Project 1 - Group 2 Write Up

This analysis was organized by Arya Maredia, Emily Wimmer, Monica Mitry, and Cecilia Rocha. We analyzed over 1,700 chocolate bars from the Kaggle Dataset, Chocolate Bar Ratings (2006-2017)[1]. The objective of this dataset was to analyze and address the research questions that we presented to determine where the best chocolate bars come from. We emphasized the relationship between which country produces the best cocoa beans and determined if those ratings utilize a relationship with cocoa solid percentage.

As chocolate lovers, we chose this dataset to determine where the best chocolate bars originate from. Chocolate is a delicacy enjoyed globally, however, different cultures influence its production, making it essential to examine the dataset to uncover unique factors and trends that highlight each country's artistry and tradition. Are there countries dominating others in terms of manufacturing? Additionally, do certain countries have a preference on how much cocoa solids are used? We refined high-level questions to help provide precise answers. We are using 3 high-level questions to address these inquiries.

1. What countries do the best cocoa beans for chocolate production originate from?
2. What country uses the highest cocoa percentage?
3. Is there a relationship between the rating of the cocoa and the cocoa solids percentage?

To answer these questions, we first examine the dataset we are working with and adjust it to align with the specific questions we aim to explore. Below is our raw dataset, which contains 9 columns providing us with clear insights for each company. After evaluating each column and data we considered which columns will help us answer the questions, bring clear visualizations, and which ones are to be dropped to ensure a clear dataset.

[1] Tibhar940. (n.d.). *Chocolate bar ratings: Python EDA & data visualization*. Kaggle. Retrieved December 4, 2024, from https://www.kaggle.com/code/tibhar940/chocolate-bar-ratings-python-eda-dataviz

```
Company \n(Maker-if known)            object
Specific Bean Origin\nor Bar Name     object
REF                                    int64
Review\nDate                           int64
Cocoa\nPercent                        object
Company\nLocation                     object
Rating                               float64
Bean\nType                            object
Broad Bean\nOrigin                    object
dtype: object
```

We started off by renaming columns to a simpler title, ensuring no extra titles overcomplicate reviewing the data columns. We also changed the “cocoa percent” column to showcase a decimal instead of a percentage to give the visuals a cleaner look with axis titles.

```
## Before we continue - rename some columns,
original_colnames = choko.columns
new_colnames = ['company', 'species', 'REF', 'review_year', 'cocoa_p',
                'company_location', 'rating', 'bean_typ', 'country']
choko = choko.rename(columns=dict(zip(original_colnames, new_colnames)))
## And modify data types
choko['cocoa_p'] = choko['cocoa_p'].str.replace('%','').astype(float)/100
choko.head()
```

For cleaning data we identified any missing values in the ‘country’ column, and we replaced these missing values with the corresponding entries from the ‘species’ column. This establishes no N/A values remain in the ‘country’ column, allowing the dataset to be cohesive and consistent for analysis.

```
## Is where any N/A values in origin country?
choko['country'].isnull().value_counts()
```

```
## Replace origin country
choko['country'] = choko['country'].fillna(choko['species'])
choko['country'].isnull().value_counts()
```

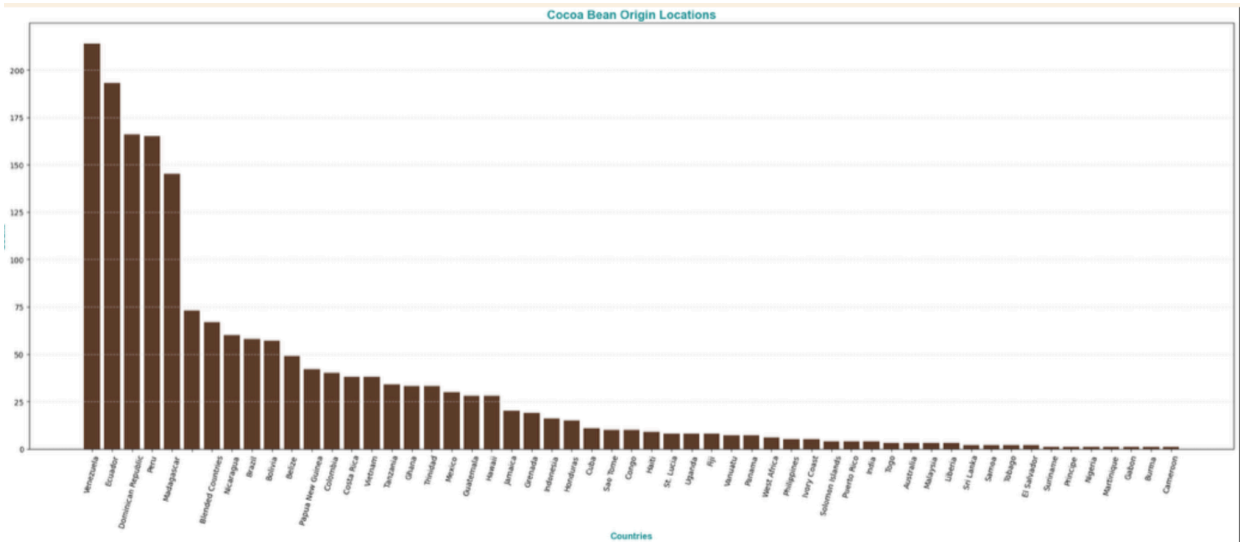
Additionally, we also decided for data cleaning to see if any words are misspelled or have any abbreviation in the ‘country’ column, by sorting and its unique values allowing us to standardize entries, allowing uniformity across the dataset.

```
## Text preparation (correction) func
def txt_prep(text):
    replacements = [
        ['-', ', '], ['/ ', ', '], ['/', ', '], ['\(', ', '], [' and', ', '], [' &', ', '], ['\)', ''],
        ['Dom Rep|DR|Domin Rep|Dominican Rep,|Domincan Republic', 'Dominican Republic'],
        ['Mad,|Mad$', 'Madagascar, '],
        ['PNG', 'Papua New Guinea, '],
        ['Guat,|Guat$', 'Guatemala, '],
        ['Ven,|Ven$|Venez,|Venez$', 'Venezuela, '],
        ['Ecu,|Ecu$|Ecuad,|Ecuad$', 'Ecuador, '],
        ['Nic,|Nic$', 'Nicaragua, '],
        ['Cost Rica', 'Costa Rica'],
        ['Mex,|Mex$', 'Mexico, '],
        ['Jam,|Jam$', 'Jamaica, '],
        ['Haw,|Haw$', 'Hawaii, '],
        ['Gre,|Gre$', 'Grenada, '],
        ['Tri,|Tri$', 'Trinidad, '],
        ['C Am', 'Central America'],
        ['S America', 'South America'],
        [', $', ''], [',  ', ', '], [', ,', ', '], ['\xa0', ' '],[',\s+', ','],
        [' Bali', ',Bali']
    ]
    for i, j in replacements:
        text = re.sub(i, j, text)
    return text
```

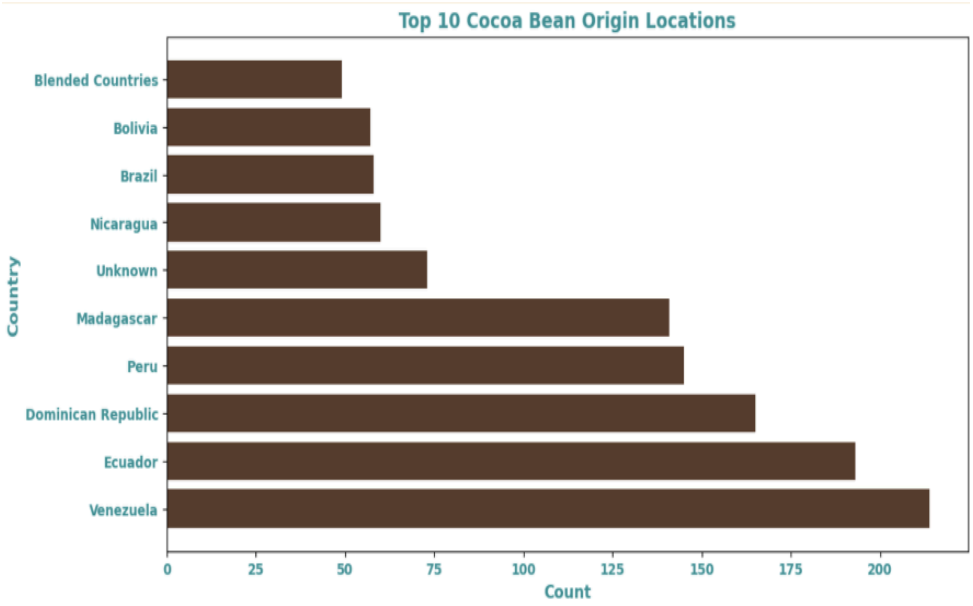
Question 1: What countries do the best cocoa beans for chocolate production originate from?

Understanding the origins of the best cocoa beans and where they are manufactured from is crucial. Locating major manufacturing companies is the key player in identifying trends in the

global chocolate industry. This specific question highlights mutuality within agriculture production and industrial marketing, illustrating how raw materials are sourced and transformed into finished products. Analyzing these factors we gain insight on regional strengths and quality standards directing the chocolate market.

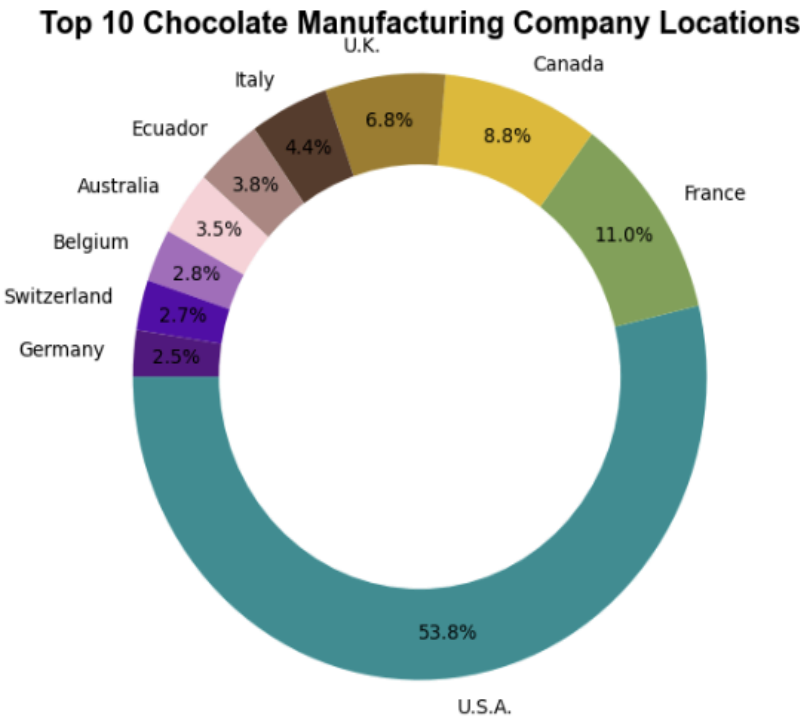


We examined two critical components: the top ten countries where cocoa beans originate and the top ten countries where chocolate manufacturing companies are located. According to the data below, Venezuela emerges as the leading source of cocoa beans, with 214 companies relying on Venezuelan beans for their production, exhibiting the country's importance of providing high-quality raw materials.



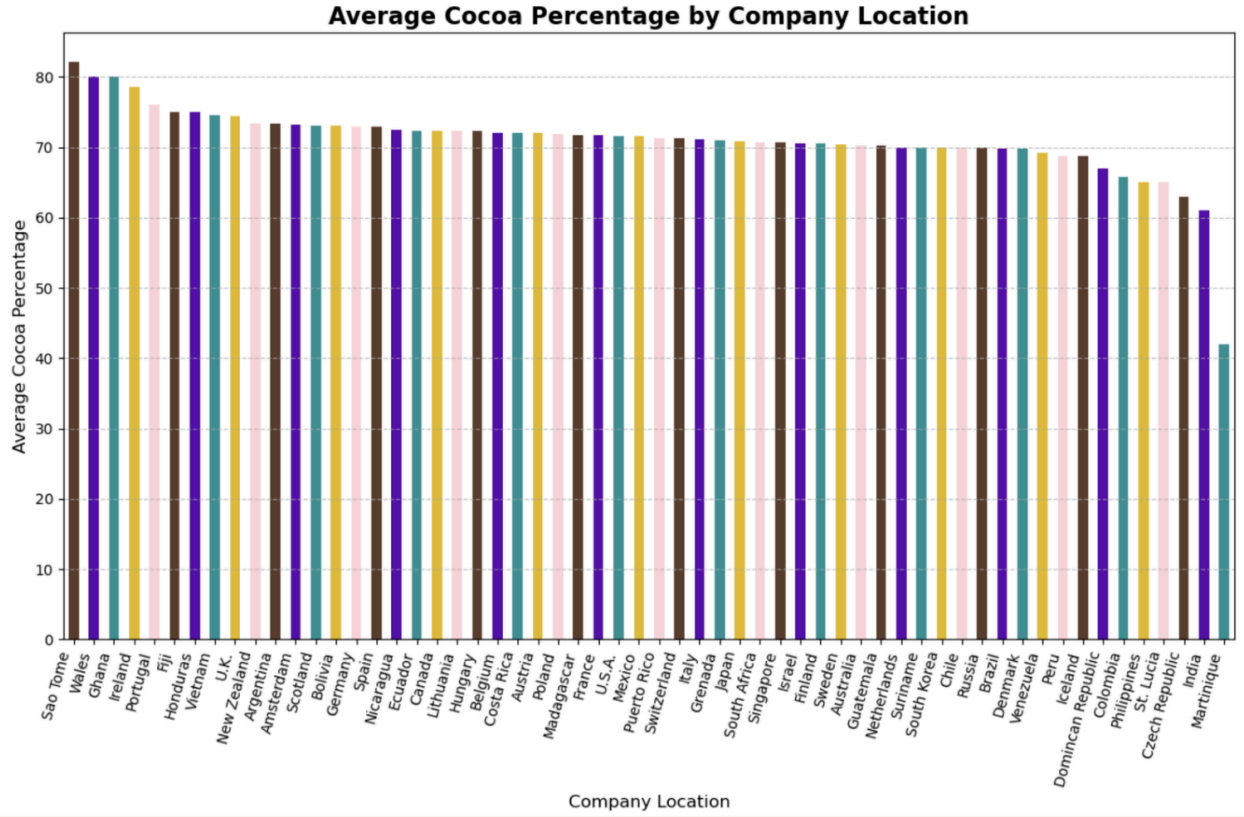
Simultaneously, the United States dominates the chocolate manufacturing division, considering it accounts for 58.3% of the total companies globally, with 763 companies in operation. This emphasizes the U.S.'s role as a key producer of chocolate bars and the focal point of innovation and consumer influence in the industry.

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

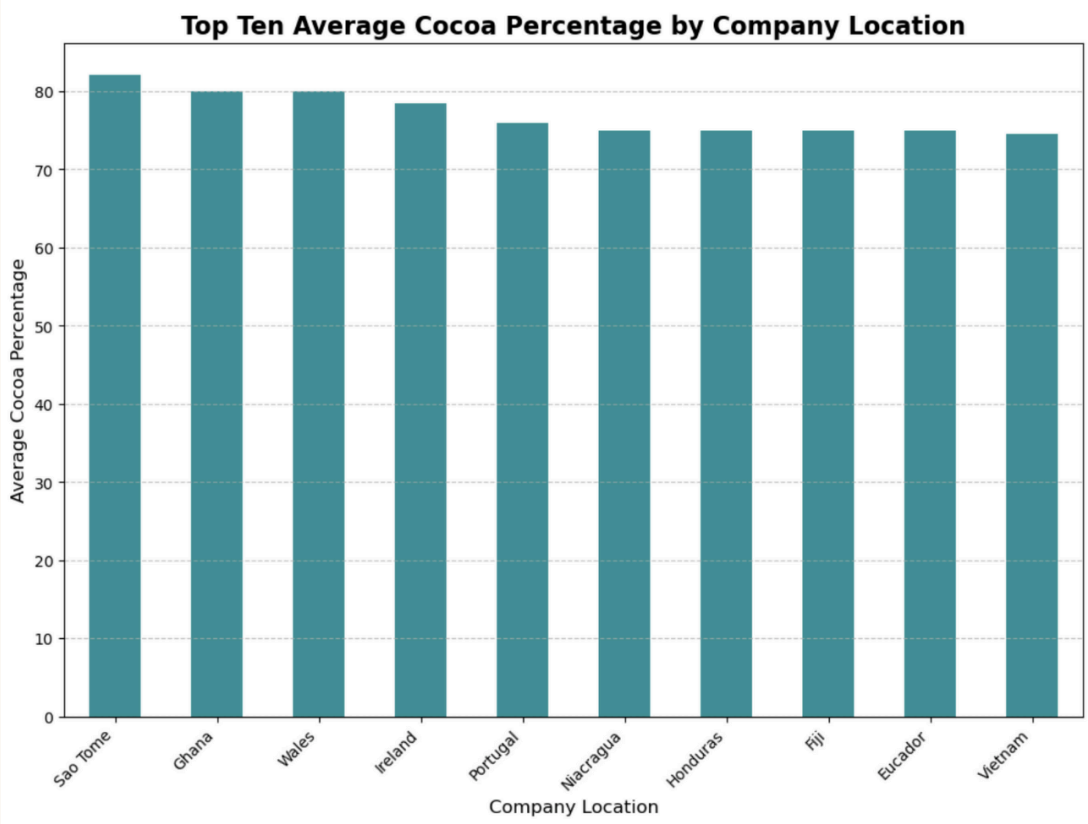


Question 2: What country uses the highest cocoa percentage?

Identifying the country with the highest cocoa percentage identifies insights into regional trends and preferences for chocolate production, specifically dark chocolate, which requires a higher cocoa percentage. To help recognize this, we calculated the average cocoa percentage utilized by companies in each country.

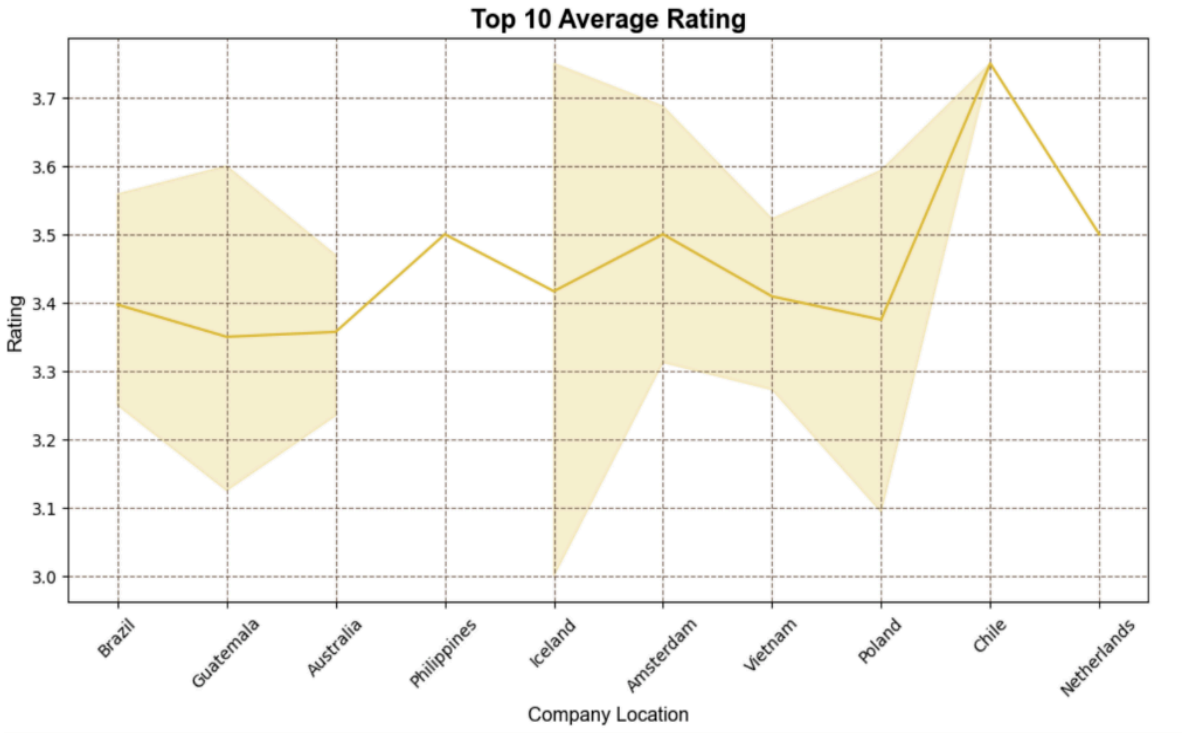


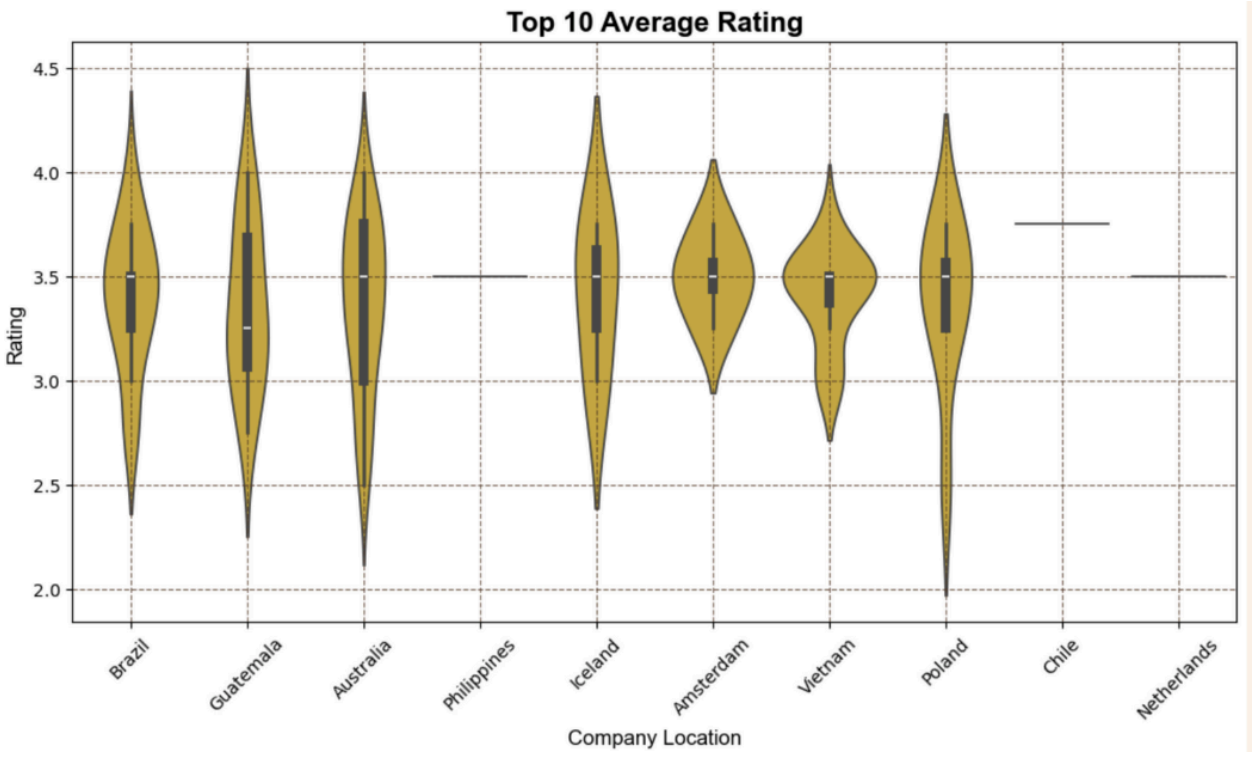
After grouping and analyzing the data, it became predominant that São Tomé has the highest average cocoa percentage at 82%, emphasizing as a leader in producing rich, dark chocolate contrary to Martinique which has the lowest average cocoa percentage at 42%, exhibiting a preference for lighter chocolate. The analysis highlights the diversity in chocolate production and regional consumer preferences across the world.



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

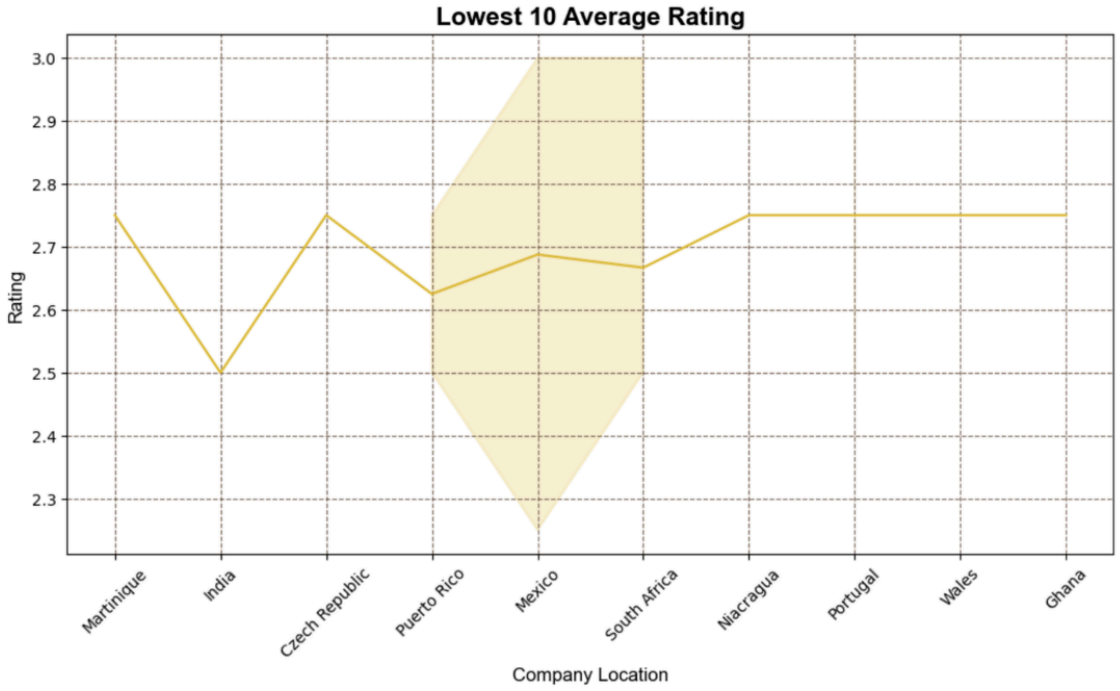
Analyzing the top 10 average chocolate bars by country provides important insights into distribution patterns and quality ratings within the chocolate industry. Countries such as Chile, The Netherlands, and the Philippines stand out the most with their higher distributions. However, identifying some countries with only one bar or country tends to achieve higher average ratings, suggesting that niche production can sometimes yield superior quality.

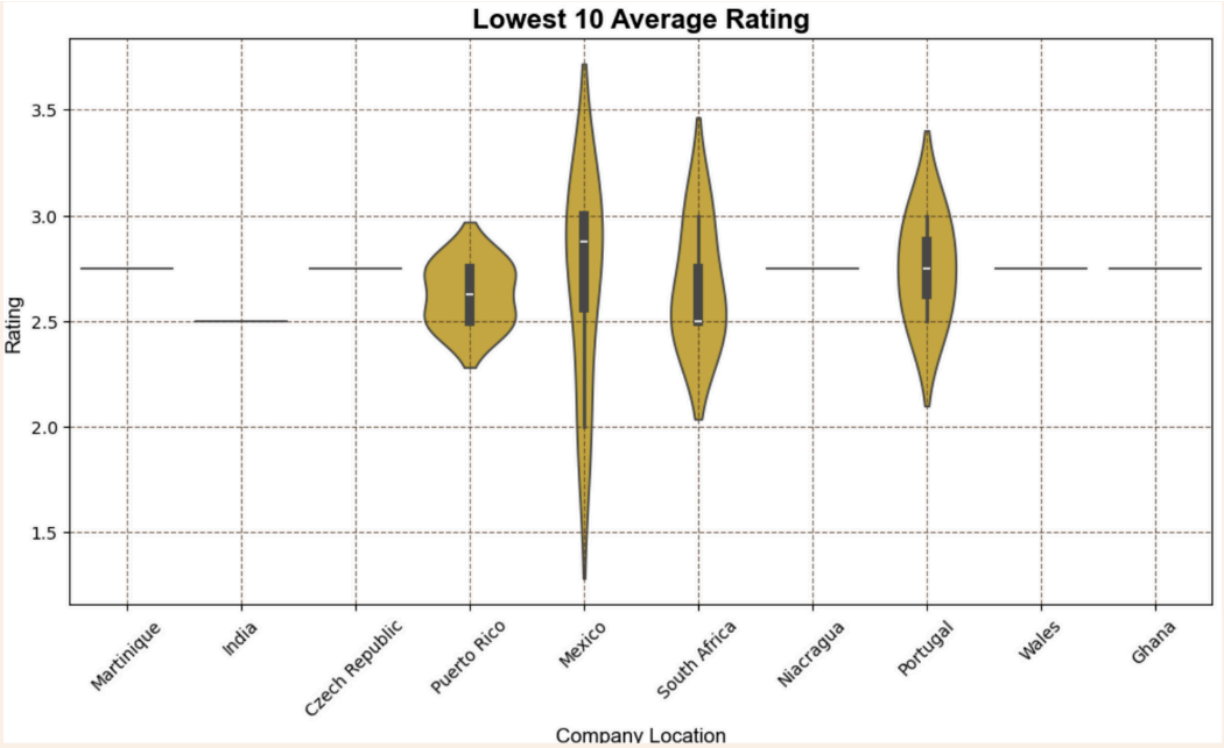




On the other end of the spectrum are countries that have lower average ratings, such as Martinique. These countries tend to have narrower distributions, with fewer offerings

contributing to their overall ratings. This analysis emphasizes the contrast between countries with broad distribution and those excelling in specialized or artisanal chocolate production, including challenges faced by regions with limited chocolate manufacturing.



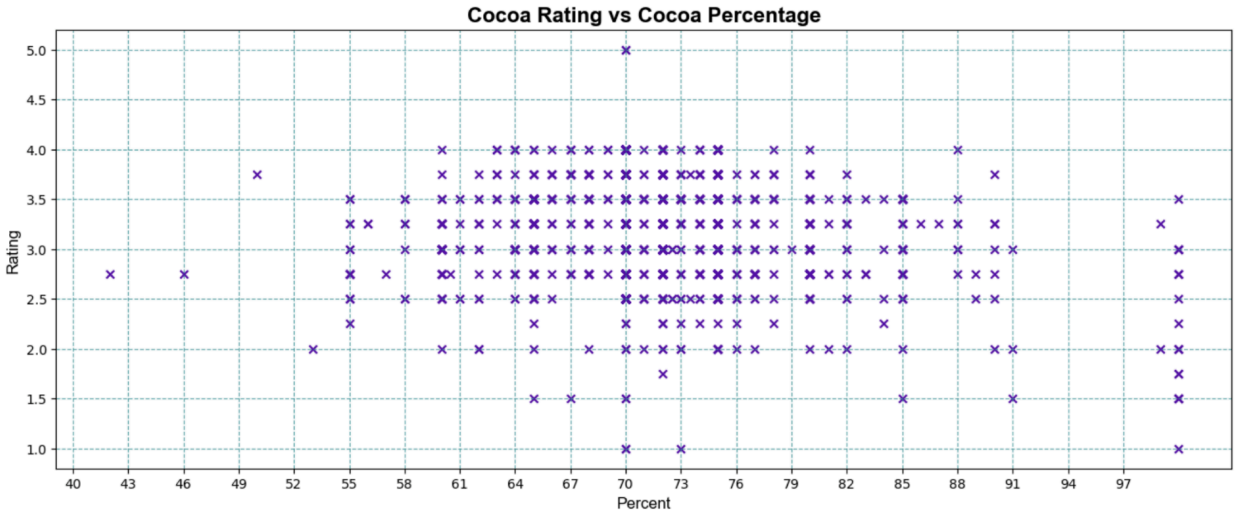


Regression

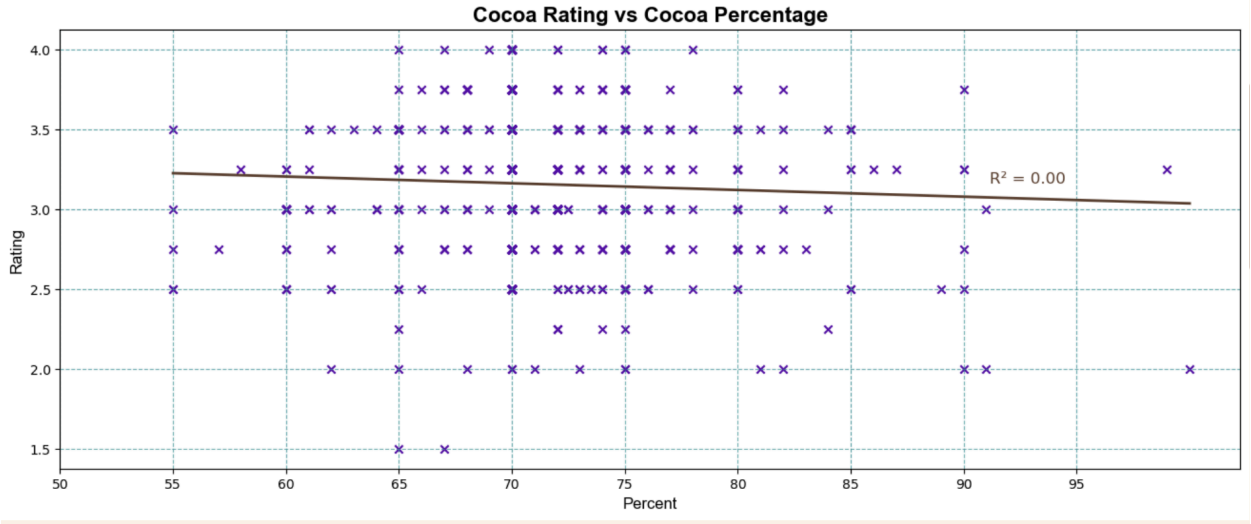
Our regression analysis helped us identify relationships between variables, with its adequacy for answering questions below involving patterns or predictions. For this case, we focused on two key questions:

1. Is there a correlation between cocoa rating and cocoa percentage?
2. Is there a negative correlation between cocoa rating and cocoa percentage in the US?

To explore these questions, we started with creating a scatter plot to visualize the relationship between continuous numerical values, such as cocoa percentage and ratings. This provides a comprehensible picture of how these variables interact and highlights any notable patterns or trends.



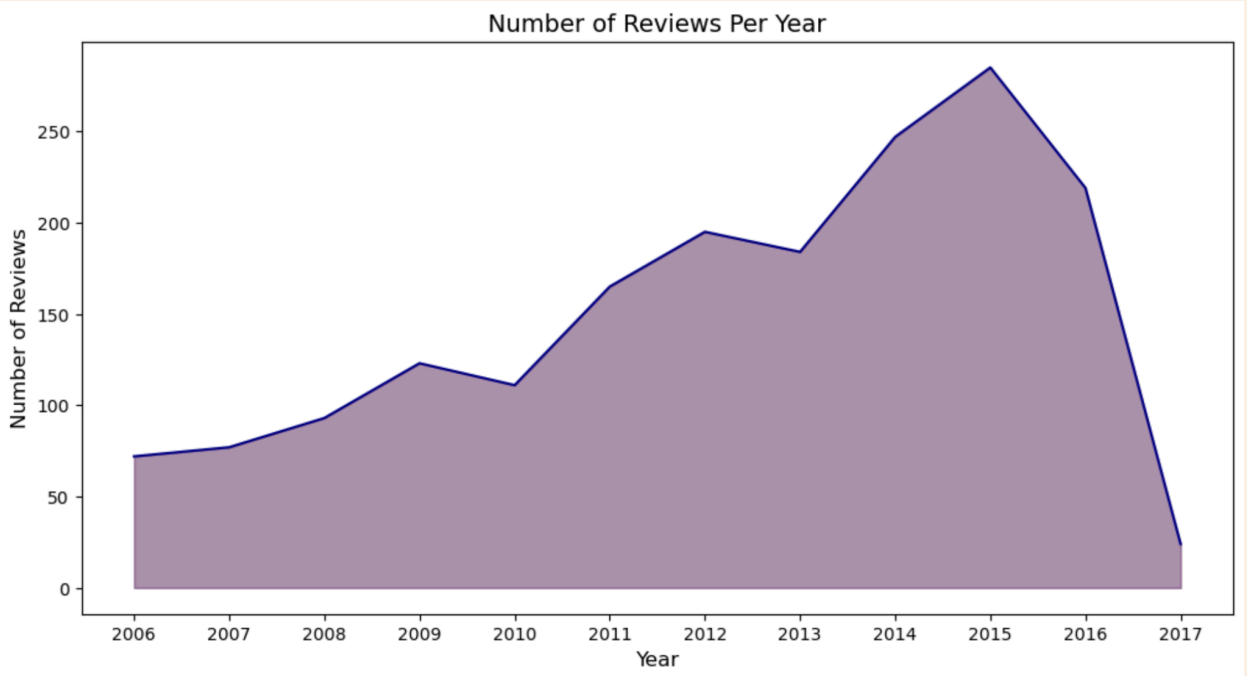
We then narrowed down the analysis to chocolate bars manufactured exclusively in the United States. Isolating this subset allowed a better understanding of how manufacturing location impacts rating and whether the U.S. made bars adhere to the border trends observed in the dataset or exhibit unique characteristics. With this focus it allowed us to approach more precise insights into the factors impacting chocolate quality and consumer preferences.



Bias and Limitations

One of the key biases in the dataset was the uneven distribution of data across years, with 2017 having significantly subsidiary data points compared to previous years. This could influence patterns and trends that depend on temporal comparisons. However, countries with one company producing a single chocolate bar tended to skew average ratings and cocoa percentages. These outliers inordinately affected the results, highlighting the impact of smaller datasets on overall statistics.

This dataset also faced some limitations, such as the “Bean type” column, which had to be dropped due to high amounts of missing data, which came to approximately 50%. This limited our ability to analyze variations based on bean type, such as patterns and trends. Similarly, the "Bean Origin" column presented challenges due to blended origins and unknown data, reducing the unrefinement of insights into the geographic impact of bean sourcing.



Despite the limited number of data points for 2017, we decided to retain this data to preserve the dataset's chronological continuity. Removing it, however, could've excluded potential insights from that year and disrupted any long term trends or patterns. By including it, we ensured a more comprehensive analysis while acknowledging the lower depictions of the subset. This approach inflicts a balance between maintaining historical context and recognizing potential bias.

Conclusion

The analysis revealed several key insights into the chocolate industry globally. The United States leads in chocolate production, possessing the highest number of manufacturing companies, highlighting it as a dominant role in the market. In contrast, the majority of cocoa beans originate from the Amazon Rainforest, emphasizing the region's critical contribution to the supply chain. Africa was revealed to be the leader in cocoa use percentage, focusing on producing high-cocoa content chocolate. Interestingly, dark chocolate bars were rated just as highly as milk chocolate bars, even in the U.S., accentuating a growing appreciation for richer flavors of dark chocolate. Chile, with its one submission, produced the highest rated chocolate bar, showcasing that unsurpassable quality can come from even a single entry. These findings

highlight the diverse factors that are shaping the chocolate industry, with factors such as raw material origin or region production strengths and consumer preferences.