Chocolate Bar Ratings

Data Visualization Project

Maxim Yugai

June 2022

Contents

1 Introduction

Within our course, for my data visualization project I have chosen a dataset about ratings of different chocolate bars from countries all over the world. At first, I thought that this particular dataset contains lots of interesting data and we could have an introductory insight into the chocolate-manufacturing industry using various visualization instruments.

2 Instruments

To complete the task I used several Python and R libraries for visualizations. Plotly, wordcloud and matplotlib in Python, gglot2 in R. Also I built a linear model in R to prove one of my hypothesis.

3 Dataset Overview

Lets have an overview of our data. Here we have a .csv file 1 that I converted to a DataFrame with pandas library. Relevant rows: name of a company, where a company is located, a cocoa bean-dealer country, cocoa percentage, of course, rating, characteristics that were mentioned in the text of a review and review date.

Since we have data about so many countries, I wanted to inspect what countries are the leaders in the production of chocolate (how many chocolate companies they have). For this purpose i built some graphs in Python. As we can see in the graph 2, the undisputed leader is U.S.A., this country has almost 250 chocolate companies. What is interesting is that the difference between the first and the second place is around 200, European countries, actually, do not have so much chocolate companies.

Next, I built a graph that shows 10 countries with the widest range of chocolate bars 3. Again, the leader is United States and the difference between the winner and the second place is huge too.

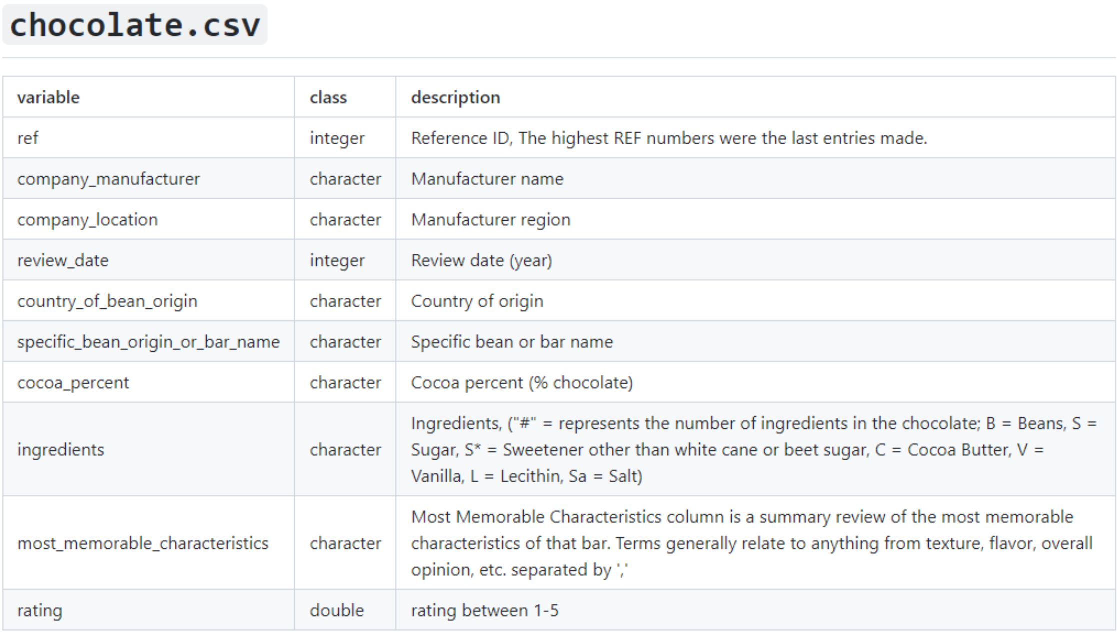
Looking on the graph 4 it becomes clear that Venezuelan beans are supplied to the largest number of companies. Almost all of the countries on this list are in South America, but there is also Madagascar and African country Tanzania.

Also, using R I investigated mean numbers of rating and cocoa content in chocolate bars (graphs 5 and 6). So, the mean number for rating is slightly above 3 on the scale from 1 to 5 and the mean value for percentage of cocoa in bars is around 70 percents.

4 Expectations and hypotheses

In this section I am going to introduce some expectations and hypotheses based on the dataset we have:

1. Switzerland will be the mean rating-leader among all countries (as it is commonly thought to produce high-quality chocolate)
2. The mean rating-leader company will be from Europe
3. There will be a relationship between cocoa percentage and rating. Less cocoa - higher rating
4. Small companies do not progress overtime, there will be no fluctuations in their ratings. For this hypothesis we picked out 4 smallest companies (big or small company - how many different bars a company produce)



chocolate.csv

variable	class	description
ref	integer	Reference ID, The highest REF numbers were the last entries made.
company_manufacturer	character	Manufacturer name
company_location	character	Manufacturer region
review_date	integer	Review date (year)
country_of_bean_origin	character	Country of origin
specific_bean_origin_or_bar_name	character	Specific bean or bar name
cocoa_percent	character	Cocoa percent (% chocolate)
ingredients	character	Ingredients, ("#" = represents the number of ingredients in the chocolate; B = Beans, S = Sugar, S* = Sweetener other than white cane or beet sugar, C = Cocoa Butter, V = Vanilla, L = Lecithin, Sa = Salt)
most_memorable_characteristics	character	Most Memorable Characteristics column is a summary review of the most memorable characteristics of that bar. Terms generally relate to anything from texture, flavor, overall opinion, etc. separated by ','
rating	double	rating between 1-5

Figure 1: Data

5 Results

5.1 Hypothesis 1

Graph 7: the fact that Vietnam has the highest mean rating among so many countries quite surprised me. And even the second place was taken by Brazil. Switzerland comes only on the 4th place.

5.2 Hypothesis 2

Graph 8: our second expectation was proved. Italian company Amedei takes the first place in chocolate rating.

5.3 Hypothesis 3

Graph 9: at first glance, it is rather difficult to say about the dependence of the rating on the percentage of cocoa content in chocolate bars. However, as you can see, the slope is negative, and the t-value (-7,4, p-value = 1,22e-13) also shows that the higher the cocoa content, the lower the rating.

5.4 Hypothesis 4

Graphs 10 and 11: it turned out that our last hypothesis was partially confirmed. Ratings of 3 of the 4 smallest companies stagnates towards the end of the time period, also 3 companies have a drop in rating (before or after stagnation). Ratings of large companies, on the other hand, often change, 2 of them show frequent fluctuations and, also, on average big companies rank higher than small companies.

5.5 Characteristics

In addition, I decided to investigate how reviewers characterize chocolate bars of 4 big and small manufacturers. Interestingly, chocolate of big companies is not described as sweet, unlike production of small companies. Plus, chocolate of big companies which ratings fluctuate is characterized as cocoa and chocolate of companies which ratings show stability is described as creamy.

6 Supportive Visualizations

This section contains all the graph (fig. 7 - 13) which helped me to analyze the data.

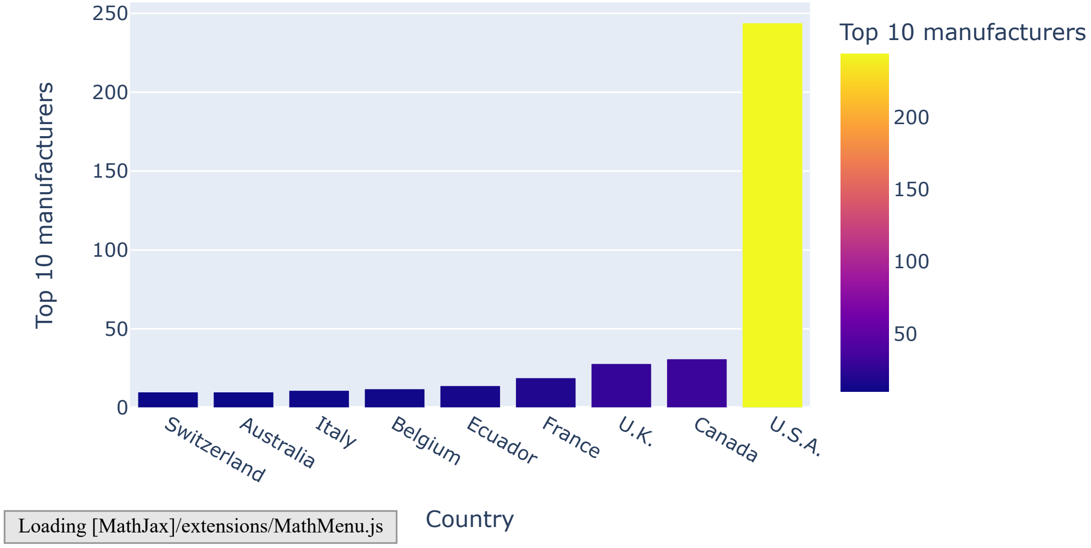
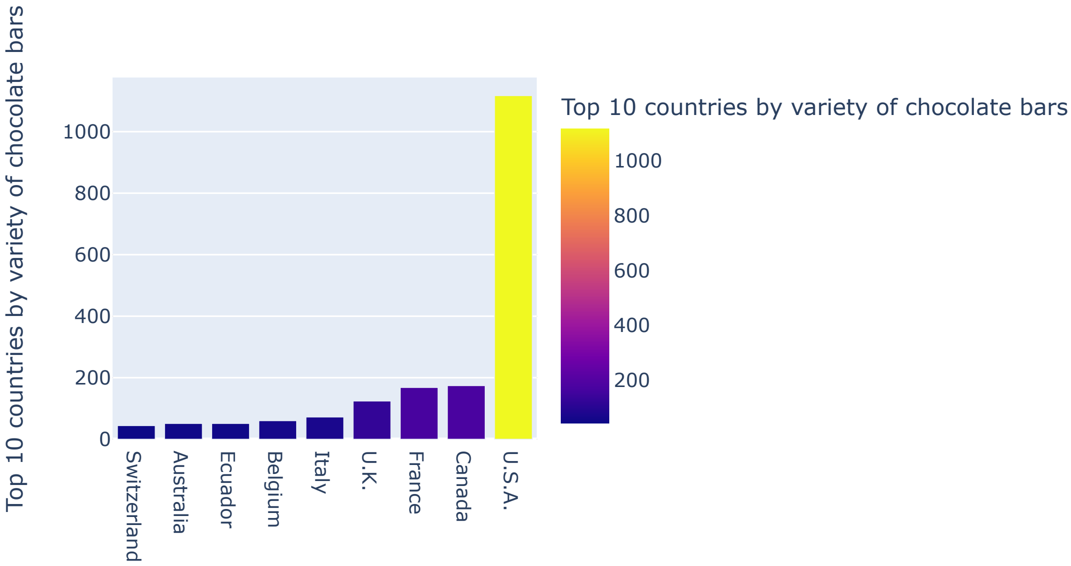
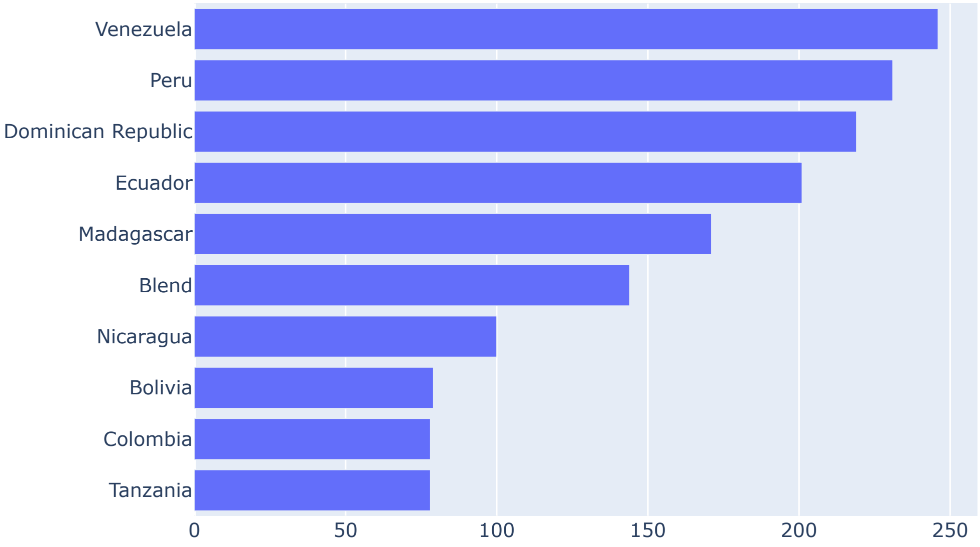


Figure 2: Top 10 countries-manufactures



Loading [MathJax]/extensions/MathMenu.js

Figure 3: Top 10 leaders of chocolate diversity



Loading [MathJax]/extensions/MathMenu.js

Figure 4: Top 10 dealers of cocoa beans

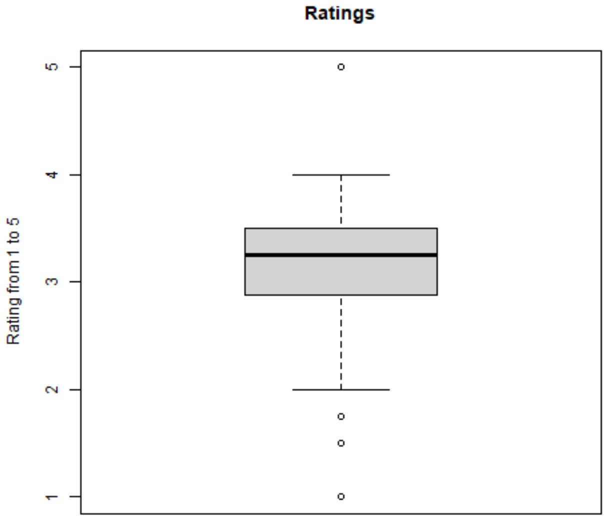


Figure 5: Mean value for ratings

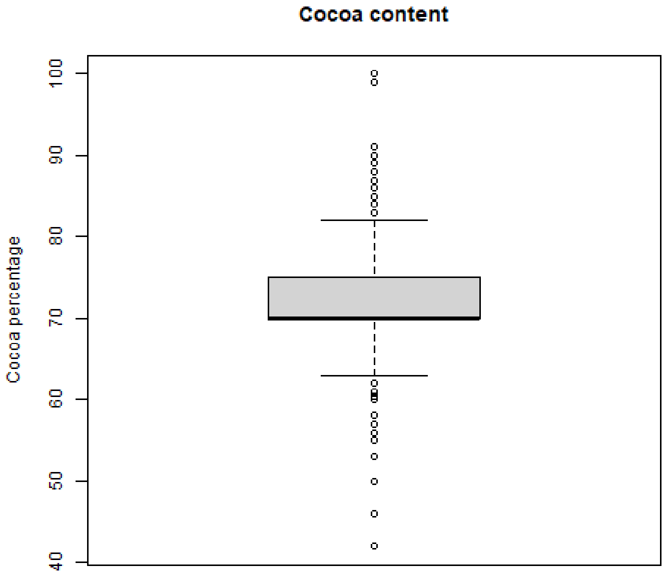


Figure 6: Mean value for cocoa content

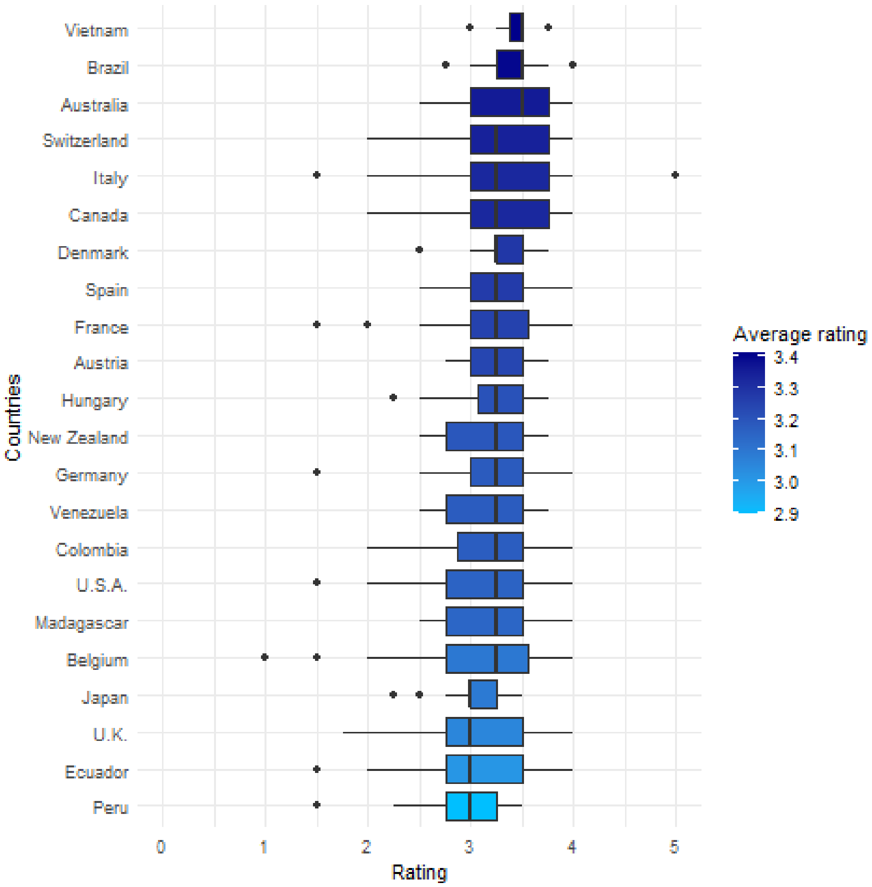


Figure 7: Mean values for rating among countries

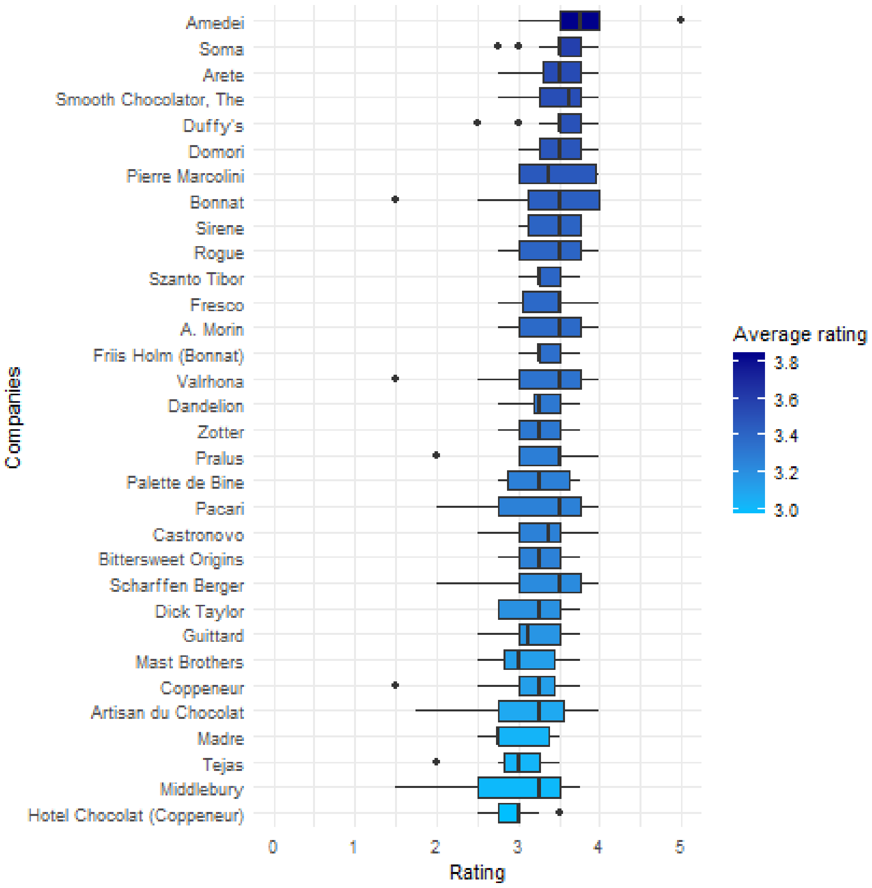


Figure 8: Mean values for rating among companies

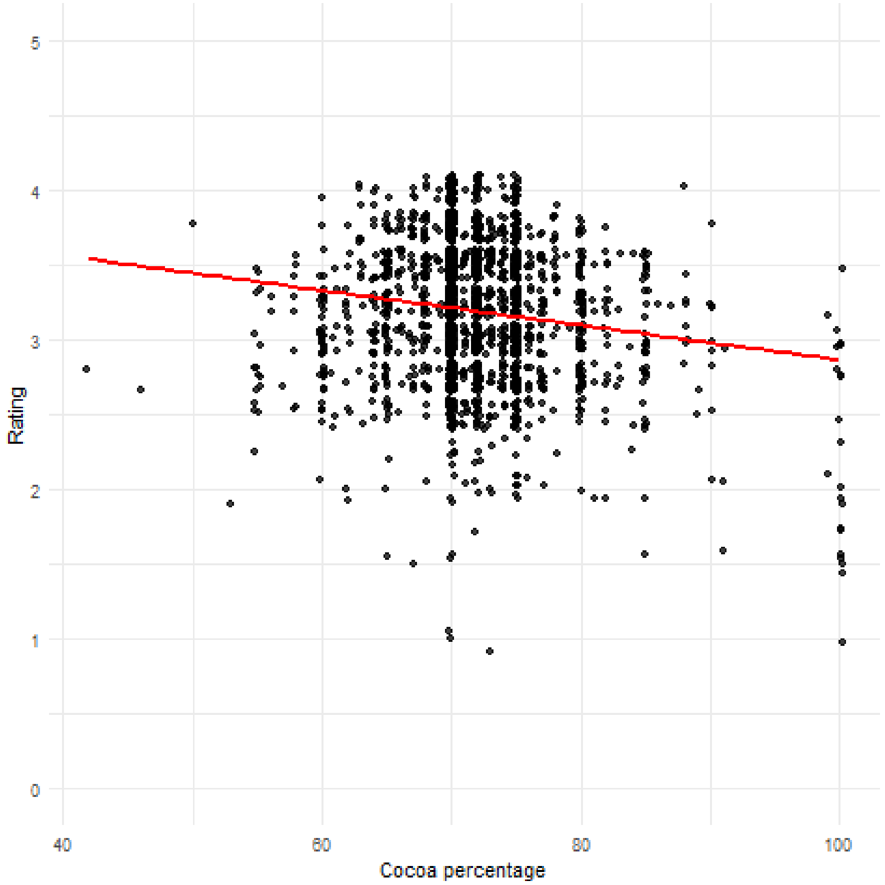


Figure 9: Rating and cocoa percentage

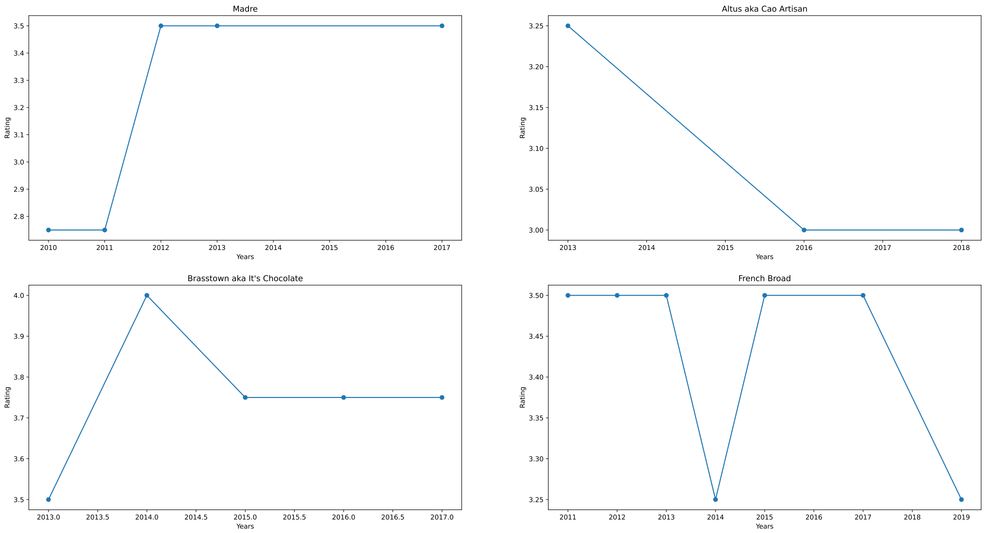


Figure 10: Rating change of 4 smallest companies

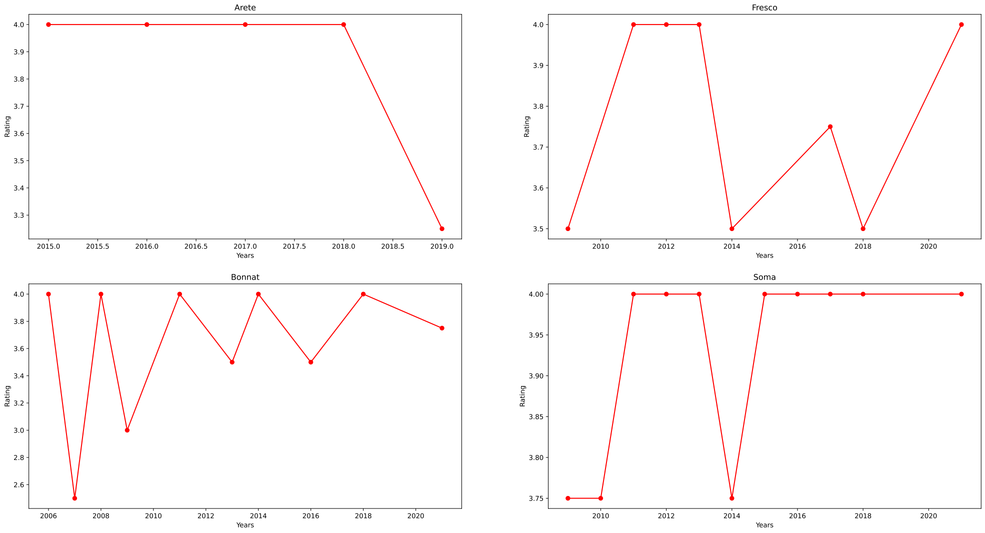


Figure 11: Rating change of 4 biggest companies

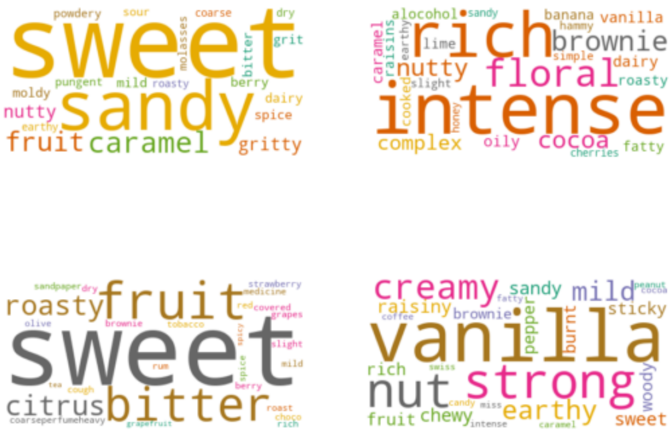


Figure 12: How the chocolate of 4 smallest companies is characterized

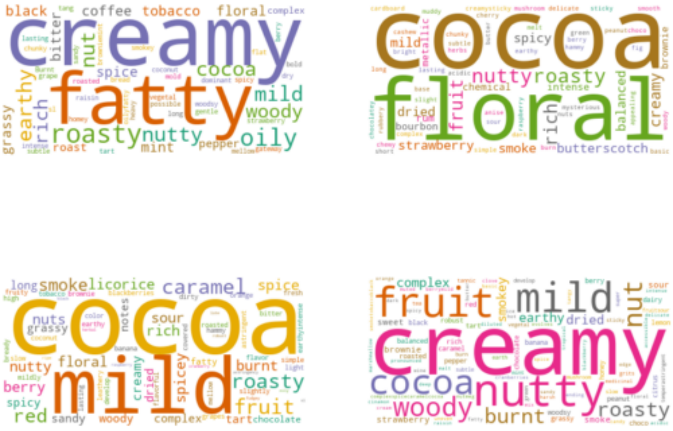


Figure 13: How the chocolate of 4 biggest companies is characterized