NOVA IMS
Information Management School

DIVING INTO CHOCOLATE

GROUP 4

BEATRIZ VIZOSO m20210666
FILIPA ALVES m20210662
HELENA OLIVEIRA r20181121
MARIA ALMEIDA m20210611

DATA VISUALIZATION PROJECT

MASTER IN DATA SCIENCE AND ADVANCED ANALYTICS

APRIL 2022

Introduction

Chocolate is a widely appreciated candy all over the world, but especially in Europe. Sixteen of the top twenty countries that consume the most chocolate are European[1] and Portugal is not an exception.

Having in mind our love for chocolate, we decided to analyse a dataset with information about different chocolates around the world. We took inspiration from previous works and from a dashboard from Observatory of Economic Complexity (OEC)[2] about Cocoa Beans. Our goal was to create a dashboard for every chocolate lover, with interactive visualizations using Plotly and Dash as software tools. In this dashboard, we wanted to make sure the user could obtain a good amount of information about chocolate to help them find their perfect chocolate, based on their choices.

We used datasets from Kaggle and other sources, which are described in the next sections. All the data provided by those datasets was displayed using various visualizations that can be updated with user interaction. We tried to explore the chocolate industry by answering questions such as "What is the perfect chocolate for me?" or "What are the main traders of Cocoa?". All the steps taken until we reached the final app result can be consulted in the Github repository provided. Finally, to make this application accessible to everyone, we deployed it to Heroku.

So, to all the fellow chocolate lovers, we hope you like it!

Link App: https://chocolate-analysis.herokuapp.com/

Link Github: https://github.com/filipacarreira/Chocolate-App

Dataset Description

This work project focuses its interest mainly on the Chocolate Bar Ratings [1] dataset from Kaggle, with information about different chocolates produced all over the world. It is possible to know which was the country of cocoa bean production, what is its respective rating (analysed by experts), ingredients (such as salt, sugar or lecithin), flavours, cocoa percentage and the country the company belongs to. A dataset [2] with the flow of importations and exportations of cocoa through the last decades was also analysed. Moreover, datasets with the coordinates of the countries were used to build the world maps further explained.

Dashboard Visualizations

During the development of each visualization the choice of marks and channels was guided by the principles of expressiveness and effectiveness.

World Map

For the first visualization in the dashboard, it was decided that an overview of the chocolates and cocoa from the dataset in the context of the world would give users a good first impression of the data.

Hence, an interactive map was created where users can choose what information they wish to see displayed, both in terms of the countries and what attributes were presented.

Table 1 was made to summarize the possible options of what information the user wishes to see displayed in the map.

[1] The World Atlas of Chocolate - https://www.sfu.ca/geog351fall03/groups-webpages/gp8/consum/consum.html

[2] Cocoa Beans in OEC - https://oec.world/en/profile/hs92/cocoa-beans

	Rating	Frequency
Country of bean origin	Countries where the cocoa beans are from are displayed according to the average rating the chocolates made with them get.	Countries where the cocoa beans are from are displayed according to the number of chocolates made from their beans.
Company location	Countries where chocolate companies are from are displayed according to the average ratings their chocolates get.	Countries where chocolate companies are from are displayed according to the number of chocolates they make.

Table 1 - Interactive options in the world map.

	World Map
Data Items	Countries
Data Attributes	Ratings, frequency, company location, country of bean origin
Visual Mark	Area
Visual Channel	Colour saturation, position in the map (country)
Encoding Rules	Ratings and frequency – colour; Country – position in the map

Table 2 - Data types and marks and channels for the World Map.

Word Cloud

Then, the user has the possibility to know which companies sell a desired chocolate, by the means of a dropdown, to select the ingredients, and a slider, to select a range of preferred cocoa percentage.

A WordCloud was used to view all the companies with chocolates that have the selected characteristics. If the names of the companies are coloured in pink, then they sell the chocolate with the highest review rating. The number of times the company takes place in the visualization corresponds to the number of chocolates owned with the given filters. The words' size is proportional to the rating of the chocolate. To mitigate the overwhelming number of companies given by some wide filter options, it was decided to spot the 15 best ranked chocolates.

The interactivity is ensured with the hover animation in order to know more details about a specific chocolate, namely its rating and company country. It is possible to zoom a specific area, so that if some companies overlap, it does not compromise the quality and understandability of the visualization.

A quicker and easier way to instantaneously see which company sells the best chocolate with the characteristics chosen is to glance at the boxes providing the company name, rating of the chocolate and company's country. In case there are multiple chocolates with the same best rating, then the company with the highest rating average is outputted (between the chocolates with the desired characteristics).

	WordCloud
Data Items	Chocolate bar
Data Attributes	Company name, Rating
Visual Mark	Points (in the form of words)
Visual Channel	Colour, Size
Encoding Rules	Rating – colour, size

Table 3 - Data types and marks and channels for the Word Cloud.

Radar Chart

After identifying the companies that are of most interest to the user, they now have the possibility to compare them in pairs. It is possible to choose two companies by the means of two dropdowns, one for each of them.

This comparison is done with the use of a radar chart, which is very useful to compare a set of quantitative attributes. The available characteristics to be compared are the level of cocoa, the rating of the company, the number of tastes and the number of flavours. All these variables are an average of the values in all the chocolates owned by the company.

The variables are in a scale from 1 to 5, to guarantee that the scale of the radar chart would be appropriate for every feature being analysed. The only variable that was changed to respect this rule was the level of cocoa, which originally was a percentage.

The interactivity is ensured by the hover action, which allows the user to know more about the point they are analysing, such as the name of the company, the value of the variable or the origin country of the company.

	Radar Chart
Data Items	Companies
Data Attributes	Level of cocoa, rating, number of ingredients, number of flavors
Visual Mark	Dot, area
Visual Channel	Colour, position in the radar axis
Encoding Rules	Company – colour; level of cocoa, rating, number of ingredients, number of flavors – radar axis

Table 4 - Data types and marks and channels for the Radar Chart.

Treemap

To enrich the analysis, the users can also get to know the countries responsible for the most quantity or USD traded in importations or exportations of cocoa, in a year of their choice between 1991 and 2019. The previous choices, except for the year, which is chosen with a slider, can be done by clicking on the buttons shown in the app.

The Treemap is an area-based visualization where the size of the rectangle represents a metric [3]. The metric to be analysed is chosen by the user and it can be the quantity of cocoa or the trade value in USD. Usually, this visualization is served with hierarchical data, however, we chose to use straight-up proportions. According to Crowdsourced Results [3], rectangle areas, such as the ones in the Treemap, have one of the highest accuracy error rates across visual channels. By using straight-up proportions we get to achieve our goal of easily seeing the main countries, regardless of where they are located (continent, region), and reducing the error rate, while getting a better-looking visualization. The user can also hover over the rectangles to see the name of the country and respective metric value.

	Treemap
Data Items	Countries
Data Attributes	Quantity/Trade (USD), flow (imports/export), year
Visual Mark	Area
Visual Channel	Colour, size (Area)
Encoding Rules	Metric (Quantity/Trade (USD)) – colour, size

Table 5 - Data types and marks and channels for the Treemap.

Bean Routes World Map

Finally, to further the knowledge on the usage of cocoa beans throughout the world, the user has the possibility to see in what country each country's beans are used, situating the routes in a map.

This visualization is done using a world map. On top of it, the main routes of cocoa beans were traced. The colour of the route represents its frequency, so darker colours represent frequent routes and light colours the least frequent.

Since there were a lot of routes in the dataset, only the 10% most frequent were plotted.

The interactivity in the visualization is ensured by the hover action, where the user can analyse the country of bean origin, the country where the company using the bean is from and the number of times this route is present in the dataset. This hover action can be done over the countries, and not over the lines. Moreover, the user can zoom in on the map, in case they want to see the routes in more detail.

	Bean Routes World Map
Data Items	Routes
Data Attributes	Origin country, destiny country, frequency
Visual Mark	Line of the route
Visual Channel	Colour, position in the map (country)
Encoding Rules	Frequency – colour; country – position in the map

Table 6 - Data types and marks and channels for the Bean Routes World Map.

Discussion and Conclusion

With the dashboard presented, we believe the main questions that fuelled the project were answered, not only in terms of the cocoa and chocolate markets, but in terms of chocolate suggestions for each person as well.

However, as everything, our work is not perfect. Regarding the dashboard visualization, it is not prepared for different screen sizes, which would be something to improve in the future, so it could be replicable for smaller devices, for example.

Another limitation faced was related to the information provided by the dataset itself. An analysis of chocolate preferences throughout the years or ingredient trends, for example, could have been made to deepen the exploration, had the dataset provided us with information about the years of the production of chocolates, for instance, instead of the review dates.

To validate our work, and as we did not have the means to conduct an extensive study of opinions and surveys regarding the usability of our dashboard, we resorted to showing it to friends and relatives, of different ages, genders and careers in order to better understand and assess the usability of the dashboard. The feedback was good (but perhaps biased).

References

[1] *Chocolate Bar Ratings*. (2017, August 12). Kaggle. https://www.kaggle.com/datasets/rtatman/chocolate-bar-ratings

[2] *Cocoa beans; whole or broken, raw or roasted exports by country |2019.* (n.d.). World Integrated Trade Solution. https://wits.worldbank.org/trade/comtrade/en/country/ALL/year/2019/tradeflow/Exports/partner/WLD/product/180100

[3] Munzner, T. (2015). *Visualization Analysis and Design (AK Peters Visualization Series)* (1st ed.) [E-book]. A K Peters/CRC Press.