Milestone Report - Kaggle Chocolate Bar Ratings

Saleha Bakht

The Problem

These ratings were compiled by Brady Brelinski, Founding Member of the Manhattan Chocolate Society. Answers to the questions below would influence his advice and recommendations to chocolate manufacturers and retailers.

The questions are as follows:

a. Where are the best cocoa beans grown?
b. Which countries produce the highest-rated bars?
c. What's the relationship between cocoa solids percentage and rating?

Cleaning the Dataset

First I replaced all the column names so that there were no line breaks in the names.

Then, I checked to see where the null values were in all of the columns.

For the columns that had no null values, I checked their content for spelling and datatype. Spelling was corrected if found. Only the Cocoa Percent columns had to change datatypes as it came in datatype Object written with percentages. It was converted to a float type to allow for easier manipulation in future explorations.

After searching for null values, the two columns found remaining with null values were the 'Bean Type' and 'Broad Bean Type' columns.

For 'Broad Bean Origin', the null values were replaced with the corresponding value from the 'Specific Bean Origin or Bar Name' column. That only provided a value for one row and still left a host of other empty cells in the 'Broad Bean Origin' column.

These cells are left empty. There are 2 options I can take: either input Blend is a value for all the empty cells or create a model that will predict accurately where the cocoa bean actually came from. I left the values as empty/null cells to avoid making assumptions that could affect the accuracy of my analysis.

Other Datasets

The analysis was performed by the data provided on the Kaggle website. Since that data was provided by the client, no other sources were appended and analysed to solve the problems.

The dataset was limited to plain dark chocolate bars so we don't have to accomodate for missing ingredient information. No peanuts or caramel appears to have been used in the making of these bars

and an investigation into those ingredients within the bars in the dataset would not influence our analysis any.

Initial Findings

We explored the dataset a bit after cleaning to see if we could find anything interesting before attempting to answer the questions posted on Kaggle. We found that Amedei is the only company in the dataset to have 5.00 rating for a chocolate bar. The last bar to be given such a rating was in 2007. No company has had a chocolate bar rated 5.00 since.

To answer the questions asked on Kaggle:

Where are the best cocoa beans grown?

For our investigation we took the term ‘best’ and defined it as bars having a rating of 4.00 or higher. It seems like the best cocoa beans are blends grown from a combination of Dominican Republic, Madagascar, Grenada, Papua New Guinea, Hawaii, Haiti, Guatemala, Peru, Venezuela, Bolivia, and Java.

Which countries produce the highest-rated bars?

This question was answered using the countries of the company location rather than using the countries where the cocoa beans came from. The answer was Amedei, located in Italy. There is no single 4.50 rating of a chocolate bar in the dataset. However, since the Amedei rating is more than a decade old, it makes more sense to say the the higher ranking bars are made by Woodblock in the U.S.A. more recently.

What’s the relationship between cocoa solids percentage and rating?

There are no obvious winners in which Cocoa Percent has a consistently higher rating. There are two instances of a 5.0 Rating and they both occur with the Cocoa Percent is 0.7 (70%). The boxplot marks those 5.0 points as outliers. Most Cocoa Percents will land a rating from 2.0 to 4.0 regardless. There seem to be a higher quantity of ratings from 0.6 to 0.78 on the Cocoa Percent scale but that range is also where most of the ratings 2.0 or less are. There is no obvious trend for which Cocoa Percent has the better rating. 0.7 Cocoa Percent had two 5.0 ratings but it also had two 1.0 ratings, also marked as outliers. There are no consistent findings.