

## Challenge

Create a report to summarize your research. Include:

1. What is the average rating by country of origin?
2. How many bars were reviewed for each of those countries?
3. Create plots to visualize findings for questions 1 and 2.
4. Is the cacao bean's origin an indicator of quality?
5. [Optional] How does cocoa content relate to rating? What is the average cocoa content for bars with higher ratings (above 3.5)?
6. [Optional 2] Your research indicates that some consumers want to avoid bars with lecithin. Compare the average rating of bars with and without lecithin (L in the ingredients).
7. Summarize your findings.

```
In [1]: ▶ import pandas as pd
import numpy as np
import seaborn as sns
from datetime import datetime
import dateutil.parser
```

```
In [2]:  choc = pd.read_csv('chocolate_bars.csv')
        choc.head()
```

Out[2]:

	id	manufacturer	company_location	year_reviewed	bean_origin	bar_name	cocoa_percent	num_ingredients	ingredients	review
0	2454	5150	U.S.A.	2019	Tanzania	Kokoa Kamili, batch 1	76.0	3.0	B,S,C	rich cocoa, fatty, bready
1	2458	5150	U.S.A.	2019	Dominican Republic	Zorzal, batch 1	76.0	3.0	B,S,C	cocoa, vegetal, savory
2	2454	5150	U.S.A.	2019	Madagascar	Bejofo Estate, batch 1	76.0	3.0	B,S,C	cocoa, blackberry, full body
3	2542	5150	U.S.A.	2021	Fiji	Matasawalevu, batch 1	68.0	3.0	B,S,C	chewy, off, rubbery
4	2546	5150	U.S.A.	2021	Venezuela	Sur del Lago, batch 1	72.0	3.0	B,S,C	fatty, earthy, moss, nutty,chalky



```
In [3]: #knowing the data  
print(choc.shape) #2530 rows, 10 columns  
print(choc.describe())  
print(choc.info())  
choc.isna().sum()
```

(2530, 11)

	id	year_reviewed	cocoa_percent	num_ingredients	rating
count	2530.000000	2530.000000	2530.000000	2443.000000	2530.000000
mean	1429.800791	2014.374308	71.639723	3.041343	3.196344
std	757.648556	3.968267	5.616724	0.913728	0.445321
min	5.000000	2006.000000	42.000000	1.000000	1.000000
25%	802.000000	2012.000000	70.000000	2.000000	3.000000
50%	1454.000000	2015.000000	70.000000	3.000000	3.250000
75%	2079.000000	2018.000000	74.000000	4.000000	3.500000
max	2712.000000	2021.000000	100.000000	6.000000	4.000000

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 2530 entries, 0 to 2529

Data columns (total 11 columns):

#	Column	Non-Null Count	Dtype
0	id	2530 non-null	int64
1	manufacturer	2530 non-null	object
2	company_location	2530 non-null	object
3	year_reviewed	2530 non-null	int64
4	bean_origin	2530 non-null	object
5	bar_name	2530 non-null	object
6	cocoa_percent	2530 non-null	float64
7	num_ingredients	2443 non-null	float64
8	ingredients	2443 non-null	object
9	review	2530 non-null	object
10	rating	2530 non-null	float64

dtypes: float64(3), int64(2), object(6)

memory usage: 217.5+ KB

None

```
Out[3]: id          0  
manufacturer      0  
company_location  0  
year_reviewed     0
```

```
bean_origin      0
bar_name          0
cocoa_percent     0
num_ingredients  87
ingredients       87
review            0
rating            0
dtype: int64
```

### ***Insights:***

This data set appears to have a time period between 2006 and 2021. The number of ingredients range from 1 to 6, the average rating of a chocolate bar is 3.2, and the average cocoa percent is roughly 71.5.

```
In [4]: ► #What is the average rating by country of origin?
choc['bean_origin'].unique()
choc.groupby('bean_origin').rating.mean().sort_values(ascending = False)
```

```
Out[4]: bean_origin
Tobago          3.625000
China           3.500000
Sao Tome & Principe 3.500000
Solomon Islands 3.450000
Congo           3.318182
...
Sierra Leone   2.750000
St.Vincent-Grenadines 2.750000
Principe        2.750000
Martinique      2.750000
Puerto Rico    2.714286
Name: rating, Length: 62, dtype: float64
```

### ***Insights:***

Tobago, China, Sao Tome & Principe, Solomon Islands and Congo are the countries with the highest rated bars.

```
In [5]: ► #How many bars were reviewed for each of those countries?
choc.groupby('bean_origin')['bar_name'].count().sort_values(ascending=False)
```

```
Out[5]: bean_origin
Venezuela          253
Peru               244
Dominican Republic 226
Ecuador           219
Madagascar        177
...
Suriname           1
Sumatra            1
Sulawesi           1
St.Vincent-Grenadines 1
Principe           1
Name: bar_name, Length: 62, dtype: int64
```

***Insights:***

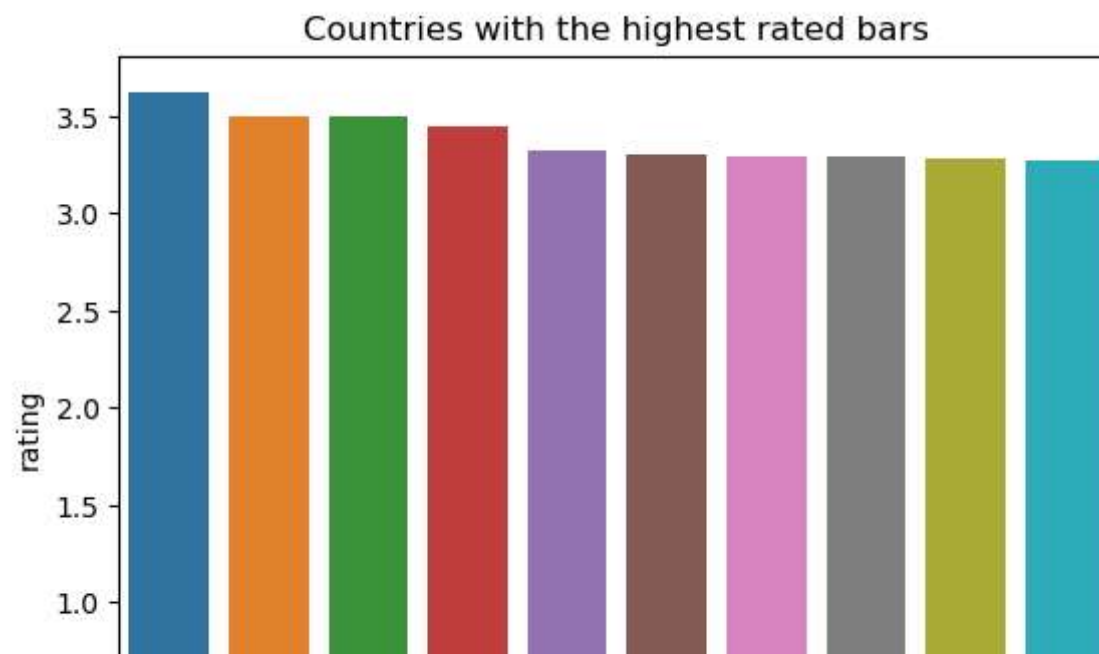
Amongst the bars reviewed, Venezuela, Peru, Dominican Republic, Ecuador and Madagascar had the highest number of bars under review.

```
In [6]: ► #question 1 viz
c = pd.DataFrame(choc.groupby('bean_origin').rating.mean().sort_values(ascending = False))
c['bean_origin'] = c.index
c
```

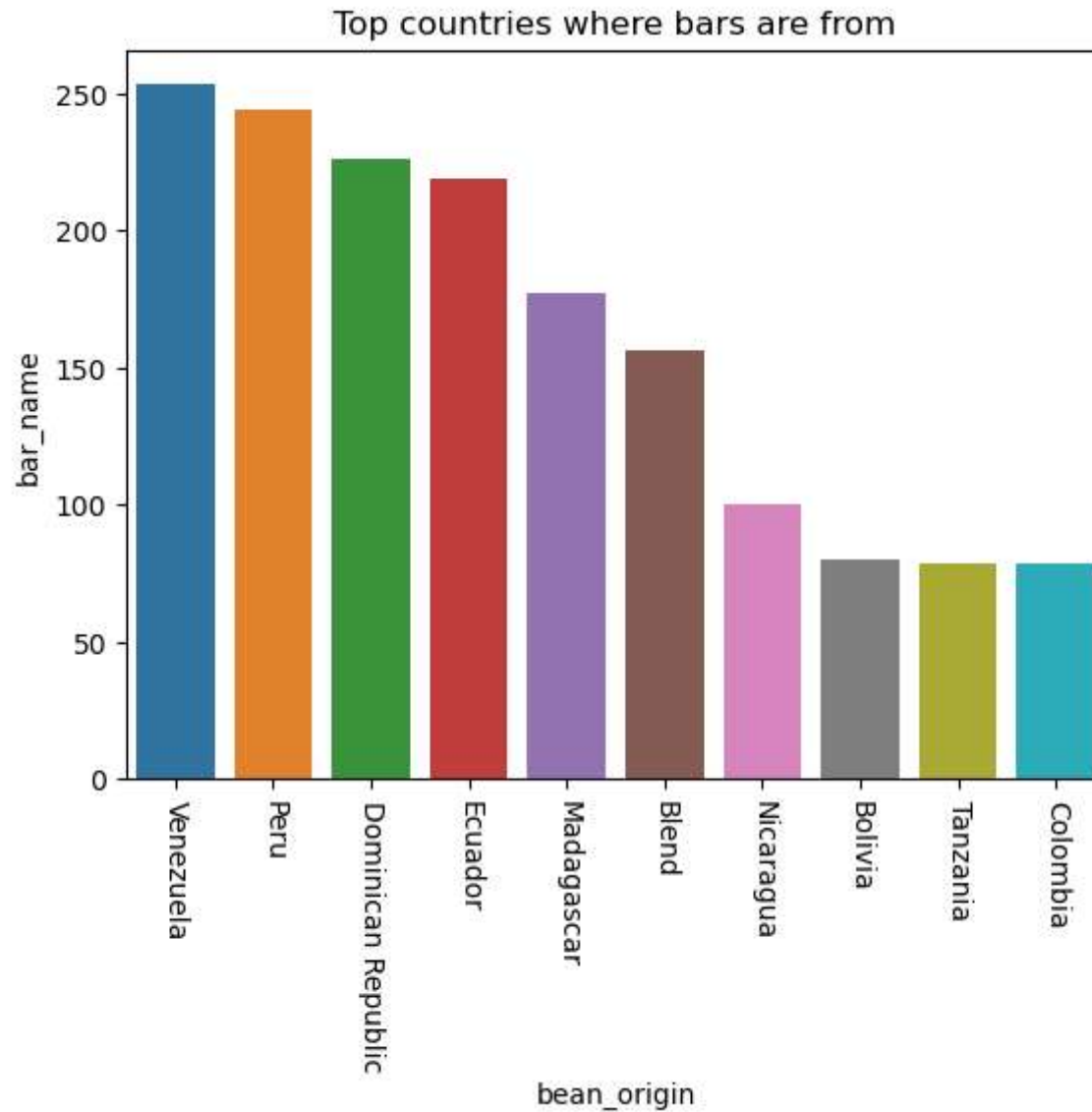
Out[6]:

	rating	bean_origin
bean_origin		
Tobago	3.625000	Tobago
China	3.500000	China
Sao Tome & Principe	3.500000	Sao Tome & Principe
Solomon Islands	3.450000	Solomon Islands
Congo	3.318182	Congo
...	...	...
Sierra Leone	2.750000	Sierra Leone
St.Vincent-Grenadines	2.750000	St.Vincent-Grenadines
Principe	2.750000	Principe
Martinique	2.750000	Martinique

```
In [7]: ► viz1 = sns.barplot(x='bean_origin', y='rating', data=c[0:10])  
viz1.set_title('Countries with the highest rated bars')  
viz1.set_xticklabels(viz1.get_xticklabels(), rotation = -90);
```



```
In [8]: ► #question 2 viz
d = pd.DataFrame(choc.groupby('bean_origin')['bar_name'].count().sort_values(ascending=False))
d['bean_origin'] = d.index
viz2 = sns.barplot(x='bean_origin', y='bar_name', data=d[0:10])
viz2.set_title('Top countries where bars are from')
viz2.set_xticklabels(viz2.get_xticklabels(), rotation = -90);
```

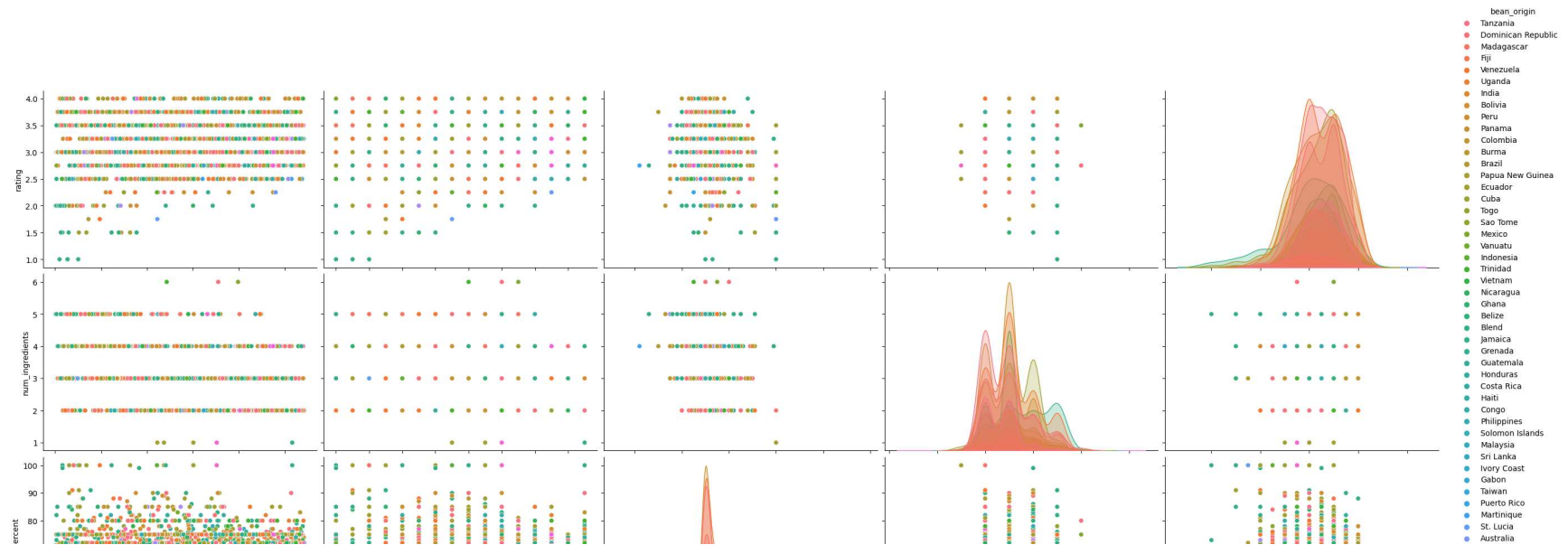






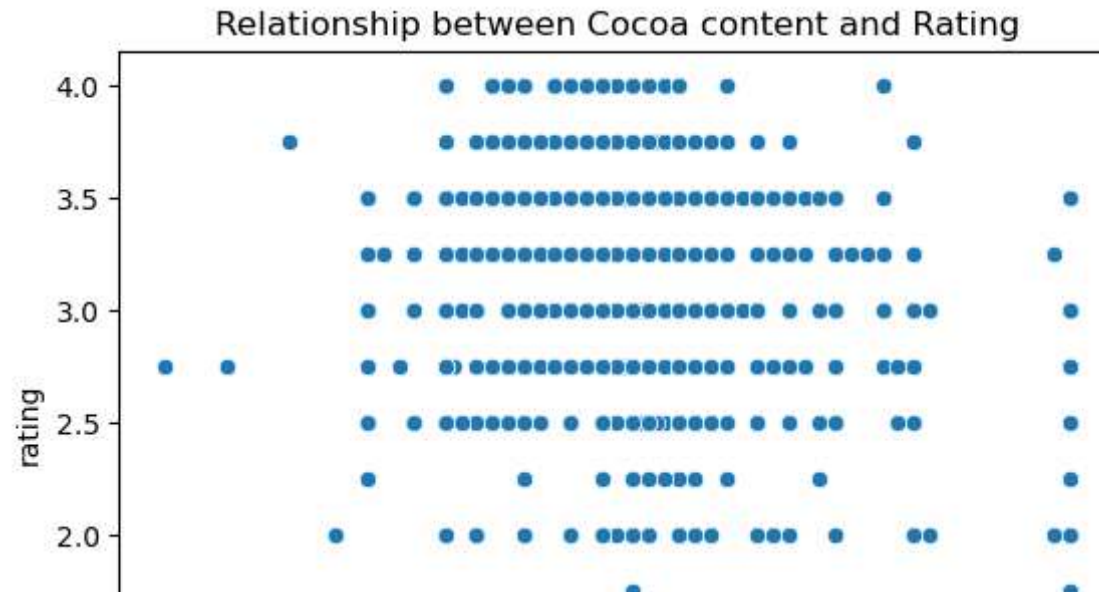
```
In [25]: ▶ #Is the cacao bean's origin an indicator of quality?
sns.pairplot(choc,hue = 'bean_origin', y_vars=['rating', 'num_ingredients', 'cocoa_percent'], height=3.5, aspect=1.5)
```

```
Out[25]: <seaborn.axisgrid.PairGrid at 0x23fd0a0dee0>
```



```
In [11]: ► #How does cocoa content relate to rating? What is the average cocoa content for bars with higher ratings (above 3.5)?
viz3 = sns.scatterplot(x='cocoa_percent', y='rating', data=choc)
viz3.set_title('Relationship between Cocoa content and Rating')
```

Out[11]: Text(0.5, 1.0, 'Relationship between Cocoa content and Rating')



```
In [23]: ► #correlation
correlation_XY = choc['cocoa_percent'].corr(choc['rating'])
print(f'The correlation between cocoa_percent and rating is {round(correlation_XY,2)}')
```

The correlation between cocoa\_percent and rating is -0.15

```
In [18]: ► #average cocoa content for bars with higher ratings >3.5
ave = (choc[choc['rating'] > 3.5].cocoa_percent.mean())
print(f'The average cocoa content for bars with higher ratings is {round(ave,2)}%')
#subset = (choc['ratings'] > 3.5)
```

The average cocoa content for bars with higher ratings is 70.95%

**Insights:**

This data does not show any clear relationship between percentage of cocoa in a bar and the rating of a bar, other factors might have an influence on rating.

```
In [20]: ▶ #[Optional 2] Your research indicates that some consumers want to avoid bars with Lecithin. Compare the average rating of bars with and without Lecithin.
#bar_lecithin = ['L']
chocc = choc.dropna(axis=0)
ave_lecithin = chocc[chocc['ingredients'].str.contains('L')].rating.mean()
print(f'The average rating of bars with lecithin is {round(ave_lecithin,2)} points')
ave_no_lecithin = chocc[~chocc['ingredients'].str.contains('L')].rating.mean()
print(f'The average rating of bars without lecithin is {round(ave_no_lecithin,2)} points')
```

```
The average rating of bars with lecithin is 3.15 points
The average rating of bars without lecithin is 3.23 points
```

## Summary

- This data set appears to have a time period between 2006 and 2021.
- The number of ingredients range from 1 to 6, the average rating of a chocolate bar is 3.2, and the average cocoa percent is roughly 71.5.
- Amongst the 2530 bars reviewed, Venezuela, Peru, Dominican Republic, Ecuador and Madagascar had the highest number of bars under review.
- This data does not show any clear relationship between percentage of cocoa in a bar and the rating of a bar, other factors might have an influence on rating.
- Bars without Lecithin on average are rated higher than bars with Lecithin, this could be related to the widely held belief that Lecithin is often used as a 'mask' for poor quality Chocolate bars, although there is a negative correlation between the percent of cocoa in a bar and what that bar is rated, so ratings might be based on factors other than cocoa content/ ingredient content.
- The pairplot does not show any obvious relationship between the origin and quality(which was determined by rating, num\_of\_ingredients and cocoa percent) and so it is concluded that there is no obvious relationship between Bean Origin and Quality.