**Step 9: Linear regression**

**1. Choose a quantitative column to predict using linear regression. Note: Linear regression may not work well for some data sets, so you will be graded on the process of performing linear regression, not on the fit of the model.**

Rating, Review. Date, REF Cocoa Percent are selected variables for regression and the Rating is used as dependent variable and other variables are independent.

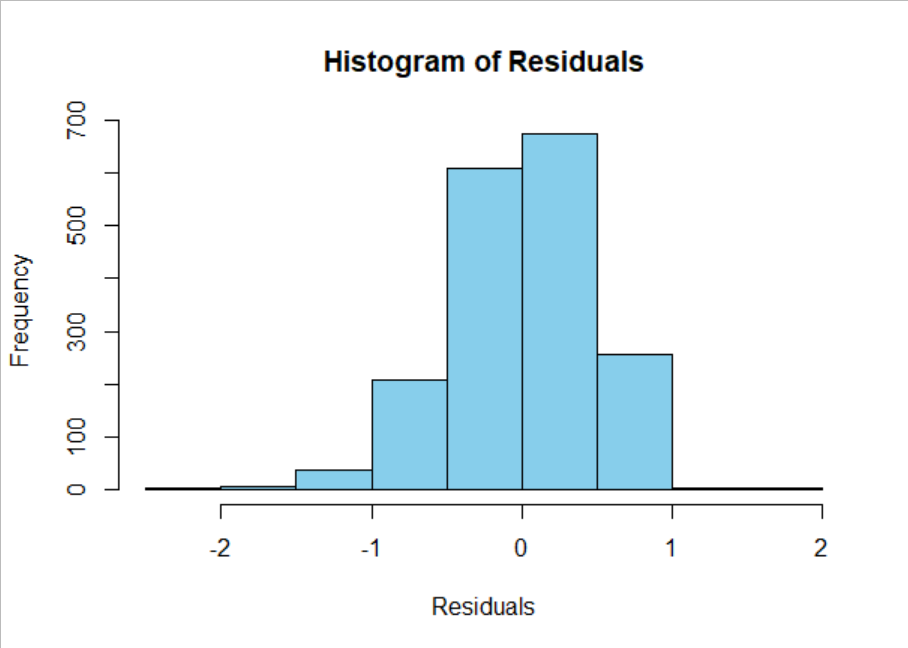
**\* Compute the linear regression model using one or more of the other data columns as the independent variable(s)**

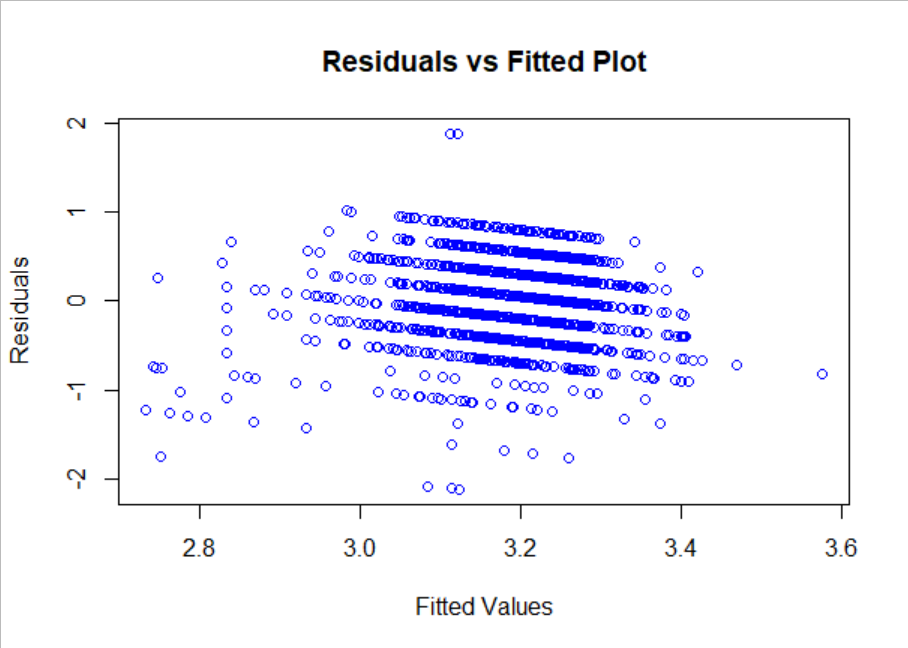
**Regression equation**

Fitted model is given below

**\* Assess the fit of the model by computing R-squared, plotting a histogram of the residuals, and plotting a scatter plot of the actual observed response value (x axis) vs. residual (y axis).**

R-squared = 0.0387





**\* Add a description of the model (including the equation) and the results of assessing the model fit to your webpage, including the plots.**

**Interpretation of Regression model**

Intercept can be interpreted as the average Ratings will be -1.35874, when REF and Cocoa Percent are zero. Review date coefficient can be interpreted as increasing one year in review data, the average rating will be increased by 0.00267 unit, REF coefficient can be interpreted as increasing one unit in REF, average Ratings will be increased by 0.00008 units, and the Cocoa Percent coefficient shows that increasing one unit in cocoa percent, the average ratings will be decreased by 0.01277 units.

**Model Fit explanation**

R-squared value shows that 3.87% variation in the dependent variable is explained by the independent variables.

Histogram of residuals shows that the residuals are approximately normally distributed. Residuals vs fitted plot not indicating any potential irregularity, and indicating a linear relationship, with constant variance.

**\* Write a few sentences above code explaining whether your linear regression model is a good fit, and why.**

The model is not good fitted, because the R-squared value is very low, and it shows that the independent variables are only explaining 3.87% variation in the dependent variable.

**Step 10: Hypothesis Testing**

**1. Come up with two different hypotheses about your data that are testable using the hypothesis tests we covered in class. For each hypothesis, do the following:**

**The two hypotheses are given below:**

**1st Hypothesis**

H0: Average rating is same for both company locations (USA and Other)

H1: Average rating is different for both company locations (USA and Other)

**Alpha =0.05**

**Test results**

Since the Leven’s Test with F-value = 3.0547 (P-value =0.08068) suggests that the variance of two groups are equal, the independent sample t test with equal variances is performed to test hypothesis.

**Calculated values**

t-calculated = 2.4301

df = 1793

p-value = 0.01519

**2nd Hypothesis**

H0: Average Cocoa. Percent is equal to than 50%

H1: Average Cocoa. Percent is greater to than 50%

**Alpha =0.05**

**Test results**

Since the Leven’s Test with F-value = 3.0547 (P-value =0.08068) suggests that the variance of two groups are equal, the independent sample t test with equal variances is performed to test hypothesis.

**Calculated values**

t-calculated = 145.39

df = 1794

p-value < 0.001

**\* Write a few sentences on your webpage stating the results of the hypothesis test and interpreting it (ex. can you reject the null hypothesis? Why or why not?).**

**1st hypothesis conclusion**

First hypothesis is conducted by converting the company location into two categories (USA and others), Independent sample t test is conducted with alpha =0.05. P-value of test indicates that the null hypothesis is rejected (as it is less than alpha), we conclude that the average rating is differnt for two different locations

**2nd hypothesis conclusion**

One sample t test is conducted to test whether the average cocoa percent is greater than 50 %. P-value of test indicates that the null hypothesis is rejected (as its less than alpha), we conclude that the coca percent is greater than 50%.

**Step 11: Your choice**

**1. Perform any 1 other analyses on your dataset. You may also perform one of the previous analysis on a different variable.**

**The hypotheses are given below:**

H0: Average cocoa is same for both company locations (USA and Other)

H1: Average cocoa is different for both company locations (USA and Other)

**Alpha =0.05**

**Test results**

Since the Leven’s Test with F-value = 13.466 (P-value =0.0005) suggests that the variance of two groups is not equal, the independent sample t test with unequal variances is performed to test hypothesis.

**Calculated values**

t-calculated = 1.0831

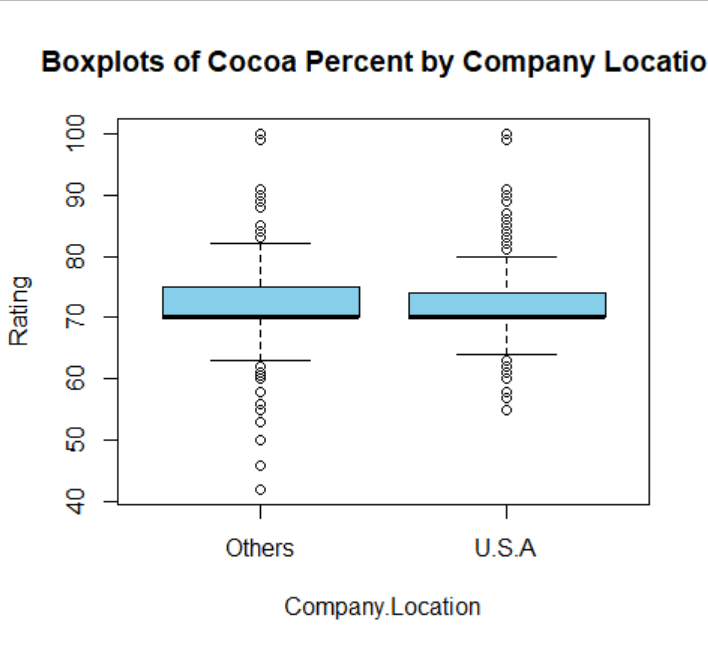
df = 1791

p-value = 0.2789

**Hypothesis conclusion**

To test the hypothesis, independent sample t test is conducted. The P-value of test indicates that the null hypothesis is not rejected (as it is greater than alpha), we conclude that the average cocoa percent is same for two different locations.

**2. Add any plots or the results of computations to your webpage.**

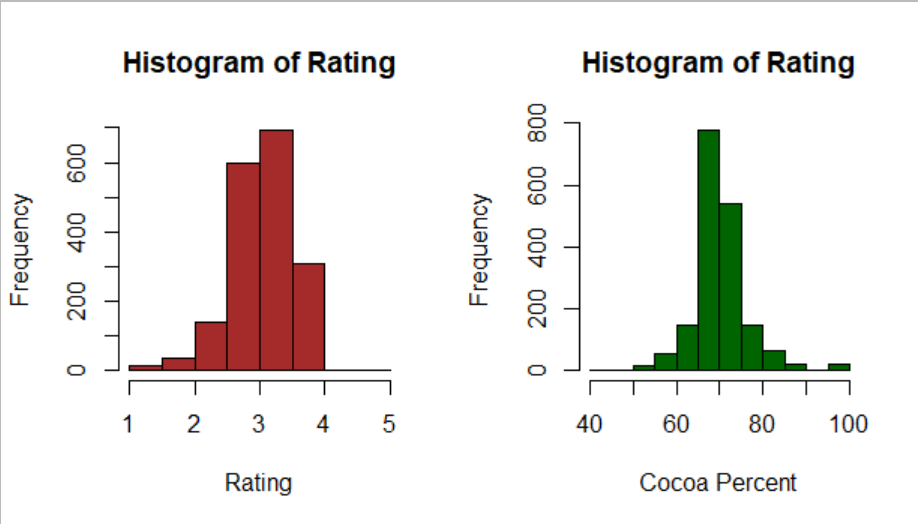


**3. Write a few sentences above code interpreting the results or plots for the reader.**

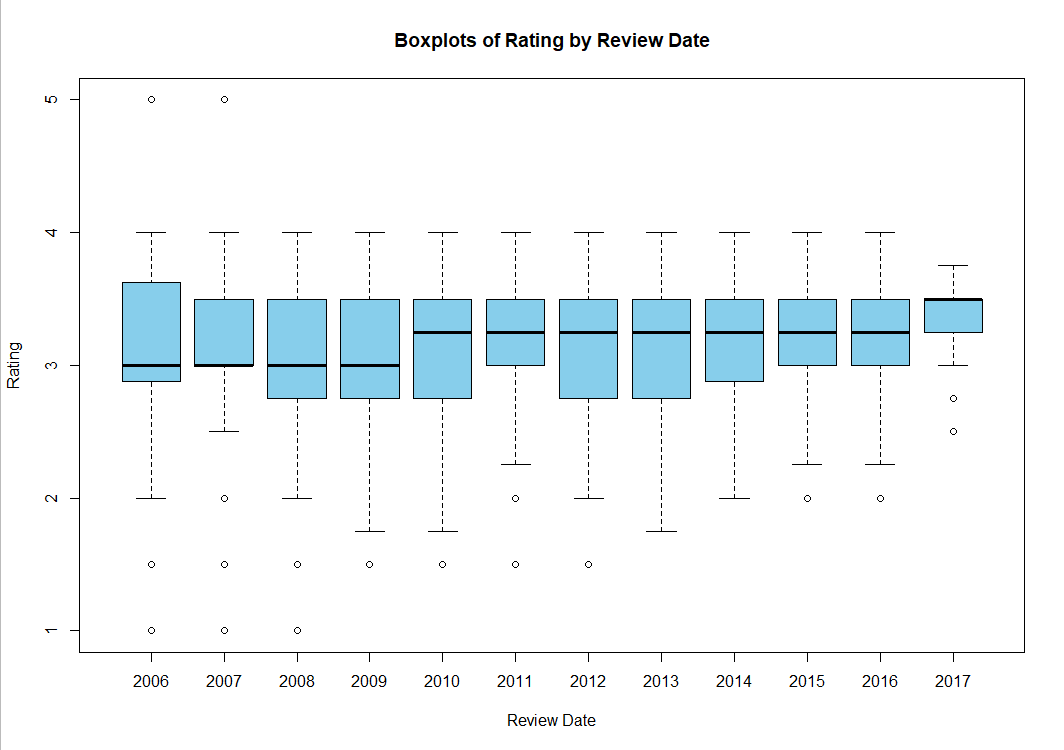
Boxplots of cocoa percent by two different locations shows that the average cocoa percent is almost similar. Similar conclusion is suggested by t test.

**Part 12: Three Extra For each of the Three extra analysis:**

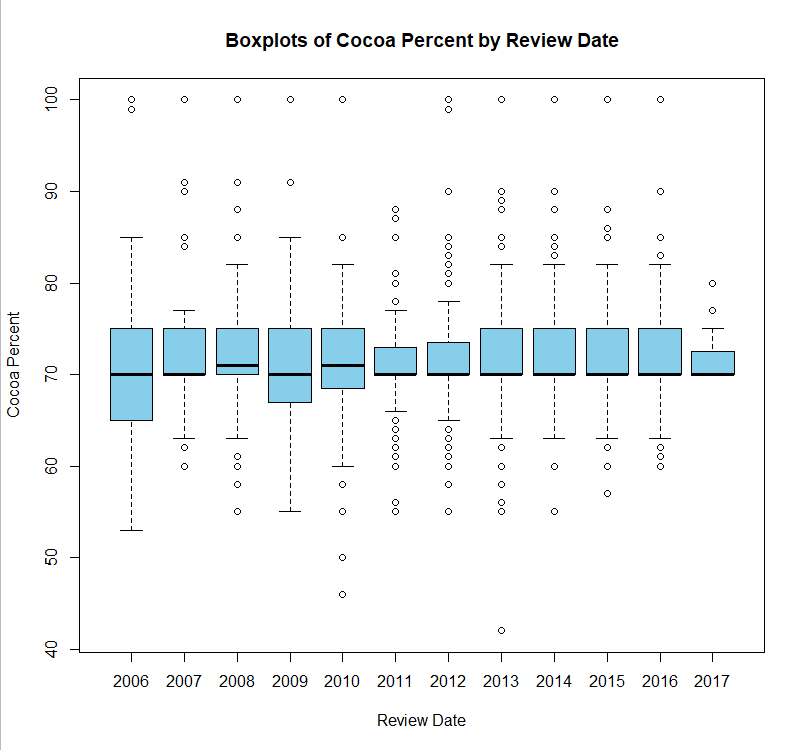
**Analysis 1: Whether the rating and cocoa percent variable have normal distribution?**



**Analysis 2: Comparing Ratings of different years**



**Analysis 3: Comparing Cocoa Percent by years**



**Add any plots or the results of computations to your webpage.**

**2. Write a few sentences above code interpreting the results or plots for the reader.**

**Analysis 1:**

Two histograms show that the rating have approximately normal distribution, and the distribution of cocoa percent is much closer to normal distribution than rating variable

**Analysis 2:**

From the boxplots, the highest rating was observed in 2017, and lowest was observed in year 2006 to 2009.

**Analysis 3:**

The highest cocoa percentage was observed in 2008 and 2010, and the lowest is observed in 2006 and 2009.