# Chocolate Bar Ratings

## Project 1 - Group 2 Write Up

Arya Maredia
Monica Mitry
Cecilia Rocha
Emily Wimmer

# Introduction

As passionate chocolate lovers, we wanted to know where the best chocolate bars originate from. Chocolate is more than just a treat, it is a delicacy enjoyed globally with each culture adding its unique touch to the experience. From the type and origin of cocoa beans to the intricate manufacturing processes and varying cocoa percentages, the journey of chocolate is as rich and diverse as its flavors.

In our quest to find the ultimate chocolate bar, we turned to data for answers. We analyzed a comprehensive dataset from Kaggle, titled *Chocolate Bar Ratings (2006-2017)*,[1]. This dataset contains 1,795 chocolate bar samples collected from the years 2006-2017. Each chocolate bar in the dataset was carefully rated based on a combination of key factors: flavor, texture, aftertaste, and overall opinion. These components offer a holistic view of what makes a chocolate bar truly stand out. By exploring the dataset, we aim to answer the most sought-after questions in the chocolate world. Our objectives, outlined below, will help uncover the secrets to discovering the finest chocolate bars on the market.

# Objectives

Our primary objectives for analyzing this dataset were: 1) to identify what countries the best cocoa beans for chocolate production originate from 2) to determine what country uses the highest cocoa percentage and 3) to explore if there is a relationship between the rating of the cocoa and the cocoa solids percentage. While we anticipated that these questions might yield distinct insights, the findings would provide valuable knowledge to help us identify the best chocolate.

[1]https://www.kaggle.com/datasets/rtatman/chocolate-bar-ratings?select=flavors_of_cacao.csv

# Data Engineering

The Chocolate Bar Ratings raw dataset contains nine columns and 1795 samples with final product formatting with column names containing multiple lines and the “Rating” column data displayed with the percent symbol for each sample (Figure 1).

```
[49]: # Data Information
      df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1795 entries, 0 to 1794
Data columns (total 9 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Company
(Maker-if known)         1795 non-null   object
 1   Specific Bean Origin
or Bar Name  1795 non-null   object
 2   REF                                 1795 non-null   int64
 3   Review
Date                        1795 non-null   int64
 4   Cocoa
Percent                     1795 non-null   object
 5   Company
Location                  1795 non-null   object
 6   Rating                              1795 non-null   float64
 7   Bean
Type                          1794 non-null   object
 8   Broad Bean
Origin                1794 non-null   object
dtypes: float64(1), int64(2), object(6)
memory usage: 126.3+ KB
```

Figure 1: Column headings from raw dataset using df.info().

Our first step in the data engineering process was to investigate the number of null values in the dataset using the df.info() function, this revealed 1 row with missing data in the “Bean Type” column. Due to the low percentage (<1%) of missing data we agreed to drop the datapoint using the df.dropna() function. Upon further investigation of the dataset in Excel it became clear that more data was missing but was undetected running a df.info() in python, as multiple samples had unicode characters expressed as “\xa0” in “Bean Type” and “Broad Bean Origin” columns. The “Bean Type” column had roughly half the samples missing and was therefore dropped from our analyses as any manipulation or replacement would have a high likelihood of being poor substitutes or skewing analyses. In an effort to preserve the “Broad Bean Origin” column as it has a low percentage of unicode and bean samples from multiple countries we created two new categories, “unknown” to replace “\xa0” and “blended” to accurately represent the number of

beans that came from multiple countries. While cleaning up the “Broad Bean Origin” column we also consolidated all spelling variations of countries to allow for complete grouping.

Column headers were renamed and shortened to omit capital letters and new lines to create more user-friendly code. The percent formatting for each sample in the “cocoa_percent” column was removed to allow us to create easier data analysis and visualizations code. Using these data engineering steps we created a reference dataframe that allowed for easier analysis.

```
[50]: # Change Column Names
df.columns = [
    'company_maker',
    'specific_bean_origin_or_bar_name',
    'ref',
    'review_date',
    'cocoa_percent',
    'company_location',
    'rating',
    'bean_type',
    'broad_bean_origin'
]

print(df.columns)
```

```
Index(['company_maker', 'specific_bean_origin_or_bar_name', 'ref',
       'review_date', 'cocoa_percent', 'company_location', 'rating',
       'bean_type', 'broad_bean_origin'],
      dtype='object')
```

Figure 2: Rename of column names for our reference dataframe.

## Question 1: From what countries do the best cocoa beans for chocolate production originate?

Understanding the origins of cocoa beans and their manufacturing processes is essential. Identifying major manufacturing companies is critical to recognizing trends in the global chocolate industry. This question highlights the intersection of agriculture production and industrial marketing, demonstrating how raw materials are sourced and transformed into finished products. By analyzing these factors, we gain insight into regional strengths, quality standards, and market direction. The figure below illustrates the numerous countries that source cocoa beans, emphasizing the need to focus on the top ten countries to provide an accurate answer.

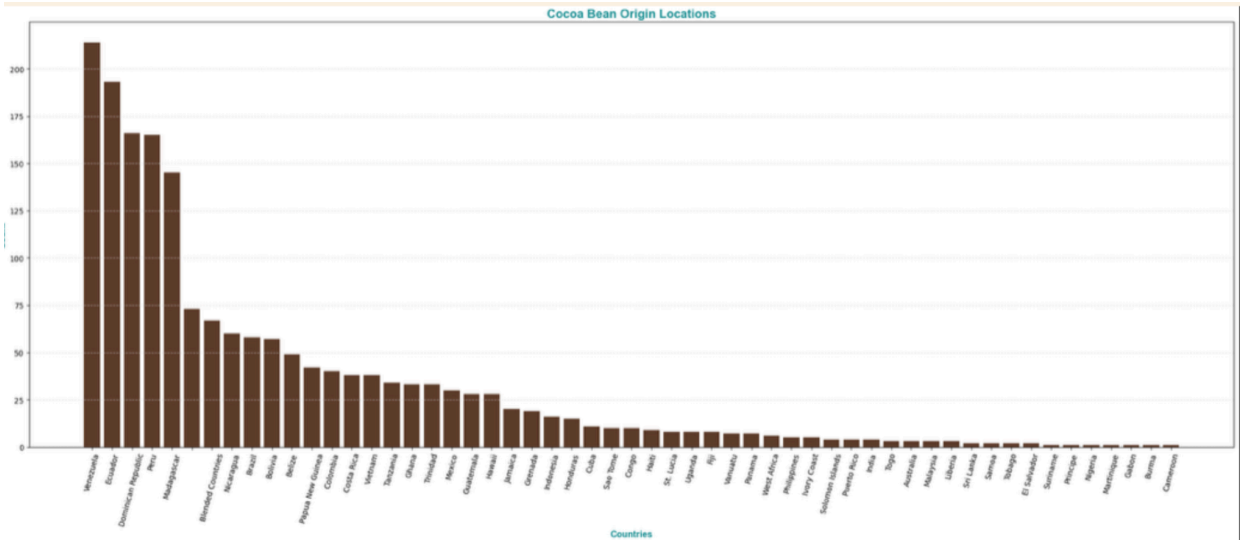


Figure 3: Bar graph depicting Cocoa Bean Origin Locations

We began by examining the top ten countries that source cocoa beans. According to the data presented below, Venezuela emerges as the leading source of cocoa beans, with 214 companies sourcing Venezuelan beans for their production, exhibiting the country's importance of providing high-quality raw materials. The figure below also reveals a mix of countries that are unknown or have been grouped together under a single category. Notably, many of the top cocoa-producing countries are nestled near the Amazon rainforest.

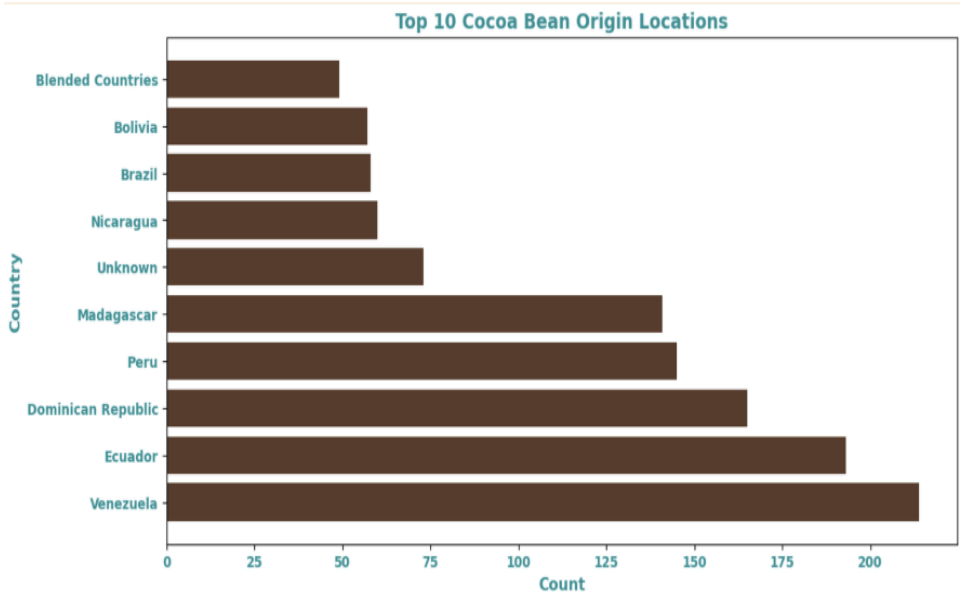


Figure 4: Bar graph depicting the Top 10 Cocoa Bean Origin Locations.

In the second figures below we see the top ten locations of chocolate manufacturing companies. The United States dominates the chocolate manufacturing division, considering it accounts for 58.3% of the total companies globally, with 763 companies in operation. This emphasizes the U.S.'s role as a key producer of chocolate bars and the focal point of innovation and consumer influence in the industry. It's also very curious to note that although France, Canada and the UK host the next top companies they only contribute 11-6%. The rest of the countries fall into a more consistent, modest range, each hosting a standard number of chocolate producing companies.

| company_location | count |
|---|---|
| U.S.A. | 763 |
| France | 156 |
| Canada | 125 |
| U.K. | 96 |
| Italy | 63 |
| Ecuador | 54 |
| Australia | 49 |
| Belgium | 40 |
| Switzerland | 38 |
| Germany | 35 |

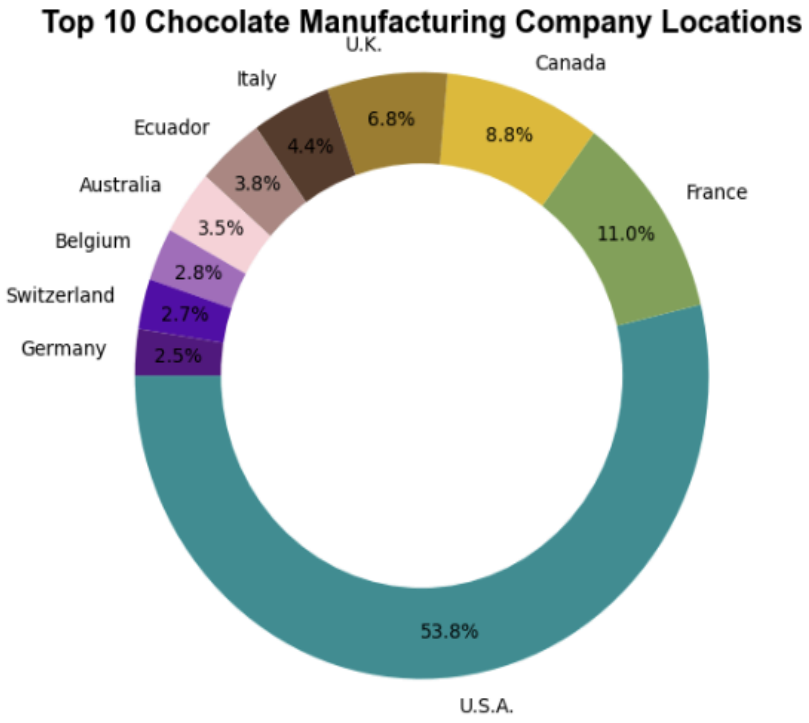


Figure 5 and 6: Depicts the Top 10 Manufacturing Company locations with a list and a donut chart.

## Question 2: What country uses the highest cocoa percentage?

Cocoa percentage can be referred to as whether the chocolate bar is richer with dark chocolate or less concentrated like milk chocolate. To begin the search for the county with the

highest cocoa percentage it was necessary to do a .groupby("company_location") and then calculate the average cocoa_percent for each company location. As mentioned earlier in data engineering, this required some cleaning with dropping the % symbol that was in each cell. After this we were able to create a bar graph that showcased the average cocoa percentage by company location.

```
[43]: avg_cocoa_percent = df_cleaned.groupby("company_location")['cocoa_percent'].mean()
      print(avg_cocoa_percent)
```

```
company_location
Argentina            73.333333
Australia            70.224490
Austria              72.000000
Belgium              72.025000
Bolivia              73.000000
Brazil               69.823529
```

Figure 7: Code calculating the average cocoa percentage by company location.

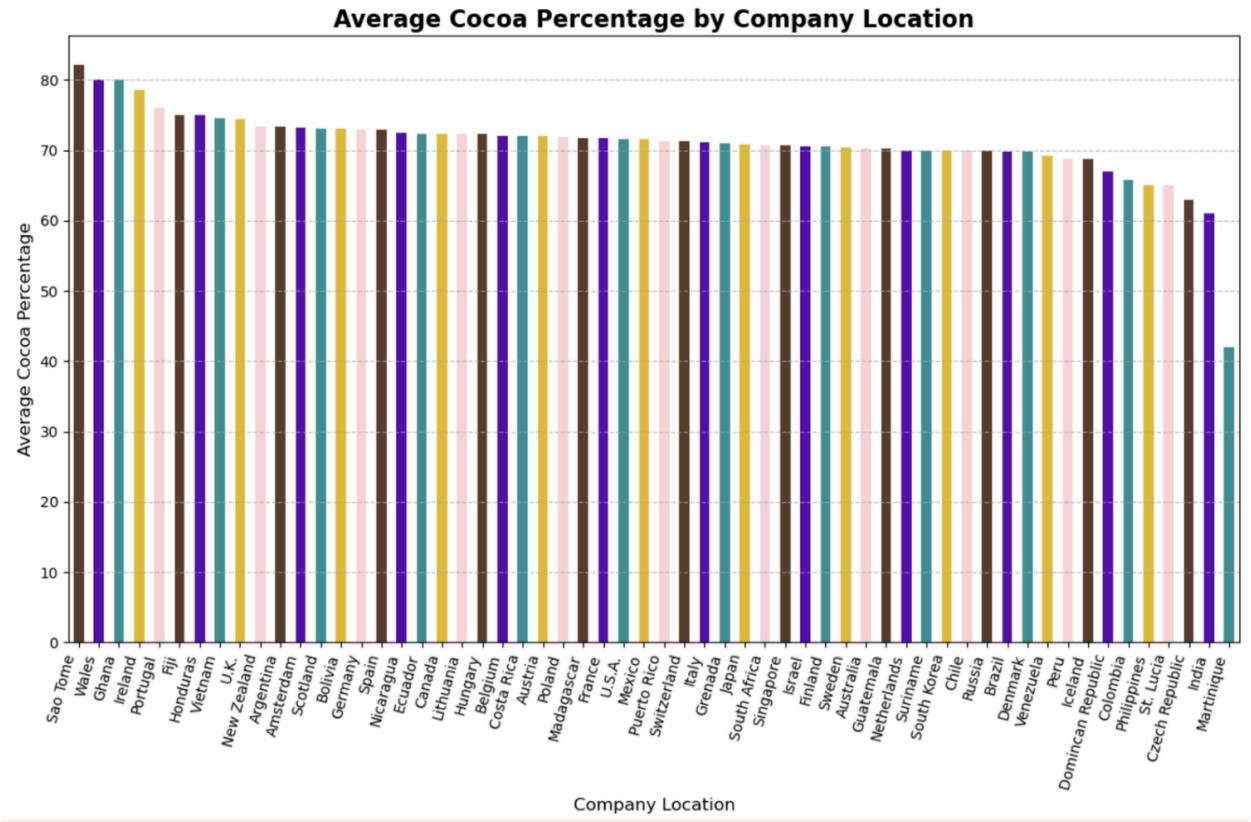


Figure 8: bar graph of average cocoa percentage by company location.

Based on the graph (Figure 8) it is clear that Sao Tome had the highest average cocoa percentage at 82%, followed by Wales and Ghana. There is a very large difference between the

highest average, Sao Tome, and the lowest average, Martinique, with a 40% difference. Overall, most countries' average cocoa percentage is between 70% and 80%, indicating that the majority prefer to use a high cocoa percentage. The top three manufacturing company locations that were mentioned previously (USA, France, and Canada) are found more within the top 30 of cocoa percentage company locations.

We can take a deeper look into the top ten average cocoa percentages with the below bar graph. There is a consistency that the top ten are all above 75%. Companies located within these countries have more of a reputation of producing highly rich chocolate products.

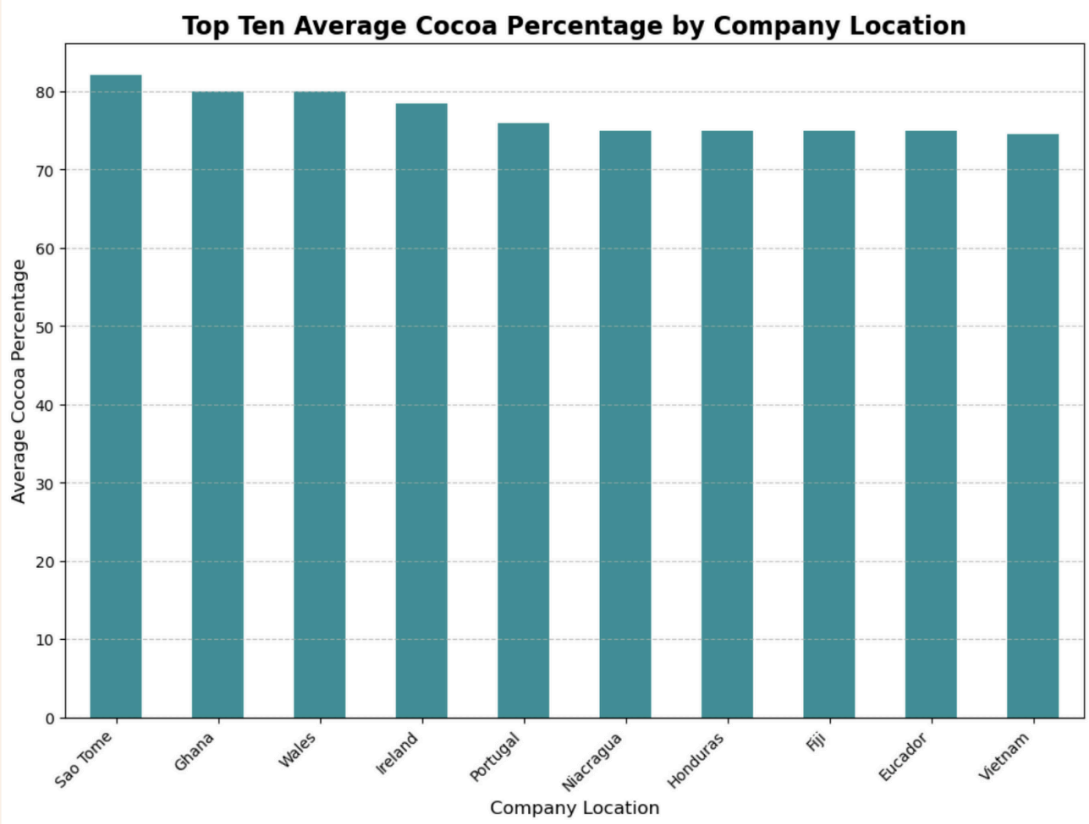


Figure 9: bar graph depicting top ten average cocoa percentage by company location.

## Question 3: Is there a relationship between the rating of the cocoa and the company location?

Understanding the relationship between company location and the cocoa rating is what will answer the main question of "where does the best chocolate bar come from?" One way to answer this question is to find the average rating by company location just like the average cocoa percentage. Very similarly to finding the average percentage, it required doing a .groupby("company_location") and finding the mean for the rating for each company location.

```
[35]: avg_ratings = df_cleaned.groupby('company_location')['rating'].mean()
      avg_ratings_sorted = avg_ratings.sort_values(ascending=False)
      print(avg_ratings_sorted)
```

```
company_location
Chile                3.750000
Philippines          3.500000
Netherlands          3.500000
Iceland              3.416667
Vietnam              3.409091
Brazil               3.397059
Poland               3.375000
Australia            3.357143
Guatemala            3.350000
```

Figure 10: Code calculating the average rating by company location.

There were many company locations within the same country and multiple chocolate bars while some countries had only one company location with one bar. These limitations would skew the average rating. For this reason it was helpful to create a line chart depicting the average rating and a violin chart that showcased the distribution of company locations. The company location with the highest average rating is Chile, however, when looking at the violin chart it is clear that Chile is one of the locations with very limited data. India is very similar to Chile with the one company in the country but with the lowest average rating.

Very similarly to the average cocoa percentage there was a lack of the top countries with the highest manufacturing rate with the higher average chocolate bar rating. However, the company location with the highest rated chocolate bar was in Italy with a rating of five. Despite that many of the company locations with the top ten average ratings were still well with a score of four.

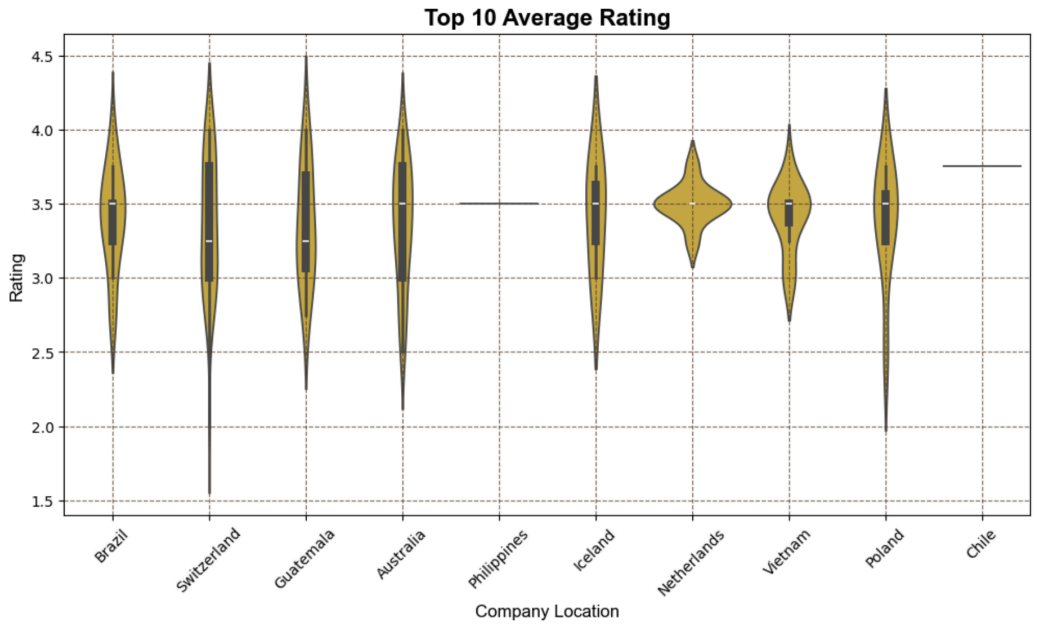


Figure 11: Violin chart depicting the distribution of the Top 10 Average Rating by Company Location

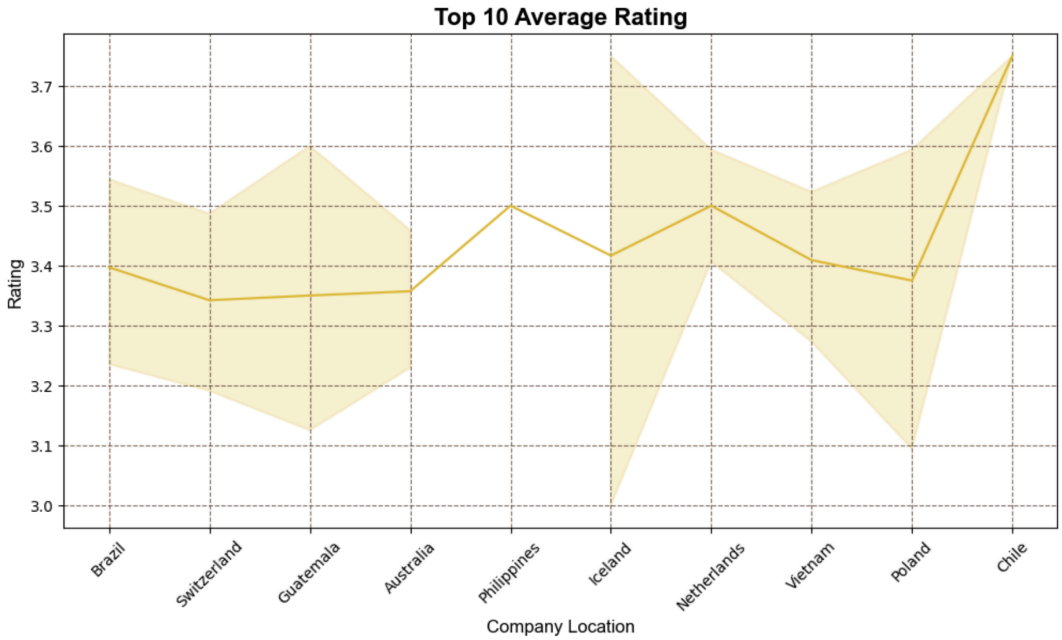


Figure 12: Line chart depicting the distribution of the Top 10 Average Rating by Company Location

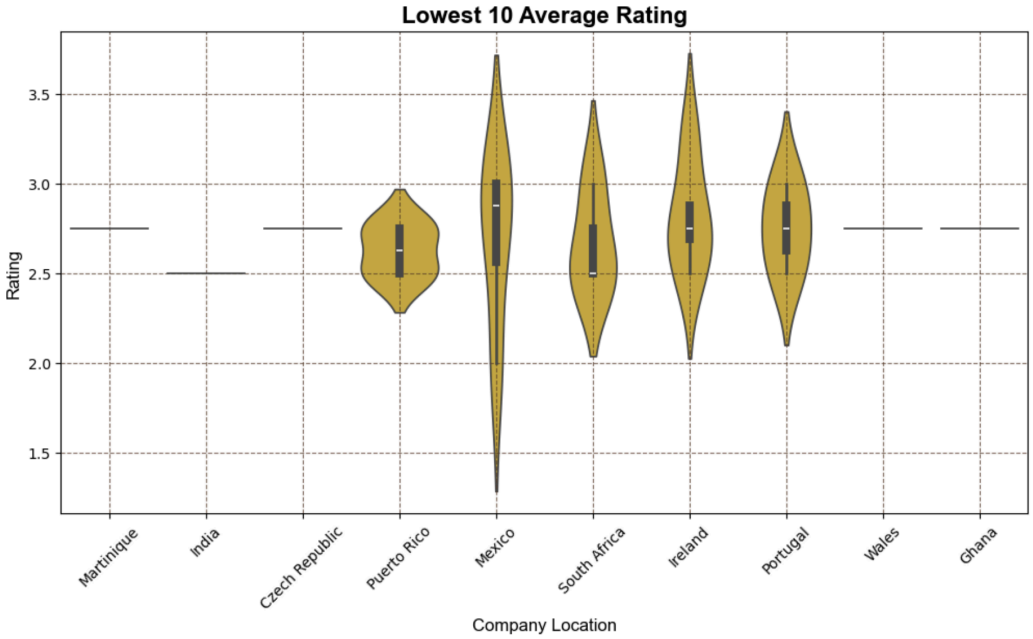


Figure 13: Violin chart depicting the distribution of the Lowest 10 Average Rating by Company Location

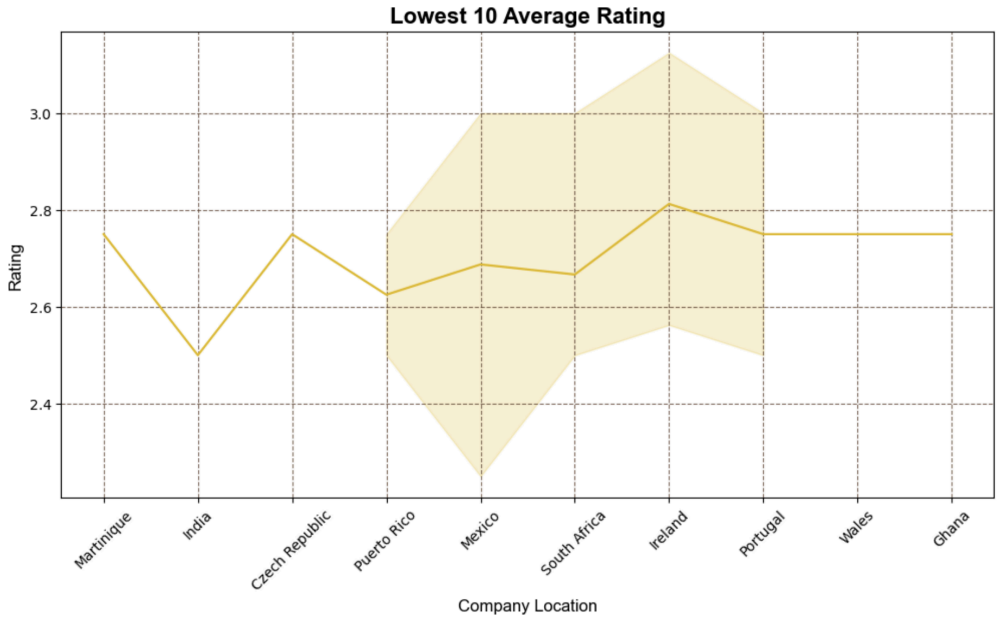


Figure 14: Line chart depicting the distribution of the Lowest 10 Average Rating by Company Location

# Regression

To determine if there is a relationship between any of the data we ran two linear regressions using the only two columns of continuous numerical data: ratings and cocoa percentage. We want to know 1) is there a correlation between cocoa rating and cocoa percentage and 2) is there a negative correlation between cocoa rating and cocoa percentage in the US?

Our first regression analyzed cocoa percent vs rating for all samples. We found there to be a slight negative correlation between cocoa rating and cocoa percentage ($R^2$=0.03; Figure ). Suggesting that the higher percentage of cocoa used the lower the rating by the reviewer.

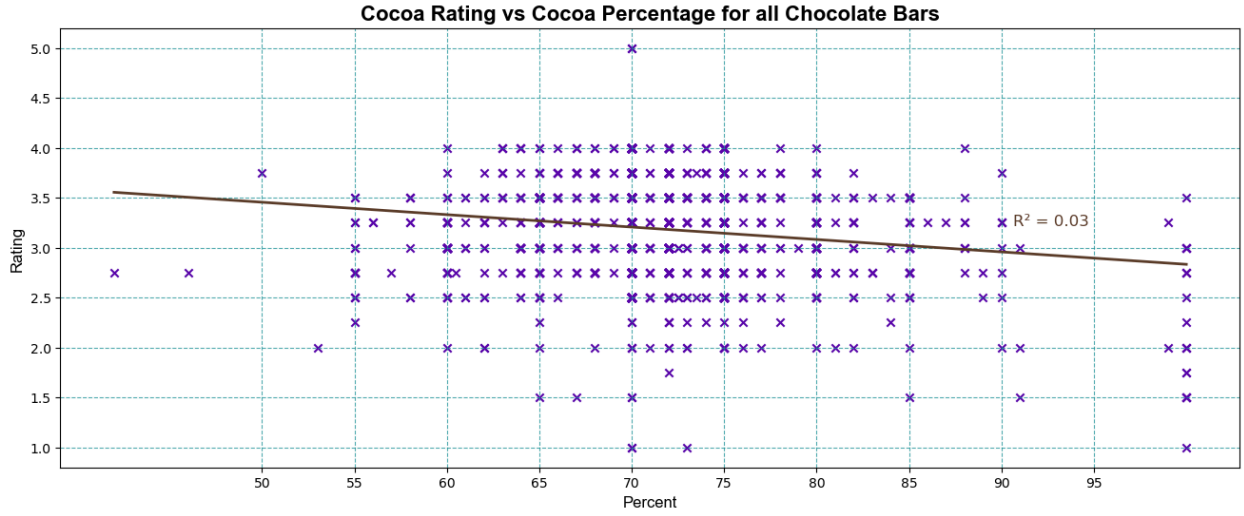


Figure 15: Cocoa percent vs Rating for all sampled chocolate bars.

We then narrowed down the analysis to chocolate bars manufactured exclusively in the United States. The United States' population tends to prefer chocolate with less cocoa percentage in it, therefore we hypothesized that ratings would decrease as cocoa percentage increased for chocolate bars manufactured in the United States, as United States based manufacturing companies may not prioritize the production of high quality, high cocoa percentage chocolate bars. We found there was no relation between cocoa percent and cocoa rating for chocolate bars

manufactured in the United States ($R^2 = 0.00$; Figure ).

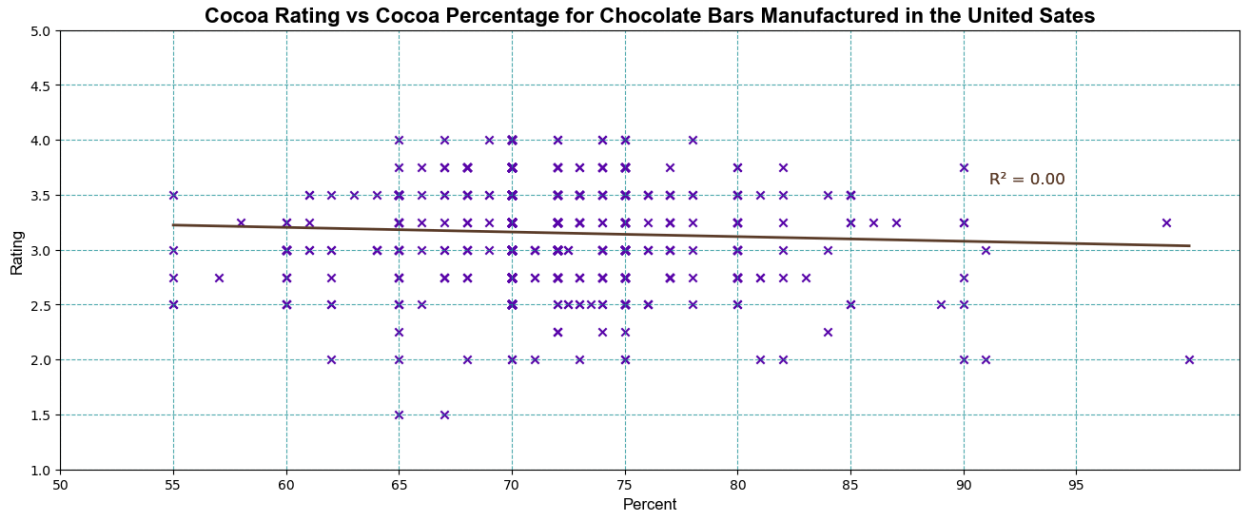


Figure 16: Cocoa percent vs Rating for chocolate bars manufactured only in the United States.

## Bias and Limitations

While this dataset was cohesive there was some non-normal data that may have caused bias and limitations to the analyses done. One bias in the dataset was the uneven distribution of data across years, with 2017 having significantly less data points compared to previous years (24; Figure). While the sample size for this review year was low we agreed to retain the integrity of the dataset as our objectives did not focus on year of review.

Not all companies produced the same number of chocolate bar samples. There were many countries that had only one sample. The sample size per country ranged from 214-1, this large range could have skewed data analyses related to country, as was evident in objective three where we found Chile to have the highest rated chocolate bar with only one submission. Chocolate bars were rated by unknown reviewers, when conducting subjective analyses such as rating or preference this dataset is subject to each reviewer's personal preference. There is potential that personal preference created biased reviews of chocolate bars based on the bitterness or cocoa percentage found in each chocolate bar sampled.

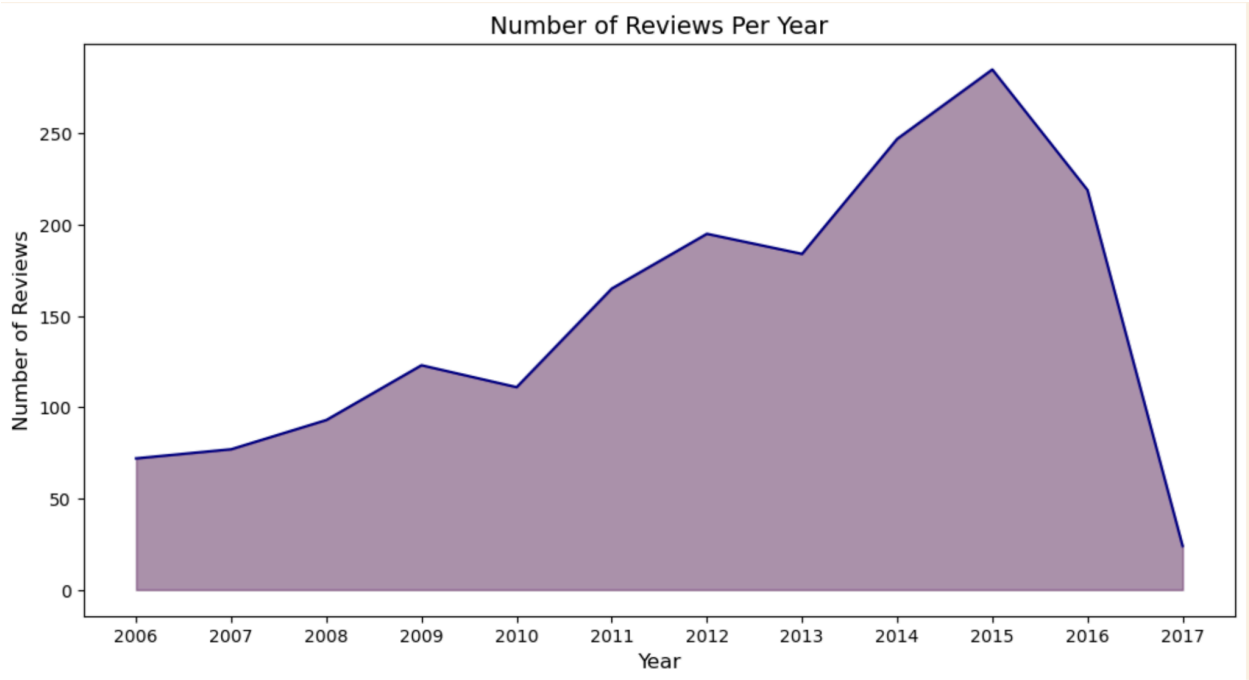


Figure 17: Number of reviews per year for the duration of the study.

This dataset also faced some limitations, such as the “Bean Type” column, which we chose to drop due to high amounts of missing data (~50%). This limited our ability to analyze trends related to bean type. Similarly, the "Bean Origin" column presented challenges due to a high sample size with either blended origins and unknown data, the creation of these broad categories reduced the accuracy of our analyses for objective one.

# Conclusion

Our analysis gave insight to where the highest rated chocolate bar is manufactured and allowed us to investigate where the best cocoa beans originate from, where most of the manufacturing of chocolate bars occurs, what countries use the highest cocoa percentage in their manufacturing and the relationship between cocoa percentage and rating.

The United States was a strong leader in manufacturing (763 locations) with 607 more locations than the next leading country (France, 156 locations). We found the majority of cocoa beans originate from countries within the Amazon Rainforest, potentially speaking to the demand of these bean types or ease to grow. Africa was revealed to be the leader in cocoa use percentage, focusing on producing high-cocoa content chocolate. Contrary to our hypothesis

there was no strong correlation between cocoa percentage and rating, even in the U.S. To see a slightly more negative correlation between cocoa rating and coco percentage among all country manufacturers of chocolate bars compared to those only manufactured in the United States was not what we were expecting. Therefore, we rejected our second hypothesis that there would be a negative correlation between cocoa rating and coco percentage of chocolate bars manufactured solely in the United States.

The highest rated chocolate bar came from Chile. Chile had a total count of one chocolate bar in the dataset. These analyses begin to investigate the multiple factors that influence the production of a chocolate bar. For future work

## Work Cited

Tatman, Rachael. "Chocolate Bar Ratings." *Kaggle*, 2017, https://www.kaggle.com/datasets/rtatman/chocolate-bar-ratings.

Malhotra, Yajat. “EDA and Cleaning Chocolate Ratings.” Kaggle, 2021, https://www.kaggle.com/code/iamyajat/eda-and-cleaning-chocolate-ratings