Chocolate Bar Rating

- A comprehensive analysis of
  chocolate taste rating influenced by various factors

Jinghan Xu, Fengshen Zhou, Yihan Shi, Zuer Zheng, Xinyi Chen

Introduction:

Chocolate, a type of sweet food mainly made of cocoa products (including cocoa butter, cocoa powder or cocoa pulp) and sugar, tastes delicate accompanied with a strong aroma. As one of the most popular leisure snacks around the world, the global consumption of chocolate products reached 137.599 billions in 2019, and is expected to exceed 182.090 billions in 2025. (Businesswire,2020). Although most people always associate chocolate with diseases including obesity, diabetes and high blood pressure, researchers proved that the key ingredient of chocolate, cocoa, is actually heart-healthy. It might also prevent cognitive decline and lower cholesterol levels. Therefore, we are more interested in doing research for the taste of chocolate in order to better enjoy delicious chocolate.

To achieve the goal of the project, which is to explore the elements that may influence the taste of chocolate, we are going to use the Chocolate *Bar 2020* dataset collected by Soroush Ghaderi which carries information from 2006 to 2020 in 66 countries about chocolate reviews and tastes. The data set contains information on where the chocolate comes from, their cocoa percentage, the type of chocolate beans used, where the chocolate beans are grown, and expert ratings compiled by Brady Brelinski, who is the founding member of the Manhattan Chocolate Society, for more than 1,700 chocolate bars. We will analyze the associations between the chocolate rating and the factors including cocoa percentage, ingredients and so on.

Question of Interest

1. What factors contribute to the highest rating chocolate bars?
    a. Will the factors cocoa beans and chocolate ingredients affect the rating?
    b. Is there any specific type of cocoa bean that is more popular?
    c. What flavor/ingredients do people prefer when they choose chocolate?
2. Which model performs better in classification? LDA or logistic model?

Background:

Our source of the dataset is from the following link:
https://www.kaggle.com/soroushghaderi/chocolate-bar-2020?select=chocolate.csv

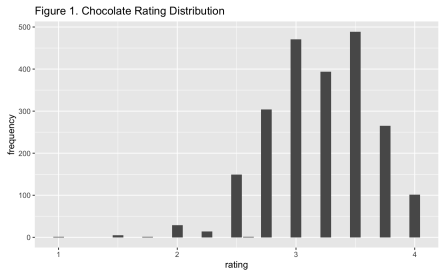
The variables in the dataset include REF(unique number for number), Company (Name of the company manufacturing the bar), Company_Location (The country where the company is located), ReviewDate (Date review for chocolate bar), Country_of_bean_origrin (Country of chocolate bean), Specific_Bean_Original_Bar Name (The specific province of chocolate bean),CocoaPercent (Cocoa percentage (darkness) of the chocolate bar being reviewed), Rating (Expert rating for chocolate bar), Counts_of_ingredients (Number of ingredients used), Beans/Cocoa_butter/Vanilia/Lecithin/Salt/Sugar (Whether chocolate has these ingredients or not), First/Second/Third/Fourth_taste (Taste of chocolate in different times).

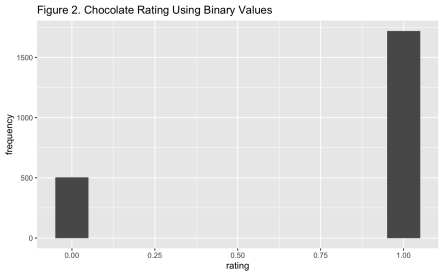
The Cocoa Rating system is divided into 4 levels to tell how great the taste is:

- 3.5 ~ 4 = Highly Recommended;
- 3.0 ~ 3.49 = Recommended;
- 2.0 ~ 2.9 = Disappointing;
- 1.0 ~ 1.9 = Unpleasant.

First we'll take a look at our rating scores' distribution as shown in Figure 1. From the figure, we can observe that most of the data are distributed between 2.75 and 4, while the weight of ratings under 2.5 and beyond 4.0 is quite light. Our distribution figure indicates that on average, gourmets are satisfied with chocolates and think they are neutral and **recommended**.

For easier readability, we will transform our rating scores into **binary** variables. If the rating score is less than 3, it will be classified as 0, meaning these chocolates are unrecommended. If the rating score is greater than or equal to 3, it will be classified as 1, meaning it is recommended or even highly recommended. The distribution of binary ratings is shown in Figure 2.





From figure 2, we can see that the proportion of high rating chocolates(binary value 1) is much higher than the low rating ones(binary value 0), which again shows that chocolates are **appealing** to people on average.

Data Analysis:

In this part, we will analyze our dataset in detail by focusing on the **potential factors that influence chocolate rating** as well as the **relationship** between chocolate rating, cocoa percent, and counts of ingredients. Time influences will also be used in order to have more precise results.

Principal factors that may influence the rating of chocolate

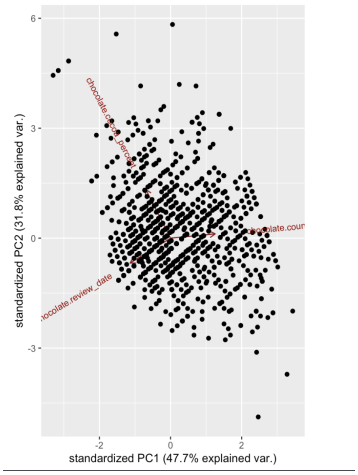
For our question of interest 1, we will make an hypothesis that the **cocoa bean cocoa percent** , **ingredients counts**, and review date have a significant effect on the chocolate rating. To test our idea, we choose to use the ****PCA** method** to find the principal components.

```
Standard deviations (1, .., p=3):
[1] 1.1967231 0.9772536 0.7828341

Rotation (n x k) = (3 x 3):
                                PC1        PC2        PC3
chocolate.cocoa_percent         -0.3796461  0.88675785 -0.2636843
chocolate.counts_of_ingredients  0.6855608  0.07828211 -0.7237944
chocolate.review_date           -0.6211886 -0.45555727 -0.6376459
```

```
Importance of components:
                                 PC1    PC2    PC3
Standard deviation     1.1967 0.9773 0.7828
Proportion of Variance 0.4774 0.3183 0.2043
Cumulative Proportion  0.4774 0.7957 1.0000
```

Considering rotations of PC1, cocoa_percent and review_date are inversely related to the first principal component(which has a negative value in rotation of PC1), when the counts_of_ingredients is positively related to it. The proportion of variance indicates how much of total variance is there in variance of a particular principal component. Hence, PC1 variability explains 47.74% of total variance of the data, when the PC2 variability explains 31.83% of total variance of the data. These first two proportion values are similar when the first proportion of variance is a bit larger.

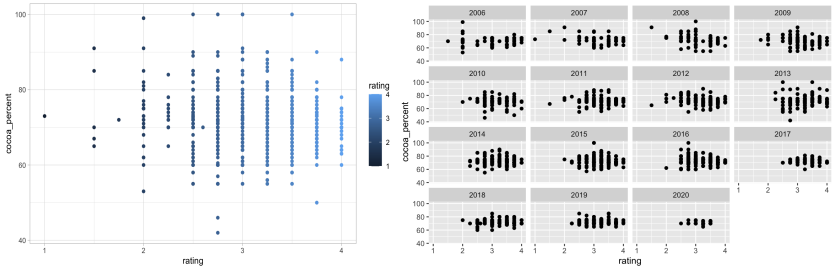
We can also obtain the ggbiplot to check out the result.



From the result of ggbiplot we can see that based on the plot of standardized PC1, we can conclude that the **cocoa_percent, counts_of_ingredients and the review_date are all relatively important effects on the first principal component.** And since all three vectors all have a large angle, they might **not have a close relationship** with each other. And we will do further analysis for cocoa_percent and counts_of_ingredients within the review_date.

a. Association **between rating and cocoa percent**

From the previous part, we obtain that there is an association between cocoa percentage and chocolate ranking. Therefore, we decided to further analyze how the cocoa percent is related to the chocolate rating. Firstly, we used a scatter plot to show the relationship between these two variables.



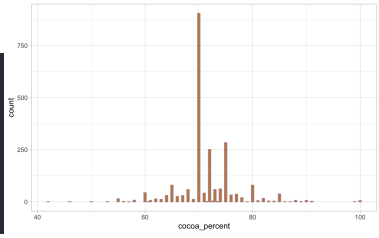
From the plots, we notice that when the most cocoa_percent of chocolate is **between 55% and 90% overall.** When rating is high, the range will be mainly 60% to 80%. Another fact we find is that **the cocoa_percent usually does not go below 50%,** and there are only 3 points that are below or equal to 50% in the data with over 2500 ratings.

In addition, we think there might be a change of rating depending on cocoa_percent as the time goes by. With the variable reviewDate, we separated the plot into several parts depending on the year. As shown in Figure 6, we can observe that from 2006 to 2020, the

cocoa percent range with high ratings is relatively stable, which is still **60% to 80%**. This agrees with our previous results.

To support what we observe in the previous plots, we will use cluster sampling procedure. We will split the chocolate ratings into 4 unsupervised clusters within rating score and cocoa percent. We then calculate their average rating and the corresponding average cocoa percent.

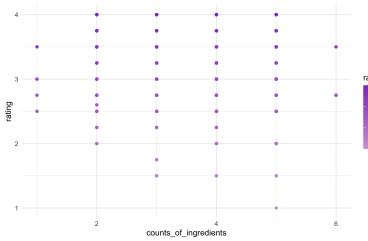
	rating	cocoa_percent
1	3.138776	62.45102
2	3.021186	83.31638
3	3.249885	70.25844
4	3.156627	74.97590



Using the cluster sampling, we get the result that the cluster cocoa_percent mean of 70.25844 has the highest cluster rating mean 3.249885. And the cluster cocoa_percent mean of 62.45102 and the cluster cocoa_percent mean of 74.97590 have the cluster rating mean 3.138776 and the cluster rating mean 3.156627 respectively, which are similar. When the cocoa_percent mean reaches 83.31638, the lowest cluster mean is 3.0211865, which is the lowest rating within these 4 clusters. Therefore, we think our result supports our statement that it is more possible for the chocolate with cocoa_percent around 70% to get a higher rating of chocolate.
Also, from the distribution of the cocoa_percent bar, we can observe that the frequency of the 70% cocoa-percent is much higher than others; meanwhile, the frequency gradually declines as the percentage of cocoa goes beyond or goes down. Based on observation, we predict that the companies may notice that the chocolate of 70% cocoa rate may be acceptable to people.

b. <u>Association between the rating and counts of ingredients</u>

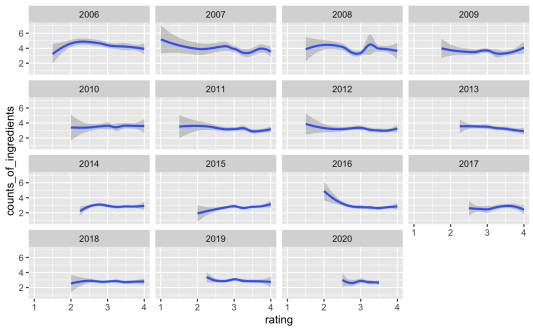
As we mentioned, the counts of ingredients are also an important component for the chocolate rating. Firstly, we use a scatter plot with ggplot2 to obtain the relationship between ingredient counts and rating. From Figure 8, we can conclude that the main range of counts_of_ingredients is from 2 to 5, no matter what rating it is. Also, we observe that when rating is high, the counts of ingredients will never be too high or too low. So we will choose a cluster sampling method to do more tests. Using the kmean function to randomly assign 10 unsupervised clusters.




```
K-means clustering with 10 clusters of sizes 58, 397, 137, 285, 337, 285, 324, 175,
126, 100

Cluster means:
      rating counts_of_ingredients
1  2.443966                5.051724
2  3.385390                3.000000
3  3.344891                5.007299
4  3.386842                1.996491
5  2.829377                3.000000
6  2.801228                1.985965
7  3.331019                4.000000
8  3.825714                3.000000
9  2.565476                4.000000
10 3.805000                2.000000
```

From the 10 cluster samplings, we can notice that the chocolate with 3 counts_of_ingredients has the highest rating mean 3.825714; the chocolate with 5 counts_of_ingredients has the lowest rating mean 2.44396. At the same time, the chocolate with 1 or 6 ingredients does not exist in our samples, which also agrees with our point that the proportion of the chocolate with range of counts_of_ingredients from 2 to 5 is large.

Additionally, a change of rating depending on count_of_ingredients may appear as the time goes by as well. We separated the plot into several parts depending on the count_of_ingredients with the variable reviewDate. The result is shown in Figure 8.



From the plot we can see that the both counts of_ingredients and the chocolate rating are stabilized. The chocolates usually use around 3 ingredients; meanwhile, the ratings of the chocolate is on a scale of 2.5 to 3.5, which indicates that the chocolate tastes are neither disappointing or fantastic.

Classification

After analyzing the cocoa percent and counts of ingredients, we will now do the classification using those two variables for prediction. We will try two different models here, the LDA model and the logistic model. We will answer our question of interest 2 here, which asks for a better classification model. Selection of the better model will be based on their accuracy. Besides, we will also find out whether both factors are necessary for classification or not.

Before we do the classification setup, we will first split the dataset with binary ratings into train data and test data with ratio 80%:20%. After splitting, the frequency of train data is:

Var1	Freq
0	403
1	1376

The frequency of test data is:

Var1	Freq
0	101
1	344

I. LDA Model

First we will create the LDA model with cocoa_percents and counts_of_ingredients as predict factors, and the Rating as response factor. The results of the LDA model is:

```
Call:
lda(factor(Rating) ~ cocoa_percent + counts_of_ingredients, data = train)

Prior probabilities of groups:
        0         1
0.2265318 0.7734682

Group means:
  cocoa_percent counts_of_ingredients
0      71.75310              3.183623
1      71.36919              3.040698

Coefficients of linear discriminants:
                            LD1
cocoa_percent         -0.110344
counts_of_ingredients -1.012864
```

By calculation, its accuracy is 0.7730337.

II. Logistic Model

The results of the logistic model is:

```
Call:  glm(formula = factor(Rating) ~ cocoa_percent + counts_of_ingredients,
    family = binomial(link = "logit"), data = train)

Coefficients:
          (Intercept)          cocoa_percent  counts_of_ingredients
              3.23046               -0.01996               -0.18453

Degrees of Freedom: 1778 Total (i.e. Null);  1776 Residual
Null Deviance:      1904
Residual Deviance: 1893         AIC: 1899
```

By calculation, its accuracy is 0.7730337.

By comparison, we can see that both counts of ingredients and cocoa percent have a greater influence on chocolate rating in the LDA model than in the logistic model. The two models' accuracy, though, are the same, so we can use both models for prediction.

Besides analyzing different classification models,we will also determine whether both predictors are necessary for the purpose of classification.

We will use the p-value calculated from the logistic model to analyze the necessity of two predictors. The outcomes are:

```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           3.23046    0.83492   3.869 0.000109 ***
cocoa_percent        -0.01996    0.01081  -1.846 0.064890 .
counts_of_ingredients -0.18453    0.06122  -3.014 0.002574 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1903.7  on 1778  degrees of freedom
Residual deviance: 1893.0  on 1776  degrees of freedom
AIC: 1899
```

We can see that the p-value of cocoa percent is 0.064890, and the p-value of counts of ingredients is 0.002574. By using a significant level of 5%, we can conclude that cocoa percent is not statistically significant at this level, while counts of ingredients is statistically significant. Thus, we will conclude that count of ingredients is necessary for classification but cocoa_percent is not.

Then, we will also take a look at the model without cocoa percent:

```
Call:  glm(formula = factor(Rating) ~ counts_of_ingredients, family = binomial(link = "logit"),
    data = train)

Coefficients:
          (Intercept)  counts_of_ingredients
                1.735                 -0.163

Degrees of Freedom: 1778 Total (i.e. Null);  1777 Residual
Null Deviance:      1904
Residual Deviance: 1896         AIC: 1900
```

The new model's accuracy is still 0.7730337 by calculation. Thus, it is not necessary to include cocoa_percent. We only need counts_of_ingredients in the classification which shows the same performance with the full model.
With this outcome, we will continue to analyze the influence of each ingredient.

Ingredients

From the previous part we have known that ingredients have an essential influence on the chocolate bar rating since it may affect the taste of chocolate. We Also observed that the less ingredients chocolate bar has led to the trend of the healthy diet. Therefore, we will also be interested in observing which ingredients that people would prefer when they choose chocolate.

1. **Beans**

i. Beans

```
table(chocolate_binary$rating, chocolate_binary$beans) %>%
  kbl() %>%
  kable_material(c("striped", "hover"))
```

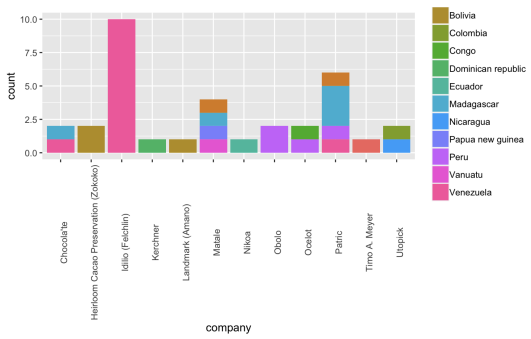
	have_bean
0	504
1	1720

```
 [1] "Chocola'te"                        "Heirloom Cacao Preservation (Zokoko)"
 [3] "Idilio (Felchlin)"                 "Kerchner"
 [5] "Landmark (Amano)"                  "Matale"
 [7] "Nikoa"                             "Obolo"
 [9] "Ocelot"                            "Patric"
[11] "Timo A. Meyer"                     "Utopick"
```

Remember that in the Background part, we have found out that the total number of 0's and 1's are 504 and 1720 respectively. Thus, we can conclude from the above table that all chocolate bars are made with cocoa beans.

However, there are quite different types of cocoa beans, and they come from many different locations. Our next step is to find out cocoa bean origins used by chocolate companies with high average ratings.

We first find out companies that have an average rating score greater than or equal to 3.75. Those companies are:



We then create a bar plot to observe where their beans are originating from. The result is shown in Figure X.

From the bar plot, we can see that there are a total of 11 bean origins. Among them, beans from Venezuela are used the most frequently. For example, all 10 chocolates from company Idilio are made with cocoa beans from Venezuela. Besides, beans from Madagascar and Peru are also used very frequently. We will thus recommend using cocoa beans from Venezuela, Madagascar, and Peru.

2. Other ingredients(cocoa_butter, vanilla, lecithin, salt, sugar, sweetener)

We can use the logistic regression to determine each ingredients' impact on chocolate bar ratings:

We will choose Rating(the binary response) as our y, and all the ingredients as our predictors.

From the glm function result we can see that the coefficient of **have not cocoa butter** is -0.26416, and the co coefficient of **have sugar** is 0.58138. This indicates that the ingredients

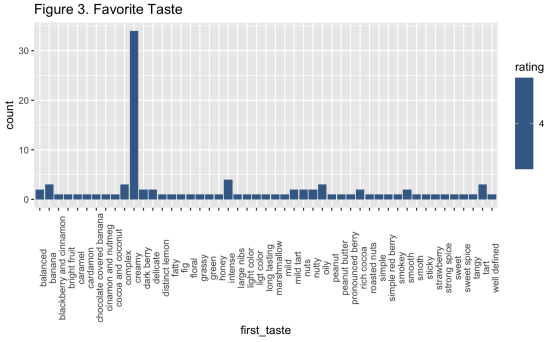
```
Call:  glm(formula = chocolate_binary$Rating ~ chocolate_binary$cocoa_butter +
    chocolate_binary$vanilla + chocolate_binary$lecithin + chocolate_binary$salt +
    chocolate_binary$sugar + chocolate_binary$sweetener_without_sugar,
    family = binomial, data = chocolate_binary)

Coefficients:
                                                  (Intercept)
                                                      0.81220
                    chocolate_binary$cocoa_butterhave_not_cocoa_butter
                                                     -0.26416
                                chocolate_binary$vanillahave_vanila
                                                     -0.74082
                           chocolate_binary$lecithinhave_not_lecithin
                                                      0.10278
                                    chocolate_binary$salthave_salt
                                                     -0.02589
                                  chocolate_binary$sugarhave_sugar
                                                      0.58138
chocolate_binary$sweetener_without_sugarhave_sweetener_without_sugar
                                                     -0.07764

Degrees of Freedom: 2223 Total (i.e. Null);  2217 Residual
Null Deviance:      2380
Residual Deviance: 2340         AIC: 2354
```

cocoa_butter and sugar have a positive correlation with the binary Rating result, and the effectiveness of sugar is larger since the coefficient is larger.

The coefficient of **have_vanilla** is -0.74082, which is a large proportion of the negative correlation coefficient for the binary Rating result. So we can conclude that adding vanilla will reduce the rating of the chocolate. The coefficient of **have_not_lecithin ** is 0.10278, **have_salt** is -0.02589 and **have_sweetness* is -0.07764, which also has a negative correlation to the binary Rating result, but these tree coefficients are quite small, we can consider lecithin, salt and sweetness might not affect the taste experience too much maybe because those are not common flavors for chocolate choosing, only those people who wants something special will choose salt or sweetness for their chocolate.

We also analyze the chocolate taste that receives the most frequent high ratings . Since there is a great loss in our data of second_taste, third_taste, and fourth_taste data, we'll mainly draw suggestive conclusions through data in the first taste. Here, we will use a bar plot to plot the frequency of receiving a rating of 4 for all tastes in **first_taste**. The results are shown in the following figure :

Figure 3. Favorite Taste



We observe that **creamy taste** receives much more rating of 4 than all other tastes. Thus, we can make a hypothesis that for most people, they prefer creamy-flavor chocolate, and adding cocoa_butter into the chocolate can improve the creamy concentration. The observation result is matching the taste survey. And it is also confirmed that there are less ingredients people prefer when they choose chocolate.

Conclusion

In general, people's attitudes towards chocolate are positive and recommondented based on the fact that there are huge amounts of rating more than 3, meaning recommended or highly recommended. In the project we focus on how the potential factors (including cocoa_percent, counts_of_ingredients and review data) and ingredients influence chocolate rating. To analyze the potential factors, according to the PCA methods and biplot, we concluded that the cocoa_percent, counts_of_ingredients are relatively important elements and related the rating within different review_date; however, these factors do not have correlationships with each other. To further analyze each factor's influence for the response variable rating, we use the scatter plot of the individual variable and response, along with the cluster sampling. We speculate that about 70% cocoa_percent is customers' inclination; meanwhile, delicious chocolate is usually made from around three ingredients. Additionally, we try to fit the LDA model and the best logistic model to predict classification with higher accuracy. We obtain that both models have the same accuracy, and only the factor of counts_of_ingredients is necessary. In ingredient analysis, we are curious about what ingredients can raise the rating of chocolate and what can reduce the rating. Firstly, since the cocoa beans are the main component in the chocolates, we analyze the chocolate beans separately from the others, suggesting that cocoa beans from Venezuela, Madagascar, and Peru are more popular. After that, we research other ingredients (cocoa_butter, vanilla, lecithin, salt, sugar, sweetener) based on the glm function. Our result shows that the best way to produce customers' favorite flavor chocolate is by using cocoa butter and sugar, while vanilla, lecithin, salt, and sweetener without sugar should not be added. Finally, we expand the research and find that the chocolate bar gains the highest points when first_taste is creamy. Thus, butter can be considered as a key factor, which contributes to the "creamy taste" with a higher rating. The summary table of our analysis is the following:

Best Way to Produce Chocolate Bar

elements	method
cocoa butter	use cocoa butter
vanillar	do not use vanilla
lecithin	do not use lecithin
salt	do not use salt
sugar	use sugar
sweetener without sugar	do not use sweetner without sugar
count of ingredients	use three ingredients: cocoa bean, butter, and sugar
cocoa beans	use cocoa beans from Venezuela, Madagascar, and Peru

Reference:

Businesswire.(2020). “Global Chocolate Market - Forecasts from 2020 to 2025”. Retrieved from https://www.businesswire.com/news/home/20201207005451/en/Global-Chocolate-Market-Report-2020-Market-to-Reach-US182.090-Billion-by-2025-Increasing-from-US137.599-Billion-in-2019---ResearchAndMarkets.com

Code Appendix

```
library(kableExtra)
library(ggplot2)
library(ggbiplot)
library(dplyr)
library(MASS)
library(devtools)
chocolate <- read.csv("~/Desktop/chocolate.csv")
attach(chocolate)
chocolate_binary <- chocolate
chocolate_binary$Rating <- ifelse(chocolate$rating>= 3,1,0)
#chocolate_binary$rating <- ifelse(chocolate$rating>= 3,1,0)
freq.rating <- data.frame(rating = c(1.00, 1.50, 1.75, 2.00, 2.25, 2.50,
                                     2.60, 2.75, 3.00, 3.25, 3.50, 3.75, 4.00),
                          frequency = c(length(chocolate$rating[chocolate$rating == 1]),
                                        length(chocolate$rating[chocolate$rating == 1.5]),
                                        length(chocolate$rating[chocolate$rating == 1.75]),
                                        length(chocolate$rating[chocolate$rating == 2]),
                                        length(chocolate$rating[chocolate$rating == 2.25]),
                                        length(chocolate$rating[chocolate$rating == 2.5]),
                                        length(chocolate$rating[chocolate$rating == 2.6]),
                                        length(chocolate$rating[chocolate$rating == 2.75]),
                                        length(chocolate$rating[chocolate$rating == 3]),
                                        length(chocolate$rating[chocolate$rating == 3.25]),
                                        length(chocolate$rating[chocolate$rating == 3.5]),
                                        length(chocolate$rating[chocolate$rating == 3.75]),
                                        length(chocolate$rating[chocolate$rating == 4])))


freq.rating_binary <- data.frame(rating = c(0,1),
                                 frequency = c(504,1720))
par(mfrow=c(1,2))
```

```
Figure1 <- ggplot(freq.rating)+
 geom_col(aes(x=rating,y=frequency))+
 ggtitle("Figure 1. Chocolate Rating Distribution")

Figure2 <- ggplot(freq.rating_binary)+
 geom_col(aes(x=rating,y=frequency),width = 0.1)+
 ggtitle("Figure 2. Chocolate Rating Using Binary Values")



gridExtra::grid.arrange(Figure1,Figure2,ncol=1)
# TASTE

# first taste
taste1 <- chocolate %>% group_by(first_taste) %>% filter(rating == 4)
ggplot(taste1,aes(x = first_taste, fill = rating)) +
 geom_bar()+
 theme(axis.text.x = element_text(angle = 90))+
 ggtitle("Figure 3. Favorite Taste")

#second taste
#taste2 <- chocolate %>% group_by(second_taste) %>% filter(rating == 4)
#ggplot(taste2,aes(x = second_taste, fill = rating)) +
#  geom_bar()+
#  theme(axis.text.x = element_text(angle = 90))

#third taste
#taste3 <- chocolate %>% group_by(third_taste) %>% filter(rating == 4)
#ggplot(taste3,aes(x = third_taste, fill = rating)) +
#  geom_bar()+
# theme(axis.text.x = element_text(angle = 90))

#fourth taste
#taste4 <- chocolate %>% group_by(fourth_taste) %>% filter(rating == 4)
#ggplot(taste4,aes(x = fourth_taste, fill = rating)) +
#  geom_bar()+
#  theme(axis.text.x = element_text(angle = 90))
chocolate_rating <- data.frame(chocolate$cocoa_percent,
                               chocolate$counts_of_ingredients,
                               chocolate$review_date)
pr.out<-prcomp(chocolate_rating ,scale=TRUE)
pr.out$rotation
summary(pr.out)
x<- chocolate_binary %>%filter(Rating=="1")%>%dplyr::select(c(counts_of_ingredients,cocoa_percent))
#install_github("vqv/ggbiplot")
ggbiplot(pr.out)+ggtitle("Figure 4. Biplot")
#rating and cocoa_ percent
```

```
Figure5 <- ggplot(data=chocolate_binary,aes(y=rating,x=cocoa_percent,color = rating))+
 geom_point()+
 coord_flip()+
 ggtitle("Figure 5. Rating vs. Chocolate Percent")

Figure6 <- ggplot(data=chocolate_binary)+
 geom_point(aes(x=rating,y=cocoa_percent))+
 facet_wrap(~review_date)+
 ggtitle("Figure 6. Rating vs. Chocolare Percent by Date")

Figure5


Figure6
#gridExtra::grid.arrange(Figure5,Figure6,ncol=1)
#cluster--rating vs. cocoa_percent
set.seed(15)
km_data<-data.frame(chocolate_binary$rating,chocolate_binary$cocoa_percent)
km.out<-kmeans(km_data,4,nstart=100)
km.out$centers

ggplot(chocolate,aes(y=rating,x=counts_of_ingredients,color = rating)) +
 geom_point()+
 ggtitle("Figure 7. Ingredients vs. Rating")+
 coord_flip()
ggplot(data=chocolate_binary)+
  geom_smooth(aes(x=rating, y=counts_of_ingredients))+
  facet_wrap(~review_date)+
 ggtitle("Figure 8. ")

set.seed(141)
km_data2 <- data.frame(chocolate_binary$rating,chocolate_binary$counts_of_ingredients)
km.out2<-kmeans(km_data2,10,nstart=100)
km.out2$centers
set.seed(1)

data1  <- chocolate_binary[chocolate_binary$Rating == 0, ]
data2  <- chocolate_binary[chocolate_binary$Rating == 1, ]

id <- sample(1:nrow(data1), 0.8 * nrow(data1))

data11 <- data1[id, ]
data12 <- data1[-id,]
id <- sample(1:nrow(data2), 0.8 * nrow(data2))

data21 <- data2[id, ]
data22 <- data2[-id, ]
```

Commented [1]: 新改好的.

Commented [2]: OKOKOKOKOKOK

```
train <- rbind(data11, data21)
test <- rbind(data12, data22)
table(train$Rating)%>%
 kbl() %>%
 kable_styling()
table(test$Rating)%>%
 kbl() %>%
 kable_styling()
#lda model setip
lda_fit <- lda(factor(Rating) ~ cocoa_percent +  counts_of_ingredients, data = train)
lda_fit
p1 <- predict(lda_fit, test)
acc <- mean(p1$class == test$Rating)
acc
# Logistic Model

logit_fit <- glm(factor(Rating) ~ cocoa_percent + counts_of_ingredients, data = train,
                 family = binomial(link = "logit"))
logit_fit
p2 <- predict(logit_fit, test, type = "response")

acc2 <- mean(round(p2) == test$Rating)
acc2
summary(logit_fit)$coefficients
logit_fit3 <- glm(factor(Rating) ~ counts_of_ingredients , data = train,
                 family = binomial(link = "logit"))
logit_fit3
p <- predict(logit_fit3, test, type = "response")

acc3 <- mean(round(p) == test$Rating)
acc3

table(chocolate_binary$Rating, chocolate_binary$beans) %>%
 kbl() %>%
 kable_material(c("striped", "hover"))
company_rating <- chocolate %>% group_by(company) %>% filter(mean(rating) >= 3.75)
unique(company_rating$company)

df <- chocolate%>% filter(company==c("Chocola'te","Idilio (Felchlin)","Landmark (Amano)",
                                     "Nikoa","Ocelot","Timo A. Meyer",
                                     "Heirloom Cacao Preservation (Zokoko)",
                                     "Kerchner","Matale","Obolo","Patric","Utopick" ))
ggplot(company_rating,aes(x=company,fill=country_of_bean_origin))+
 geom_bar() +
 theme(axis.text.x = element_text(angle = 90))+
 theme(legend.title = element_text(size = 0.001))
n.butter <- table(chocolate_binary$Rating, chocolate_binary$cocoa_butter)
n.butter%>%
```

```
 kbl() %>%
 kable_material(c("striped", "hover"))
plot(n.butter)
n.vanilla <- table(chocolate_binary$Rating, chocolate_binary$vanilla)
n.vanilla %>%
 kbl() %>%
 kable_material(c("striped", "hover"))
plot(n.vanilla)
n.lecithin <- table(chocolate_binary$Rating, chocolate_binary$lecithin)
n.lecithin %>%
 kbl() %>%
 kable_material(c("striped", "hover"))
plot(n.lecithin)
n.salt <- table(chocolate_binary$Rating, chocolate_binary$salt)
n.salt %>%
 kbl() %>%
 kable_material(c("striped", "hover"))
plot(n.salt)
n.sugar <- table(chocolate_binary$Rating, chocolate_binary$sugar)
n.sugar %>%
 kbl() %>%
 kable_material(c("striped", "hover"))
plot(n.sugar)
n.sweetener <- table(chocolate_binary$Rating, chocolate_binary$sweetener_without_sugar)
n.sweetener %>%
 kbl() %>%
 kable_material(c("striped", "hover"))
plot(n.sweetener)

glm(chocolate_binary$Rating ~ chocolate_binary$cocoa_butter
   +chocolate_binary$vanilla
   +chocolate_binary$lecithin
   +chocolate_binary$salt
   +chocolate_binary$sugar
   +chocolate_binary$sweetener_without_sugar,
   data = chocolate_binary, family = binomial)
best_com <- data.frame(elements = c("cocoa butter",
                                    "vanillar",
                                    "lecithin",
                                    "salt",
                                    "sugar",
                                    "sweetener without sugar",
                                    "count of ingredients",
                                    "cocoa beans"),
                       method = c("use cocoa butter",
                                  "do not use vanilla",
                                  "do not use lecithin",
                                  "do not use salt",
```

```
                    "use sugar",
                    "do not use sweetner without sugar",
                    "use three ingredients: cocoa bean, butter, and sugar",
                    "use cocoa beans from Venezuela, Madagascar, and Peru"))
best_com %>%
 kbl(caption = "Best Way to Produce Chocolate Bar") %>%
 kable_material(c("striped", "hover"))
best_com%>%
 kbl(caption = "Best Way to Produce Chocolate Bar") %>%
 kable_classic(full_width = F, html_font = "Cambria")
```