Capstone 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.

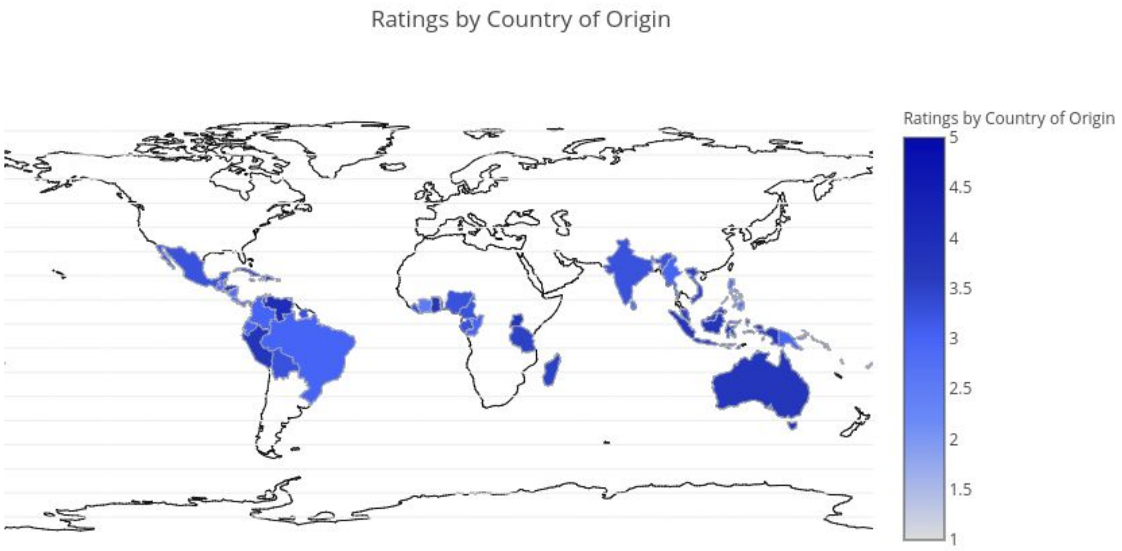
These cells are left empty. There are 2 options I can take: either input Blend is a value for all the empty cells or create a model that will predict accurately where the cocoa bean actually came from. I left the values as empty/null cells to avoid making assumptions that could affect the accuracy of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

In the third section of the code, titled ‘Answering Unasked Questions (EDA)’, a map is made from the gathered data. This map is cross-referenced with a map of the world maps to see if there a gradient patterns in the ratings of chocolate bars and the water source responsible for the growth of the beans.



We can see beans mostly originate from the Southern Hemisphere. The higher rated beans come from Australia and some western regions of South America. We can cross-reference this map with a map of the world's oceans to see which water sources are responsible for the growth of cocoa beans in those coastal countries.

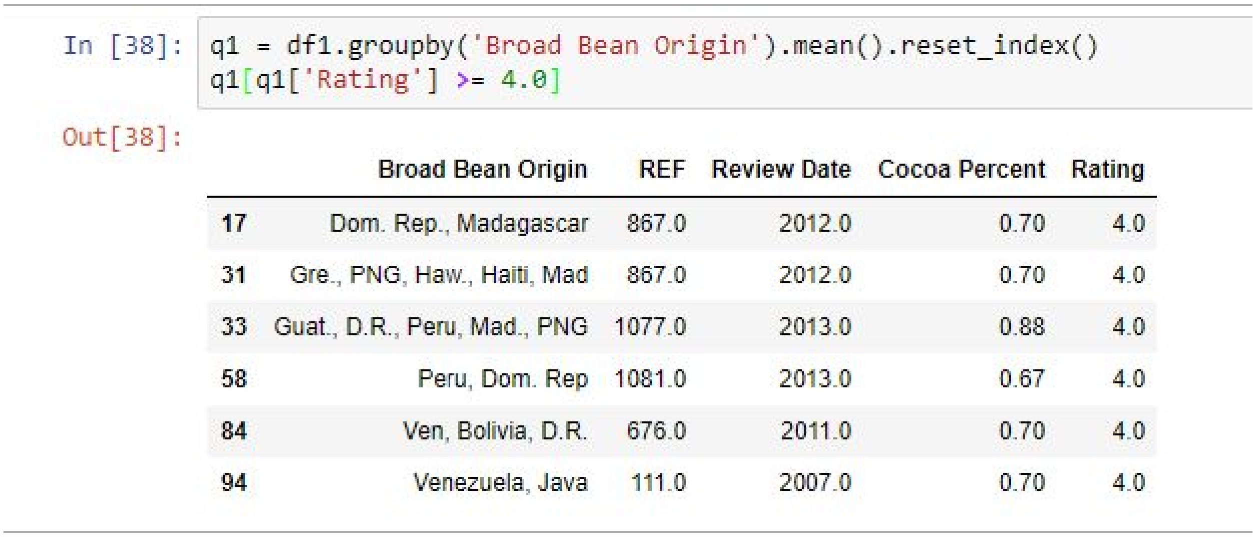
Initial Findings

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.



```
In [38]: q1 = df1.groupby('Broad Bean Origin').mean().reset_index()
         q1[q1['Rating'] >= 4.0]
```

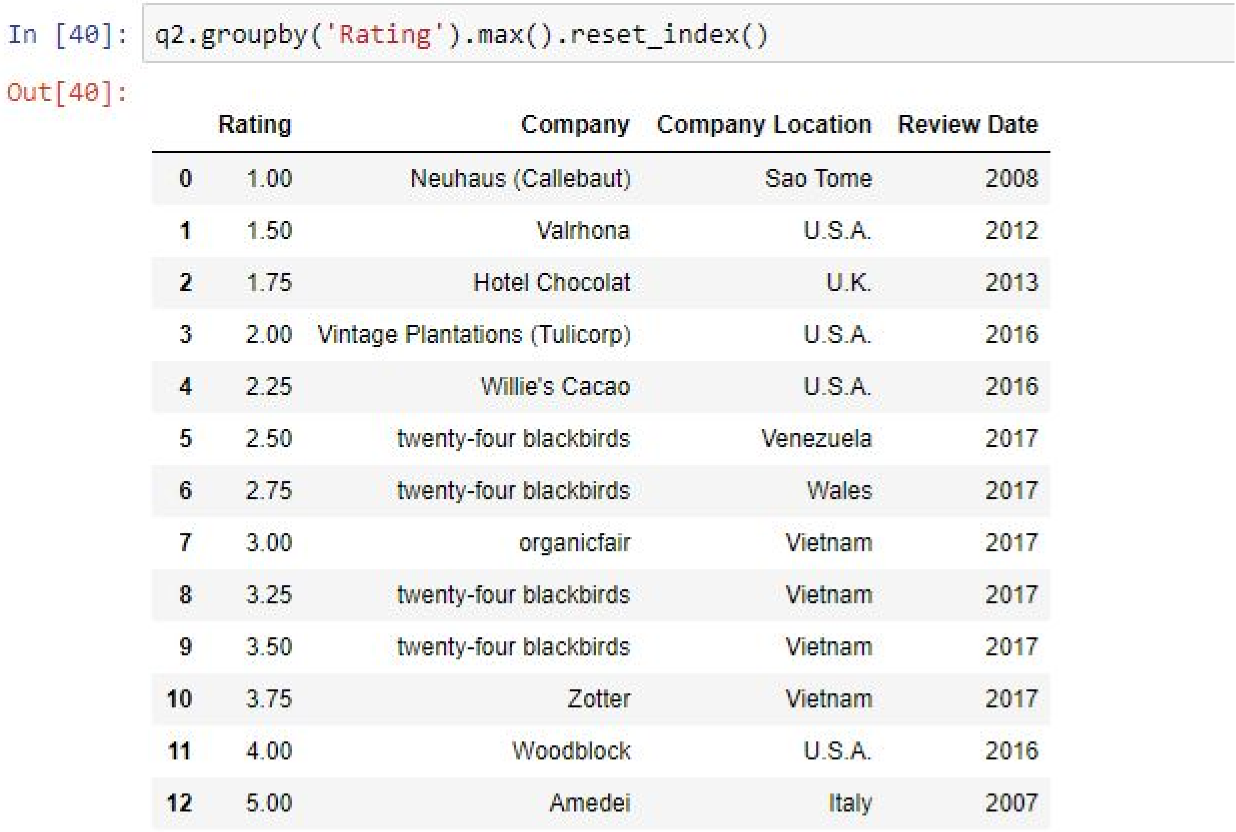
```
Out[38]:
```

	Broad Bean Origin	REF	Review Date	Cocoa Percent	Rating
17	Dom. Rep., Madagascar	867.0	2012.0	0.70	4.0
31	Gre., PNG, Haw., Haiti, Mad	867.0	2012.0	0.70	4.0
33	Guat., D.R., Peru, Mad., PNG	1077.0	2013.0	0.88	4.0
58	Peru, Dom. Rep	1081.0	2013.0	0.67	4.0
84	Ven, Bolivia, D.R.	676.0	2011.0	0.70	4.0
94	Venezuela, Java	111.0	2007.0	0.70	4.0

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.



```
In [40]: q2.groupby('Rating').max().reset_index()
```

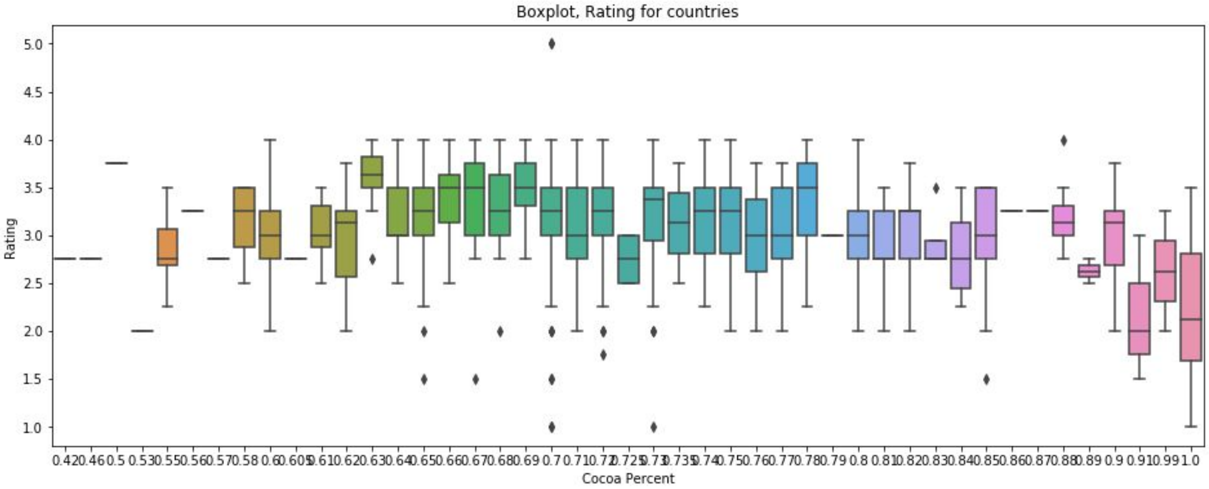
Out[40]:

	Rating	Company	Company Location	Review Date
0	1.00	Neuhaus (Callebaut)	Sao Tome	2008
1	1.50	Valrhona	U.S.A.	2012
2	1.75	Hotel Chocolat	U.K.	2013
3	2.00	Vintage Plantations (Tulicorp)	U.S.A.	2016
4	2.25	Willie's Cacao	U.S.A.	2016
5	2.50	twenty-four blackbirds	Venezuela	2017
6	2.75	twenty-four blackbirds	Wales	2017
7	3.00	organicfair	Vietnam	2017
8	3.25	twenty-four blackbirds	Vietnam	2017
9	3.50	twenty-four blackbirds	Vietnam	2017
10	3.75	Zotter	Vietnam	2017
11	4.00	Woodblock	U.S.A.	2016
12	5.00	Amedei	Italy	2007

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?

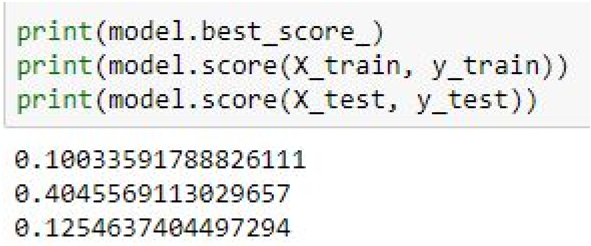
A boxplot was used to view distinctions in the graph not otherwise clear or delineated in a swarmplot.



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.

Machine Learning

One round of RandomForestRegressor() was used to attempt to predict the Ratings value of chocolate bar. This function did have some level of success as it came with a high score. We could not use a 'bag of words' model since we did not have a consistent vocabulary. 'Amedei' did not always mean 5.0 rating and neither did 'Italy'. We could not supply positive or negative consistent connotations with the strings in the dataframe. Therefore, we removed the strings from the dataframe that we used to predict the ratings of the chocolate bars.

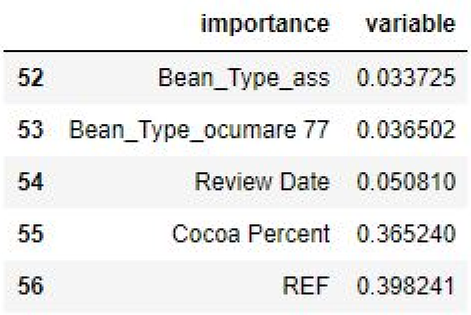


```
print(model.best_score_)
print(model.score(X_train, y_train))
print(model.score(X_test, y_test))
```

```
0.10033591788826111
0.4045569113029657
0.1254637404497294
```

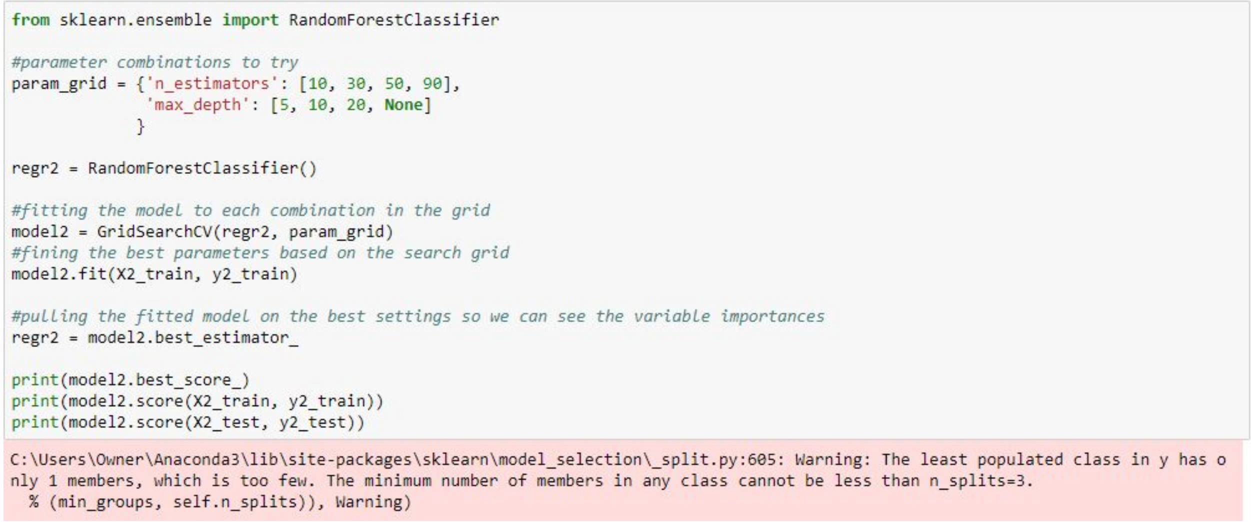
The scores provided are of the metric R^2. The highest possible score is 1. We can see that the accuracy of the test set did was greater than accuracy of the best score. However, it did not do half as well as the training set performed.

A variable importance chart was made for the variables involved in the classifier.



	importance	variable
52	Bean_Type_ass	0.033725
53	Bean_Type_ocumare 77	0.036502
54	Review Date	0.050810
55	Cocoa Percent	0.365240
56	REF	0.398241

As was mentioned in the first section of this report, in ‘Cleaning the Dataset’, an attempt was made to determine the missing Broad Bean Origin values using a predictive model. This was unsuccessful as sci-kit learn determined that enough information for the endeavour was not available.



```
from sklearn.ensemble import RandomForestClassifier

#parameter combinations to try
param_grid = {'n_estimators': [10, 30, 50, 90],
              'max_depth': [5, 10, 20, None]
             }

regr2 = RandomForestClassifier()

#fitting the model to each combination in the grid
model2 = GridSearchCV(regr2, param_grid)
#fining the best parameters based on the search grid
model2.fit(X2_train, y2_train)

#pulling the fitted model on the best settings so we can see the variable importances
regr2 = model2.best_estimator_

print(model2.best_score_)
print(model2.score(X2_train, y2_train))
print(model2.score(X2_test, y2_test))
```

```
C:\Users\Owner\Anaconda3\lib\site-packages\sklearn\model_selection\_split.py:605: Warning: The least populated class in y has o
nly 1 members, which is too few. The minimum number of members in any class cannot be less than n_splits=3.
  % (min_groups, self.n_splits)), Warning)
```

This was the extent of the exploration into the Chocolate Bar Rating dataset as found on Kaggle.

Conclusion

1- In terms of Cocoa Percentage 70-80%s seem to be best. 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. There are no consistent findings.

2- Indian Ocean and South Pacific Ocean tend to be the oceans around which origin coastal countries with the highest rated chocolate bars are.

3- Amedei was the most successful chocolate bar manufacturer in the dataset but that was over a decade ago. More recently, Woodstock in the U.S.A. has had the highest rated chocolate bars. In terms of investment and looking for the next best-selling bar, attention should be paid to Woodstock rather than Amedei.