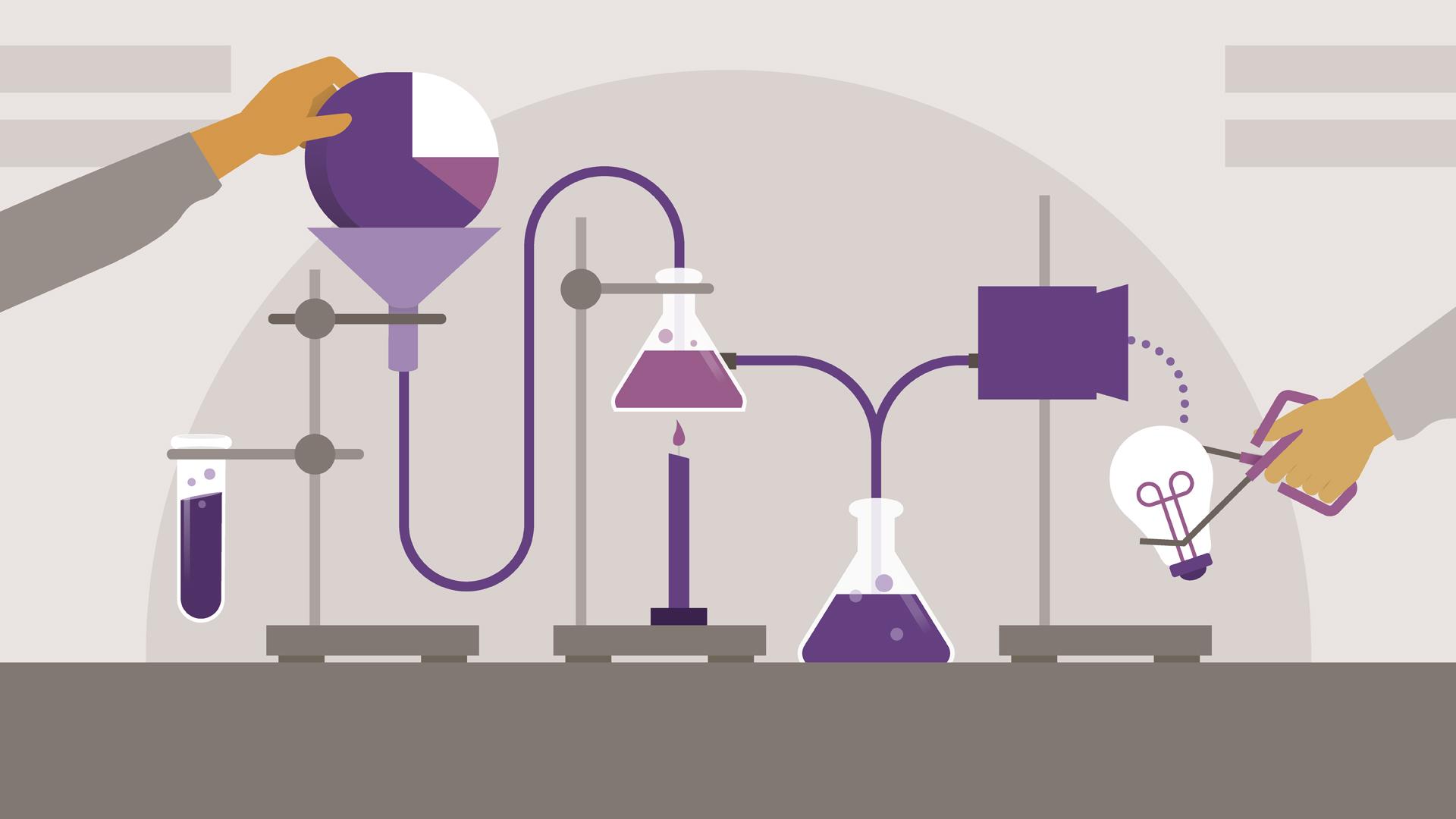
REPORT ON EXPLORATORY DATA ANALYSIS FOR CHOCOLATE BAR RATINGS DATASET

****

**BERLINDA ASIEDU**

**BARBARA ASIAMAH**

**BENEDICT NARTEY**

**BELINDA OSEI-TUTU**

**Azubi Africa, June 2020 Cohort.**

**TABLE OF CONTENT**

[**INTRODUCTION**](#_16kltn1h00pw) **3**

[**SOURCE OF THE DATASET**](#_vfcm1o5wj9k7) **3**

[**TOOLS USED**](#_kod3uvc28if6) **3**

[**PYTHON LIBRARIES USED**](#_6kmzfdnjfinz) **3**

[**DATA EXPLORATORY ANALYSIS PROCESS**](#_ukesj1uuvimz) **4**

[**· UNDERSTANDING THE DATA**](#_4lnux2ou307c) **5**

[**· QUESTIONS**](#_97rd8s9uzja0) **6**

[**· HYPOTHESIS**](#_5kbn4159kqck) **7**

[**· DATA CLEANING**](#_impvkn7mumn5) **8**

[**· ANALYSIS**](#_k0sj0crkdo7) **9**

[**FINDINGS & OBSERVATIONS**](#_xrsubxse0m4u) **17**

[**· RECOMMENDATIONS**](#_srfyf2sjaei) **18**

[**CONCLUSION**](#_63g32cwl5rwu) **18**

### **INTRODUCTION**

Exploratory Data Analysis(EDA) is one of the most crucial steps in a Data Science project. EDA in basic terms is a way of "Understanding the data with the help of visualizations and descriptive statistics".

Chocolate is a preparation of roasted and ground [cacao](https://en.wikipedia.org/wiki/Theobroma_cacao) seeds that is made in the form of a liquid, paste, or in a block, which may also be used as a flavoring ingredient in other food. Chocolate is a worldwide loved candy for both kids and adults. Chocolate is basically made from cacao and a few other additives.

The exploration of this data will be based on the data analysis cycle: understanding the data, posing questions, drawing hypotheses, data cleaning, data analysis, findings, and recommendations. This report spells out in detail the steps and procedures used in analysing the Chocolate Bar Ratings dataset. This is an analysis of 1795 chocolate bars from around the world. From the data, we will be looking at the chocolate bars’ various manufacturing locations, country of origin of the cacao beans, and overall rating, among a variety of other miscellaneous information. The data was gathered from Brady Brelinski of the Manhattan Chocolate Society.

### **SOURCE OF THE DATASET**

* The dataset used here was given by the Azubi Africa, June 2020 Cohort Program.
* The original ratings were compiled by Brady Brelinski, Founding Member of the Manhattan Chocolate Society.

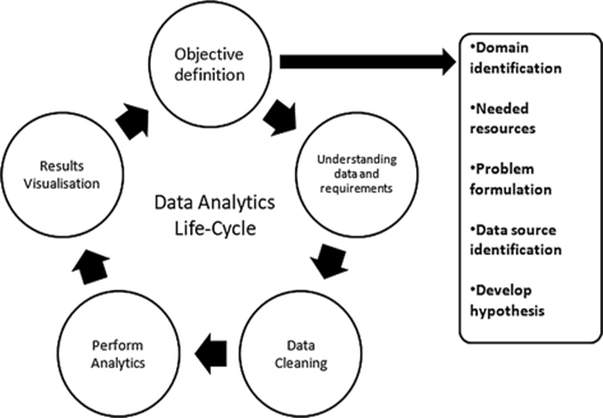
### **TOOLS USED**

* [Jupyter Notebook via th](https://jupyter.org/)e cocalc, Collaborative Calculation and Data Science - For analysis and code
* [Google Sheets](https://www.google.com/sheets/about/) - For exploratory data analysis report
* Google Slides - For Presentation

### **PYTHON LIBRARIES USED**

* pandas
* matplotlib
* Seaborn
* Numpy

### **DATA EXPLORATORY ANALYSIS PROCESS**



### **·** **UNDERSTANDING THE DATA**



Analysis of data can only be effectively conducted when the data scientist is well informed and understands the data given. To achieve this, data profiling is done on datasets before any analysis is made. Data profiling involves summarizing a dataset through descriptive statistics, obtained in this project using Python’s describe function. After successfully completing profiling of the dataset, the results show that the dataset can be worked on after a few corrections.This dataset contains expert ratings of 1,795 individual chocolate bars from 2006-2017 along with information on their regional origin, percentage of cocoa, the variety of chocolate beans used and where the beans were grown. A rating here only represents an experience with one bar from one batch. Batch numbers, vintages and review dates are included in the database when known.

A simple skim through the data can show that there are chocolates with rating, 1,0.9 and so on. However this is not enough to draw conclusions, thus the need for the data analysis.

A thorough look at the data revealed 9 columns and 1796 rows including column headings.The column heads cover the chocolate rating, the years under consideration, the specific cocoa bean type used to produce the chocolate, the companies producing the chocolate and more. The specific bean talks about the specific geo-region of origin for the bar.The column REF refers to a value linked to when the review was entered in the database. Under this the higher the REF, the more recent the rating. The review date is the date of publication of the review. The column cocoa percent talks about the cocoa percentage or the darkness of the chocolate bar being reviewed. The company location refers to the manufacturer’s base country. The rating refers to the expert rating for the bar and the bean type is the variety or breed of bean used, if provided. Lastly, the bean origin refers to the broad geo-region of origin for the bean.

### **·** **QUESTIONS**



After going through the data and understanding it we came up with a number of questions that will guide us in our analysis.

* Which countries are the top 10 producers of beans used for chocolate bars?
* What is the rating with the highest frequency?
* How many chocolate bars have a 100% rating?
* Which countries are the top 10 bean suppliers?
* Which 10 companies have the highest ratings?
* What is the average rating for each year?
* Is there a relationship between the origin of a bean and the average rating of bars made from it?
* Which chocolate beans have the highest ratings?
* What is the average rating of chocolate bars made up of 100% cocoa?
* What is the average rating of chocolate bars made up of 70% cocoa?

These questions were posed to help us make a good analysis of our data. They are the directional points in undertaking the exploratory data analysis.

### **·** **HYPOTHESIS**



From the questions asked above, we drew our hypothesis as stated below.

Null hypotheses

* There is no relationship between the percentage of cocoa in a chocolate bar and its rating.
* Less than 50% of raters will give chocolate bars ratings higher than 2.5.
* The average percentage of cocoa in bars does not change as the years go by.
* The average rating of chocolate bars decreases over the years.

Alternate hypotheses

* There is a relationship between the percentage of cocoa in a chocolate bar and its rating.
* More than 50% of experts will give chocolate bar ratings higher than 2.5.
* The average percentage of cocoa in bars changes as the years go by.
* The average rating of chocolate bars increases over the years.

### **·** **DATA CLEANING**



Data cleaning is important . This is because you cannot analyze data with errors and expect to have a true analysis of the data. Thus we started by checking for and taking out all the null values. Under the specific bean type column, we did not take out any null values. This is because in the data description given us it was stated that they may be columns that will not be provided for. The reason for this is unknown.

We also checked for outliers. We found some outliers in the rating column. We did not take them out because they had no significant bearing on the data.

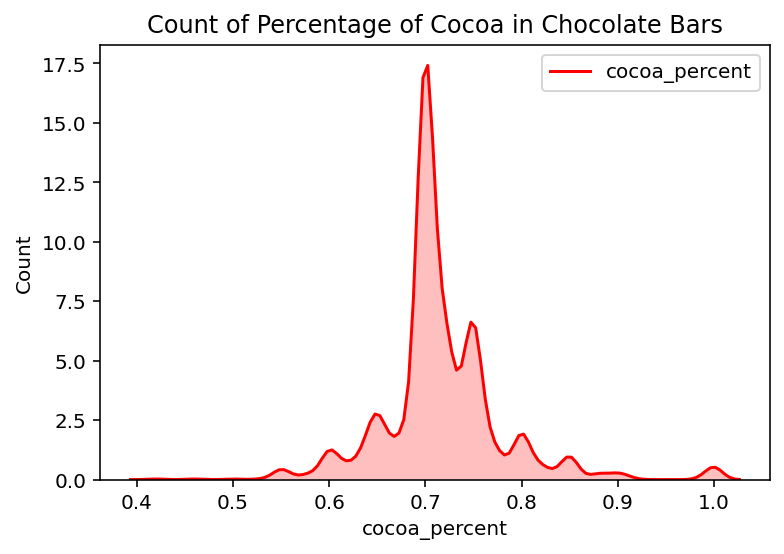
We renamed the columns in order to be able to work with them in Python. All these were done under data cleaning in order to make the data useful for analyses.

### **·ANALYSIS**

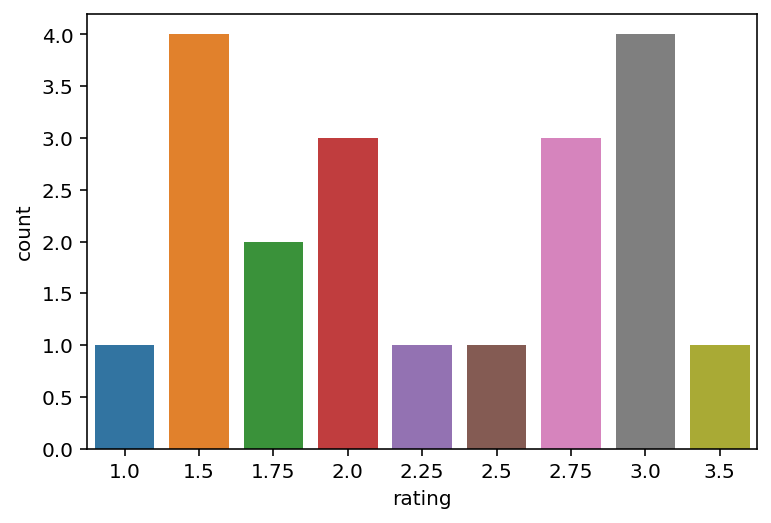
Our analysis looks at how to prove or disprove our hypothesis and answer the questions we posed. The univariate and bivariate virtualization methods were used in this analysis. Univariate virtualization was used to provide summary statistics for each field in the data set and the bivariate virtualization was conducted to find the relationships that existed between two variables in the dataset. We used graphical representations for better understanding.

1. **Count of Percentage of Cocoa in Chocolate Bars.**

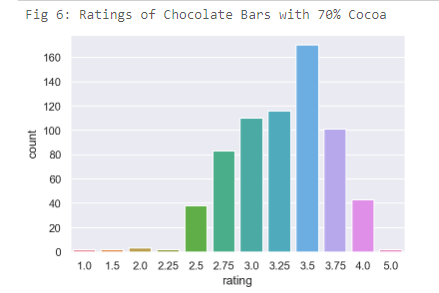
As figure 1 shows, a lot of chocolate bars are made with about 70%-75% percentage of cocoa. From the graph, it can be seen that the highest protruding end covers the cocoa count of between 0.7-0.75 representing 70%-75% of cocoa percentage. It is however interesting to note that chocolate bars with cocoa percentage of 0.9-1.0 or 90%-100% are almost recording as low ratings as chocolate with 0.1-0.5 or 10%-50% of cocoa percentage. Does the data tell us anything about how consumers are rating chocolate bars based on the percentage of cocoa that is used in each bar? The following figures aim to observe if there is a relationship between the percentage of cocoa used, and consumer ratings.



*Figure 1: Frequency of percentages of cocoa in bars.*



*Figure 2: Ratings of Chocolate Bars with 100% Cocoa*



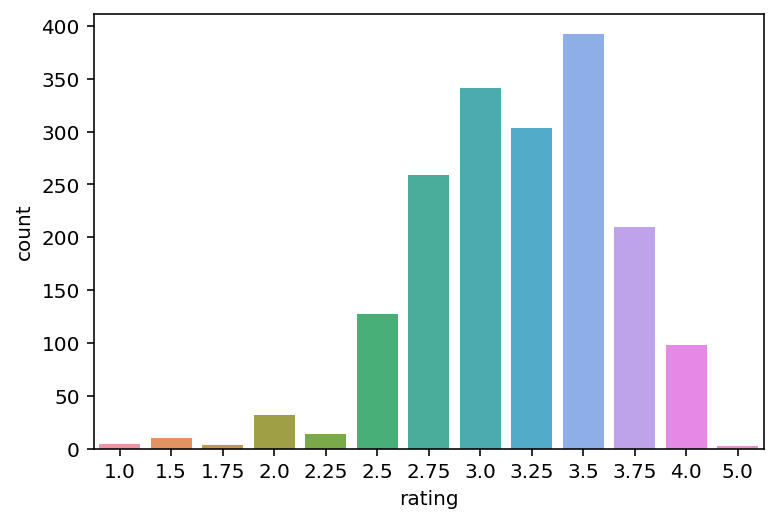
*Figure 3: Ratings of Chocolate Bars with 70% cocoa*

This comparison was most interesting. First, using the data, we found out how many chocolate bars had 100 % cocoa, which was the maximum amount of cocoa used in this survey, and in chocolate bars in general. Around 20 chocolate bars had a percentage of 100% cocoa, all originating from different parts of the world. The reason to do so was to find out how consumers are rating chocolate bars that have 100% cocoa. These would be the most rich, and perhaps the most darkest and bitter bars. As figure 2 shows us, most consumers do not prefer chocolate bars that have 100% cocoa in them. At most, they are rating it at around 3.0 (satisfactory), while most individuals felt that these chocolate bars were around a 1.5 - palatable but unpleasant. There are many factors that would give this result, one being if this result is due to the bitterness of the chocolate, or, is it because of the type of bean that is used? More experiments would have to follow to solve these questions.

Moving forward, it was then observed that the mean amount of cocoa percent that was used is 70%. We already observed in Figure 1 that most individuals rate chocolate bars with 70% cocoa with a rating of 3.5 (above satisfactory, approaching pleasurable), and when taking a closer look in Figure 3, the result matches. Most individuals are preferring chocolate bars that have just enough cocoa, but not enough to be overbearing. This analysis results in a rejection of the null hypothesis that the amount of cocoa in a chocolate bar has no relationship with the rating it is given.

1. **Frequency of Rating**

It is important to know just how many chocolate bars were rated. In total, around 1795 chocolate bars were rated from 63 countries. The most prevalent number of ratings given by experts was between 3.0 to 3.5, with the highest being 3.5 as seen in Fig 5 below. This shows that most people are rating their chocolate just a little high above satisfactory. Interestingly, we can barely see any rating for 5.0. The frequency shows that most experts gave a rating higher than 2.5, leading to a rejection of the null hypothesis stating that ‘less than 50% of raters will give chocolate bars ratings higher than 2.5.

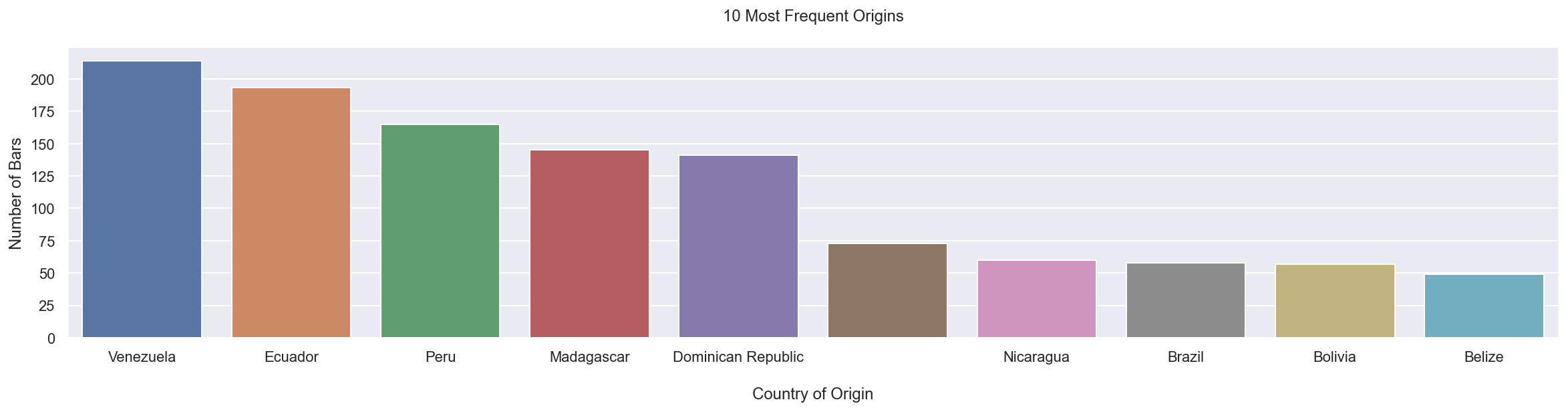


*Figure 4: Frequency of ratings*

This data is mainly due to customer preference. There is a lot of bias at play as ratings are based on factors pertaining to the individual such as pallets instead of solely based on factors reported in the dataset. However, one of the most important ingredients of a chocolate bar is the cocoa bean. Where the cocoa bean comes from tells us a couple different things: for once, it will tell us which countries produce the premium chocolate bars, and which region of the world are they located.

1. **Country with the highest rating for bean type**

In Figure 6, the data shows us that the beans with the highest rating of chocolate bars are coming from Venezuela.. All the other countries have beans coming from the South American continent as well as Africa, and the Carribean countries.

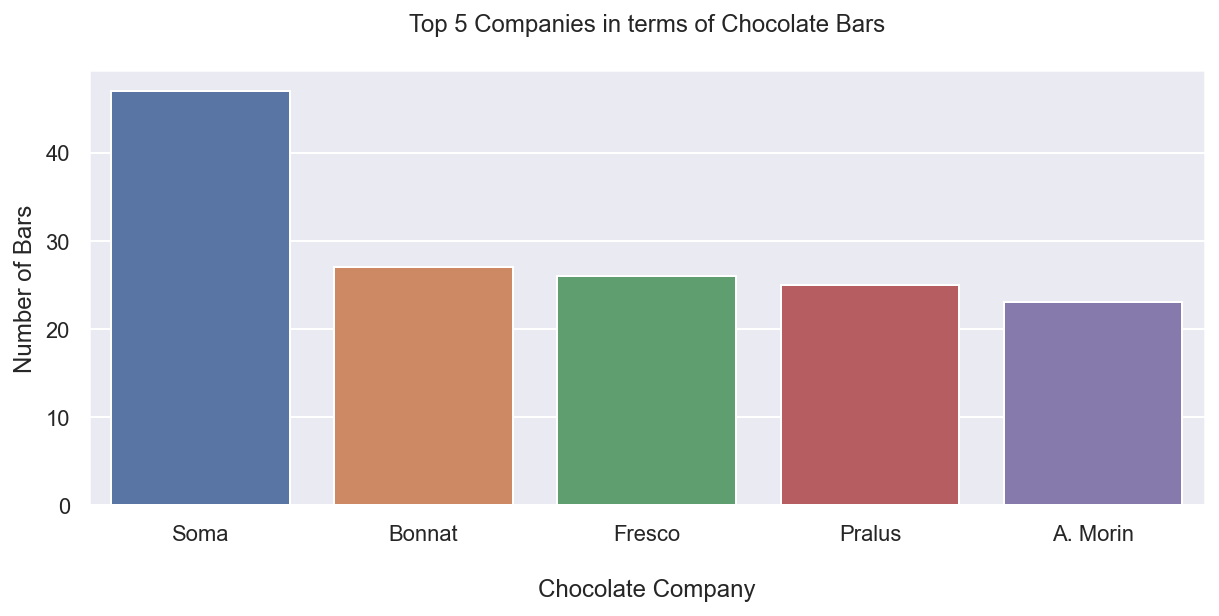
**

*Figure 5: Top ten countries whose beans are mostly used*

1. **Top 5 companies producing the highest number of chocolate bars**

Soma has the highest number of chocolate bars in this dataset with 47.

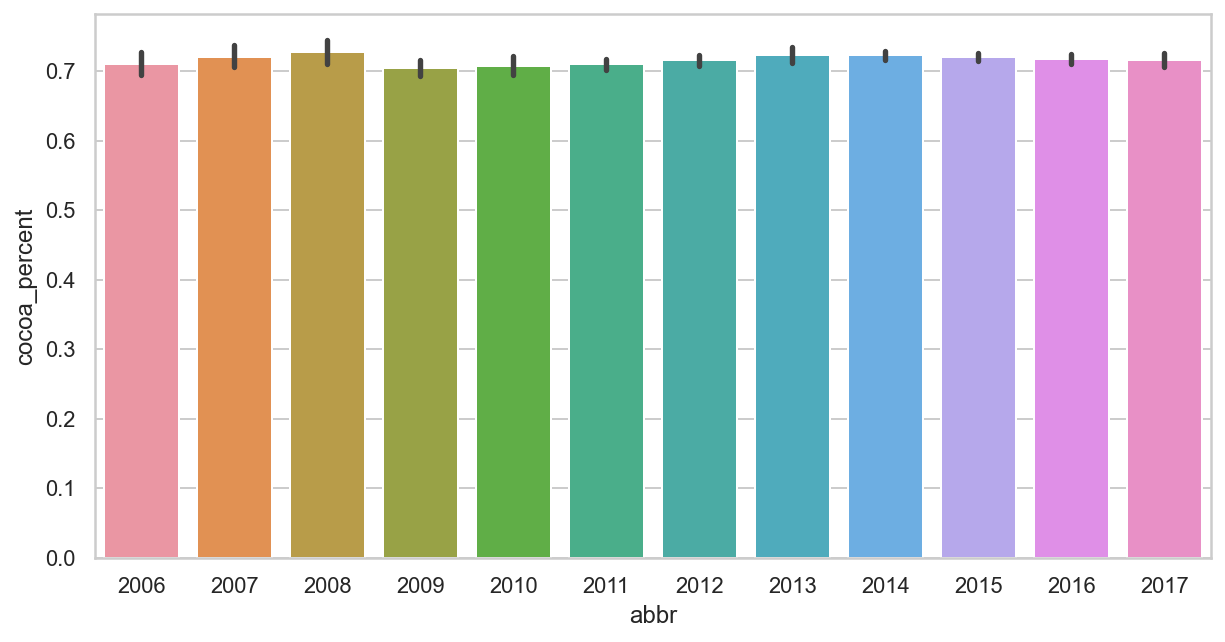
The next closest competitor, Bonnat falls short of the leader by 20 bars. The other companies are in close competition with one another producing almost about the same number of chocolate bars.

**

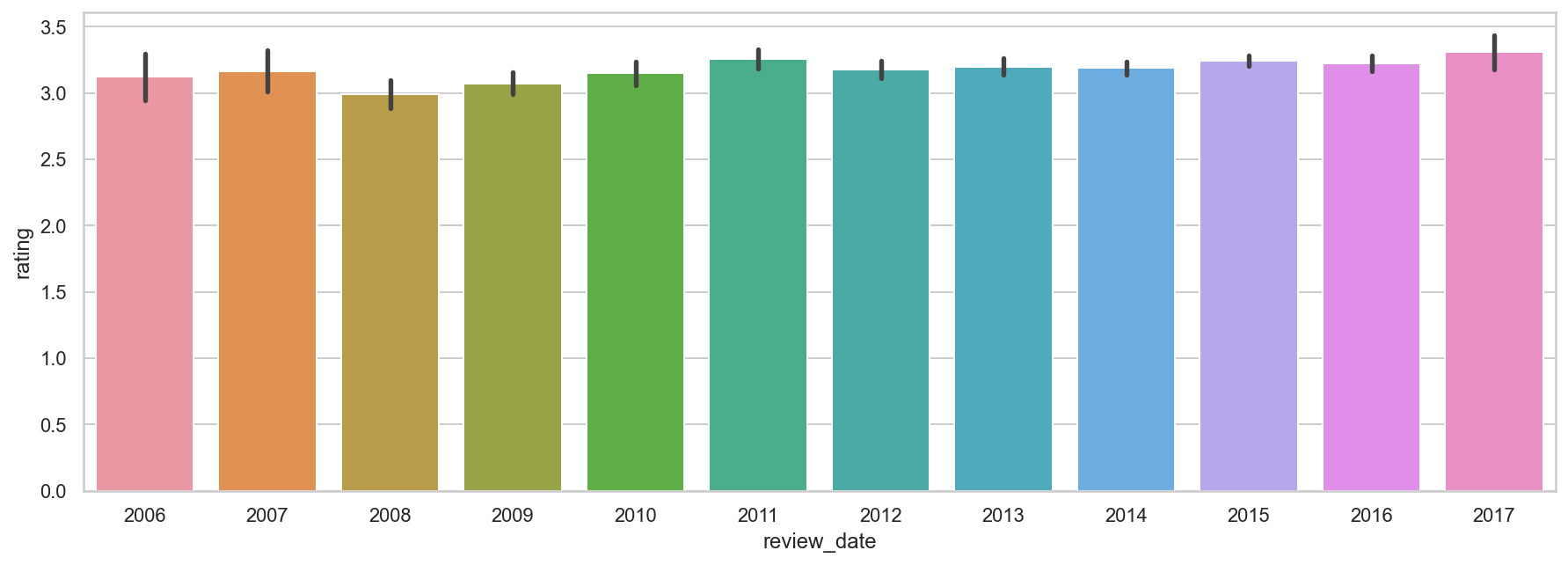
*Figure 6: Top 5 companies with the most consumers*

1. **Measure of Cocoa percentage used over the years under review**

This chart below shows the percentage of cocoa used in the various years for chocolate production. We would like to see if this has any bearing on the rating that were given to the chocolates in those same years. It can be observed that the year 2008 has the highest number of cocoa percentages used in chocolate production. This is closely followed by the years, 2007, 2013, 2014, 2015. The year with the lowest use of cocoa in chocolate production is 2009, followed by 2010 and 2011.

*Figure 7:Percentage of cocoa used for chocolate over the years*

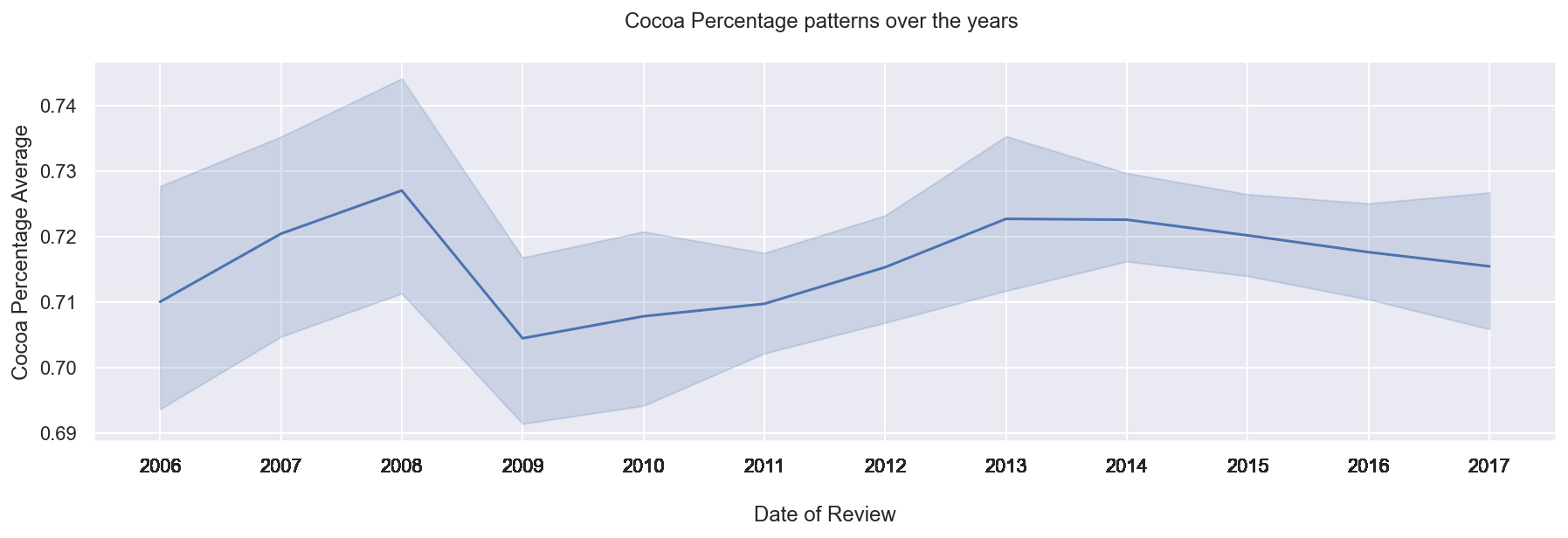
This graph below goes to support the above graph on whether the coca percent used in the various years has any bearings on the ratings that were given in those years. From this graph we can see that 2017 has the highest rating. However, it has one of the lowest cocoa percentages used in figure 7 above. On the other hand, 2008 which has one of the highest cocoa percentages used in figure 7 above has one of the lowest ratings in figure 8. We can see clearly that the higher the percentage of cocoa used in the chocolate bars, the lower the rating. This explains why the year 2008 has a very low rating whilst 2017 and 2011 have very high ratings.

**

*Figure 8: Ratings of bars over the years.*

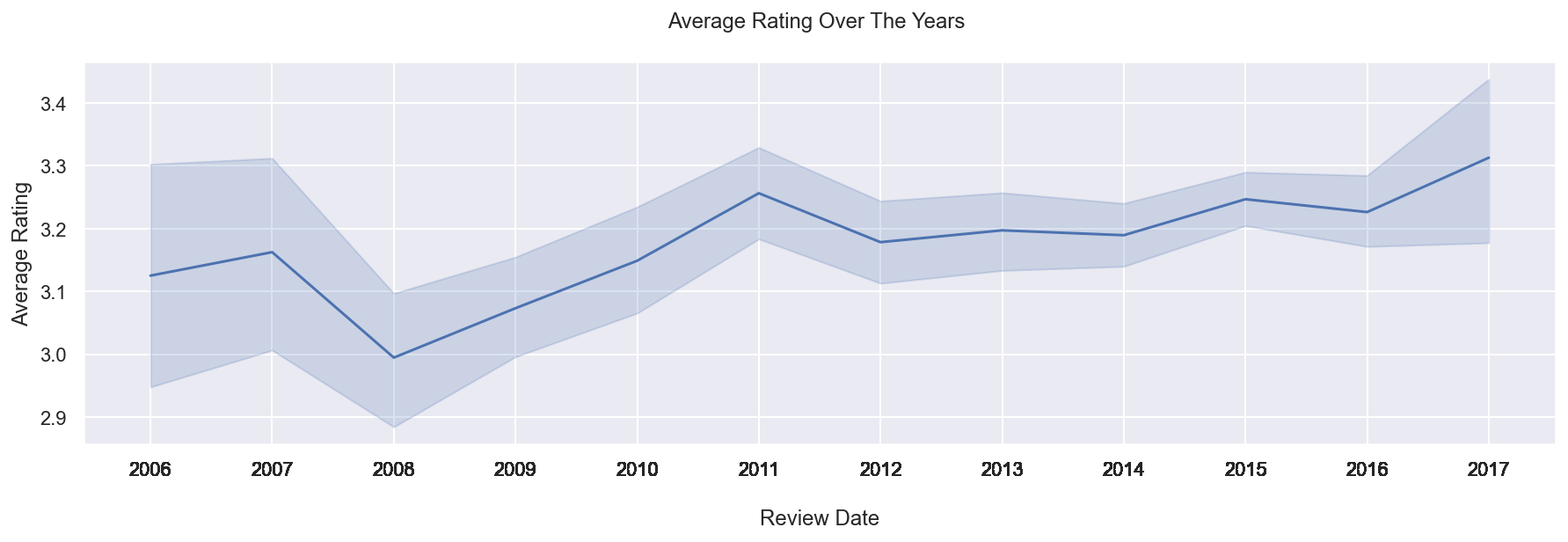
1. **Trends over the years.**

Over the years, the percentage of cocoa in a chocolate bar has seen rises and falls. In studying this trend, the average percentage per year is considered. In the year 2008, the highest percentage average of cocoa in chocolate bars was recorded at approximately 73%. This was after a continuous increase from 2006 through 2007 and in 2008. However, in 2009 there was a decrease in the average percentage of cocoa, making the new average of 70%. Observing the graph as shown in Figure 11, the lowest average was recorded in 2009. From 2009, there was a rise in the amount of cocoa used in chocolate bars up until 2013. By this time, the average had hit a 72.4%, almost reaching the peak value in 2008.There was a steady fall in the average percentage from 2014, reaching 71.6% in 2017. After this analysis is made, the null hypothesis that states that ‘The average percentage of cocoa in bars does not change as the years go by’ is rejected. Hence, it is concluded that the average percentage of cocoa in chocolate bars varies with year.

**

*Figure 9 : Average percentage of cocoa in bars over the years.*

After observing the changes in the amount of cocoa used in chocolate bars over the years, we sought to analyze the average ratings for each year. It is assumed that as time increases, new recipes introduced, new ingredients that enhance taste are discovered and advancements are made in the production of chocolate. This will increase quality and hence, rating of chocolate bars. To test this hypothesis, a graph is generated, with review dates against average ratings. From figure 12, it is observed that average rating increased from 3.14 in 2006 to 3.18 in 2007. However, ratings dropped to 3 in 2008. The ratings increased since then to 2011, when it reached 3.26. After this average, there have been fluctuations in the ratings over the years. In 2017, the average ratings reached an apex of 3.31. This analysis results in a conclusion that the average ratings of chocolate bars has increased, rejecting the null hypothesis.

**

*Figure 10: Average rating over the years.*

### **FINDINGS & OBSERVATIONS**



Performing an Exploratory Data analysis on the Chocolate Bar Ratings Dataset has provided an insight into the different variables in the dataset, their relationships and how they can be used to make business decisions in the cocoa bean and chocolate production. This analysis has made it possible to use the data available to answer questions, test assumptions and validate hypotheses as stated in the introductory section of this report. After thorough analysis of the data, the following are some findings that were made;

* Chocolate bars made with 70% cocoa have a higher rating than those made up of 100%.
* The highest ratings given to chocolate bars fell between 3 and 3.5 with 3.5 being the rating with the highest frequency.
* The highest average rating of chocolate bars is 3.31, recorded in the year, 2017.
* The lowest average rating, 3, was recorded in 2008.
* The highest percentage of cocoa in chocolate bars was recorded in 2008.
* The top 5 broad bean origins are Venezuela, Ecuador, Peru, madagascar and Dominican Republic.
* Venezuela supplies the largest amount of cocoa beans for chocolate production.
* Over time average rating isn’t increasing rapidly but the cases of extreme ratings are decreasing.

### **·** **RECOMMENDATIONS**

We recommend that in order to get a good analysis of chocolate bar ratings,

* Based on the analysis and conclusions, we recommend that for a company to maximise profit, 70% of cocoa beans in chocolate will give the best flavour.
* Majority of the beans come from the South American continent and so it may be cheaper getting cocoa beans from such areas.
* For a company to maximise chocolate production, techniques used in the Soma company can be adapted to suit the local company.
* More data needs to be collected for further studies into particular bean type and percentage required to gain higher ratings

### **CONCLUSION**

This analysis has helped bring some insight into consumer patterns in the chocolate industry. It is possible to perhaps see a pattern in the data that could lead us to observe why a specific amount of cocoa is used in chocolate bars and how it affects consumer rating. Finally, it has shown us how cocoa percentage has an impact on the consumption of chocolate, and where the highest rated chocolate and cocoa beans come from.

In conclusion, for chocolate industry players, this is a very useful data that can help make business decisions. One very profound factor that can be taken into consideration is the consumer preference. This can be very crucial in determining the demand of a particular chocolate bar. For instance, if the sale is low, will it be because the bar contains too much cocoa? Or the chocolate bar is expensive because the bean used in its production is being imported. These factors and many more such help determine customer behaviour and can help us predict the sale rate of a chocolate bar in the near future. The findings and observations shown above in the graphs represented can be used for chocolate business advisory. This data can be used to pitch ideas to chocolate manufacturing companies.

This data can also be used by such companies to increase sales or alter consumer preference of certain chocolate bars.