Final Report: Chocolate Bar Ratings

Saleha Bakht • 08.08.2018

Overview

The dataset analyzed was compiled by Brady Brelinski, Founding Member of the Manhattan Chocolate Society. Answers to his positted questions would influence his advice and recommendations to chocolate manufacturers and retailers.

Progress - Brady Brelinski’s Questions

- Where are the best cocoa beans grown?
- Which countries produce the highest-rated bars?
- What is the relationship between cocoa solids percentage and rating?

Progress - Machine Learning

- Predicting the Ratings value for the chocolate bars using a RandomForestRegressor.

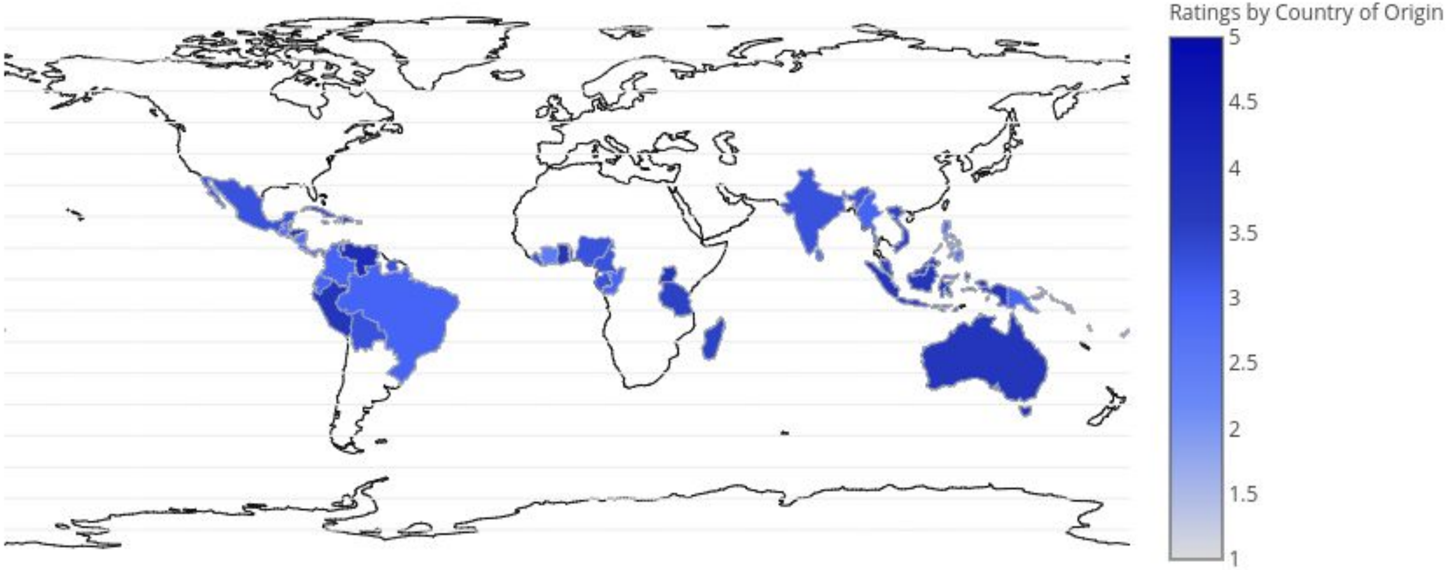
- Predicting the Broad Bean Origin values for the chocolate bars using a RandomForestClassifier.

EDA - Geographical Locations

Beans mostly originate from the Southern Hemisphere. The higher rated beans come from Australia and some western regions of South America.

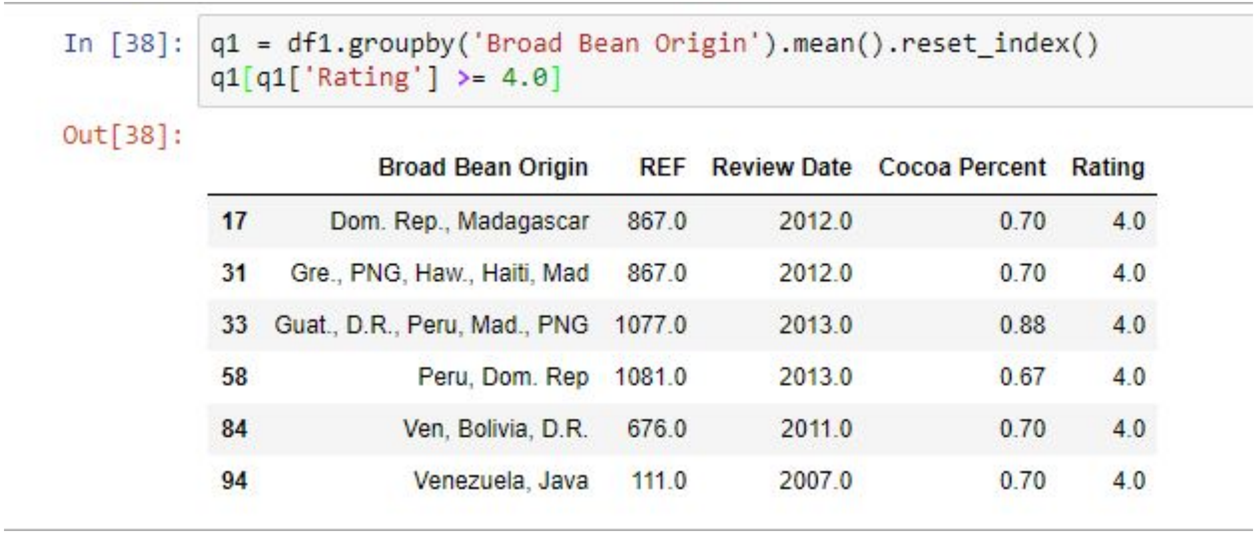
The Indian Ocean and South Pacific Ocean tend to be the oceans around which coastal countries with the highest rated chocolate beans are. Most of the cocoa bean growing countries appear to be around the same strip of the globe around the equator.

Ratings by Country of Origin



Posed Questions

Where are the best cocoa beans grown?



In [38]:

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

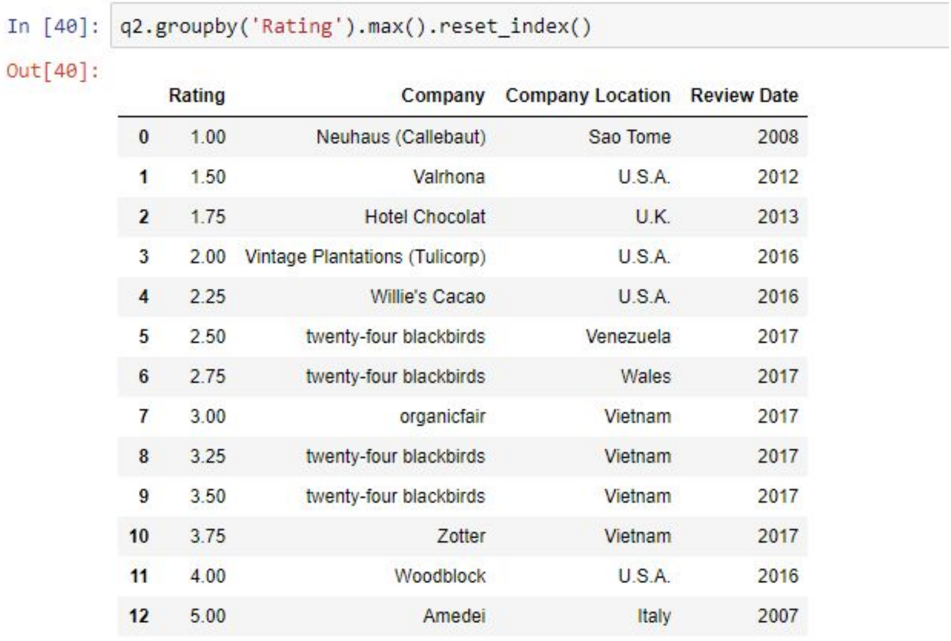
Out[38]:

	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

We took the term ‘best’ and defined it as bars having a rating of 4.00 or higher. 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 provide the highest-ranked bars?

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 more recently the higher ranking bars are made by Woodblock in the U.S.A..

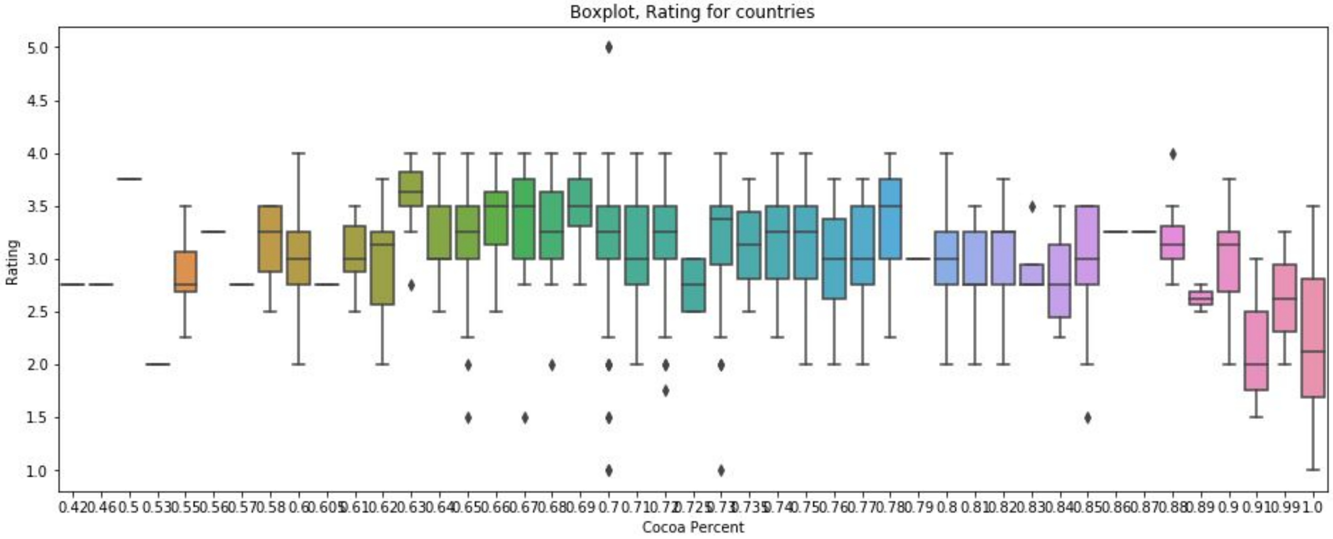


```
In [40]: q2.groupby('Rating').max().reset_index()
```

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

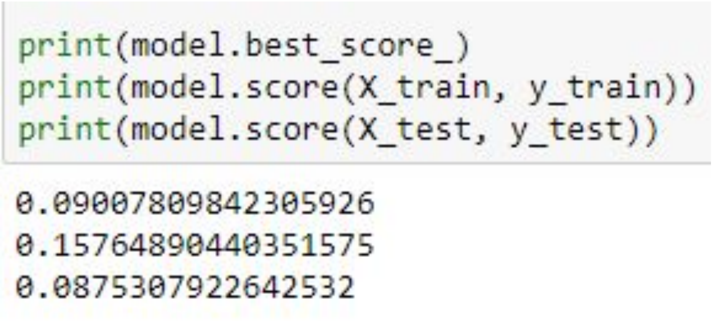
What is the relationship between cocoa solids percentage and ratings?



There is no obvious trend for which Cocoa Percent has the better rating. 0.7 Cocoa Percent had two 5.00 ratings but it also had two 1.00 ratings, also marked as outliers. There are no consistent findings.

Machine Learning

RandomForestRegressor() to predict Rating

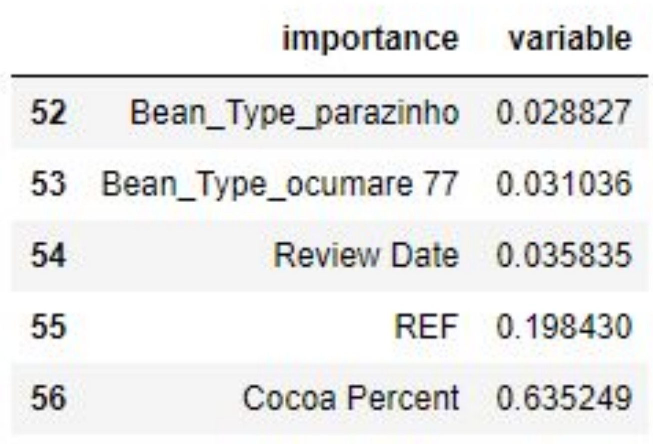


```
print(model.best_score_)
print(model.score(X_train, y_train))
print(model.score(X_test, y_test))
```

```
0.09007809842305926
0.15764890440351575
0.0875307922642532
```

One round of RandomForestRegressor() was used to attempt to predict the Ratings value of the chocolate bars. 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 was only 0.003 less than the accuracy of the best score. However, it did more than half as well as the training set performed.

Variable Importance

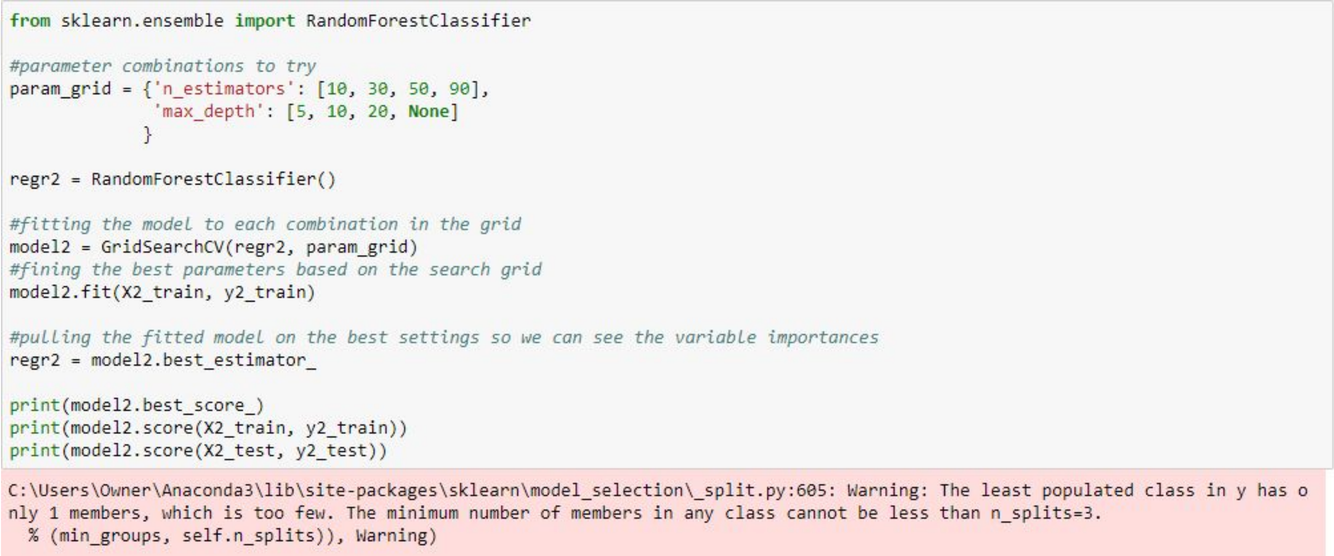


	importance	variable
52	Bean_Type_parazinho	0.028827
53	Bean_Type_ocumare 77	0.031036
54	Review Date	0.035835
55	REF	0.198430
56	Cocoa Percent	0.635249

As can be imagined, the Cocoa Percent values of the bars played the highest role in determining the rating of the chocolate bar even though, recall, there were no consistent ratings for each cocoa percentage.

RandomForestClassifier() to predict Broad Bean Orig

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)
```

Suggestions

1 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.

2 Indian Ocean and South Pacific Ocean tend to be the oceans around which coastal countries with the highest rated chocolate beans are. Good investment potential: Horn of Africa.

Thank you