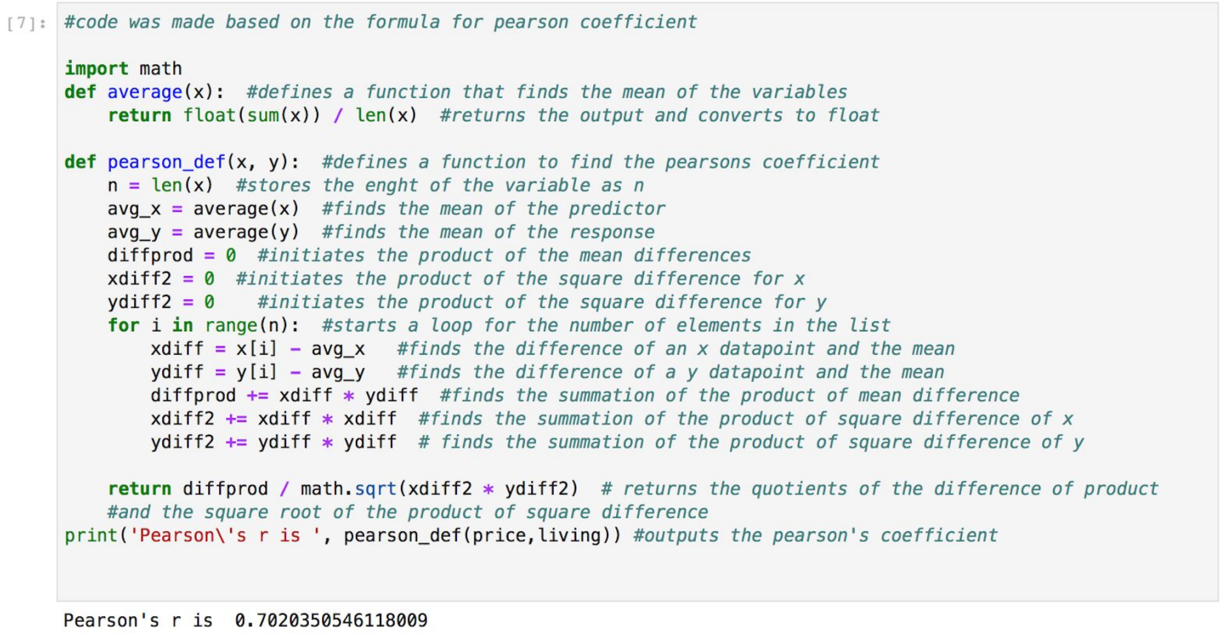
OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.493
Model:	OLS	**Adj. R-squared:**	0.493
Method:	Least Squares	**F-statistic:**	2.100e+04
Date:	Sat, 01 Feb 2020	**Prob (F-statistic):**	0.00
Time:	23:51:22	**Log-Likelihood:**	-3.0027e+05
No. Observations:	21613	**AIC:**	6.005e+05
Df Residuals:	21611	**BIC:**	6.006e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-4.358e+04	4402.690	-9.899	0.000	-5.22e+04	-3.5e+04
sqft_living	280.6236	1.936	144.920	0.000	276.828	284.419

Omnibus:	14832.490	**Durbin-Watson:**	1.983
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	546444.713
Skew:	2.824	**Prob(JB):**	0.00
Kurtosis:	26.977	**Cond. No.**	5.63e+03

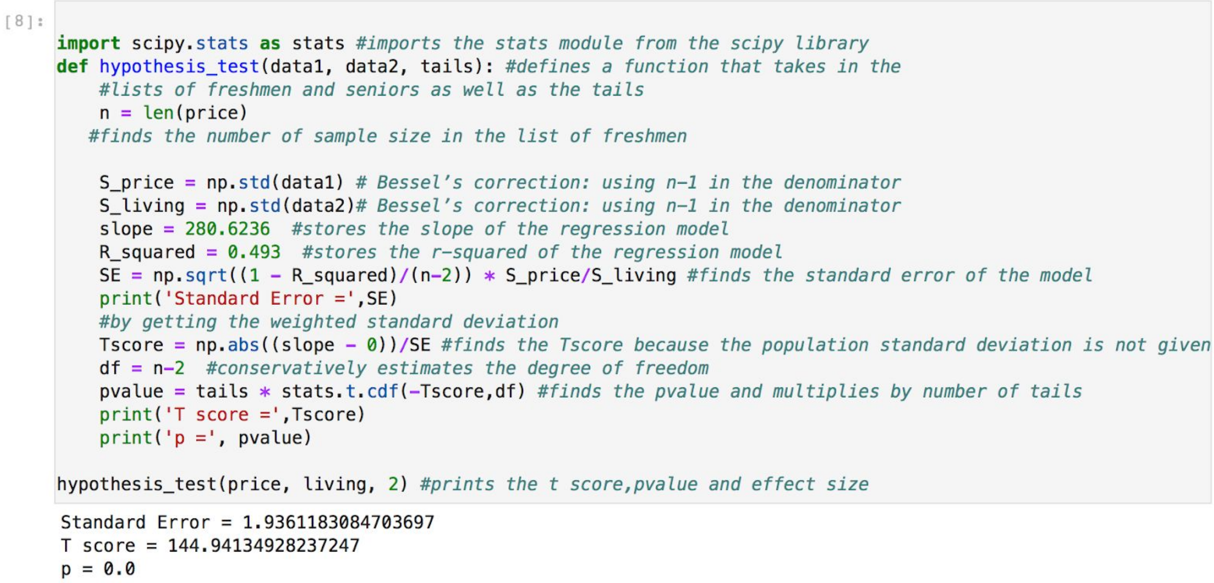
Appendix D: Pearson’s correlation coefficient



```
#code was made based on the formula for pearson coefficient

import math
def average(x):  #defines a function that finds the mean of the variables
    return float(sum(x)) / len(x)  #returns the output and converts to float

def pearson_def(x, y):  #defines a function to find the pearsons coefficient
    n = len(x)  #stores the enght of the variable as n
    avg_x = average(x)  #finds the mean of the predictor
    avg_y = average(y)  #finds the mean of the response
    diffprod = 0  #initiates the product of the mean differences
    xdiff2 = 0  #initiates the product of the square difference for x
    ydiff2 = 0    #initiates the product of the square difference for y
    for i in range(n):  #starts a loop for the number of elements in the list
        xdiff = x[i] - avg_x   #finds the difference of an x datapoint and the mean
        ydiff = y[i] - avg_y   #finds the difference of a y datapoint and the mean
        diffprod += xdiff * ydiff  #finds the summation of the product of mean difference
        xdiff2 += xdiff * xdiff  #finds the summation of the product of square difference of x
        ydiff2 += ydiff * ydiff  # finds the summation of the product of square difference of y

    return diffprod / math.sqrt(xdiff2 * ydiff2)  # returns the quotients of the difference of product
    #and the square root of the product of square difference
print('Pearson\'s r is ', pearson_def(price,living)) #outputs the pearson's coefficient
```

```
Pearson's r is  0.7020350546118009
```

Appendix E: Significance testing



[8]:

```
import scipy.stats as stats #imports the stats module from the scipy library
def hypothesis_test(data1, data2, tails): #defines a function that takes in the
    #lists of freshmen and seniors as well as the tails
    n = len(price)
   #finds the number of sample size in the list of freshmen

    S_price = np.std(data1) # Bessel’s correction: using n−1 in the denominator
    S_living = np.std(data2)# Bessel’s correction: using n−1 in the denominator
    slope = 280.6236  #stores the slope of the regression model
    R_squared = 0.493  #stores the r-squared of the regression model
    SE = np.sqrt((1 - R_squared)/(n-2)) * S_price/S_living #finds the standard error of the model
    print('Standard Error =',SE)
    #by getting the weighted standard deviation
    Tscore = np.abs((slope - 0))/SE #finds the Tscore because the population standard deviation is not given
    df = n-2  #conservatively estimates the degree of freedom
    pvalue = tails * stats.t.cdf(-Tscore,df) #finds the pvalue and multiplies by number of tails
    print('T score =',Tscore)
    print('p =', pvalue)

hypothesis_test(price, living, 2) #prints the t score,pvalue and effect size
```

```
Standard Error = 1.9361183084703697
T score = 144.94134928237247
p = 0.0
```