Drexel University

Final Report of Chocolate Around the World

BSAN - 460 - 001

Professor Muge Capan

Team 2

Otis McCullough

Maybellyn Yap

Salvi Patel

Alissa Shargorodsky

Emily Nhan

Table of Contents

Executive Summary:

Premium chocolate bars are a growing industry globally. Our team's aim is to analyze data provided to us containing information on chocolate bar reviews. Our objectives are to:

- conduct a market research on chocolate sales and trends in different countries worldwide as well as commonly used ingredients in chocolate production,
- use descriptive analytics using the given data to explore patterns and relationships in chocolate features, such as distribution of overall ratings, distribution of ratings by country and by ingredient count, which countries produce the highest-rated bars, what is the favorite taste, what combination of first and second taste result in high ratings, and which companies have the highest ratings, etc.,
- predict rating as function of features available in the data using linear regression analysis and compare different models' performance to find the best fitting model,
- determine which features and threshold values result in high rankings using advanced machine learning methods such as Random Forests,
- determine different categories/clusters of chocolate bars that have similar features within the cluster/group using Clustering methods, and
- interpret key findings and translate findings into business insights.

We hope to utilize statistical methods covered in class to perform these objectives for our clients. This report will demonstrate a holistic approach to this project in detail. It will highlight the methods, the market research, the data, the results, and the recommendations.

Introduction:

Scope and Purpose

Our team was hired by a top international culinary consulting firm to predicate how to produce the best rated chocolate bar, while saving money. Chocolate is the most popular candy in the world, where United States residents consume around 2.8 billion pounds. Chocolate is a highly competitive industry, in which many companies are trying to become the best rated chocolate. The main ingredients used to create chocolate consists of cocoa butter, lecithin, vanilla, salt, and sugar. Our team received a dataset with over 2,221 observations. To provide a better representation of the data, there were a few notable features included to assist our team with the analysis. These included the company name, bean origins, cocoa percentage, ingredients, rating, and taste. Our data focused more on dark chocolate, so our team wanted to figure out the best cocoa percentage for it.

Goals

Our goal is to provide our client with adequate information to produce a high quality chocolate bar at the lowest price. In order to achieve this goal, our team created an issue tree that included the following: main study question, sub questions, components of sub questions, and analysis that were performed. We wanted to predicate consumers' rating of a chocolate bar, with the given features, for our main study question. To further dissect the information, we investigated the relationship between cocoa percentage/ rating, cocoa percentage/ counts of ingredients, specific bean origin/ rating, and individual ingredient/ rating. Our group also wanted to find the difference between the first and second taste. The methods used to test our questions were descriptive analytics, data manipulation, linear regression, chi-square test, clustering, origin analysis, and ANOVA. Our issue tree is further analyzed later in our report. We backed our findings with market research, which provided our team enough information to give recommendations to our client.

Literature/Industry Review:

Chocolate's history has been present for quite a long time, stemming from the first discovery in ancient Mesoamerica, present day Mexico, almost 4,000 years ago. Chocolate at this time was being used for ritualistic, medical, and edible purposes. In 2021, the global and cocoa chocolate market size was valued at $46.61 billion. The market is now projected to grow from $48.29 billion in 2022 to $67.88 billion by 2029, displaying a CAGR of 4.98% during this projected period. The market is portraying such progressive growth due to the growth and development of the chocolate confectionary industry. According to an article from Fortune Business Insights, "There has been a significant rise in the origin share of global grindings, which has boosted the overall consumption of cocoa-based ingredients…Chocolate has remained the leading flavor in the new launches in beverage, bakery, and confectionary items. It has also remained one of the widely used ingredients in the sweets and beverages sector" (Fortune Business Insights, 2022).

Cocoa butter, cocoa powder, and cocoa liquor demand is expected to increase across numerous industries. This upward trend is set to feed market growth in the future. This is relevant to the project because our study questions are heavily centered on the impact of ingredients used in a quality chocolate bar.

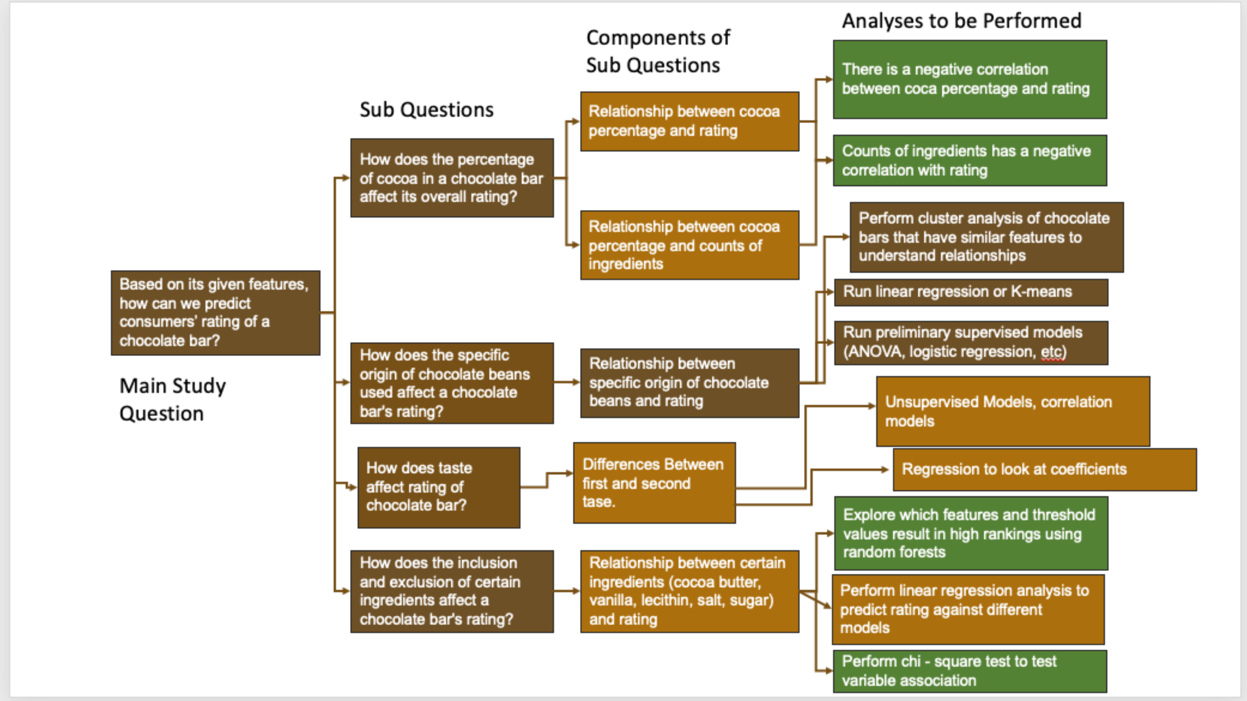
With all this demand for cocoa and chocolate, there are present and foreseeable concerns. One of which is sourcing of cocoa beans. The most prominent issues in cocoa supply chains include the use of child labor as well as expansion of cocoa production into protected reserves. Most cocoa is grown throughout the tropics. Although the crop originated in Latin America, almost 80% of it is currently produced in African Countries. Most of the world's cocoa is grown in two West African Countries: Cote d'Ivoire and Ghana. The processing and manufacturing of the cocoa products then usually occurs in Europe. According to a 2015 U.S Labor Department report, more than two million children were engaged in dangerous labor in cocoa growing regions. When asked in the spring of 2019, Hershey, Mars, and Nestle could not guarantee that any of their chocolate was produced without engaged child labor (Whoriskey, Siegel, 2019). To this day, chocolate companies cannot identify the farms where their cocoa is produced, let alone whether child labor was involved. Companies are inducing more of a strict tolerance to this aspect of cocoa production, and seek to produce cocoa within a pure production quota. There has been an observed increase in consumer preference for awareness of the origin of all the ingredients used in the chocolate. This is related to the project because it was asked by our client to recommend actionable business strategies, and one of our recommendations which will further be discussed, it to implement fair-trade practices in their chocolate bar products in order to not contribute towards slave labor where children and adult workers are exploited.

Another concern for the production of chocolate in the future, which is the rapid changing demand for certain products, including those with substitutes, equivalents, and chocolate products with healthier options. Within the last decade, more and more people have transitioned towards healthier eating habits and lifestyles. This can be due to allergies to particular components in the chocolate, or perhaps a preference to be dairy-free, vegan, etc. According to *Allied Market Research*, "the low calorie chocolate market is expected to reach $2.11 billion by 2030." This is relevant to the project scope because a recommendation that will be later discussed, focuses on providing healthier chocolate bar options for this healthy chocolate market that has experienced significant growth over the past decade.

Proposed Approach:

We hope to provide business insights to our clients by using statistical methods we have learned in Business Analytics classes. Ideally, we will answer questions we have proposed that will help the client improve their business. Our issue tree below highlights our proposed analysis topics.

Figure 1: Issue Tree v.3


Main Study Question
Sub Questions
Components of Sub Questions
Analyses to be Performed
Based on its given features, how can we predict consumers' rating of a chocolate bar?
How does the percentage of cocoa in a chocolate bar affect its overall rating?
How does the specific origin of chocolate beans used affect a chocolate bar's rating?
How does taste affect rating of chocolate bar?
How does the inclusion and exclusion of certain ingredients affect a chocolate bar's rating?
Relationship between cocoa percentage and rating
Relationship between cocoa percentage and counts of ingredients
Relationship between specific origin of chocolate beans and rating
Differences Between first and second tase.
Relationship between certain ingredients (cocoa butter, vanilla, lecithin, salt, sugar) and rating
There is a negative correlation between coca percentage and rating
Counts of ingredients has a negative correlation with rating
Perform cluster analysis of chocolate bars that have similar features to understand relationships
Run linear regression or K-means
Run preliminary supervised models (ANOVA, logistic regression, etc)
Unsupervised Models, correlation models
Regression to look at coefficients
Explore which features and threshold values result in high rankings using random forests
Perform linear regression analysis to predict rating against different models
Perform chi - square test to test variable association

In our issue tree, our overarching objective is to predict a chocolate bar that will maximize rating for the client. The purpose of this is so they could have the highest rated chocolate bar on the market. This would give them a competitive advantage in the chocolate bar market by having the most desirable chocolate bar. Stemming from this overall, questions come from sub-components of what determines the rating we hope to look at a majority of these variables to see what is impactful on over rating. Once we have determined that impactful on rating, we would then optimize these factors to produce the maximum rating. Do note that the component that is not included in our issue tree is that we determined cost was a very important factor to incorporate for our clients. We discovered from our research that ingredients vary in cost. We aim to minimize cost for our clients while still maintaining an above average rating overall

Data Elements:

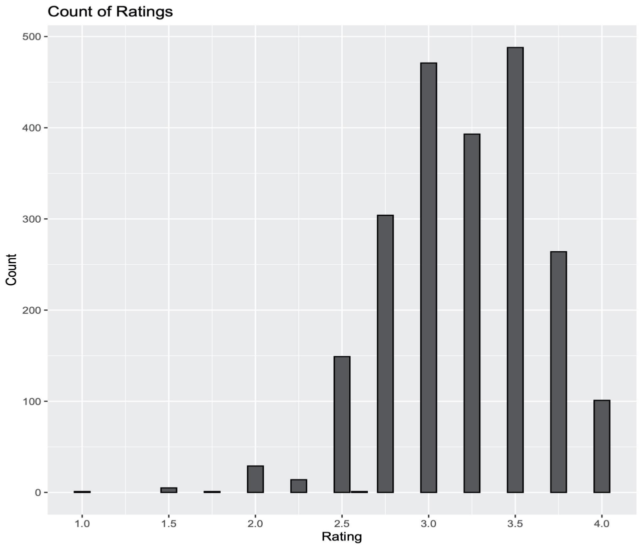
Our dataset consists of 2,221 observations with 15 variables. There are no missing or duplicated values. The composition of the variables are three numerical variables. The other 12 are categorical. Five of those 12 could be classified as binary variables. The variables specific to the company that are included in these reviews are, the company that manufactured the chocolate bar, the location of the company. The two variables relating to the cocoa bean are country of bean origin and specific bean origin or bar name. There are four variables that have to do with the rating itself. The first is the score on a scale of one through five. Figure 1.2 depicts the rating scale.

Figure 1.2: Table of Ratings

Score	**Description**
4.0 - 5.0	Outstanding
3.5 - 3.9	Highly Recommended
3.0 - 3.49	Recommended
2.0 - 2.9	Disappointing
1.0 - 1.9	Unpleasant

Another variable describing the rating is, year of review, this is the year the review was taken. This variable has a range of years from 2012 to 2020. The last two variables describing the rating are first and second taste. This variable takes a wide number of inputs and examines the participants' description of the taste of the chocolate. The rest of the variable describes the composition of the chocolate bar. These include the cocoa percentage of the chocolate bar and the number of ingredients. The last of the variables describing the chocolate bar tell if it contains sugar, vanilla, cocoa butter, salt, and lecithin. In addition, no imputations or removing of data were done as there are no missing values and duplications.

Descriptive Analytics



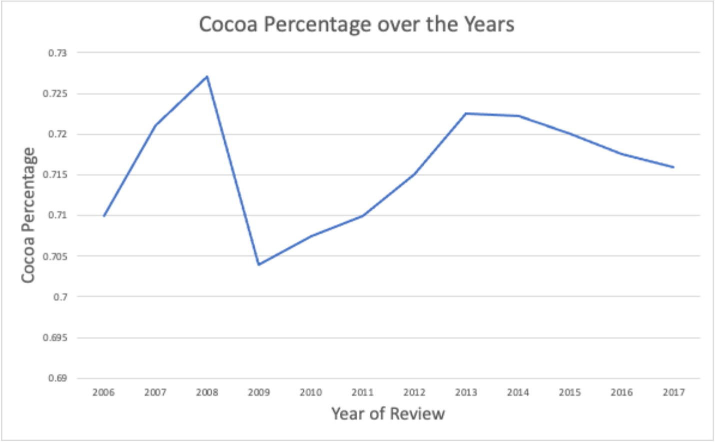


Figure 1.4: Cocoa Percentage Trend Over the Years

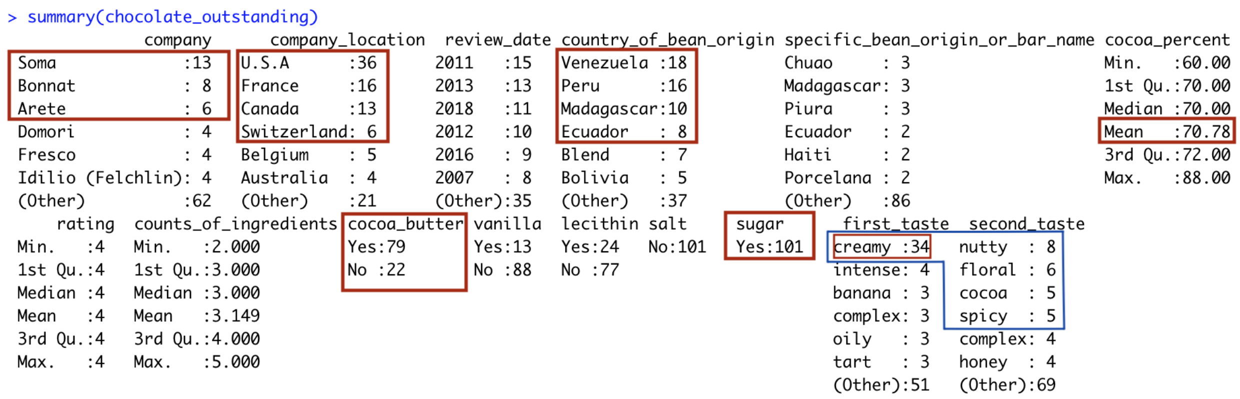
Figure 1.3: Distribution of Overall Ratings

Data Manipulation

By conducting data manipulation, we were able to answer the most questions asked by our clients. We filtered out the ratings based on the score category as listed in Figure 1.5.

Results

We provide results from outstanding rated chocolate bars for this part of the report and the rest of the R code will be inserted in the Appendix section. Figure 1.5 presents the results obtained for all outstanding rated chocolate bars.



```
> summary(chocolate_outstanding)
              company      company_location  review_date country_of_bean_origin specific_bean_origin_or_bar_name cocoa_percent
 Soma             :13   U.S.A      :36      2011   :15   Venezuela :18          Chuao     : 3                    Min.   :60.00
 Bonnat           : 8   France     :16      2013   :13   Peru      :16          Madagascar: 3                    1st Qu.:70.00
 Arete            : 6   Canada     :13      2018   :11   Madagascar:10          Piura     : 3                    Median :70.00
 Domori           : 4   Switzerland: 6      2012   :10   Ecuador   : 8          Ecuador   : 2                    Mean   :70.78
 Fresco           : 4   Belgium    : 5      2016   : 9   Blend     : 7          Haiti     : 2                    3rd Qu.:72.00
 Idilio (Felchlin): 4   Australia  : 4      2007   : 8   Bolivia   : 5          Porcelana : 2                    Max.   :88.00
 (Other)          :62   (Other)    :21      (Other):35   (Other)   :37          (Other)   :86
     rating  counts_of_ingredients cocoa_butter vanilla  lecithin salt     sugar      first_taste  second_taste
 Min.   :4   Min.   :2.000         Yes:79       Yes:13   Yes:24   No:101   Yes:101   creamy :34   nutty   : 8
 1st Qu.:4   1st Qu.:3.000         No :22       No :88   No :77                      intense: 4   floral  : 6
 Median :4   Median :3.000                                                           banana : 3   cocoa   : 5
 Mean   :4   Mean   :3.149                                                           complex: 3   spicy   : 5
 3rd Qu.:4   3rd Qu.:4.000                                                           oily   : 3   complex: 4
 Max.   :4   Max.   :5.000                                                           tart   : 3   honey   : 4
                                                                                     (Other):51   (Other):69
```

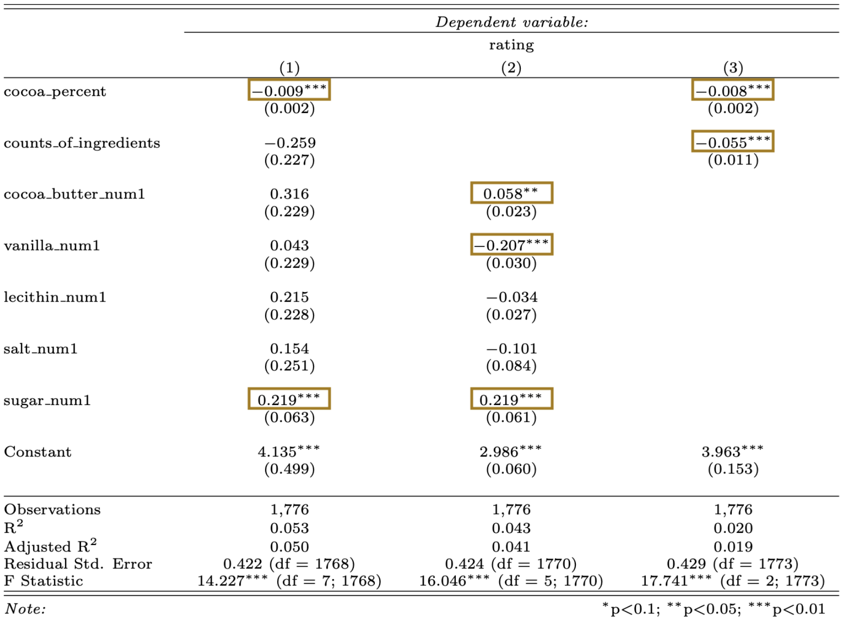
Figure 1.5

Soma, Bonnat, and Arete are the top 3 companies that were able to create outstanding chocolate bars. Next, outstanding rated chocolate bars are created through companies located in the United States, France, Canada, and Switzerland, with emphasis on cocoa beans originated from Venezuela, Peru, Madagascar, and Ecuador. From our overall dataset, an average of 71.49% of cocoa percentage is the most famous/observations, with 70.78% being the most observed in our outstanding rated bracket. We found out that out of 101 observations in the outstanding bracket, 79 of the chocolate bars contain cocoa butter, 13 contain vanilla, 24 contain lecithin. Furthermore, all 101 observations contained sugar but did not contain any salt. Lastly, the taste of "creamy" tops for the most popular taste, followed by nutty, floral, cocoa, and spicy. We deduce these mentioned tastes are the optimal combination of what form the best rated chocolate bar.

Linear Regression

To predict the ratings of a chocolate bar, we ran multiple linear regression models to compare the model results and retained the best fitting model for our rating prediction analysis. Figure 1.6 presents the models that we ran.

Results



	Dependent variable:		
	rating		
	(1)	(2)	(3)
cocoa_percent	−0.009***		−0.008***
	(0.002)		(0.002)
counts_of_ingredients	−0.259		−0.055***
	(0.227)		(0.011)
cocoa_butter_num1	0.316	0.058**	
	(0.229)	(0.023)	
vanilla_num1	0.043	−0.207***	
	(0.229)	(0.030)	
lecithin_num1	0.215	−0.034	
	(0.228)	(0.027)	
salt_num1	0.154	−0.101	
	(0.251)	(0.084)	
sugar_num1	0.219***	0.219***	
	(0.063)	(0.061)	
Constant	4.135***	2.986***	3.963***
	(0.499)	(0.060)	(0.153)
Observations	1,776	1,776	1,776
R^2	0.053	0.043	0.020
Adjusted R^2	0.050	0.041	0.019
Residual Std. Error	0.422 (df = 1768)	0.424 (df = 1770)	0.429 (df = 1773)
F Statistic	14.227*** (df = 7; 1768)	16.046*** (df = 5; 1770)	17.741*** (df = 2; 1773)
Note:			*p<0.1; **p<0.05; ***p<0.01

Figure 1.6

Like many datasets out there, obtaining a high adjusted R^2 value would be rare, therefore the low adjusted R^2 values reflected provided an insight that the variables only explained a certain amount of variation in the models. We obtained an adjusted R^2 of 0.05 for the first model, which was the highest among the three. We deduce this may be due to its amount of variables in the model being substantial than the rest. With the first model, we carried out a prediction on the recommended combination of cocoa percentage and whether or not to contain sugar. We retrieved the result of rating 3.35 for 42% cocoa and contains sugar, and rating 3.22 for 70% cocoa and contains sugar. We provide full predictions in our Appendix section.

Cluster Analysis

In addition to analyzing our ingredients variables, we ran a cluster analysis on outstanding rated bean origin and companies in hopes of providing insights on the similarities between the components in our dataset. Figure 1.7 presents outstanding rated companies, while figure 1.8 presents outstanding rated bean origin.

Results

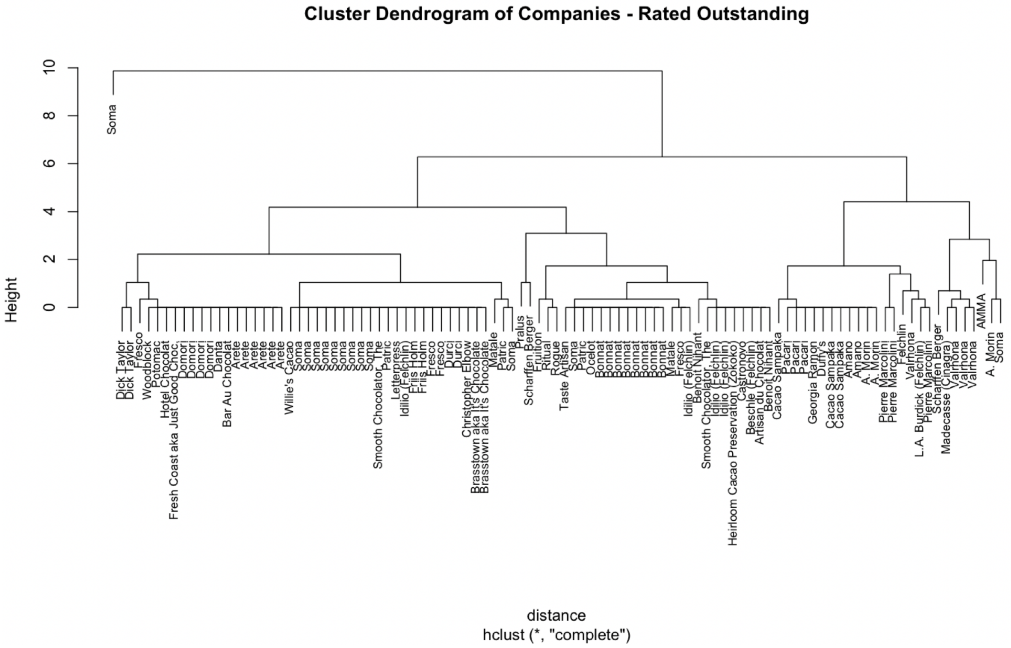


Figure 1.7

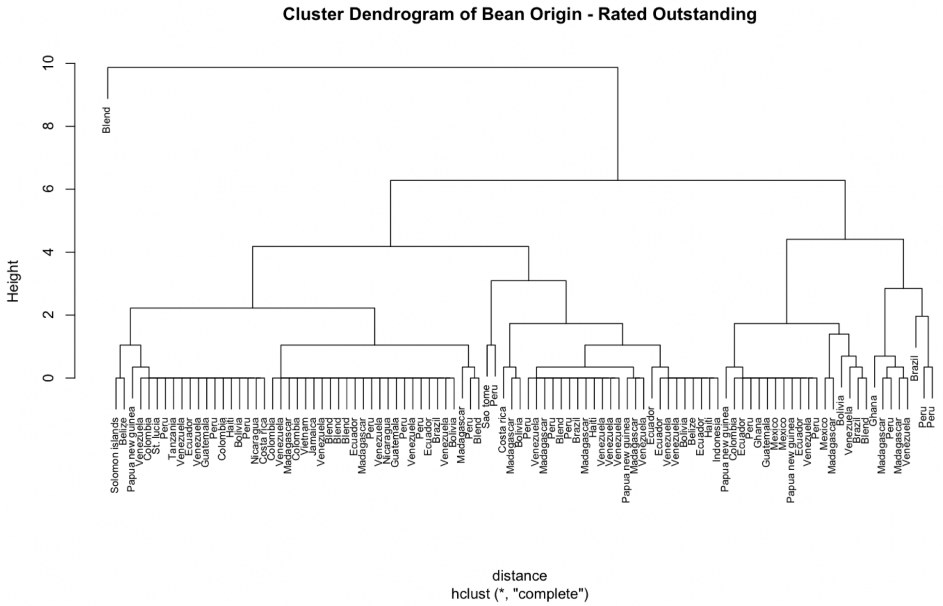


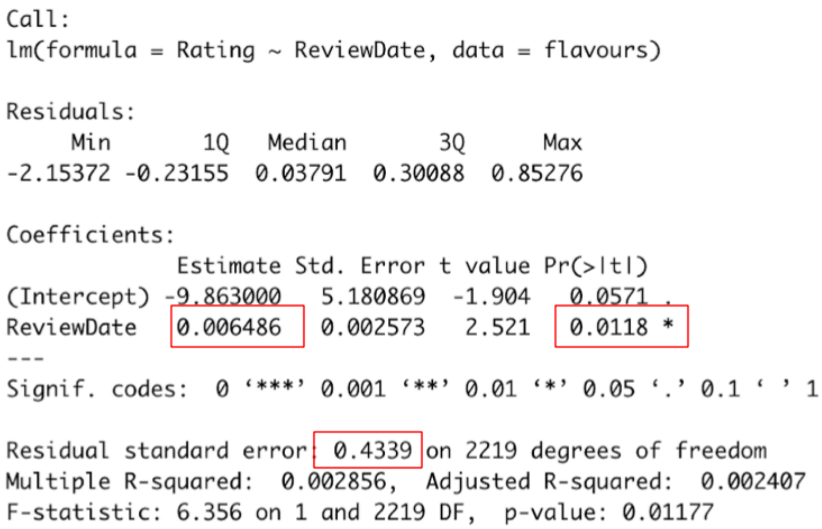
Figure 1.8

Through both of our cluster analysis, we obtained results that the big difference between clusters is cluster soma and the rest of the companies, and cluster Blend and the rest of the bean origin. Specifically in the bean origin, we see similarities between origins such as Venezuela, Ecuador, Peru, and Madagascar, of which appeared to be the top origins that produce top rated chocolate bars.

Linear Regression 2

To analyze if the date of a review had an effect on the rating score of the chocolate bar, linear regression was utilized. Simple linear regression was chosen because the outcome variable was numerical quantitative and there was just one one numerical predictor variable. Simple linear regression was also an appropriate choice because by looking at the coefficient it could be determined not only if the year of rating was statistically significant, but if the effect was positive or negative. See figure 1.9 below for the results.

Figure 1.9: Linear Regression for Review Date on Rating



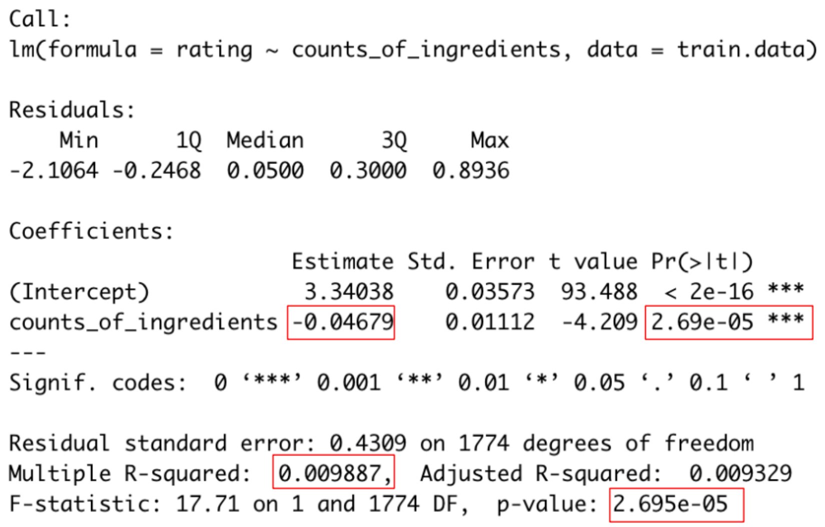
```
Call:
lm(formula = Rating ~ ReviewDate, data = flavours)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.15372 -0.23155  0.03791  0.30088  0.85276 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept) -9.863000   5.180869  -1.904   0.0571 .
ReviewDate   0.006486   0.002573   2.521   0.0118 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4339 on 2219 degrees of freedom
Multiple R-squared:  0.002856,	Adjusted R-squared:  0.002407 
F-statistic: 6.356 on 1 and 2219 DF,  p-value: 0.01177
```

After viewing the results, the R^2 is observable that the model does not fit particularly well. However, the residual standard error (RSE) is significantly better. This RSE shows our predictions will be stronger. Moving onto the coefficient line, the p-value is statistically significant given a confidence interval of 95%. The coefficient value is small but positive. This shows a positive correlation between review date and rating. As time increases, the average rating of the chocolate should improve.

Linear Regression 3:

Simple linear regression was used again to look at counts of ingredients. This was another variable that was important to explore for a cost analysis benefit. See figure below for results.

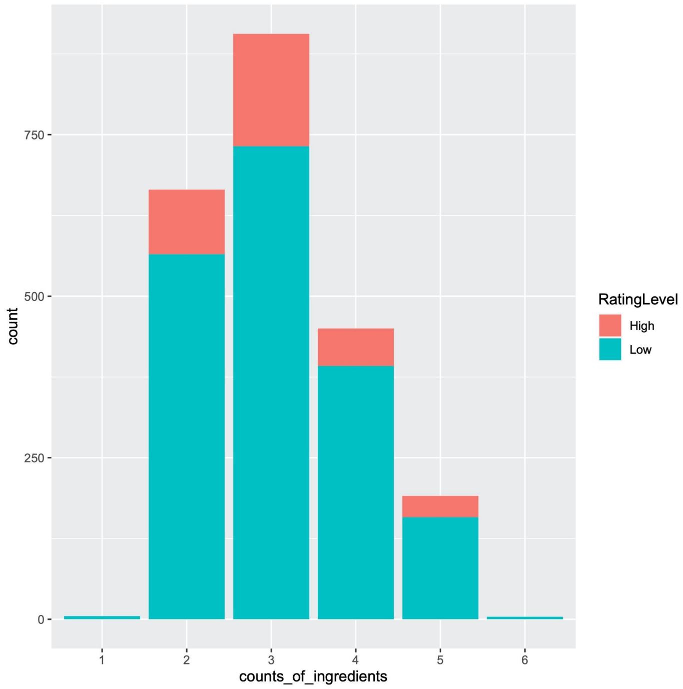
Figure 2.0: Simple Linear Regression Counts of Ingredients on Rating



```
Call:
lm(formula = rating ~ counts_of_ingredients, data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.1064 -0.2468  0.0500  0.3000  0.8936 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)            3.34038    0.03573  93.488  < 2e-16 ***
counts_of_ingredients -0.04679    0.01112  -4.209 2.69e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4309 on 1774 degrees of freedom
Multiple R-squared:  0.009887,	Adjusted R-squared:  0.009329 
F-statistic: 17.71 on 1 and 1774 DF,  p-value: 2.695e-05
```

As it is visible from our results the R^2 value is not particularly high. The p-value is very statistically significant, however. The coefficient value shows a negative correlation with counts of ingredients on rating. This means as more ingredients are added to the chocolate bar, the rating will decrease. From doing a descriptive analysis on these findings it can be determined that three ingredients is the optimal number of ingredients to include on the chocolate bar. See the figure below for the graph.

Figure 2.1: Counts of Ingredients on Rating

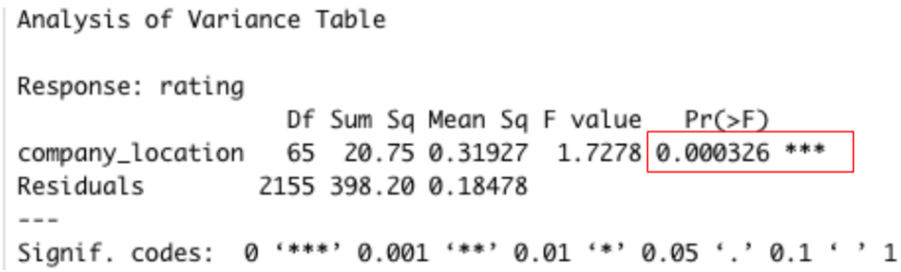


By dividing rating into two categorical variables 2.9 and below as low and 3.0 and higher as high rating it is evident that three ingredients has the highest number of highly rated chocolate bars by a large margin.

ANOVA

An Analysis of Variance test (ANOVA) was performed on the country of bean origin and rating to see if there was a difference in rating based on the origin of the bean. An ANOVA was also performed on the type of bean to see if there was a difference between these two variables. The ANOVA for this type of bean proved statistically insignificant. However the hypothesis testing for bean origin yielded results. Our null hypothesis (HO) was: “The means of the rating values for the chocolate bars produced in each country in the dataset are equal.” Our Alternative hypothesis (HA) was: “At least one country has a mean for the rating values of their chocolate bars that is significantly different from the rest”. Based on the results below (See figure 2.2) we were able to reject the null hypothesis and conclude that was an impact on rating based on bean origin.

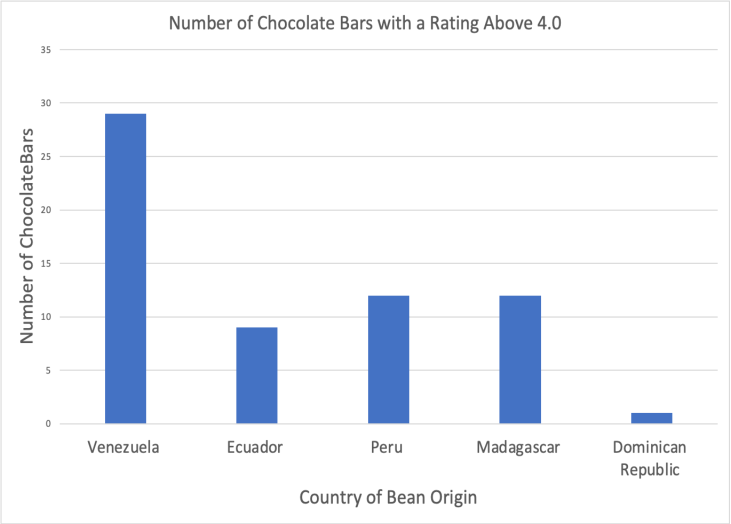
Figure 2.2: Anova Results



```
Analysis of Variance Table

Response: rating
                   Df Sum Sq Mean Sq F value    Pr(>F)
company_location   65  20.75 0.31927  1.7278 0.000326 ***
Residuals        2155 398.20 0.18478
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Taking this step further by using descriptive analytics, a graph was made showing top countries of bean origin with a rating of 4.0 or higher. Please see figure 2.3 below.

Figure 2.3: Graph of Bean Origin with a Rating 4.0 or higher



Based on these results it is recommended that our clients source their beans from one of these countries: Venezuela, Ecuador, Peru, or Madagascar. The Dominican Republic just has one count so this country is not so significant.

Recommendations:

Ingredients

As we dive deeper into the analysis that we’ve performed, we arrive at some discussions regarding recommendations that can be made based on our results. With cocoa percentage being statistically significant in predicting our rating, we recommend including at least 70% cocoa in chocolate production. Do note that 70% cocoa is considered as the baseline for a chocolate bar to be considered as dark chocolate. Furthermore, containing sugar in chocolate bars would also have a positive influence on rating. The patterns that we see in outstanding rated chocolate bars are many of them containing cocoa butter, cocoa nibs, organic cane sugar, which contributed to the creaminess and nutty flavor that consumers experienced in chocolate bars.

As discussed in the method's section, it is recommended that our client produces a chocolate bar with 3 ingredients. Two of the required ingredients from our analysis are sugar and vanilla. From Pearson's Chi-Squared test it was evident that sugar and vanilla are the only two ingredients that have a statistical significance on the rating. After performing the linear regression model, it was shown that sugar and vanilla have a positive correlation on the rating.

Alternative Ingredients

The main ingredients that can be added to create chocolate are cacao beans, cocoa butter, lecithin, vanilla, salt, and sugar. Cocoa butter and vanilla beans are expensive ingredients because there is a high demand for these items, but not enough supplies. Vanilla beans are only grown in three locations: Madagascar, Mexico, and Tahiti. This inherently makes it difficult to supply as many beans as possible to satisfy the demand. The plant takes three to four years to mature and is only able to be harvested a few days out of the year. The reason cocoa butter is expensive is because there is also a demand for it, but a lack of supplies. It is made from cacao beans, but the trees also take around two to four years to mature. There are not enough cacao beans to satisfy the demand. An alternate ingredient that can be used instead of cocoa butter is lecithin. Essentially, they both are similar to help smooth out the chocolate. Only 0.5% of lecithin is needed compared to 3 to 4% of cocoa butter, for a thin coating. The average price in 2022 in the United States for lecithin is $1 to $2 per kilogram, where cocoa butter is $4.02 to $4.14 per kilogram. An alternative for vanilla beans is vanilla extract. One vanilla bean will equal one teaspoon of vanilla extract. The average price in 2022 in the United States for vanilla extract is about $16 to $20 for a 4-ounce bottle compared to $34.18 to $44.88 per kilogram. Our team recommends two combinations to create a two-ounce 70% chocolate bar. The first option is to use 70% of cacao beans, 30% of sugar, and 0.5% of vanilla. The second option is to use 70%

of cacao beans, 30% of sugar, and 0.5% of lecithin. Our team believes that these two options would be a good combination for a dark chocolate bar.

Healthier Chocolate Bar Option + Fair Trade Sourcing

Most mainstream and popular chocolate bars are created by large companies like Hershey's, Nestlé, etc. Oftentimes, the chocolate bars are made from poor-quality, artificial, and filler ingredients. This recommendation focuses on using cleaner and higher-quality ingredients in creating chocolate bars that are on the healthier side, such as dairy-free, less or no sugar, vegan, etc., while still being transparent about cost. Cleaner and high-quality ingredients will run to be more expensive than the production costs of mass-manufactured chocolate bars, however given that the healthy chocolate bar market is continuously growing, this is a highly profitable opportunity that will help the client stand out from competitors. The average costs of production for a healthier chocolate bar may vary depending on the region, supplier, etc.

Another recommendation is to source ingredients ethically and become fair-trade certified. Most cocoa beans are sourced from developing countries where oftentimes, laborers are not paid for fairly for their hard work and contribution in the supply chain process. Not only do fair-trade initiatives make cocoa farming a more sustainable practice, but they also ensure that children are not being exploited in child labor, while farmers and workers are being paid respectfully and fairly. If the client's chocolate bar product were to become fair-trade certified, then every time a consumer buys the chocolate bar, they are putting money back into the hands of farmers who grew and harvested these crops, which ultimately helps them build their futures.

Maximizing Profit by Investing in Single Origin Cocoa

An upward trend that can be seen lately is the rising emphasis on single-origin cocoa which drives significant demand in for specialty chocolate and premiumize offerings. This shift in trend is specifically seen in consumer behavior as part of Covid-19's impact where personal health and wealthness was prioritized, as well as their society and environment. Single-origin cocoa trends propel innovations and new product development which leads to higher demand for high-quality cocoa powder and cocoa butter, in which the premiumize offerings that were mentioned before. Through our cluster analysis, we recommend clients to refer to dendrogram that depicts the bean origins that resulted in similarities in sourcing their beans. In addition, we recommend clients opting for cocoa sustainability programs, partnering with other companies to build cocoa processing farms in South America that use advanced farming practices to promote biodiversity and carbon capture. Investing in single origin cocoa beans is crucial as a specific variety of cocoa harvested in one region will take on the characteristics of the region where it's grown, attributing to many unique tastes to chocolate bars, and thus increasing the sales of chocolate production.

Conclusion:

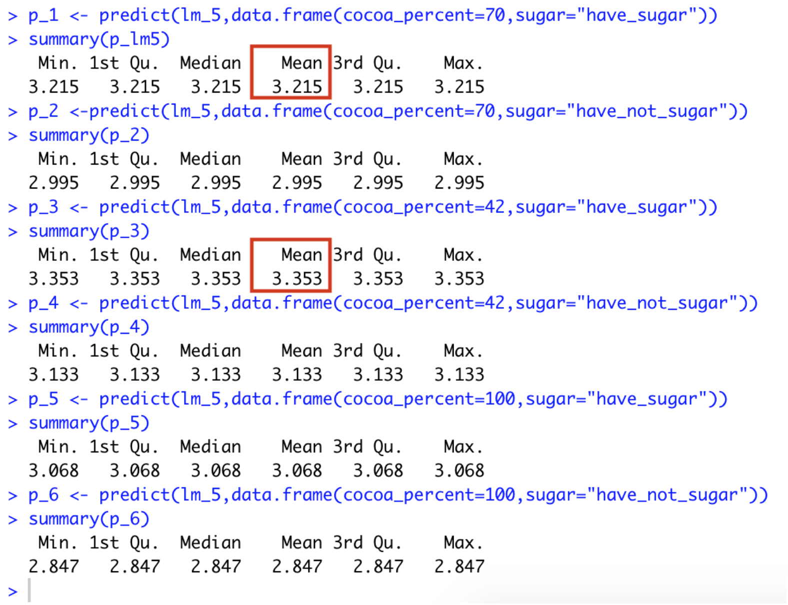
Our group's goal was to provide actionable business insights to our clients, the top fine dark chocolate bar producing companies. We were able to do this, using our insights from our descriptive and predictive analytics. Overall, we have determined our clients should produce chocolate bars with three ingredients. Two of the ingredients should be vanilla and sugar. Cocoa percentage should be right around 70%. Our clients need to take advantage of the new technology that develops each passing year. Our client also needs to focus on sourcing beans from a single high quality source in South America. We believe that this combination of insights will lead our clients to have a competitive edge in the market and lead them to long term success.

(Extra Credit) Company Analysis:

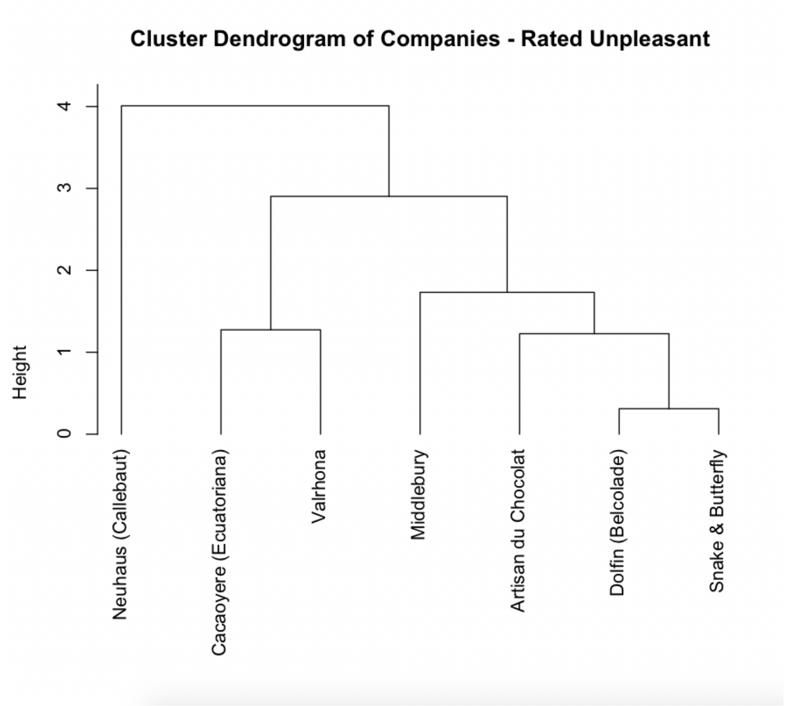
From the result of our cluster analysis, the origin of the beans that companies sourced are mainly from developing countries. We researched the trends of leaders in the industry and noted that key players often gain competitive advantage by adopting fair-trade certification. The increasing awareness regarding labor welfare is expected to fuel demand for fair trade cocoa in foreseeable years, in which fair trade standards ensure workers or farmers are paid fair price for cocoa beans. From our dataset, the company Soma stood out among its peers because of their effort in upholding fair-trade standards. While it is a private company, its annual sales reach $2.7 million according to Dun & Bradstreet. Soma sourced their beans from countries including Ecuador, Peru, and Venezuela. According to the company, beans sourced from the Dominican Republic are always fair-trade, and beans sourced from Madagascar are always organic. Furthermore, the company pays 4 or 5 times the ICCO (International Cocoa Organization) daily price for cocoa beans, while founding the Heirloom Cacao Preservation Initiative that protects the natural reproduction of fine flavored cacao. Taking the company as a reference, we would recommend adopting strategies such as the bean-to-bar concept against companies like Soma. The concept of bean-to-bar is a way for chocolate companies to distinguish their product and mass produced chocolate. Bean-to-bar concept involves small batches of chocolate production, ensuring the highest quality possible of a chocolate bar. “To achieve these results, makers have to purchase small amounts of ingredients at a time, and bootstrap most of their equipment. This makes bean to bar chocolate more expensive than industrial chocolate, so small makers have had to educate their customers as to why their chocolate tastes so different and costs so much more than other “premium” chocolates” (brandcocoa).

Appendix

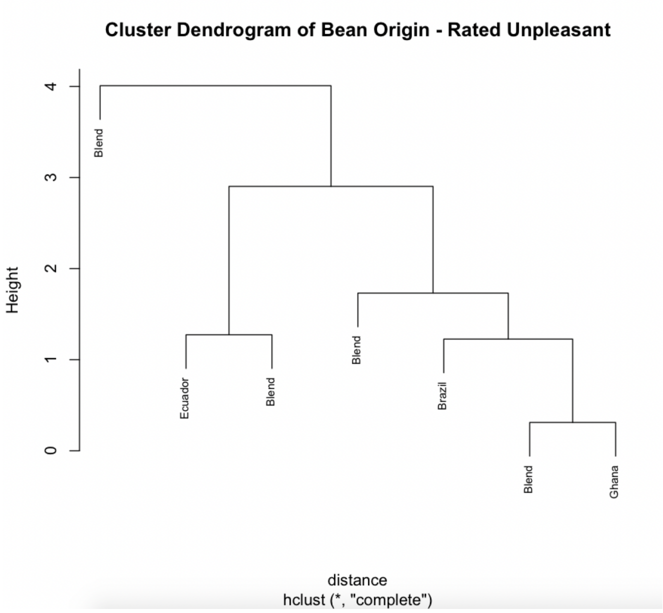
Results of linear regression prediction:



```
> p_1 <- predict(lm_5,data.frame(cocoa_percent=70,sugar="have_sugar"))
> summary(p_lm5)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  3.215   3.215   3.215   3.215   3.215   3.215 
> p_2 <-predict(lm_5,data.frame(cocoa_percent=70,sugar="have_not_sugar"))
> summary(p_2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.995   2.995   2.995   2.995   2.995   2.995 
> p_3 <- predict(lm_5,data.frame(cocoa_percent=42,sugar="have_sugar"))
> summary(p_3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  3.353   3.353   3.353   3.353   3.353   3.353 
> p_4 <- predict(lm_5,data.frame(cocoa_percent=42,sugar="have_not_sugar"))
> summary(p_4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  3.133   3.133   3.133   3.133   3.133   3.133 
> p_5 <- predict(lm_5,data.frame(cocoa_percent=100,sugar="have_sugar"))
> summary(p_5)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  3.068   3.068   3.068   3.068   3.068   3.068 
> p_6 <- predict(lm_5,data.frame(cocoa_percent=100,sugar="have_not_sugar"))
> summary(p_6)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.847   2.847   2.847   2.847   2.847   2.847 
> 
```



Cluster Dendrogram of Companies - Rated Unpleasant



References (needs to be APA style)

DeBra, C. (1970, January 1). *Arete - Brasil 64% Organic Dark Milk Bar - Oct. 28, 2017*. Arete - Brasil 64% Organic Dark Milk bar - Oct. 28, 2017. Retrieved June 3, 2022, from http://www.chocolatebanquet.com/2017/10/arete-brasil-64-organic-dark-milk-bar.html

Chocolat bonnat madre de dios peru 75% dark chocolate bar. World Wide Chocolate. (2022, May 20). Retrieved June 3, 2022, from https://www.worldwidechocolate.com/shop/chocolat-bonnat/madre-de-dios-peru-dark-chocolate-bar/

Soma guasare venezuela 70%. Bar & Cocoa. (n.d.). Retrieved June 3, 2022, from https://barandcocoa.com/collections/soma/products/soma-guasare-venezuela-70

Cocoa and chocolate market size, trends: Growth Report, 2029. Cocoa and Chocolate Market Size, Trends | Growth Report, 2029. (n.d.). Retrieved June 3, 2022, from https://www.fortunebusinessinsights.com/industry-reports/cocoa-and-chocolate-market-100075

WP Company. (2019, June 5). *Hershey, Nestle and Mars won't promise their chocolate is free of Child labor*. The Washington Post. Retrieved June 3, 2022, from https://www.washingtonpost.com/graphics/2019/business/hershey-nestle-mars-chocolate-child-labor-west-africa/

API, S. (2020, September 13). *What is Bean to bar chocolate?* Bar & Cocoa. Retrieved June 3, 2022, from https://barandcocoa.com/pages/what-is-bean-to-bar-chocolate

Allied Market Research. (2021, November 16). *Global Low Calorie Chocolate Market is expected to reach $2.11 billion by 2030: Says Amr*. GlobeNewswire News Room. Retrieved June 3, 2022, from https://www.globenewswire.com/news-release/2021/11/16/2335449/0/en/Global-Low-Calorie-Chocolate-Market-Is-Expected-to-Reach-2-11-Billion-by-2030-Says-AMR.html

R Code

```
#cols <- c("company", "company_location", "review_date", "country_of_bean_origin",
"specific_bean_origin_or_bar_name", "first_taste", "second_taste")
#chocolate[cols] <- lapply(chocolate[cols], factor)

#chocolate_unpleasant <- chocolate %>% filter(rating>=1.0&rating<=1.9)
#chocolate_disappointing <- chocolate %>% filter(rating>=2.0&rating<=2.9)
#chocolate_recommend <- chocolate %>% filter(rating>=3.0&rating<=3.49)
#chocolate_highly_recommend <- chocolate %>% filter(rating>=3.5&rating<=3.9)
#chocolate_outstanding <- chocolate %>% filter(rating>=4.0&rating<=5.0)

#chocolate_unpleasant <- chocolate %>% filter(rating>=1.0&rating<=1.9)
#cols <- c("company", "company_location", "review_date", "country_of_bean_origin",
"specific_bean_origin_or_bar_name","cocoa_butter","vanilla","lecithin","salt","sugar","first_tast
e","second_taste")
#chocolate_unpleasant[ ,cols] <- lapply(chocolate_unpleasant[ ,cols] , factor)

#chocolate_disappointing <- chocolate %>% filter(rating>=2.0&rating<=2.9)
#cols <- c("company", "company_location", "review_date", "country_of_bean_origin",
"specific_bean_origin_or_bar_name","cocoa_butter","vanilla","lecithin","salt","sugar","first_tast
e","second_taste")
#chocolate_disappointing[ ,cols] <- lapply(chocolate_disappointing[ ,cols] , factor)

#chocolate_recommend <- chocolate %>% filter(rating>=3.0&rating<=3.49)
#cols <- c("company", "company_location", "review_date", "country_of_bean_origin",
"specific_bean_origin_or_bar_name","cocoa_butter","vanilla","lecithin","salt","sugar","first_tast
e","second_taste")
#chocolate_recommend[ ,cols] <- lapply(chocolate_recommend[ ,cols] , factor)

#chocolate_highly_recommend <- chocolate %>% filter(rating>=3.5&rating<=3.9)
```

```
#cols <- c("company", "company_location", "review_date", "country_of_bean_origin",
"specific_bean_origin_or_bar_name","cocoa_butter","vanilla","lecithin","salt","sugar","first_taste","second_taste")
#chocolate_highly_recommend[ ,cols] <- lapply(chocolate_highly_recommend[ ,cols] , factor)

#chocolate_outstanding <- chocolate %>% filter(rating>=4.0&rating<=5.0)
#cols <- c("company", "company_location", "review_date", "country_of_bean_origin",
"specific_bean_origin_or_bar_name","cocoa_butter","vanilla","lecithin","salt","sugar","first_taste","second_taste")
#chocolate_outstanding[ ,cols] <- lapply(chocolate_outstanding[ ,cols] , factor)

#chocolate[  , "cocoa_butter_num"] = NA
#for (i in 1:nrow(chocolate)) {
  if (chocolate$cocoa_butter[i] == "have_cocoa_butter") {
    chocolate$cocoa_butter_num[i] <- 1
  }
  else {
    chocolate$cocoa_butter_num[i] <- 0
  }
}

#chocolate[  , "vanilla_num"] = NA
#for (i in 1:nrow(chocolate)) {
  if (chocolate$vanilla[i] == "have_vanila") {
    chocolate$vanilla_num[i] <- 1
  }
  else {
    chocolate$vanilla_num[i] <- 0
  }
}
```

```
#chocolate[  , "lecithin_num"] = NA
#for (i in 1:nrow(chocolate)) {
  if (chocolate$lecithin[i] == "have_lecithin") {
    chocolate$lecithin_num[i] <- 1
  }
  else {
    chocolate$lecithin_num[i] <- 0
  }
}

#chocolate[  , "salt_num"] = NA
#for (i in 1:nrow(chocolate)) {
  if (chocolate$salt[i] == "have_salt") {
    chocolate$salt_num[i] <- 1
  }
  else {
    chocolate$salt_num[i] <- 0
  }
}

#chocolate[  , "sugar_num"] = NA
#for (i in 1:nrow(chocolate)) {
  if (chocolate$sugar[i] == "have_sugar") {
    chocolate$sugar_num[i] <- 1
  }
  else {
    chocolate$sugar_num[i] <- 0
  }
}
```

```
#cols <- c("cocoa_butter_num", "vanilla_num", "lecithin_num", "salt_num", "sugar_num",
"first_taste","second_taste", "cocoa_butter","vanilla","lecithin","salt","sugar")
#chocolate[cols] <- lapply(chocolate[cols], factor)

#training.samples <- sample(nrow(chocolate),0.80*nrow(chocolate))
#train.data  <- chocolate[training.samples, ]
#test.data <- chocolate[-training.samples, ]
#lm_1 <- lm (rating~cocoa_percent+ counts_of_ingredients+ cocoa_butter_num+ vanilla_num+
lecithin_num+salt_num+sugar_num,data=train.data)
#lm_2 <- lm (rating~cocoa_butter_num+ vanilla_num+ lecithin_num+ salt_num+ sugar_num,
data=train.data)
#lm_3 <- lm (rating~cocoa_percent+counts_of_ingredients,data=train.data)
#stargazer(lm_1, lm_2, lm_3, title="Results")

#p_1 <- predict(lm_5,data.frame(cocoa_percent=70,sugar="have_sugar"))
#p_2 <-predict(lm_5,data.frame(cocoa_percent=70,sugar="have_not_sugar"))
#p_3 <- predict(lm_5,data.frame(cocoa_percent=42,sugar="have_sugar"))
#p_4 <- predict(lm_5,data.frame(cocoa_percent=42,sugar="have_not_sugar"))
#p_5 <- predict(lm_5,data.frame(cocoa_percent=100,sugar="have_sugar"))
#p_6 <- predict(lm_5,data.frame(cocoa_percent=100,sugar="have_not_sugar"))

#library(cluster)
#normalization
#z <- chocolate_unpleasant[,c(6:8)]
#m <- apply(z,2,mean)
#s <- apply(z,2,sd)
#z <- scale(z,m,s)

#euclidean distance
#distance <- dist(z)
#print(distance,digits=3)
```

```
#cluster dendrogram with complete linkage
#hc.c <- hclust(distance)
#plot(hc.c,cex=0.7,labels=chocolate_outstanding$country_of_bean_origin,main="Cluster
Dendrogram of Bean Origin - Rated Outstanding")
#plot(hc.c,hang=-1)

#cluster dendrogram with complete linkage
#hc.c <- hclust(distance)
#plot(hc.c,cex=0.7,labels=chocolate_outstanding$company,main="Cluster Dendrogram of
Companies - Rated Outstanding")
#plot(hc.c,hang=-1)

######### does origin matter

mod = lm(rating~country_of_bean_origin,chocolate)
#code below checks standard deviation assumption of One Way Anova Test
locations = unique(data$country_of_bean_origin)
deviations = c()
for (i in 1:length(locations)) { deviations[i] = sd(data[data$country_of_bean_origin ==
locations[i], ]$rating);}
deviations
#line below checks normality assumption of One Way Anova Test
plot(qqnorm(aov(mod)$residuals))
anova(mod)
ggplot(data = meanRatingByYear, aes(x = ReviewDate,  y = Rating)) +
  geom_line() +
  scale_x_continuous(breaks = c(2006:2019)) +
  labs(title = "Mean Rating of Chocolate Bars over Years",
       x = "Review Year",
       y = "Mean Rating")
```

```
#### count of rating
ggplot(flavours, aes(x=Rating, fill=Rating,color=Blue)) +
  geom_bar(color="black") +
  scale_x_continuous(breaks=seq(0,5,0.5))+
  labs (x = "Rating",
        y = "Count",
        title = "Count of Ratings")

####

##fitting of linear model

model1 <- lm(Rating~ReviewDate,data=flavours)
sm1<-summary(model1)
sm1
#printing RMSE for model 1
mean(sm1$residuals^2)
#plotting residual plots for model 1
par(mfrow=c(2,2))
plot(model1)



ggplot(data=model1,
       aes(x=.fitted, y=.resid)) +
  geom_point( ) +
  geom_hline(yintercept=0) +
  geom_smooth(se=TRUE, method="loess",
              method.args=list(degree=1, family="symmetric")) +
  labs(x="fitted Values", y="Residuals",title="Residuals vs Fitted plot for predicting Rating given
ReviewDate")


####plot

meanRatingByYear <- flavours %>%
  group_by(ReviewDate) %>%
  summarise(Rating = mean(Rating))
print(meanRatingByYear)

ggplot(data = meanRatingByYear, aes(x = ReviewDate, y = Rating)) +
  geom_line() +
  scale_x_continuous(breaks = c(2006:2017)) +
  labs(title = "Mean Rating of Chocolate Bars over Years",
       x = "Review Year",
       y = "Mean Rating")

ggplot(flavours, aes(x=Rating, fill=Rating,color=Blue)) +
  geom_bar(color="black") +
  scale_x_continuous(breaks=seq(0,5,0.5))+
  labs (x = "Rating",
        y = "Count",
        title = "Count of Ratings")
```