Bean-to-Bar Chocolate Analysis

DATS 6101, Project 1

By Kate Jones, Purvi Thakor, and Akshay Kamath

Group name: **Wubba lubba dub dub**

Index

Summary:

Chocolate means different things to different people: it can be a special treat, a guilty pleasure, or a delicacy to be thought over and evaluated much like wine and beer. But for many people around the world, it is also a serious industry. It is well known that cacao originated in Central America over 5000 years ago, but its popularity and production has spread globally. Chocolate is produced by companies around the world, leading to almost 2000 different types of chocolate bars.

Bean-to-Bar is a process which involves numerous steps to transform a cacao bean into chocolate. These bars have been rated by experts on a scale from 1 to 5. There are many potential influences on this rating, including the origin of the chocolate, the percentage of cacao it contains, and the origin of the beans used in production.

We obtained a Bean-to-Bar chocolate dataset which was compiled by Brady Brelinski who is the founding member of the Manhattan Chocolate Society. This dataset included the ratings of 1795 chocolate bars. We had data on 9 variables for each bar, as seen by finding the dimensions of the dataset.

Flavors of Cacao Rating System:

- 5= Elite (Transcending beyond the ordinary limits)
- 4= Premium (Superior flavor development, character and style)
- 3= Satisfactory (3.0) to praiseworthy(3.75) (well made with special qualities)
- 2= Disappointing (Passable but contains at least one significant flaw)
- 1= Unpleasant (mostly unpalatable)

The users were asked to rate the chocolate bars on a scale of 1 to 5, with 1 being unpleasant & 5 being excellent, on a set of parameters - Flavor, Texture, After Melt & Overall. Not all the bars in each range are considered equal, so to show variance from bars in the same range rating with .25, .50 or .75 have been assigned.

Development of our research question:

After examining the different variables in the data, we were interested to see if there were factors which contributed to a chocolate bar receiving a higher rating. We decided potential influences could include the percentage of cacao in the chocolate bar, the origin of the cacao beans used in production, the company of production, and the country of origin.

Our question focused in on the country of origin, and how that determines the quality of a chocolate bar. We often hear of Belgian chocolates being the best, but we were curious, is this actually true, based on the data? This led us to settle on the question, *Is there a relationship between a chocolate bar's country of origin and its rating?*

When answering this question, we also considered a few related questions, to ensure our results were accurate. Should we consider all countries, or only those who produce a large volume of chocolates? How do other factors, like cacao percentage, influence rating as well, and what is the interaction between these two factors? Does the origin of the cacao matter? By addressing these concerns throughout our exploratory data analysis, we were able to further hone in on answering our original question, about the relationship between origin and rating.

Our question was formulated to meet the standards of a SMART goal:

Specific: The question focuses on the relationship between two specific variables- where chocolate is from, and how high its quality is.

Measurable: We have data on expert ratings and country, so we can measure how the two are related. Since the ratings are numeric, it is possible to use them as a dependent variable in ANOVA analysis, and determine if the differences between countries are statistically significant.

Answerable: Our question can either be answered yes (if the two variables have a high correlation, based on ANOVA analysis) or no (if there is no relationship).

Relevant: Pricing and trade of chocolates is largely based on the perceived quality of certain regions (ie Belgium), so determining if these quality differences are legitimate is an important consideration for the chocolate market.

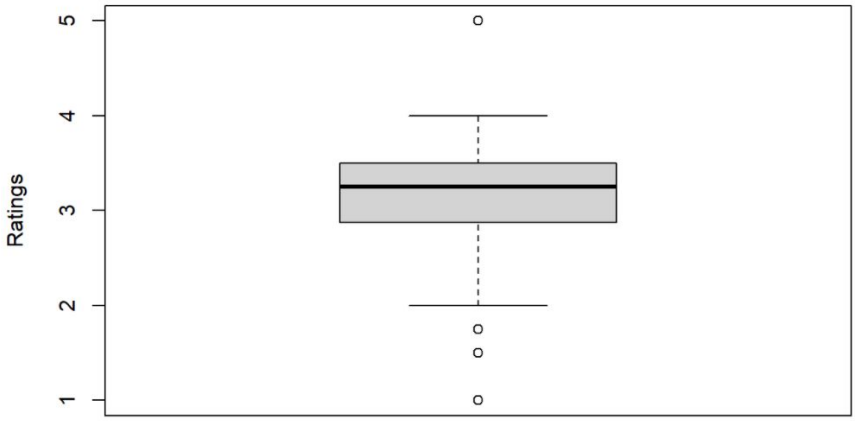
Timely: Data is already collected, so the question can be answered quickly and effectively.

Descriptive Statistics & Exploratory Data Analysis:

After establishing a question, we set about conducting exploratory data analysis (EDA) in R. Data comprises of 1795 observations from across 416 different companies which are at 60 different locations and there are 101 different regions from which these companies get their cacao beans from. According to basic EDA, we found that the most (764) types of chocolates came from the US. Soma, the Canadian chocolate company, has the highest number of products (47) rated in this survey. 57 companies (the highest) get their cacao beans from Madagascar.

Since we were interested in looking at how ratings are affected, we took a basic boxplot which displayed that our data did not have a lot of outliers.

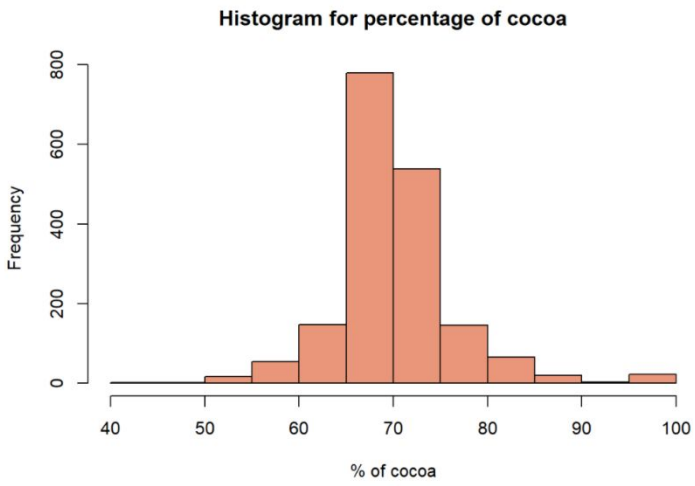
```
boxplot(chocolate$Rating, ylab = "Ratings", col="light gray")
```



Also, on an average, most of the chocolates were rated at just around 3. One more thing that we noted was that chocolates above the rating of 3.5 are not too frequent, which means that these are genuine ratings and not overwhelmingly common.

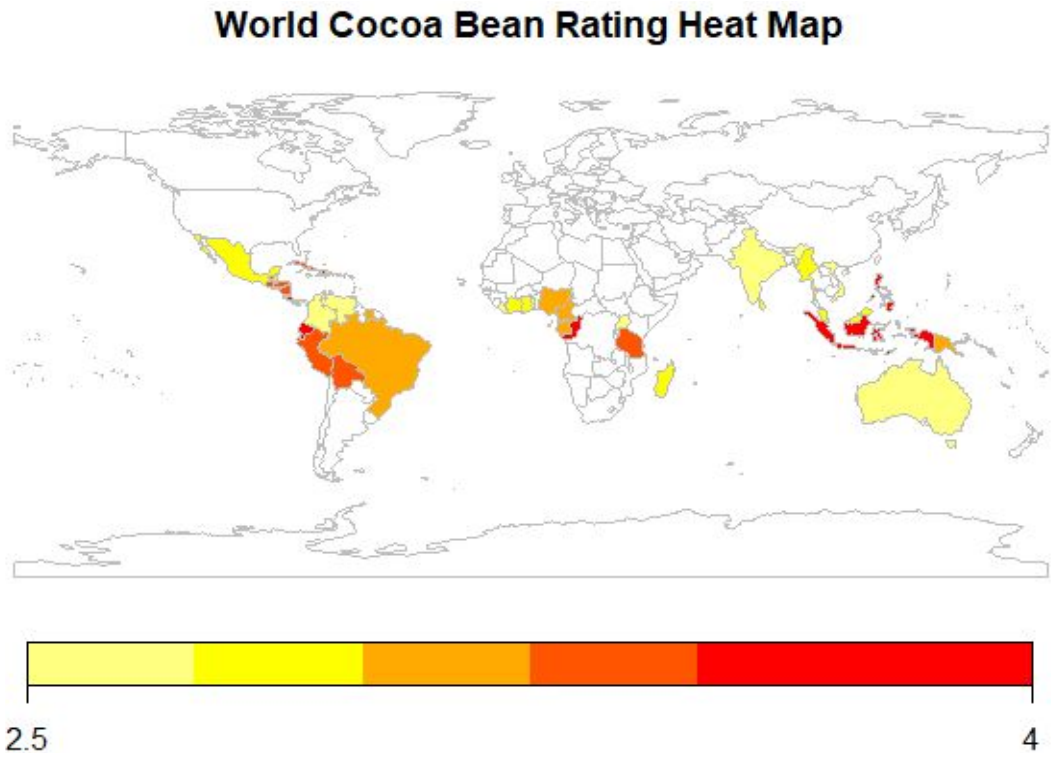
We then plotted a histogram to check where our data was concentrated.

```
#Producing a histogram for cocoa percentage
Cocoa_Percentage<-100*chocolate$Cocoa_percent
hist(Cocoa_Percentage, main = "Histogram for percentage of cocoa", xlab = "% of cocoa", col= "darksalmon", border = "black")
```



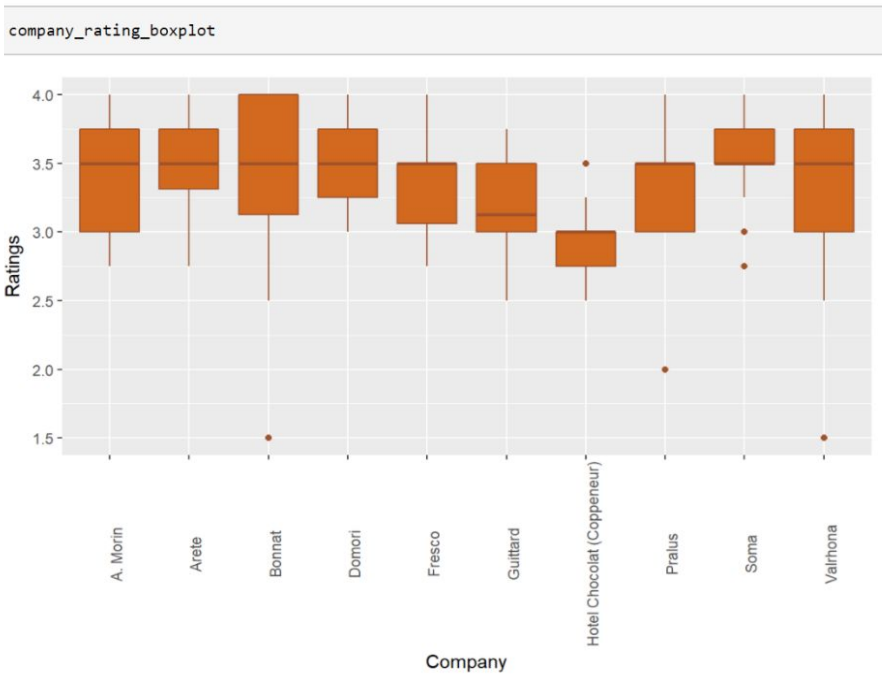
We observed that most of the chocolates in the survey consisted of around 70% of dark chocolate.

We then wanted to see if there was any relationship between where beans were sourced from, and the rating received. So, we created a world map.



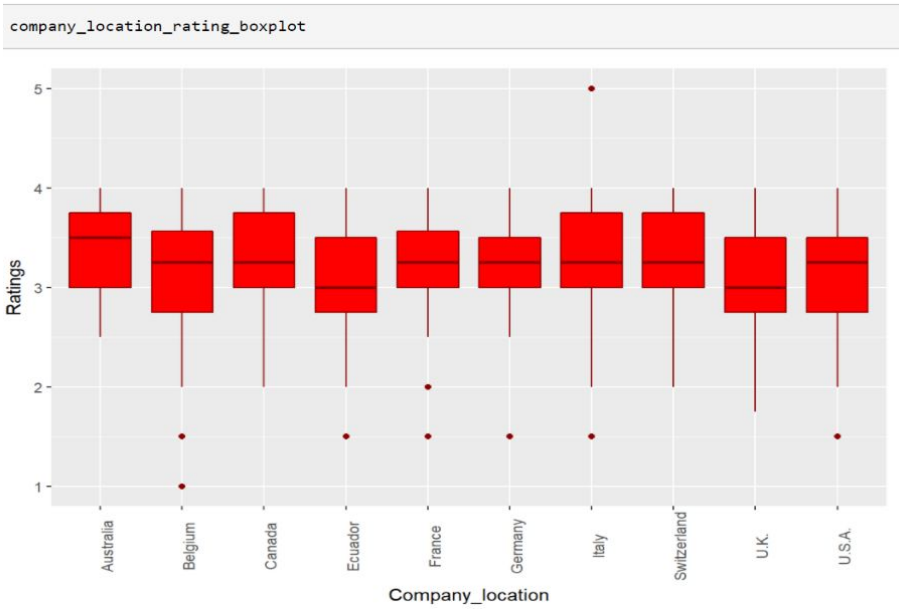
As per the map, we can see that the highest rated chocolates have beans originating from Indonesia in Asia, Republic of Congo in Africa & Ecuador in South America.

To see how the ratings were distributed across companies which produced the chocolates, we took a list of companies with most number of chocolates rated and plotted a box plot of ratings against the companies.



All ten chocolate companies have more or less the same median rating (surprisingly!). However, Bonnat, the French chocolatier company, has 27 chocolates rated & 8 of their chocolates have received a 4 rating.

We also wanted to see how the ratings were distributed across Countries from where the chocolates were produced. We took a list of countries with most number of chocolates rated in this survey and plotted a box plot of ratings against the countries.



Australia has the highest median rating according to this box plot with an median rating of 3.5.

We then wanted to determine whether the relationship between rating and location was statistically significant. Hence, we used a One-way ANOVA for checking whether this relationship was significant. We can see that the p-value was statistically significant, as it was well below an alpha-level of 5%. This implies that there is a significant relationship between the Location Factor and how well a chocolate bar rates. But, we wanted to see whether it was worth considering how other factors may have contributed to this rating. Hence, we did a Two-way ANOVA to see if cacao Percentage could also contribute to this rating. Again, we got results which indicated that there is a statistical significance of location, even after controlling for another factor (cacao percentage) and their interaction. We also found it interesting to note that the impact of cacao Percent on Rating appears to be statistically significant. However, a quick check of correlation finds almost no relationship between Percent and Rating!

Limitations of the dataset

- Data in question has almost 1800 entries. However, we do not have information about the number of people who rated each chocolate.
- Data has a lot of missing values or incorrect values in the Bean Origin & Type of Bean columns.

Conclusion

- Our initial question, about whether a chocolate's origin influences its quality, was answered using ANOVA-- it appears that location does have a significant impact on chocolate rating.
- We also found that other factors can be influential as well, including origin of the beans, as seen in our heat map.
- Finally, we find that percentage of cacao and location are both statistically significant in a two-way ANOVA, implying that percentage of cacao explains some of the variance in rating.
- However, it's worth noting that percentage of cacao on its own does not appear to influence rating, since there's almost no correlation between the two.