CORRELATION & REGRESSION:

ASSIGNMENT REPORT

CS51: Formal Analysis
Minerva Schools at KGI

Khanh Nguyen
February 2019

CORRELATION & REGRESSION

i. Introduction

This report is built based on a dataset about energy's consumption of a low-energy building. The research question of interest is: "How much can the variation in energy use of appliances in the building be explained by the energy consumption of the light fixtures?"

To answer the research question, I use simple regression to explore the relationship between two variables: the energy use of appliances and energy use of the light fixtures. The regression method employs the use of Pearson's correlation coefficient, the coefficient of determination, the regression model and the evaluation of statistical significance of the slope of the regression equation.

ii. Dataset

This dataset is obtained from Kaggle database with information centered around energy consumption and temperature in different areas of the low-energy building. These areas can be the kitchen, living room, or office room. For the purpose of this research, I particularly examine the energy use of appliances and energy use of light fixtures.

The energy use of appliances and energy use of light fixtures are in watt-hour (Wh) unit. They are numerical continuous variables because both of them can take numbers as value and these numbers are not necessary to be discrete (whole numbers). Even though these are theoretical continuous variables, the values in the dataset are rounded to the nearest 10. An assumption can be made is that the measurement techniques were not precise enough to obtain more digits. Based on the research question, the energy consumption of light fixtures is the predictor (independent) variable. The energy use of appliances is the response (dependent) variable[1].

With a large dataset of almost 20,000 data points for each variable, I use a Python function to clean null data points.

iii. Methods

The software I used to analyze the dataset are Excel and Python. Excel is used for importing data to Python, and Python is used to perform the mathematical operations and data visualization. In addition, for Python, I used library such as pandas to import the data, numpy to calculate relevant descriptive statistics, matplotlib to generate visual graph, and statsmodels to calculate the regression line.

[1] #variables: This is my description of the predictor and response variables.

a) Preliminary analysis

Here is an overview description of the statistics I'm exploring[2] (table 1). Full calculation can be found in appendix B.

	Energy use of light fixtures	Energy use of appliances
Count	n = 19735	n = 19735
Mean	$\bar{x}_{light}$ = 3.802	$\bar{x}_{appliances}$ = 97.695
Median	0	60
Mode	0	50
Standard deviation	s_{light} = 7.936	$s_{appliances}$ = 102.522
Range	70	1070

Table 1: Overview description of energy consumption of light fixtures and of total appliances of the building

The energy use of light fixtures has a large sample size (n = 19735) with a right-skewed distribution. While the median and more are both 0, the mean is slightly shifted to the right ($\bar{x}_{light}$ = 3.802), and the range of this data is high[3][4] (range = 70). There are a few super large energy users (outliers) that are heavily skewing this distribution (figure 1 - can be found in appendix C).

[2] #descriptivestats: This is the relevant description of the statistics of 2 variables I'm exploring: the sample size, mean, median, mode, standard deviation and range of the dataset.
[3] #distributions: This is the interpretation from the histogram of figure 1.
[4] #dataviz: This is the histogram of distribution of energy uses by the light fixtures.

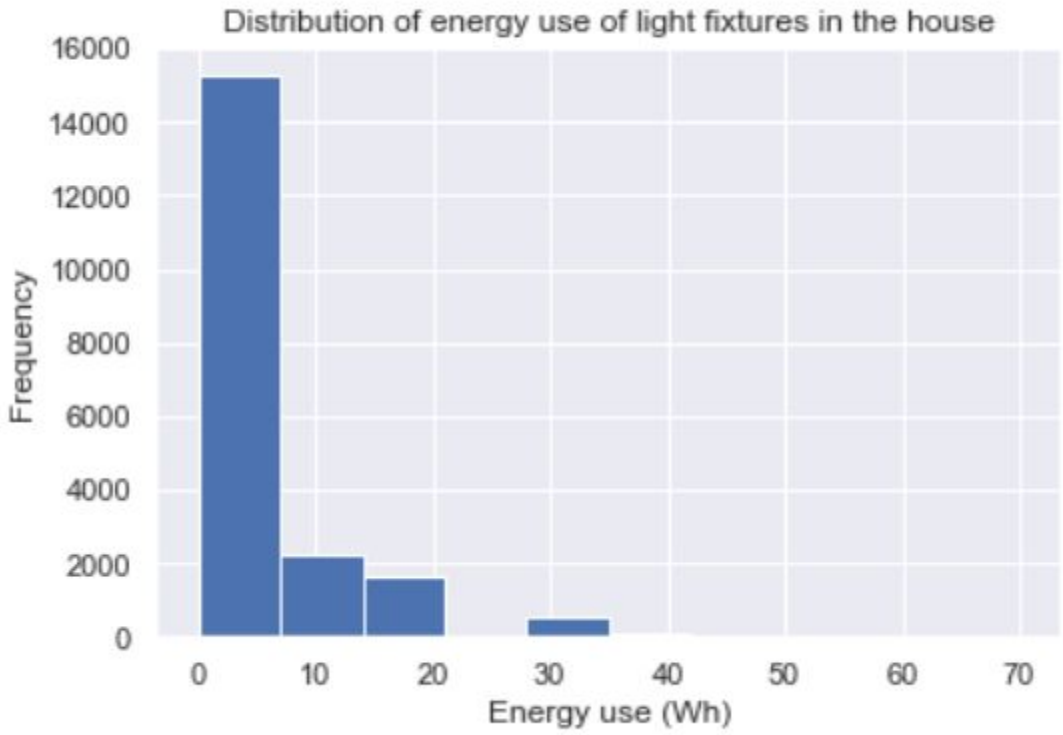


Figure 1: Distribution of energy use by the light fixtures

The distribution of the energy use of total appliances is similar to the energy use of light fixtures above. The distribution is right-skewed with large sample size (n = 19735). It has the median and the mode of 60 and 50, respectively. With the long right tail, it has big range (range = 1070) and standard deviation[5][6] ($s_{appliances}$ = 102.522) (figure 2 - can be found in appendix C).

[5] #distribution: This is the description of the histogram of figure 2.
[6] #dataviz: This is the histogram of distribution of energy uses by the appliances.

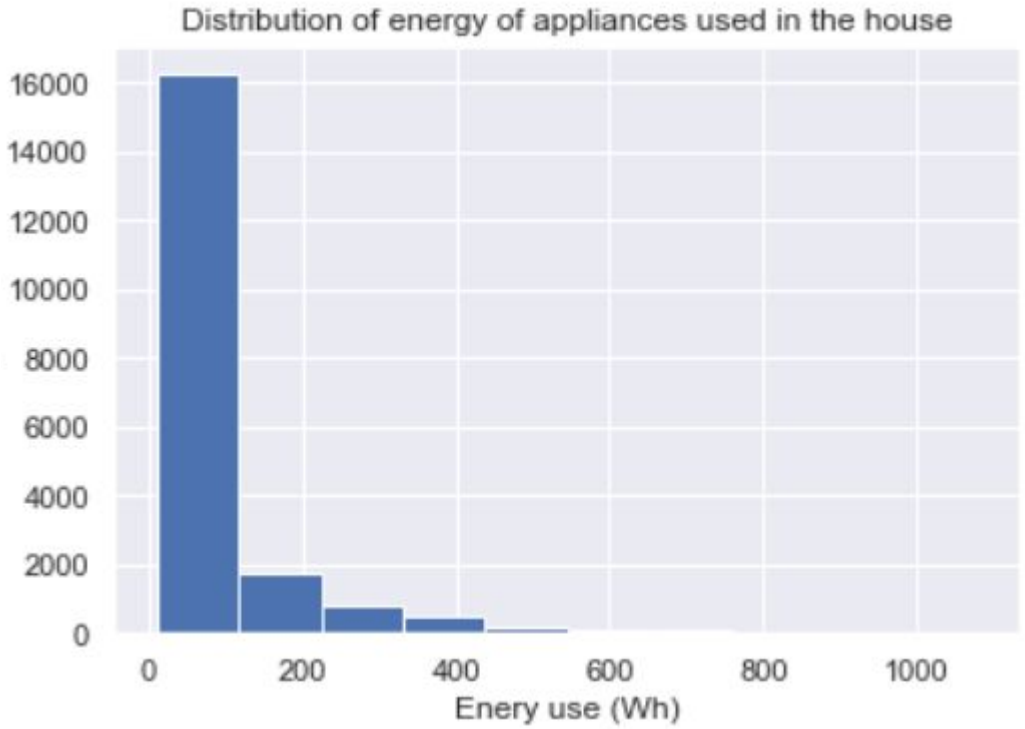


Figure 2: Distribution of energy use by the appliances in total

To estimate the degree of interdependence of the values of these two variables, I use the Pearson's correlation coefficient (r). It is calculated according to this formula:

$$r = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \, \sigma_y}$$

The formula primarily concerns with the mean and standard deviation of the data points. In other words, the formula is obtained by firstly taking sum of the product of how many standard deviation points that each data point is away from the mean, then secondly divide them by the total number of data points.

Detailed calculation of this Pearson's r can be found in appendix D.

b) Full-scale analysis

To further explain the variation of energy use by the appliances based on the energy consumption of light fixtures, I employ the simple regression statistical method. With this method, I compute the coefficient of determination (R^2) and the regression equation. The regression equation provides the least squares line to represent the relationship between the two variables.

The coefficient of determination (R^2) is computed by multiple ways:

- It is the squares of the Pearson's correlation coefficient
- Or it can be computed by this formula:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

$$SSTO = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$\boxed{R^2 = 1 - \frac{SSE}{SSTO}}$$

The least squares line takes the following general formula:

$$\hat{y} = b_o + b_1 x$$

Whereas: $\hat{y}$ is the response value, which is the prediction of y according to x. b_o is the y-intercept, which is the predicted value of y supposed then x is equal to 0. b_1 is the slope, which indicates the degree to which y can be predicted by x.

However, it seems like the least squares line is not the best method to evaluate the relationship of these two variables. The scatter plot and residual plot do not strongly support these 4 conditions of fitting the least squares line:

- Linearity: Figure 3 is the scatter plot and the least squares line of the two variables. Unfortunately, the two variables do not show a strong linear relationship with most clusters concentrated to the left of the graph, and generally decreases to the right.
- Nearly normal residuals: Both figure 4 and 5 shows that the distribution of the residuals are not normal. The distribution in figure 5 shows a high peak value with a long right tail.

- Constant variability: The variability of points around the least squares line do not appear to be constant, with the variation degrades gradually from the left to the right in figure 3
- Independent observations: As this is a secondary source, we cannot conclude that these observations are not interrelated. However, we still can somewhat accept the observations are independent, as long as they are less than 10% of the real population.

The full calculation of the coefficient of determination (R^2) and the regression equation can be found at appendix E

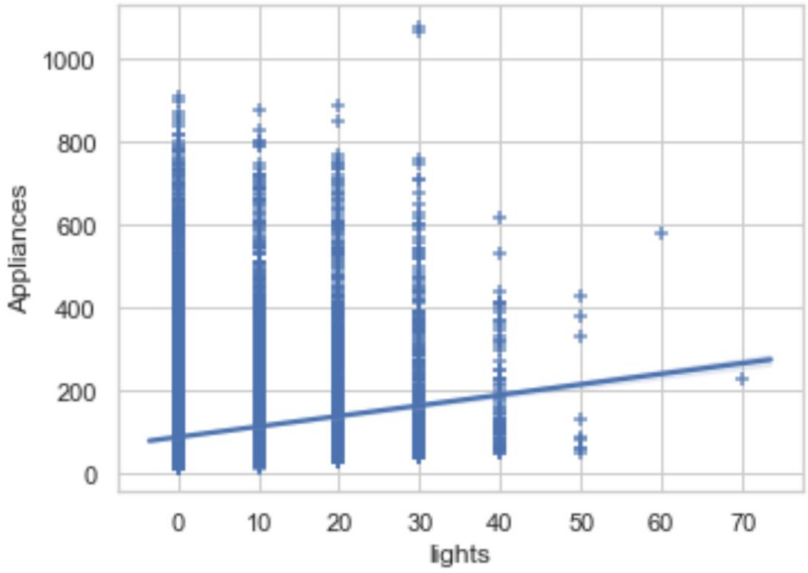


Figure 3: Scatter plot of the energy use of the light fixtures and appliances and the least squares line

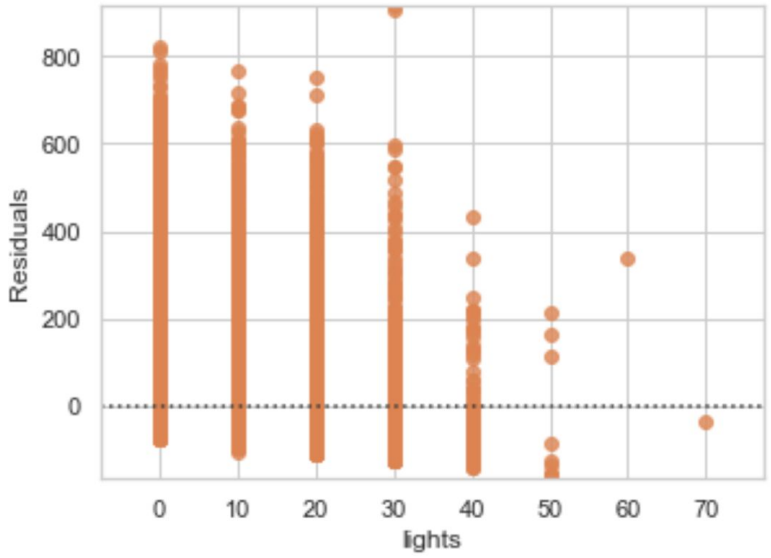


Figure 4: Residual plot of the energy use of the light fixtures

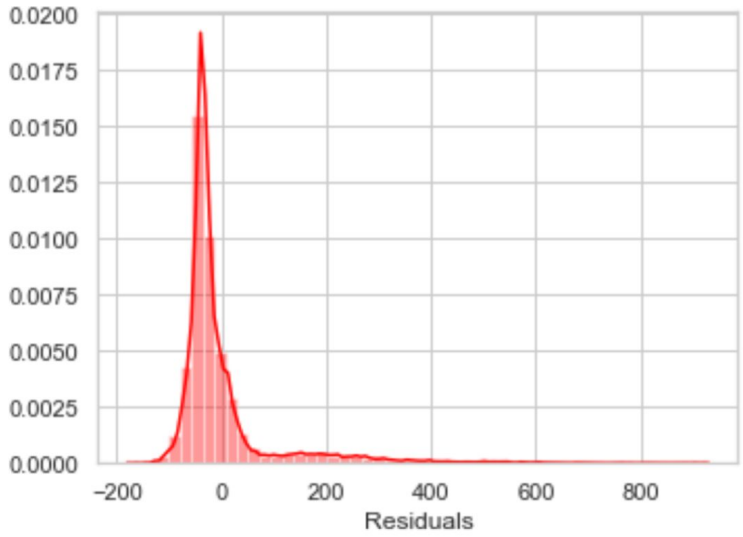


Figure 5: The distribution of the residuals of the energy use of the light fixtures

Lastly, I perform the statistical significance test of the slope of the regression equation. It is determined by the p-value with the significance level of 0.05.

The null and alternative hypothesis is as follow:

- Null hypothesis: There is no linear relationship between the energy use of light fixtures and energy use of appliances in the population: $\beta_1 = 0$

- Alternative hypothesis: There is some linear relationship between the energy use of light fixtures and energy use of appliances in the population: $\beta_1 \neq 0$

If the p-value is larger than the significance level of 0.05, then we will accept the null hypothesis. However, if the p-value is smaller than the significance level, then we will reject the null hypothesis and accept the alternative hypothesis.

The full calculation of the p-value can be found at appendix E.

iv. Results and Conclusions

a) Correlation:

From the formula, I obtain the Pearson's r value of 0.197, out of the possible values from -1 to 1, indicates a somewhat weak and positive correlation between these two variables. However, even with the strong correlation, we cannot conclude the causal relationship between them. Detailed calculation of this Pearson's r can be found in appendix D[7].

b) Regression:

The R-squared is 0.039, which means that 3.9% of the variation in the energy use of appliances is explained by the energy use of light.

The regression equation is appliances = 2.549 * lights + 88.005. The slope of 2.549 indicates that, for each incremental change in the energy use of light fixtures, there would be an 2.549 times incremental change accordingly in the energy use of appliances. The y-intercept of 88.005 indicates that even with 0 use of light fixtures, the energy use of appliances would still be 88.005 Wh. The Full calculation of this can be found at appendix E[8].

c) Significance

With the standard error of 0.09 and t-statistics of 28.27, we have the p-value of less than 0.00001, which is significantly smaller than the significance level I set (0.05). Because of the p-value lower than the significance level, we are rejecting the null hypothesis and accepting the alternative hypothesis: there is a linear relationship between the two variables. A summary of these values can also be found at appendix E[9].

The result of the Pearson's r value and significance test suggests there is evidence of the linear relationship between the energy use of the light fixtures and appliances. However, the regression

[7] #correlation: This is the interpretation of the result of Pearson's r correlation coefficient.
[8] #regression: This is the interpretation of the result of Pearson's r correlation coefficient.
[9] #significance: This is the interpretation from the hypothesis test on the slope of the regression model.

R-squared provides little explanation of the variation in the energy use of appliances from the energy use of the light fixtures.

As most research, this research is also prone to the influence of the extraneous variables. Some extraneous variables can be the different type of energy meters that lead to different measurement of energy.

v. References

Kaggle. (n.d.). Food choices and preferences of college students. Retrieved from https://www.kaggle.com/borapajo/food-choices

vi. Appendix

Appendix A: Import data and packages to Python. Create subset

In [1]:
```
#import needed packages
import pandas as pd
import numpy as np
import statistics
import matplotlib.pyplot as plt
import matplotlib
%matplotlib inline
import scipy
from scipy import stats
import statsmodels.api as statsmodels
import seaborn as sns
sns.set(color_codes=True)

#import data
energydata_df = pd.read_csv("KAG_energydata_complete.csv") #import the data from csv file
energydata_df = energydata_df.dropna(subset = ['lights','appliances'])

lights = energydata_df['lights']
appliances = energydata_df['appliances']

#Convert the data into list for easier manipulation
list_lights = list(lights)
list_appliances = list(appliances)

energydata_df.head(10) #shows the first 10 rows of the data
```

Out[1]:

	appliances	lights
0	60	30
1	60	30
2	50	30
3	50	40
4	60	40
5	50	40
6	60	50
7	60	50
8	60	40
9	70	40

Appendix B: Descriptive statistics for the dataset

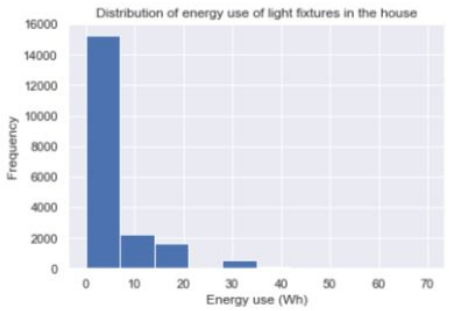
Appendix B: Descriptive statistics for the dataset

In [6]:
```
# Descriptive statistics for the dataset
# Build a function to calculate descriptive statistics (mean, mode, range, and standard deviation) based on Python inbuilt funct
def descriptive_stats(data):
    print("Count: ", len(data))
    print("Mean: ", round(np.mean(data),3))
    print("Median: ", np.median(data))
    print("Mode: ", stats.mode(data)[0][0])
    print("Standard deviation: ", round(np.std(data),3))
    print("Range: ", max(data)-min(data))

print("Descriptive stats for energy use of light fixtures in the house")
descriptive_stats(list_lights)

print("\nDescriptive stats for energy use of the appliances")
descriptive_stats(list_appliances)
```

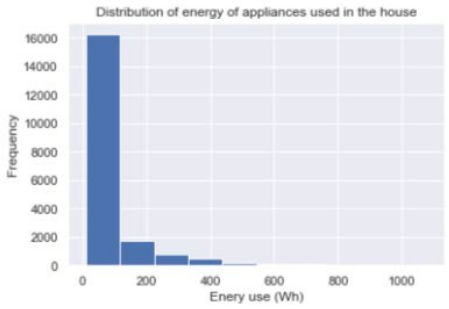
```
Descriptive stats for energy use of light fixtures in the house
Count:  19735
Mean:  3.802
Median:  0.0
Mode:  0
Standard deviation:  7.936
Range:  70

Descriptive stats for energy use of the appliances
Count:  19735
Mean:  97.695
Median:  60.0
Mode:  50
Standard deviation:  102.522
Range:  1070
```

Appendix C: Data visualization of the dataset

```
In [5]: #Data visualization of the dataset
        plt.hist(list_lights)
        plt.title("Distribution of energy use of light fixtures in the house")
        plt.xlabel("Energy use (Wh)")
Out[5]: Text(0,0.5,'Frequency')
```




```
In [6]: plt.hist(list_appliances)
        plt.title("Distribution of energy of appliances used in the house")
        plt.xlabel("Enery use (Wh)")
Out[6]: Text(0,0.5,'Frequency')
```



Appendix D: Confidence Pearson's correlation coefficient

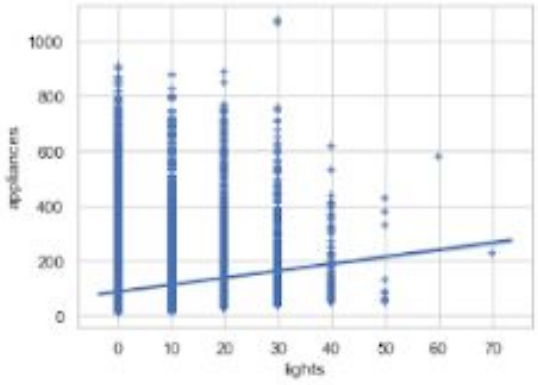
```
[7]: #Function to calculate the pearson's correlation value for parameters
def pcorr(x, y):
    print("The pearson's r value comparing", x , "to", y , "is:")
    print(energydata_df[x].corr(energydata_df[y]))
    print("")
```

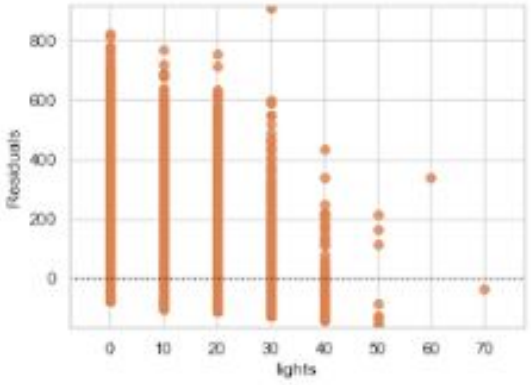
```
The pearson's r value comparing appliances to lights is:
0.1972775602062427
```

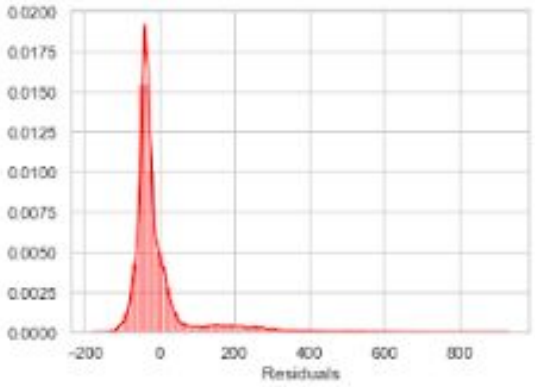
Appendix E: Hypothesis test of the slope of the regression equation and compute coefficient of determination

```
In [11]: 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 equation

    #fit the regression line using "statsmodels" library:
    X = statsmodels.add_constant(energydata_df[column_x])
    Y = energydata_df[column_y]
    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)
    slope = round(regressionmodel.params[1],3)
    intercept = round(regressionmodel.params[0],3)

    #make plots:
    sns.set_style("whitegrid")
    fig, (ax1, ax2) = plt.subplots(ncols=2, sharex=True, figsize=(12,4))
    sns.regplot(x=column_x, y=column_y, data=energydata_df, marker="+", ax=ax1) #scatter plot
    sns.residplot(x=column_x, y=column_y, data=energydata_df, ax=ax2) #residual plot
    ax2.set(ylabel='Residuals')
    ax2.set_ylim(min(regressionmodel.resid)-1,max(regressionmodel.resid)+1)
    plt.figure(figsize=(5.5,4)) #histogram
    sns.distplot(regressionmodel.resid, kde=True, axlabel='Residuals', color='red') #histogram

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

regression_model('lights', 'appliances')
```

```
R-squared =  0.039
Regression equation: appliances =  2.549 * lights +  88.005
                            OLS Regression Results
==============================================================================
Dep. Variable:             appliances   R-squared:                       0.039
Model:                            OLS   Adj. R-squared:                  0.039
Method:                 Least Squares   F-statistic:                     799.1
Date:                Thu, 22 Aug 2019   Prob (F-statistic):          2.31e-172
Time:                        08:00:31   Log-Likelihood:            -1.1899e+05
No. Observations:               19735   AIC:                         2.380e+05
Df Residuals:                   19733   BIC:                         2.380e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         88.0054      0.793    110.928      0.000      86.450      89.560
lights         2.5486      0.090     28.268      0.000       2.372       2.725
==============================================================================
Omnibus:                    14002.597   Durbin-Watson:                   0.524
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           194013.201
Skew:                           3.377   Prob(JB):                         0.00
Kurtosis:                      16.796   Cond. No.                         9.78
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```


appliances
lights


Residuals
lights


Residuals