PREDICTORS OF CHOCOLATE RATINGS

Caroline Hussey

Codecademy Data Science Project

June 2021

PROJECT OBJECTIVES

- Aquire data from source and format to pandas dataframe
- Analyse how ratings have changed over time
- Assess the impact cocoa percentage may have on product rating
- For each feature, identify the top 10 chocolate products
- Find out which features are associated with the product rating
- Test different machine learning models to see which can most accurately predict new product rating

METHODS AND LANGUAGES

- Jupyter Notebook
- Python
- Beautiful Soup
- Pandas
- NumPy
- Data Visualisations
- Hypothesis Testing – Chi-Squared
- Machine Learning – Logistic Regression, K-nearest Neighbours

DATA ACQUISITION

Cacao Ratings

Compiled ratings of over 1700 Chocolate bars
Ratings are from 1-5

Company (Maker-if known)	Specific Bean Origin or Bar Name	REF	Review Date	Cocoa Percent	Company Location	Rating	Bean Type	Broad Bean Origin
A. Morin	Agua Grande	1876	2016	63%	France	3.75		Sao Tome
A. Morin	Kpime	1676	2015	70%	France	2.75		Togo
A. Morin	Atsane	1676	2015	70%	France	3		Togo
A. Morin	Akata	1680	2015	70%	France	3.5		Togo
A. Morin	Quilla	1704	2015	70%	France	3.5		Peru
A. Morin	Carenero	1315	2014	70%	France	2.75	Criollo	Venezuela
A. Morin	Cuba	1315	2014	70%	France	3.5		Cuba
A. Morin	Sur del Lago	1315	2014	70%	France	3.5	Criollo	Venezuela
A. Morin	Puerto Cabello	1319	2014	70%	France	3.75	Criollo	Venezuela

Extraction Method

- Beautiful soup to parse html
- The target content is labelled with class names
- .select(class): this method takes the class name as parameter and returns all html elements (including code) with that class name.
- .get_text(): extracts text from the html element
- Here we will loop through all elements with that class name, use the get_text() method on each iteration, and append the results to an array.
- Repeat for each feature in the html table.
- Combine all arrays to pandas dataframe.

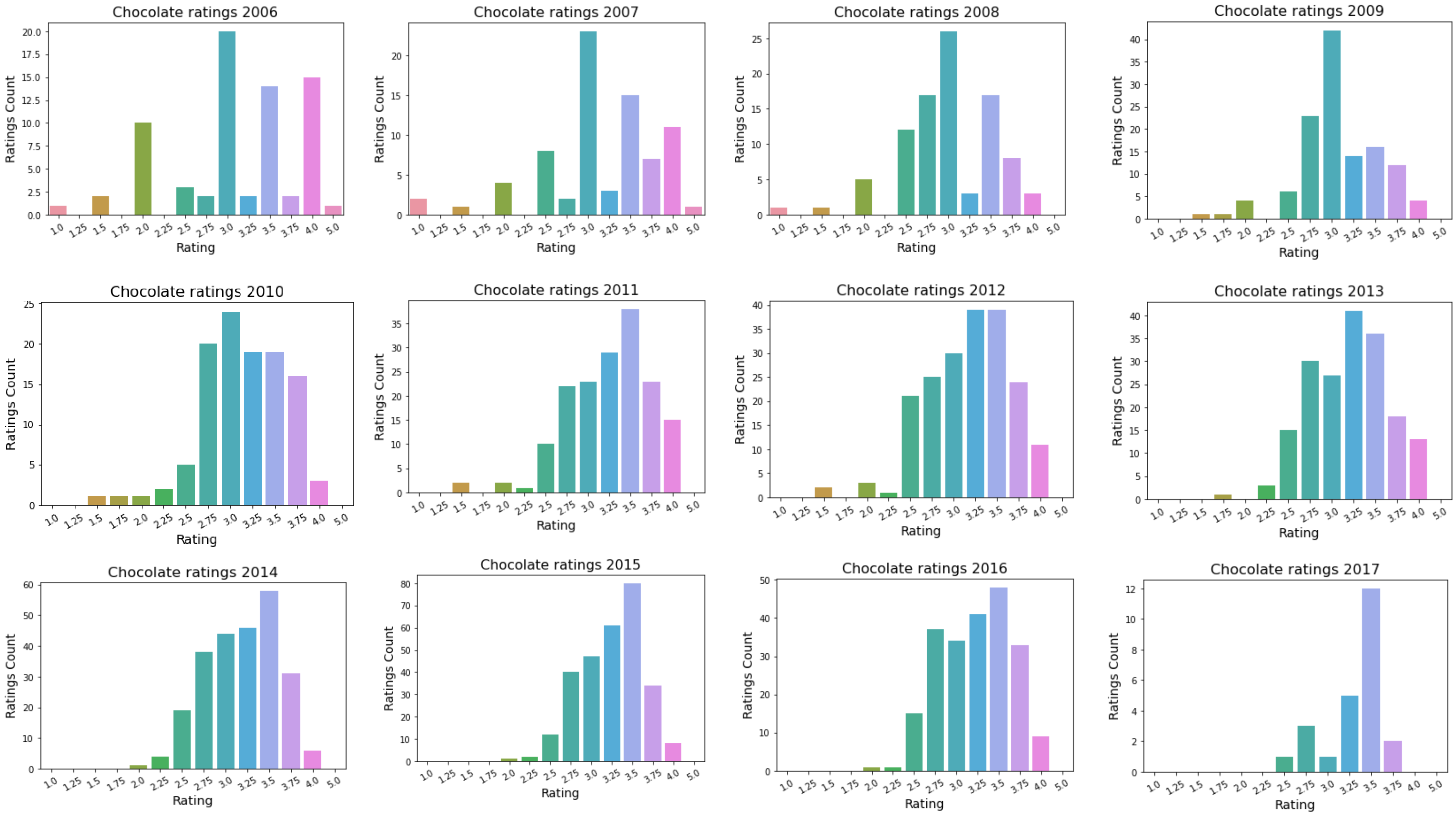
HTML Structure

<tr>
<td class="Company">**A. Morin**</td>
<td class="Origin">**Agua Grande**</td>
<td class="REF">**1876**</td>
<td class="ReviewDate">**2016**</td>
<td class="CocoaPercent">**63%**</td>
<td class="CompanyLocation">**France**</td>
<td class="Rating">**3.75**</td>
<td class="BeanType"> </td>
<td class="BroadBeanOrigin">**Sao Tome**</td>
</tr>

EXPLORATORY DATA ANALYSIS

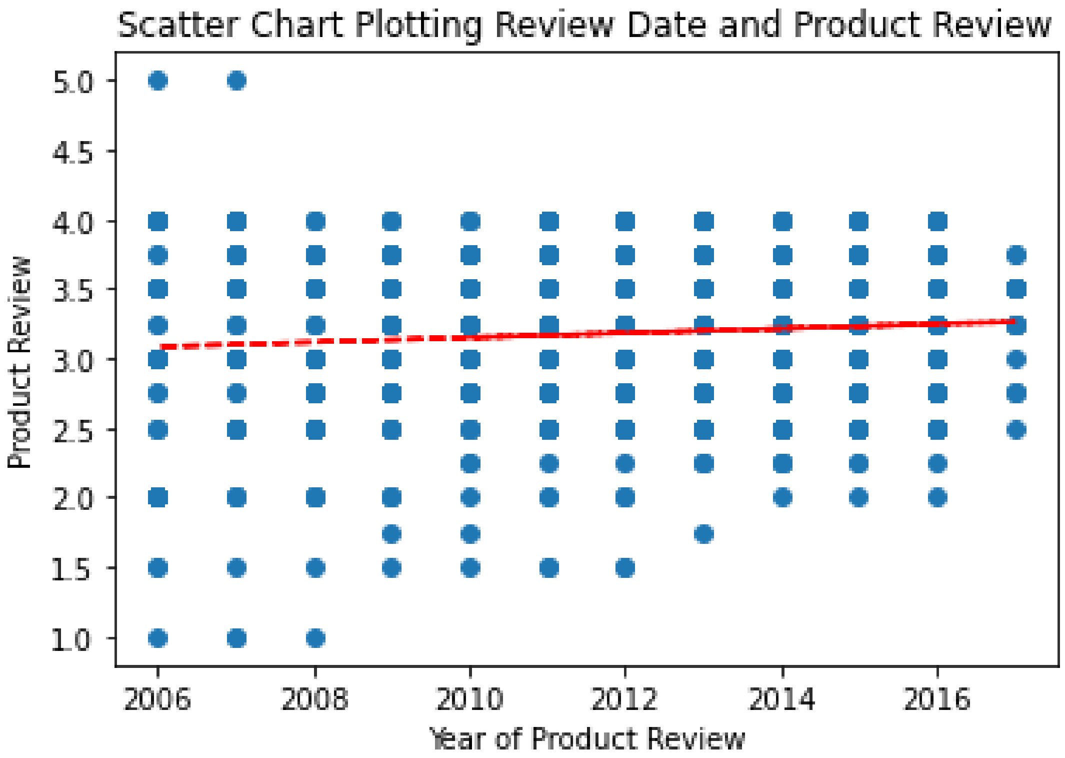
- Pandas .info() and .describe() to view summary statistics, data types and size of dataframe
- Split dataframe into sections based on ratings boundaries for clearer visualisations (eg. Rated under 1.5 to plot lower rated products, rated above 3.5 to plot higher rated products)
- Create a separate dataframe for each year to visualise ratings per year
- Split into a separate dataframe for cocoa percentage boundaries to visualise ratings based on cocoa percentage
- Remove null values for more accurate assessment of analyses focusing on Bean Types and Broad Bean Origin (the two series containing null values)

CHOCOLATE RATINGS BY YEAR

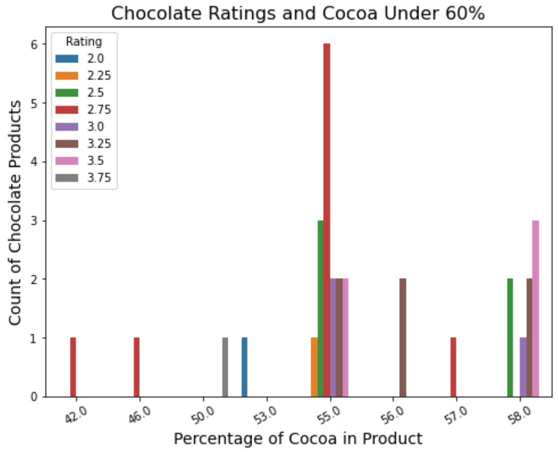


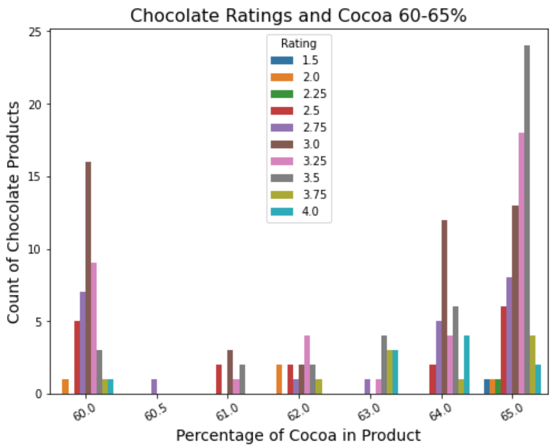
CHOCOLATE RATINGS BY YEAR

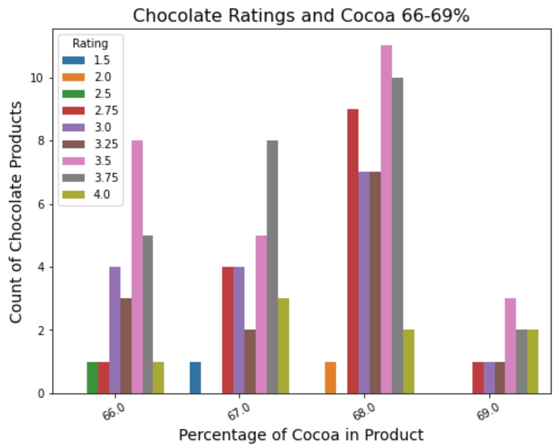
- Annual ratings show a greater range, but lower number of ratings before 2010.
- 1.0 and 5.0 ratings are only given in 2006, 2007 and 2008.
- Higher ratings (3.5 +) were more common from 2010.
- Chocolate products increase steadily from 2010.
- Scatter plot shows a linear correlation between the product review and the year it was given.
- The sample size from 2017 appears smaller in 2017 with a much lower number of products rated (possibly the data was collected part way through the year)

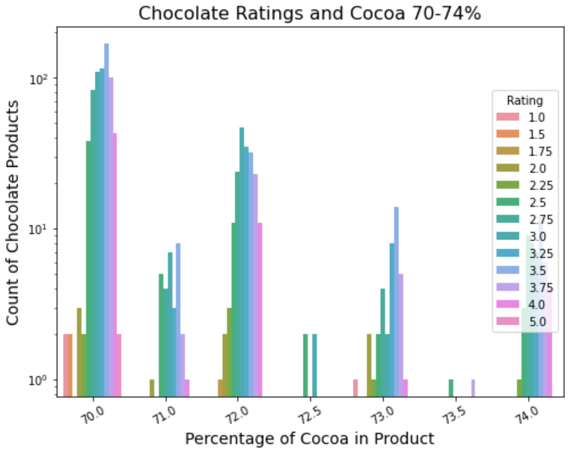


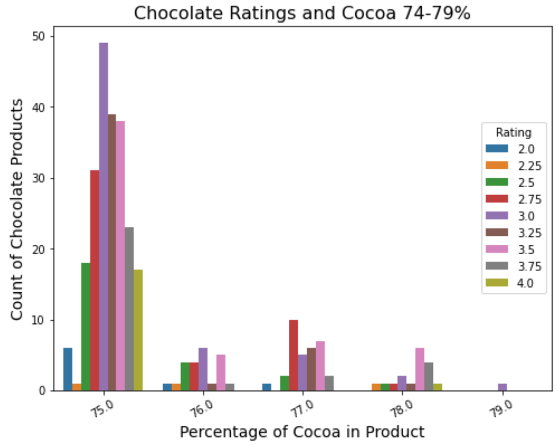
CHOCOLATE RATINGS BY COCOA PERCENTAGE

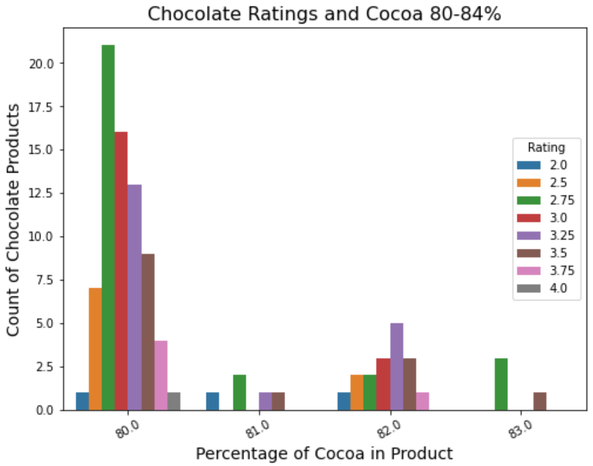


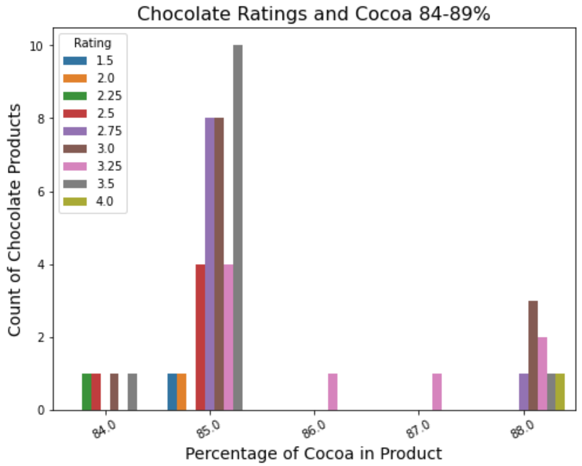


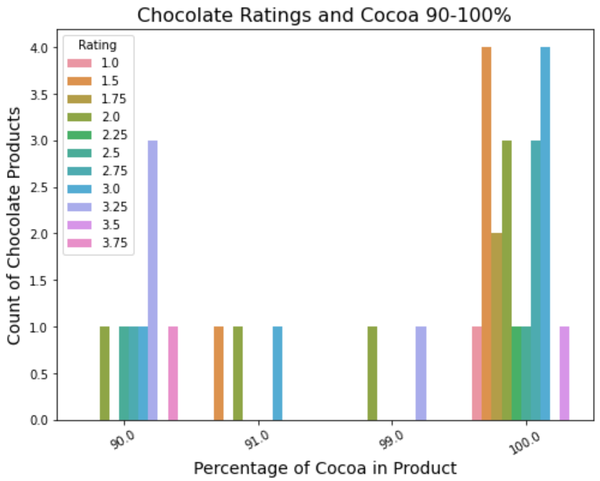






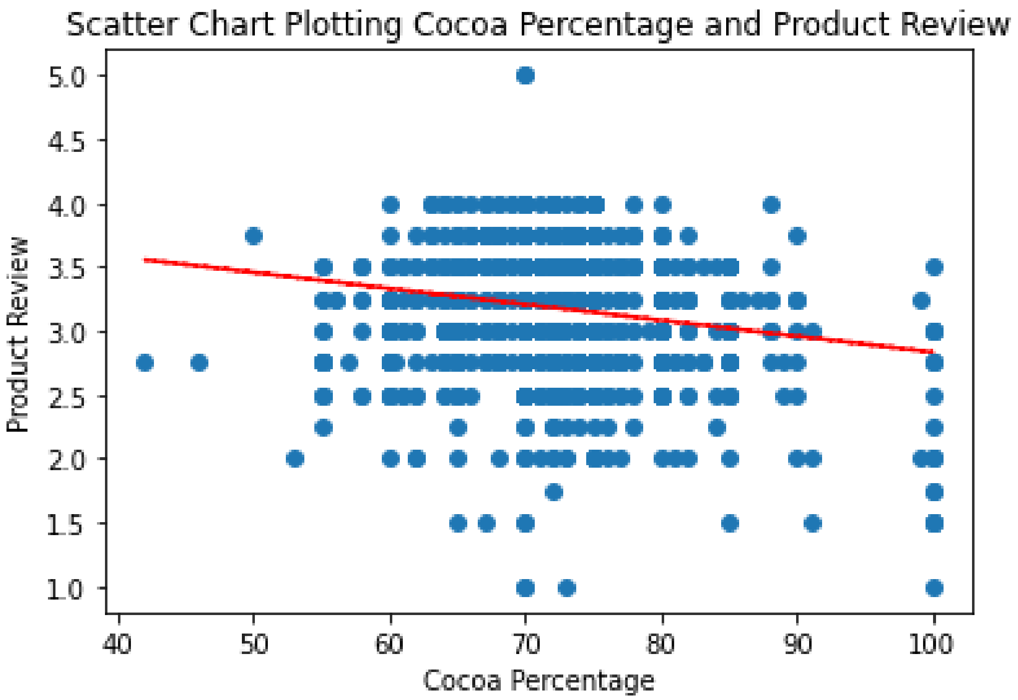






CHOCOLATE RATINGS BY COCOA PERCENTAGE

- Barplots show a greater number of reviews at specific ranges of Cocoa Percentage
- Most chocolate products contain between 70-75% of cocoa (this graph is scaled using a log scale for easier viewing)
- It is possible the higher number of reviews in these ranges is due to the higher number of products produced with that percentage of cocoa
- Scatter plot shows a linear correlation between the product review and the percentage of cocoa.



TOP 10 AVERAGE RATINGS FOR BEAN ORIGIN GROUPED BY YEAR

- Bean origin reports the local origin of the cocoa product. This can refer to a region in a country, a farm, or a co-operative.
- The highest average rated chocolate product based on bean origin is Chuao, with an average rating of 5.0. in 2007.
- A closer look at products whose origin is Chuao show that beans originating from here were marketed throughout the timescale covered in this dataset. These products received good ratings (3.5+) in 2006, 2007, 2011 and 2015, but ratings on other years showed average or less than average ratings.

Origin	Review Date	Review
Chuao	2007	5
Toscano Black	2006	4.5
ABOCFA Coop	2015	4
Alto Beni, Cru Savage	2006	4
Asante	2009	4
Bali, Sukrama Bros. Farm, Melaya,	2011	4
Bellavista Coop, #225, LR, MC, CG	2013	4
Cabosse	2007	4
Carenero Superior, Urrutia, Barlo	2011	4
Chuao	2006	4

TOP 10 AVERAGE RATINGS FOR BROAD BEAN ORIGIN

- Broad Bean Origin reports the wider origins of the cocoa product.
- Usually refers to the country of origin
- Top Rated chocolate products from broad bean origins average review is 4.0.
- Top rated products based on broad bean origin include mixed origin products
- Dominican Republic/Madagascar, Gre./PNG/Hawaii/Haiti/Madagascar, Tobago, Venezuela/ Bolivia/Dominicn Republic, and Venezuela/ Java are the highest rated broad bean origins in the products rated.

Broad Bean Origin	Review
Dom. Rep., Madagascar	4
Gre., PNG, Haw., Haiti, Mad	4
Tobago	4
Ven, Bolivia, D.R.	4
Venezuela, Java	4
DR, Ecuador, Peru	3.75
Dominican Rep., Bali	3.75
PNG, Vanuatu, Mad	3.75
Peru, Belize	3.75
South America	3.75

TOP 10 AVERAGE RATINGS FOR BEAN TYPE GROUPED BY YEAR

- A significant number of null values in this dataset, so null values were removed prior to analysis
- Bean Types analysed both overall and grouped by year to compare the ratings of bean types in these categories.
- Criollo – both wild and Ocumare 67 - Bean Type shows the highest average rating amongst bean types both overall and in 2006 and 2007 with an average rating of 4.0.
- A closer look shows that overall products with Criollo beans have an average overall rating (3.2), and when grouped by year the average rating each year is also about average.

Bean Type (Overall)	Review
Criollo (Ocumare 67)	4
Criollo (Wild)	4
Trinitario (85% Criollo)	3.875
Amazon mix	3.75
Blend-Forastero,Criollo	3.75
Criollo (Ocumare 77)	3.75
Forastero (Amelonado)	3.75
Trinitario, Nacional	3.75
Trinitario, TCGA	3.75
Amazon, ICS	3.625

Bean Type	Review Year	Rating
Criollo (Ocumare 67)	2007	4
Criollo (Wild)	2006	4
Criollo, Trinitario	2006	4
	2007	4
Beniano	2016	3.875
Trinitario (85% Criollo)	2007	3.875
Amazon mix	2016	3.75
Amazon, ICS	2016	3.75
Blend	2010	3.75
Blend-Forastero,Criollo	2008	3.75

TOP 10 AVERAGE RATINGS FOR COMPANY GROUPED BY YEAR

- Tobago Estate (Pralus) shows the highest average of ratings amongst companies with an average rating of 4.0.
- A closer look shows only one product review for Tobago Estate (Pralus) – in 2012.
- Heirloom Cacao Preservation (Zokoko) and Ocelot both have only two product entries (Heirloom in 2016 and Ocelot in 2015). Both companies received a rating of 3.75 and a second of 4.0, giving an average rating of 3.875 for both companies.

Company	Review
Tobago Estate (Pralus)	4
Heirloom Cacao Preservation (Zokoko)	3.875
Ocelot	3.875
Amedei	3.846154
Matale	3.8125
Patric	3.791667
Idilio (Felchlin)	3.775
Acalli	3.75
Chocola'te	3.75
Christopher Morel (Felchlin)	3.75

TOP 10 AVERAGE RATINGS FOR COMPANY LOCATION GROUPED BY YEAR

- Companies based in Equador show the highest rating of cocoa products in 2016.
- Belgium, Australia and Scotland show high ratings in 2011, 2013 and 2015 respectively.
- The overall average rating for companies whose location is Ecuador is 3.009.
- Ratings for companies whose location is Ecuador show a steady improvement in ratings over time, with lower ratings received prior to 2010 and average – high ratings received since.

Company Location	Review Date	Review
Ecuador	2016	4
Belgium	2011	3.875
Australia	2013	3.8125
Scotland	2015	3.8125
Bolivia	2011	3.75
Canada	2010	3.75
Chile	2015	3.75
Colombia	2015	3.75
Iceland	2016	3.75
Italy	2006	3.75

Ecuador Mean Ratings over time	
2007	3.416667
2008	2.822917
2009	2.725
2010	3
2011	3.375
2012	3.375
2014	3.357143
2015	3.416667
2016	4

CHI SQUARED – STATISTICAL SIGNIFICANCE OF FEATURES OF CHOCOLATE PRODUCTION

- The rating of each product is compared against it's feature to test the hypothesis that that feature has a significant association with the product rating.
- Calculation is carried out using SciPy's chi2 _contingency (from SciPy's stats module)
- Significance threshold: 0.05. Pval above that value is not significantly different from the other. Anything under that is significantly different.
- N/A values were removed from Bean Type and Broad Bean Origin prior to testing.
- Features that do not show a statistically significant association with the rating are highlighted in green.
- The chi-squared contingency test indicated there is no association between the origin of a cocoa bean and their rating.
- Features that show a statistically significant association with the rating are highlighted in red.
- This indicates there is an association between the company, company location, bean type, cocoa percentage and the year the product was reviewed, and their rating.

Feature	Pval
Origin	0.8878288022134178
Broad Bean Origin	0.9999999999999815
Bean Type	0.004829483755625918
Company	1.5671040278099777e-25
Company Location	0.00010569966557185
Review Year	2.180736007035077e-25
Cocoa Percentage	8.550113133375805e-26

MACHINE LEARNING – PREDICT RATINGS OF CHOCOLATE PRODUCTS

Linear Regression

- A simple model for continuous datatypes
- Model is trained using cocoa percentage and date of review features.
- The label is product rating.
- Data is scaled using sklearn's standard scaler and scored with .score()
- Linear Regression shows a very low accuracy score, suggesting this model is not the best for this dataset.

K-Nearest Neighbours

- A new column is added to the dataframe to label 'Ratings' as either 1 for good, or 0 for bad.
- n/a values are removed from the dataframe prior to testing
- Features selected are those that showed a positive association with chocolate rating: Company, company location, bean type, cocoa percentage and year of review.
- Label encoder and onehotendoder are utilised to convert each feature to numeric and binary format so that it can be processed by machine learning models.
- K-Nearest Neighbours shows the highest accuracy score (82%).

MACHINE LEARNING – PREDICT RATINGS OF CHOCOLATE PRODUCTS

Logistic Regression

- The same method to select and transform features for K-Nearest Neighbours is applied to the logistic regression model.
- The model is trained and tested using sklearn's built in LogisticRegression() method.
- Logistic Regression shows a good accuracy score (80%).
- 'New' data is created to test-predict logistic regression model
- Sklearn's .predict() method is used to predict the rating of new test products
- Sklearn's .predict_proba() method is used to show the probability that the new test products will be rated either good or bad. The output is two dimensional array showing the pair probability for eachthe possible ratings for each product, with the first probability being for 1, or good, and the second for 0, or bad.

Test Product 1 : [[9.99794560e-01 2.05439761e-04]; result = bad

Test Product 2 : [3.70502592e-04 9.99629497e-01]; result = good

Test Product 3 :[0.00000000e+00 1.00000000e+00]]; result = good

- The model shows that there is no doubt that the third test product is good and little doubt that the first is bad and the second is good.

PROJECT FINDINGS

- *Aquire data from source and format to pandas dataframe*

Data was scraped from a website using the python beautiful soup library. Data was converted into a pandas dataframe and split for various analyses.

- *Find out which features are associated with the product rating*

Chi squared contingency testing shows that there is an association between Bean Type, Company, Company Location and Year of Review, and the rating of the product.

- *For each feature, identify the top 10 chocolate products*

The top ten chocolate products for each feature are detailed on individual slides.

PROJECT FINDINGS

- *Analyse how ratings have changed over time*

Annual ratings show a greater range, but lower number of ratings before 2010.

Lowest rating (1.0) and highest rating (5.0) are only given in 2006, 2007 and 2008.

From 2010 the number of Chocolate products increase steadily. Higher ratings (3.5 +) were also more common from this year.

Chi squared contingency testing shows an association between year of rating and rating received.

- *Assess the impact cocoa percentage may have on product rating*

Most chocolate products contain between 70-75% of cocoa (this graph is scaled using a log scale for easier viewing)

Chi squared contingency testing shows an association between cocoa percentage and rating received.

- *Test different machine learning models to see which can most accurately predict new product rating*

Linear Regression is the least reliable at predicting chocolate product rating.

Logistic regression and K-Nearest neighbours showed similar accuracy scoring, with K-Nearest Neighbours being slightly more accurate at 82%.