Analysis of Chocolate Bars Produced Around the World

I found an interesting dataset on kaggle called the flavors_of_cacao dataset which describes the expert ratings for 1,795 different chocolate bars along with other information that detail specifics on how the chocolate bar was created. As a simple and fun side project I decided to conduct some analysis on the dataset to learn more about the impact certain variables had in chocolate bars and discover certain trends that the chocolate bars in the dataset show.

To load the dataset into R I had to download the csv file for the dataset from this link on kaggle: https://www.kaggle.com/rtatman/chocolate-bar-ratings. Then I had to set the working directory to the location where the csv file was located on my local computer and load it into R with the read.csv function.

```
> setwd("~/Desktop")
> data = read.csv('flavors_of_cacao.csv')
```

Then I filtered the data to only include the variables of interest with this R statement below.

```
> data = data[, c(1,2,5,6,7,8,9)]
```

Introduction to Dataset

The variables of interest in this dataset are Company...Maker.if.known., Specific.Bean.Origin.or.Bar.Name, Cocoa.Percent, Company.Location, Rating, Bean.Type, and Broad.Bean.Origin. The Company...Maker.if.known. variable states the company that created a certain chocolate bar if the company that created the chocolate bar is known. The Specific.Bean.Origin.or.Bar.Name variable states the origin of the specific bean that the chocolate bar is based on or the name of the chocolate bar. The Cocoa.Percent variable states the percentage of cocoa that a certain chocolate bar has. The Company.Location variable states the location of the company that created a certain chocolate bar. The Rating variable states the expert rating that was given to a certain chocolate bar. The expert ratings are a measure from 1 to 5 based on flavor, texture, and aftermelt. A rating of 1 means that the chocolate bar was very unpleasant whereas a rating of 5 means that the chocolate bar was elite. The Bean.Type variable states the primary type of bean used to create a certain chocolate bar. The Broad.Bean.Origin variable states the country where a chocolate bar's Bean.Type variable is primarily found. The following code shows summary statistics about the data. I had to convert the Cocoa.Percent variable values into numeric values before doing this.

```
> data$Cocoa.Percent = as.numeric(sub("%", "", data$Cocoa.Percent))
> summary(data)
```

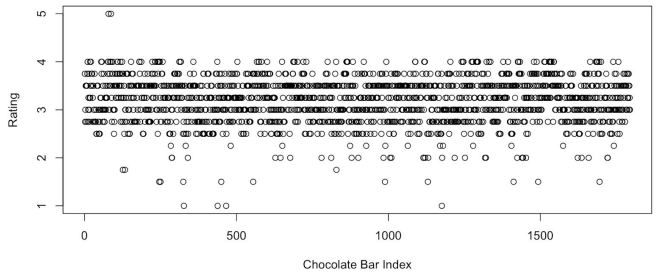
```
 Company...Maker.if.known. Specific.Bean.Origin.or.Bar.Name Cocoa.Percent   Company.Location      Rating       Bean.Type
 Length:1795               Length:1795                      Min.   : 42.0   Length:1795        Min.   :1.000   Length:1795
 Class :character          Class :character                 1st Qu.: 70.0   Class :character   1st Qu.:2.875   Class :character
 Mode  :character          Mode  :character                 Median : 70.0   Mode  :character   Median :3.250   Mode  :character
                                                            Mean   : 71.7                      Mean   :3.186
                                                            3rd Qu.: 75.0                      3rd Qu.:3.500
                                                            Max.   :100.0                      Max.   :5.000

 Broad.Bean.Origin
 Length:1795
 Class :character
 Mode  :character
```

The summary function was only able to provide useful statistics for the Rating and Cocoa.Percent variables since they are the only quantitative variables in the dataset.

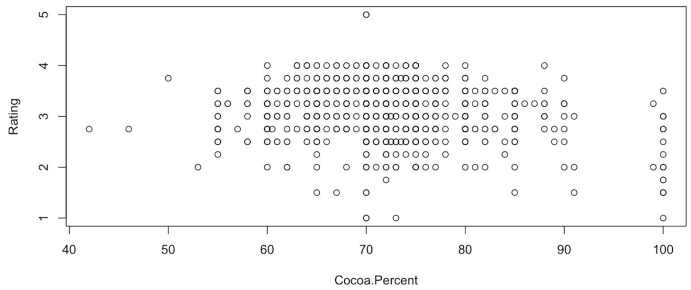
Since the only quantitative variables are Cocoa.Percent and Rating, these are the only introductory graphs I could come up with without diving into the data in too much depth which will be done in the later portion of this project that shows up later in the report.

```
plot(x = c(1:1795),y = data$Rating, ylab="Rating", xlab="Chocolate Bar Index")
```



Based on this graph, the majority of the chocolate bars in the dataset tend to have Rating values between around 2.6 and 4. There aren't a lot of cases where a chocolate bar has a Rating value outside that range.

```
> plot(x = data$Cocoa.Percent,y = data$Rating, ylab="Rating", xlab="Cocoa.Percent")
```



Based on this graph, a majority of the chocolate bars in the dataset have Cocoa.Percent values between around 55 and 85. It doesn't look like the Cocoa.Percent value can really tell us too much about the Rating value of a certain chocolate bar from this graph as there's no indicator of certain Cocoa.Percent values indicating a lower or higher Rating value than other Cocoa.Percent values. However, further analysis on this relationship will be done later to see what the data is really saying here.

Analysis

I have developed some research questions to get a better idea of what the possible trends/patterns in the dataset are and if there is enough evidence concluding that they are valid.

```
> attach(data)
```

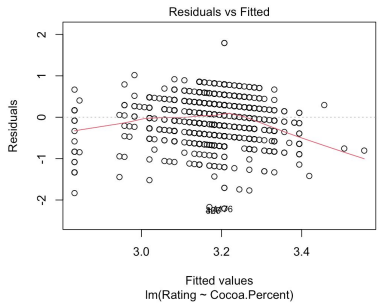
1. Is there a significant relationship between the Cocoa.Percent value and the Rating value for a certain chocolate bar?

We previously graphed the relationship between Cocoa.Percent and Rating values for all chocolate bars in the dataset. However, the relationship between the two variables was unclear just by looking at the graph. So let's conduct an investigation to find out. A Simple Linear Regression Test would be the most appropriate test for this investigation as it will analyze the relationship between the two variables and output many different statistics that can help us determine what to conclude about the relationship.

```
> mod = lm(Rating~Cocoa.Percent)
```

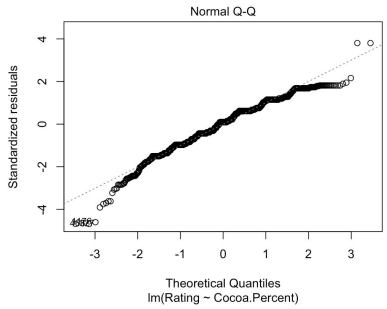
The assumptions that we are making for this test are that the Rating values(or residual errors) are independent, the Rating values can be expressed as a linear function of the Cocoa.Percent variable, the variation of the observations around the regression line is constant, and for any given Cocoa.Percent value, the Rating values(or residual errors) are Normally distributed. Since the Rating values for each row are for a different chocolate bar, they are independent so we can assume that the independence assumption is met. From looking at this Residuals vs. Fitted values plot below, the variation seems constant for the most part. The line in the Residuals vs. Fitted plot only deviates from 0 by a lot in the places where there are a small amount of data points so we can still assume that the assumptions(the variation of the observations around the regression line is constant and the Rating values can be expressed as a linear function of the Cocoa.Percent variable) are reasonably met.

```
> plot(mod)
```



From looking at the Normal Q-Q plot below, the Standardized Residuals line looks approximately linear for the most part and the line only deviates from linearity in the places where there are a small amount of data points so we can still assume that the assumption (the Rating values(or residual errors) are Normally distributed) is reasonably met.

```
> plot(mod)
```



Since all the assumptions for a Simple Linear Regression Test were reasonably met, we can safely conduct the test. The Hypotheses for this test are:
H0: β = 0
Ha: β ≠ 0
Where β = population slope for relationship between Cocoa.Percent and Rating value of a Chocolate Bar

```
> summary(mod)

Call:
lm(formula = Rating ~ Cocoa.Percent)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2071 -0.3196  0.0429  0.3178  1.7929

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.079388   0.126757  32.183  < 2e-16 ***
Cocoa.Percent -0.012461   0.001761  -7.076 2.12e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4717 on 1793 degrees of freedom
Multiple R-squared:  0.02717,           Adjusted R-squared:  0.02662
F-statistic: 50.07 on 1 and 1793 DF,  p-value: 2.122e-12
```

According to the p-value for the slope shown in the yellow highlighted line of the output, we should reject the null hypothesis stating that the population slope between Cocoa.Percent and Rating values is 0 and conclude that there is convincing evidence of the alternative hypothesis stating that the population slope between Cocoa.Percent and Rating values is significant(not zero). At the same time, we should be really wary of how we use the result of this test for future analysis of the data since the R-Squared values are really low. Even though the relationship between Cocoa.Percent and Rating value of a chocolate bar is significant, Cocoa.Percent only

explains a very small amount of the variance in Rating values which indicates that Cocoa.Percent by itself is not really a very good predictor of Rating value.

2. Is there a certain country that produces chocolate bars with a significantly different average rating than the other countries?

I thought this would be interesting to investigate because I know many people including me believe that the best chocolate comes from countries in Europe like Belgium or France. However, I only believe this based on past experience from the chocolate that I grew up eating and the chocolate that was available in grocery stores. I have never really made the initiative to find out if certain countries produced better or worse chocolate than others which is why I thought this would be a good question of interest for this data. I will be using the One Way Anova Test for this investigation as this investigation involves comparing means from many different groups of data based on one categorical variable. In our case the groups of data are the Rating values of the chocolate bars for each country in the dataset that has companies that produce chocolate bars.

```
> mod = lm(Rating~Company.Location)
```

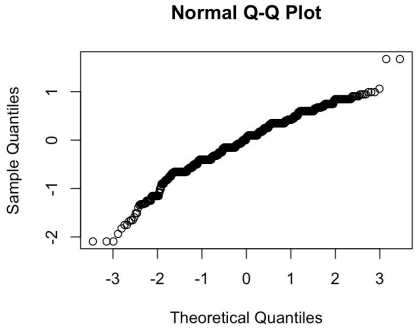
The assumptions that we are making for this test are that each sample is random and independent from the others, each population is normally distributed, and the standard deviations of the populations are similar. Since all the Rating values are for a different chocolate bar, we can be sure that each sample is random and independent from the others. In the code below, I calculated the standard deviations of all the Rating Values for Chocolate Bars from each individual country in the Company.Location variable column. The NA outputs correspond to countries with only 1 chocolate bar in the dataset. Apart from that, the standard deviations look really similar throughout the output below. There are a few instances where the standard deviation for one group might seem a bit extreme but they are still close enough to the general trend to assume that the assumption(the standard deviations of the populations are similar) is met.

```
> locations = unique(Company.Location)
> deviations = c()
> for (i in 1:length(locations)) { deviations[i] = sd(data[data$Company.Location == locations[i],
]$Rating);}
> deviations
 [1] 0.5466148 0.4419656 0.3535534 0.5683536 0.4732424 0.4665176 0.0000000 0.4325313
 [9] 0.5158338 0.4236268 0.5984437 0.3429972 0.4936480 0.4177070        NA 0.8178448
[17] 0.4757789        NA 0.1443376 0.4450015 0.4733511 0.3529029 0.3700656 0.3560002
[25] 0.2738613 0.2041241 0.5407043        NA 1.3597641 0.3486083 0.3944053 0.2886751
[33] 0.7071068 0.3535534 0.2500000 0.5204165 0.2968084 0.2311041 0.3818813 0.4506939
[41]        NA        NA 0.2091650 0.3535534 0.4424614        NA 0.2738613 0.4008919
[49] 0.3277253 0.2922613 0.2500000 0.5163978        NA 0.0000000        NA 0.3818813
[57]        NA 0.4407540        NA 0.3145764
```

From looking at the Normal Q-Q plot below, the Sample Quantities vs. Theoretical Quantities line looks approximately linear for the most part and the line only deviates from linearity a lot in

the places where there's only a small amount of data points so we can assume that the normality assumption is reasonably met.

```
> qqnorm(aov(mod)$residuals)
```



Since all the assumptions for a One Way Anova Test were reasonably met, we can safely conduct the test. The Hypotheses for this test are:

H0: The means of the Rating Values for chocolate bars produced in each individual country in the dataset are equal

Ha: At least one country has a mean for the Rating Values of their chocolate bars that is significantly different from the rest

```
> anova(mod)
Analysis of Variance Table

Response: Rating
                 Df Sum Sq Mean Sq F value    Pr(>F)
Company.Location   59  23.75 0.40255  1.8082 0.0002041 ***
Residuals        1735 386.26 0.22263
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the p-value for the F statistic shown in the yellow highlighted line of the output, we should reject the null hypothesis stating that the means of the Rating Values for chocolate bars produced in each individual country in the dataset are equal and conclude that there is convincing evidence of the alternative hypothesis stating that at least one country has a mean for the Rating Values of their chocolate bars that is significantly different from the rest. It may be more useful to find which country or countries deviate from the rest when it comes to their mean of the chocolate bar Rating values. However, this would be really hard to find since there's 60 different countries and I do not believe there is a specific test designed to discover this, at least from the tests that I have learned in college.

3. Is there a certain type of Bean that produces chocolate bars with a significantly different average Rating than the other types of Beans?

I thought it would be interesting to investigate because I think it would be useful to see if certain bean types produce better or worse chocolate bars than others. This type of information could be useful for chocolate producing companies when deciding what ingredients to use in their

chocolate bars. I also don't have much knowledge about the bean types used in making chocolate as the only thing I know is that Cacao/Cocoa plays a strong role in the creation of chocolate which is another reason for why I thought investigating the Bean types would be interesting. I will be using the One Way Anova Test for this investigation as this investigation involves comparing means from many different groups of data based on one categorical variable. In our case the groups of data are the Rating values of the chocolate bars for each Bean Type in the dataset.

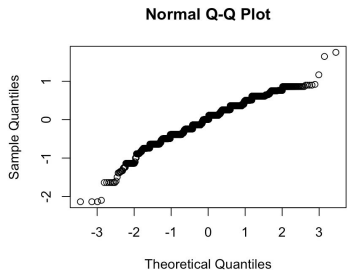
```
> mod = lm(Rating~Bean.Type)
```

The assumptions that we are making for this test are that each sample is random and independent from the others, each population is normally distributed, and the standard deviations of the populations are similar. Since all the Rating values are for a different chocolate bar, we can be sure that each sample is random and independent from the others. In the code below, I calculated the standard deviations of all the Rating Values for Chocolate Bars for each individual Bean.Type in the Bean.Type variable column. The NA outputs correspond to bean types with only 1 chocolate bar in the dataset. Apart from that, the standard deviations look really similar throughout the output below. There are a few instances where the standard deviation for one group might seem a bit extreme but they are still close enough to the general trend to assume that the assumption(the standard deviations of the populations are similar) is met.

```
> bean_types = unique(Bean.Type)
> deviations = c()
> for (i in 1:length(bean_types)) { deviations[i] = sd(data[data$Bean.Type == bean_types[i],
]$Rating);}
> deviations
 [1] 0.4830301 0.4587260 0.4212654 0.6039628 0.5438073 0.4368059 0.4009520 0.5034602
 [9] 0.5561464 0.1767767 0.1767767 0.3644345 0.5220818        NA        NA 0.7071068
[17] 0.0000000        NA        NA        NA 0.5204165 0.3535534 0.0000000 0.5419871
[25]        NA        NA 0.1767767 0.1443376        NA        NA 0.1767767 0.3818813
[33]        NA 0.5000000        NA        NA        NA        NA        NA        NA
[41]        NA 0.3535534
```

From looking at the Normal Q-Q plot here, the Sample Quantities vs. Theoretical Quantities line looks approximately linear for the most part and the line only deviates from linearity a lot in the places where there's a small amount of data points so we can still assume that the normality assumption is reasonably met.

```
> qqnorm(aov(mod)$residuals)
```



Since all the assumptions for a One Way Anova Test were reasonably met, we can safely conduct the test. The Hypotheses for this test are:
H0: The Means of the Rating Values for chocolate bars of each individual bean type in the dataset are equal
Ha: At least one bean type has a mean for the Rating Values of its chocolate bars that is significantly different from the rest

```
> anova(mod)
Analysis of Variance Table

Response: Rating
           Df Sum Sq Mean Sq F value    Pr(>F)
Bean.Type   41  21.24 0.51795  2.3355 4.468e-06 ***
Residuals 1753 388.77 0.22177
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

According to the p-value for the F-statistic shown in the yellow highlighted line of the output, we should reject the null hypothesis stating that the means of the Rating Values for chocolate bars of each individual bean type in the dataset are equal and conclude that there is convincing evidence of the alternative hypothesis stating that at least one bean type has a mean for the Rating Values of its chocolate bars that is significantly different from the rest. It may be more useful to find which Bean Type or Types deviate from the rest when it comes to their mean of the chocolate bar Rating values. However, this would be really hard to find since there’s 42 different Bean Types and I do not believe there is a specific test designed to discover this, at least from the tests that I have learned in college.