**Read me**

**Instructions to recreate code and visualisations**

To start this analysis, you will need to download this dataset as a csv file: [Chocolate Bar Ratings](https://www.kaggle.com/datasets/rtatman/chocolate-bar-ratings).

You will then have to open the dataset in Microsoft Excel to do some data preparation.

I wanted to add a column named Region, in which each country in the column “Company Location” would be filtered into a distinct region. In a new sheet, I used the Excel function ‘Remove Duplicates’ on the column A “Company Location” to obtain the unique values (approximately 60) and then manually assigned each of them a region in column B. I decided to use the region values of Asia; Europe; North America; South America; Oceania & Africa. I then created the Region column in the original dataset by using an Excel vlookup function. This function matches a value from one column with a dataset of >1 column to find a corresponding value. This column will be very helpful for some more general analysis.

I also made both interactive maps in Excel.

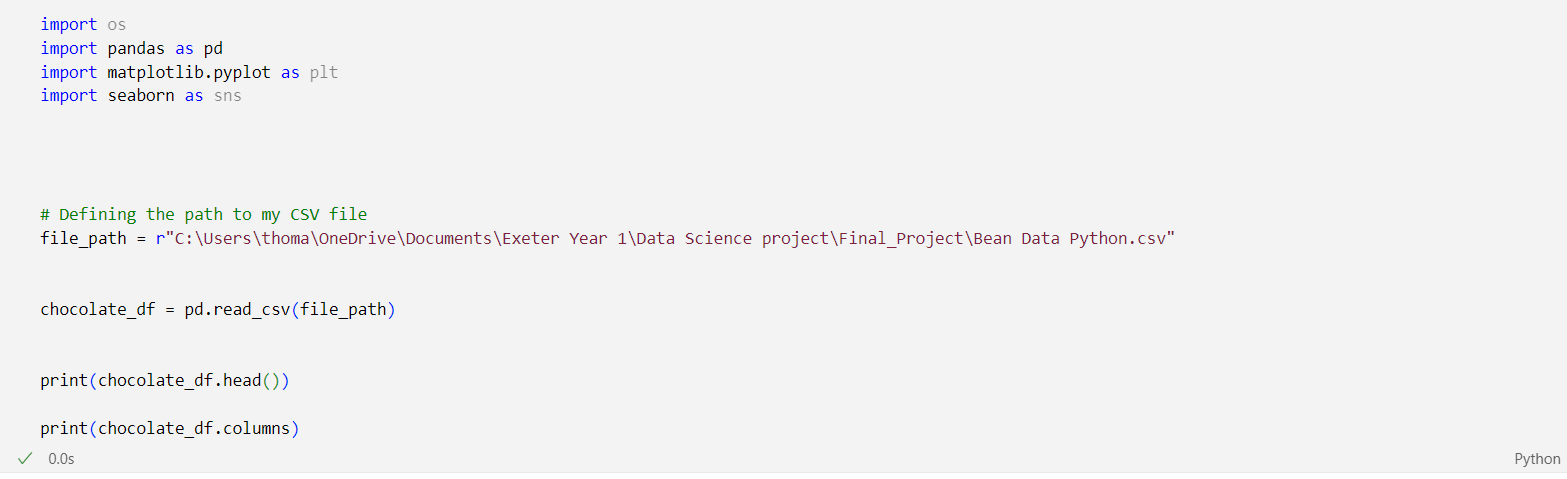
For the first map, “Interactive map of average rating by country”, we need to create a pivot table selecting Company Location as columns and Rating as Values. Once the pivot table is created, to get the Average of rating you will have to select averages by going into Value Field Settings on the pivot table. From here to create an interactive map you will need to copy the values into another sheet, pasting the values there. Then clicking on insert, you should see a Map icon which will create the map. You can then change the style to make it look like the one in my blog. Sadly, when extracting the map into VS code, it can no longer be interactive (when hovering over a country you see the average rating).

For the second map: We will need to do the same steps, however inserting Broad Bean Origin into the pivot table instead of Company Region. Following on from here, the same steps will need to be taken.

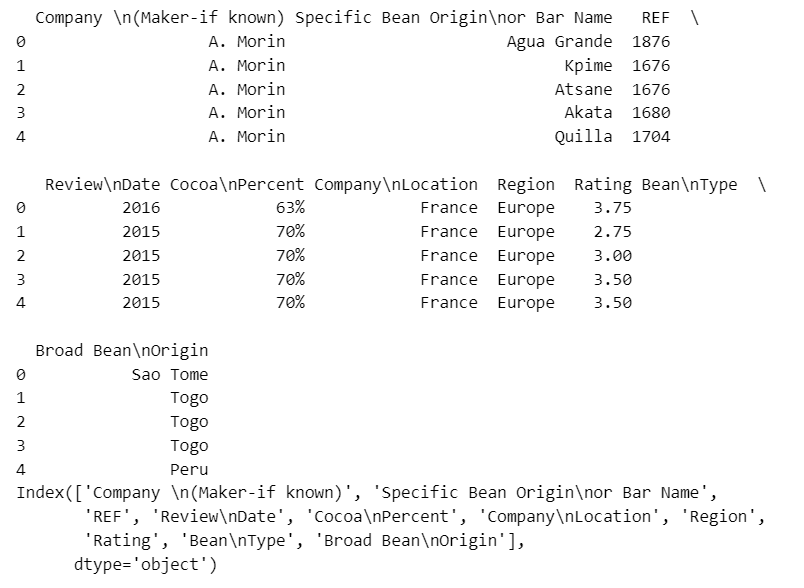
The highest rated chocolate bar can also easily be found by doing filter and find values. There is only one bar rated 5.

Loading the data into VS code:

You will need to run this code- inserting the adequate path to your csv file:



This should be the output:



Now our data is loaded into VSCode!

We can now create visualisations.

For the first plot, the distribution plot, you will need to execute this code:

A screenshot of a computer program

AI-generated content may be incorrect.

This should give this output:

A graph of chocolate bars

AI-generated content may be incorrect.

For the second bar chart, average bar rating by region, you will need to execute the following code:

A screenshot of a computer program

AI-generated content may be incorrect.

If the data preparation was done sufficiently well, you will not need lines 27 to 30. I however, named some South American countries as Latin America so had to combine them in VS Code. Using the groupby and mean functions, we can therefore create the average bar rating by region.

The average values all lay between 3.0 and 3.5, so on the y-axis I modified the range. This I understand can be misleading to the differences in regions average, however I believe it is more visually appealing and still gets the general point across as to which region has the highest average rated dark chocolate.

This should be the seen output:

A graph of a bar graph

AI-generated content may be incorrect.

To further the analysis and add some additional data on cacao bean prices, we will need to do some web scraping. The data I found is from this website: <https://www.indexmundi.com/commodities/?commodity=cocoa-beans&months=240>

To successfully web scrape the data and load it into VS Code as a dataframe you will need to run this code:

A screenshot of a computer program

AI-generated content may be incorrect.

The output should be the following:

A screenshot of a computer screen

AI-generated content may be incorrect.

I then wanted to transfer this data into an SQL database to be able to manipulate it there and create clear tables for my blog.

To do so I used the following code which loads a dataframe into an sqllite database table called ‘Prices’.

A screenshot of a computer

AI-generated content may be incorrect.

You will need to modify the path to the correct one on your computer. Here, the double back slashes are important. Similarly, index=False is very important as when importing to SQL it automatically adds a column named index, this can easily block the code from working. You should now see the table in SQL like so:

A screenshot of a computer

AI-generated content may be incorrect.

In order to obtain the two tables in the blog you will need to run the following SQL statements:

A screenshot of a computer program

AI-generated content may be incorrect.

The first SQL uses the substring function to create a column with just the years, taking the year as the first four characters of the date. I used a column alias ‘Year’ to tidy up the column title. It then creates another column average\_price and orders them in descending order from 2024 downwards. This should output the following:

A screenshot of a computer

AI-generated content may be incorrect.

The second part selects the month with the maximum percentage increase. This should output:

A screenshot of a computer

AI-generated content may be incorrect.

Lastly, I wanted to create a predictive model for future prices. To replicate this, you will need to recreate the following code in VS Code (it is split as it was too long):

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screen shot of a computer code

AI-generated content may be incorrect.

This should create the predictive random forest model, and plot the following graph:

A graph with blue and orange lines

AI-generated content may be incorrect.

**Interesting learnings:**

The first interesting learning to recall came from finding adequate data. I initially was interested in fitness, health and nutrition datasets. I found a dataset on Kaggle, however when performing basic exploratory analysis in Excel to find relationships and correlation between data the data was not adequate. I found this using numerous other datasets, the data was not rich enough to perform interesting analysis. I therefore settled for the chocolate bar rating dataset which has sufficient variables and values to perform interesting analysis.

My second important learning is to remember to create a git repository right from the start of the project. I had already performed most of my analysis and written a lot of my blog before doing so. This led to my git repository not tracking the entirety of my project.

The next learning, probably the one I found most interesting, was focused on the web scraping and loading data into SQL. I had to reflect on what aspects my analysis could improve and what data would further the findings of my project to find adequate data to scrape. I decided to scrape data on cacao bean prices as it could give an interesting angle to the blog. After the web scraping was complete and the data frame was in VS code, I had to learn how to load it into sqllite which is something I have never done. This was interesting and opened my eyes as to what is possible using multiple platforms and coding languages.

I found the idea and creation of my predictive model particularly interesting. I believe this is a tool that could be implemented into the actual chocolate industry and would be greatly helpful. Creating such a tool, even if it is a rather simple version, was very interesting and taught me a lot.

Lastly, in an aim to tidy up my blog post I learnt numerous skills. The first one being how to export from VS code into HTML file leaving just the output and hiding the code blocks. This I believe was key in creating a clean blog post finish. I then learnt how to modify my HTML file in Notepad++, removing some automatically created texts, correcting the header font sizes and adding bold tags. This enabled me to have a satisfactory clean finish on my blog.