Task A3: Exploring Data using Python and Plotly

Shubham Gupta
A0225160U

February 15, 2022

1 Introduction

- For this task, we aim to explore the Chocolate dataset.
- Through visualizations, we aim to answer a few questions.
- You can access the notebook by launching it in binder here
- Incase you want to setup the notebook locally, the repository is available here.

2 Implementation

2.1 Q1: Who makes the best chocolate with Ecudorian beans?

- In this visualization, we aim to check which countries make the best chocolate from Ecudatorian beans. The scoring is done based on the 'rating' of beans produced by each company.
- The country is identified by the location in of the company. In the visualization, we show the top 10 countries based on the number of ratings we have for each country.
- The visualization consists of a **Horizontal Violin Plot**. We choose this plot to demonstrate the range of ratings for the given to the chocolate made in the countries.
- The visualization is plotted for all breeds, and has a color coding signifying the "Coat Length" of each dog breed.

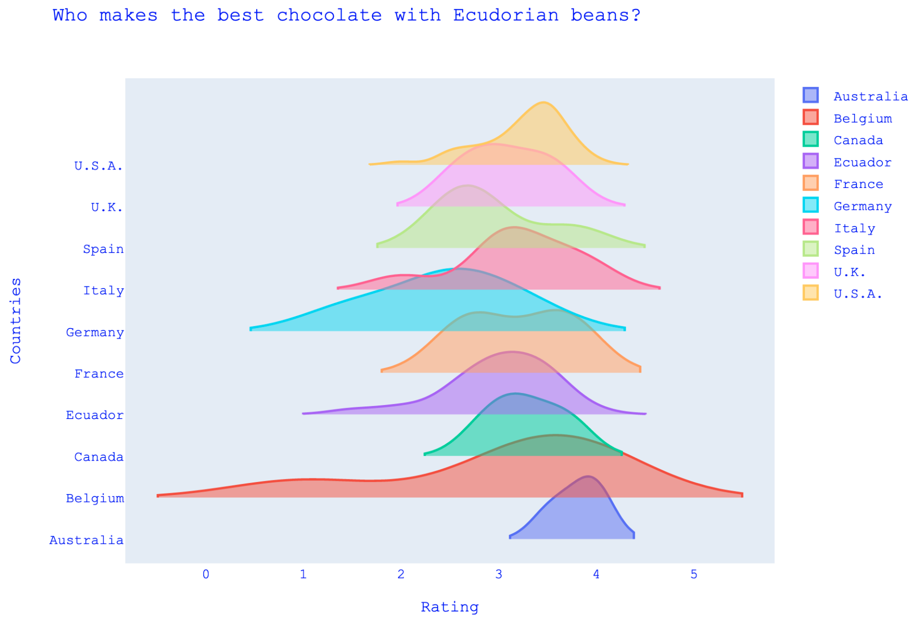

Figure 1: Who makes the best chocolate with Ecudorian beans?

- From the visualization, we see that the maximum rating for is 4 for the countries Belgium, Ecuador, USA and Italy, with Australia having the best ratings.

2.2 Q2: What words are used to describe chocolates?

- In this visualization, we aim to see the distribution of words used to describe chocolates. We compare the words used with the rating given for each chocolate.
- We use a **Box Plot** here, and only show the top 10 words as there are a total of 2455 unique words.

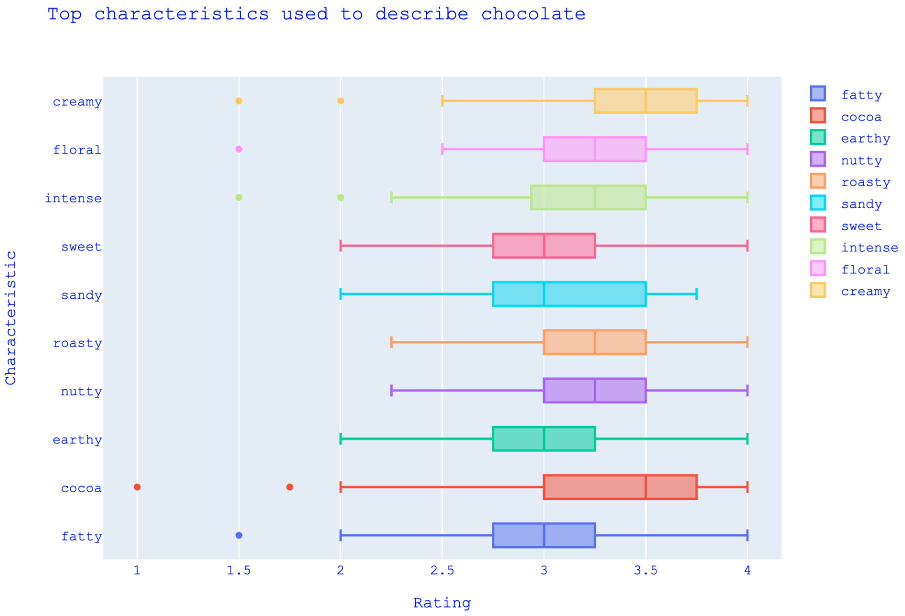

Figure 2: What words are used to describe chocolates?

- We see that ”cocoa” and ”creamy” are the words used to describe highly rated chocolates. Manufacturers should create chocolates with these two ingredients to have the best ratings!

2.3 Q3: Dark Chocolate: Where does it come from and where does it go?

- For this visualization, we aim to visualize the movement of dark chocolate from the source to the destination country.
- By definition, we select only choclates that have a cocoa percentage of above 85% as ”Dark Chocolate”. Note that similar to the previous question, the ”destination” country is decided by the location of the ”company” that produces the chocolate.
- We also preprocess the dataset to label encode both the source and the destination countries.
- We use a **Sankey Diagram** for the visualization, demonstrating the flow of dark chocolate from one region to another.

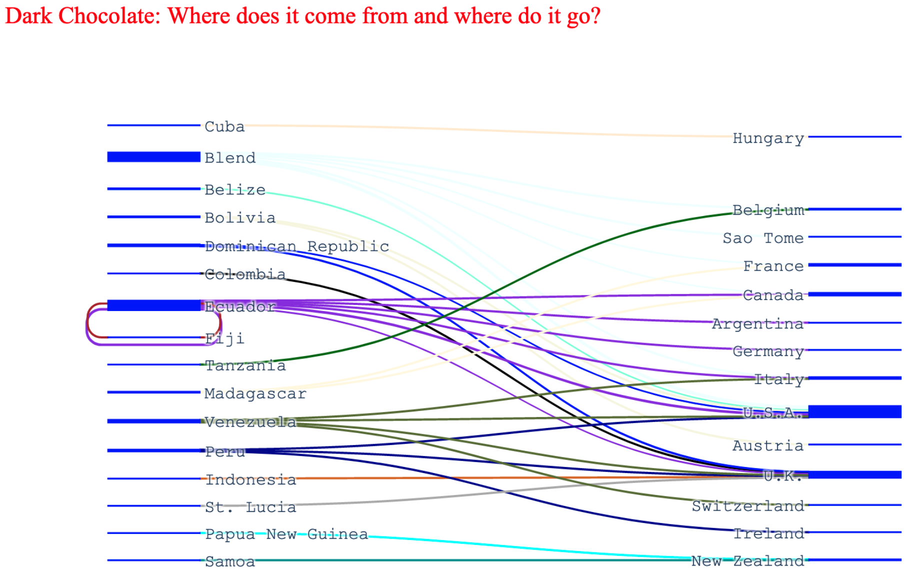

Figure 3: Dark Chocolate: Where does it come from and where does it go?

- From the visualization, it's clear that the country that exports the most amount of chocolate is "Ecuador" and "Blend". The biggest consumers of dark chocolate are the "U.S.A" and the "UK", followed by "Canada".
- Interestingly, we also see some self-loops for "Ecuador" and "Fiji", meaning that they are both the producer and consumer of dark chocolate!