Brian Avila

6.1 Sourcing Open Data

**Chocolate Flavor Data Set**

**Data Source**

Source: The source of this data set came from Kaggle. After further research, I identified the data came directly from Manhattan Chocolate Society - Home (flavorsofcacao.com). These ratings were compiled by Brady Brelinski, Founding Member of the Manhattan Chocolate Society. Although the data I pulled is from Kaggle, I believe the data can be deemed reliable.

**Data Collection:**

This data was collected using administrative data and survey responses gathered from different chocolate tasting events.

Administrative data is compiled based on the opinions of the Manhattan Chocolate Society Members as to the given rating.

This dataset contains information on the company (maker of the chocolate bar), its country of location, chocolate bar name or bean origin, Percent of cocoa in the chocolate, Bean type, Bean origin and its rating and date of review.

There are 9 columns and 1512 rows in the dataset.

**Data Limitations:**

A limitation of the data is that the background of the consumers who rated the chocolate is unknown. Some consumers can be biased towards certain criteria more than others. The chocolate rating is the result of an educated opinion about chocolate. The data started to become available in 2007.

**Ethics:**

This information is displayed on the source's website Manhattan Chocolate Society - Home (flavorsofcacao.com) as being available to the public. There does not appear to be any ethical concerns with the data.

**Relevance:**

I believe this data set meets the necessary requirements for this project as it is open source,

includes a geospatial component and meets the size and variable requirements. While not all data collected is recent, this data set was provided in the Career Foundry Project Brief. I enjoy chocolate and thought it would be interesting to dive in to find out more about this delicacy and ratings from consumers.

**Data Profile**

| variables | data type | | | |
|---|---|---|---|---|
| Company Name | qualitative | nominal | structured | time invariant |
| Bar Name | qualitative | nominal | structured | time invariant |
| REF | quantitative | discrete | structured | time variant |
| Review Date | quantitative | cont | structured | time variant |
| Cocoa % | quantitative | discrete | structured | time variant |
| Company Location | qualitative | nominal | structured | time invariant |
| Rating | quantitative | discrete | structured | time variant |
| Bean type | qualitative | nominal | structured | time invariant |
| Bean origin | qualitative | nominal | structured | time invariant |

**Data Cleaning Measures**

**Data Wrangling**

| Columns dropped | Columns Renamed | reason |
|---|---|---|
| REF | | Unnecessary for analysis |
| | Company to Company Name | More descriptive |
| | Broad Bean Origin to Bean Origin | More concise |
| | Location to Company Location | More descriptive |

**Consistency Checks**

| Consistency Checks | | | | |
|---|---|---|---|---|
| | | | | |
| Merged Sao Tome and Principe data as it is the same country | | | | |
| Changed Amsterdam to Netherlands (Location consists of country names, no cities) | | | | |
| Corrected mispelled locations | | | | |
| Removed Bean subtypes, all merged according to type | | | | |
| Omitted/removed small number of blanks and error values from bean origin, and bean type columns | | | | |

Questions to Explore the Analysis:

1. Which top 5 chocolate bars had the best rating?
2. Which company manufactures the highest rated chocolate bar?
3. What is the cocoa % on the top 5 chocolate bars?
4. What bean type had the highest ratings?
5. Where is most chocolate originated from?
6. Where does the highest rated chocolate stem from?
7. How many of the chocolate companies are in USA?