Artisan Dark Chocolate Bar Ratings Predictor

Team Candy
Bruce Jilek, Tahereh Javaheripour, Yan Kin Pang, Travis Loseke

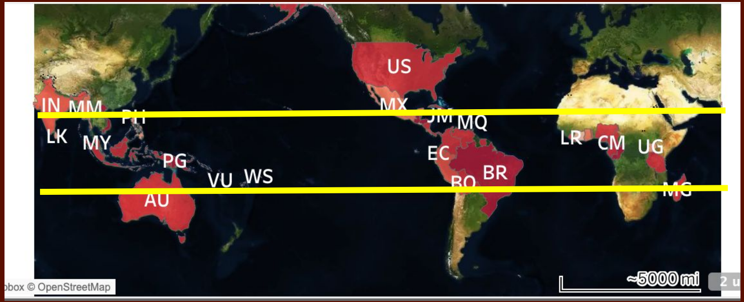
Topic Selection

Predicting Artisan chocolate bar ratings from variables such as:

- **Chocolatier**
- **Bean Type**
- **Bean Source**
- **Cocoa Percent**

A.MORIN

SoMA chocolatemaker



Reason for Topic Selection

- Chocolate was universally loved by the team
- Dataset was expansive enough to cover final project requirement
- Team became genuinely interested in if cocoa bean types and their source drive chocolate quality
- Intriguing documentary on the dark side of the cocoa bean market

Data sources

- Kaggle- Chocolate Bar ratings
- Kaggle - Countries and States Lat Long
- Flavors of Cacao website
- UN FAO - Database

Questions to answer

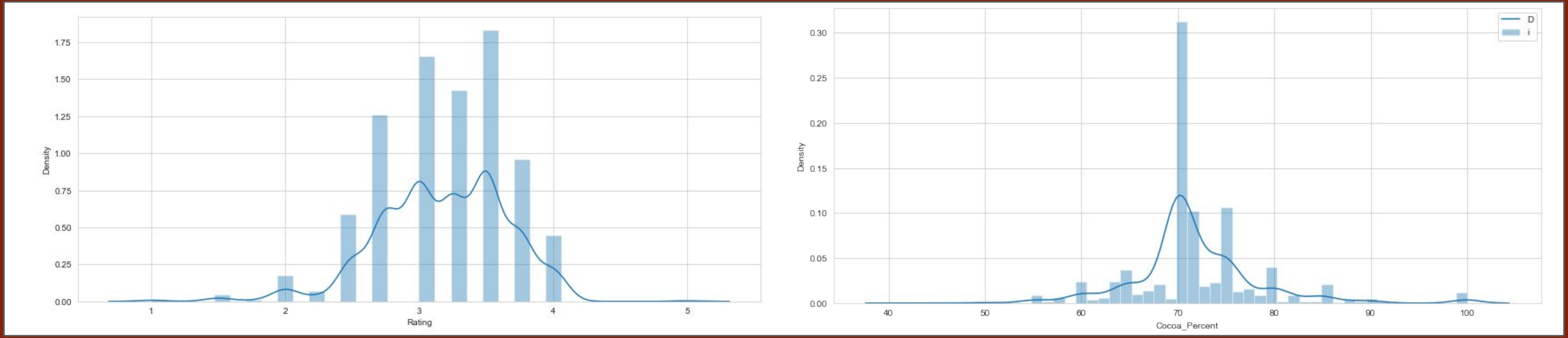
- Can we predict which chocolate bars will be rated in the top 15% (i.e. Rating >= 3.75, one SD above the Mean), based on:
  - Review date
  - Cocoa Percent
  - Bean Type: Criollo, Trinitario, or Forastero
  - Broad Bean Origin: Cocoa beans source country

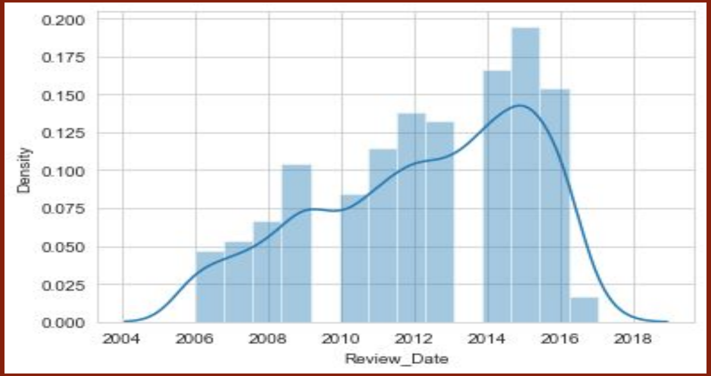
- Where are the best cocoa beans grown?
- Which countries produce the highest-rated bars?
- What's the relationship between cocoa solids percentage and rating?

Data exploration phase

- Pandas, Matplotlib, and Seaborn libraries were used to review, evaluate and clean datasets
    - Examine null counts
    - Data continuity
    - Data types
- Tableau was used to make initial map plots to look at geographic distribution issues related to typographical errors and level of detail

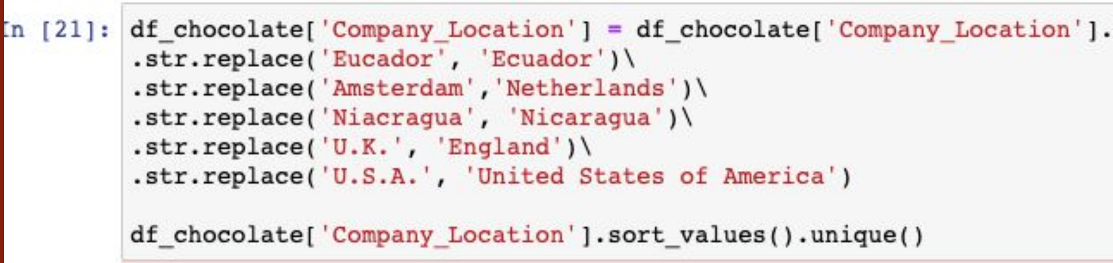
EDA distributions



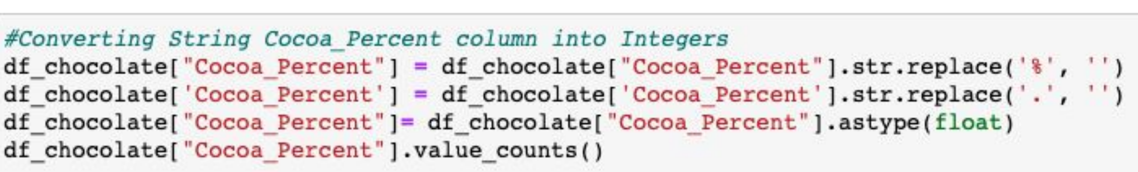


Data Preprocessing and Cleaning

- Corrected Typos (Countries)
- Formatted numerical values
- Add Continent/Region and joined Ingredients, and Most Memorable columns
- Handled outliers in Review Data and Cocoa Percent
- Dropped or replace null values



```
In [21]: df_chocolate['Company_Location'] = df_chocolate['Company_Location'].
         .str.replace('Eucador', 'Ecuador')\
         .str.replace('Amsterdam','Netherlands')\
         .str.replace('Niacragua', 'Nicaragua')\
         .str.replace('U.K.', 'England')\
         .str.replace('U.S.A.', 'United States of America')

         df_chocolate['Company_Location'].sort_values().unique()
```

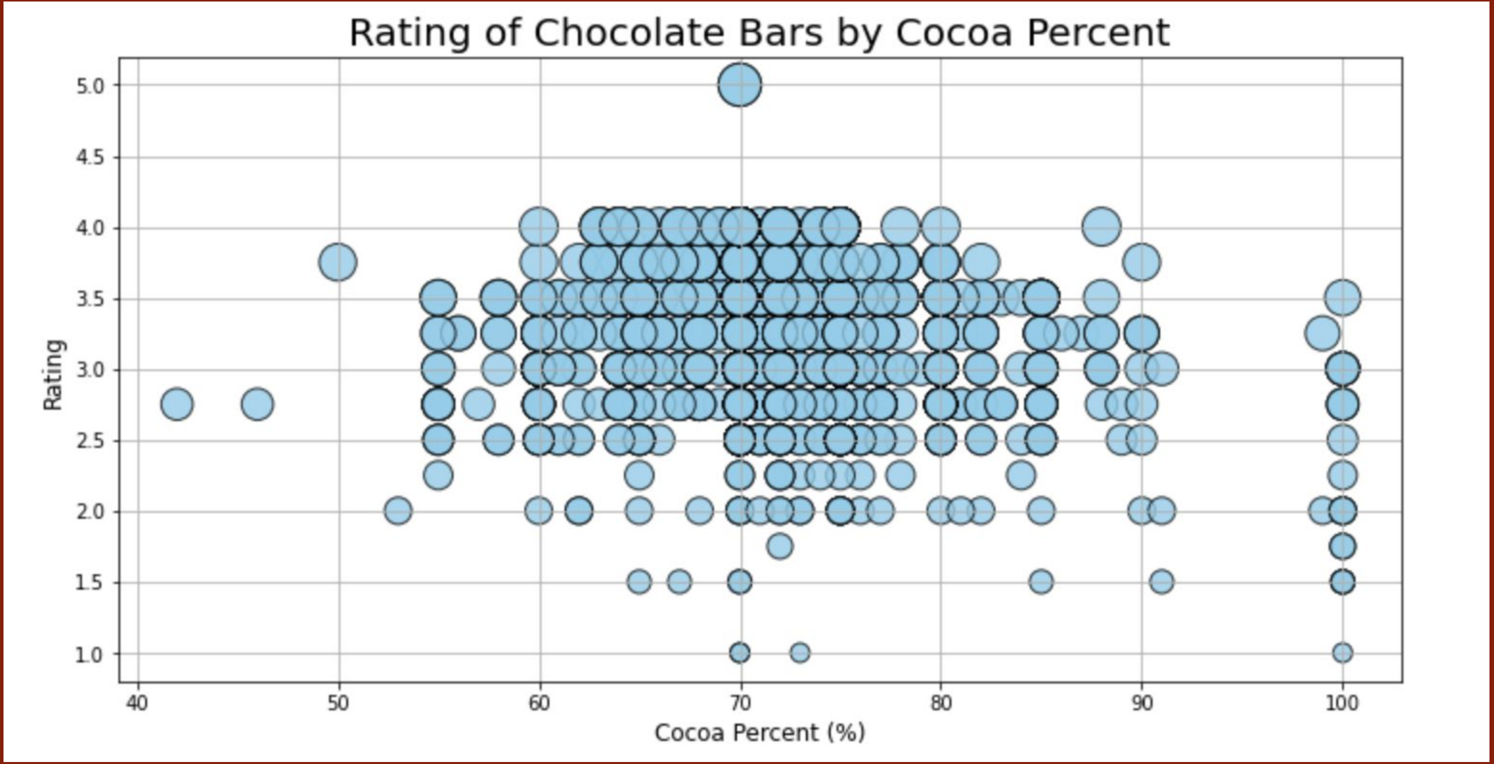


```
#Converting String Cocoa_Percent column into Integers
df_chocolate["Cocoa_Percent"] = df_chocolate["Cocoa_Percent"].str.replace('%', '')
df_chocolate['Cocoa_Percent'] = df_chocolate['Cocoa_Percent'].str.replace('.', '')
df_chocolate["Cocoa_Percent"]= df_chocolate["Cocoa_Percent"].astype(float)
df_chocolate["Cocoa_Percent"].value_counts()
```

Data Analysis phase

- Pandas and Matplot lib were used to review, evaluate and clean datasets
  - Initial plots of rating trends vs other features
  - Examine plots to look for skewness in the features and dataset
  - Quicklook plots for any trends or significant outlier anomalies
- Tableau was used to make initial map plots to look at geographic distribution of cocoa bean sources and data bias in geographic space

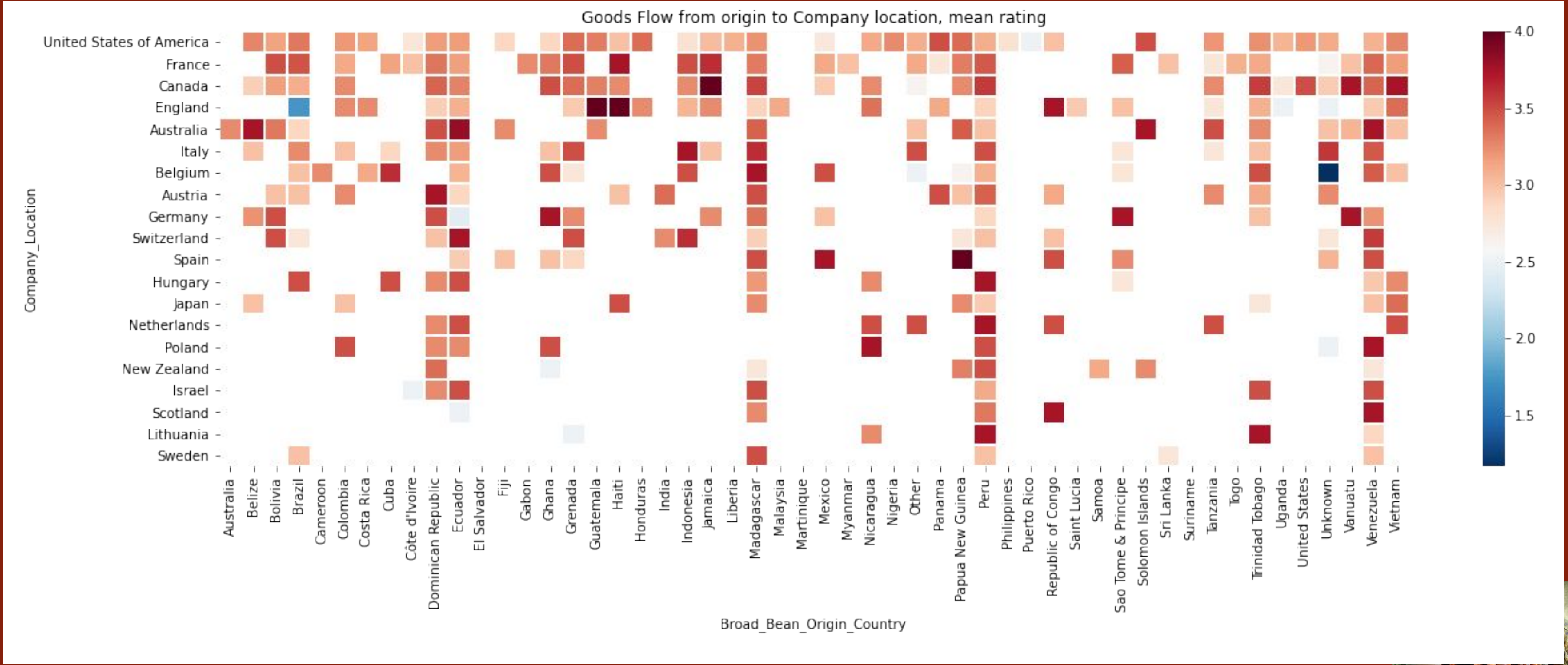
Analysis Example:Does Cocoa % lead to better bar rating



- Rating peak near ~70% cocoa and fall off at higher and lower values

Chocolates global trade

Chocolate Artisan Countries and their bean sources



Machine Learning Model

1.

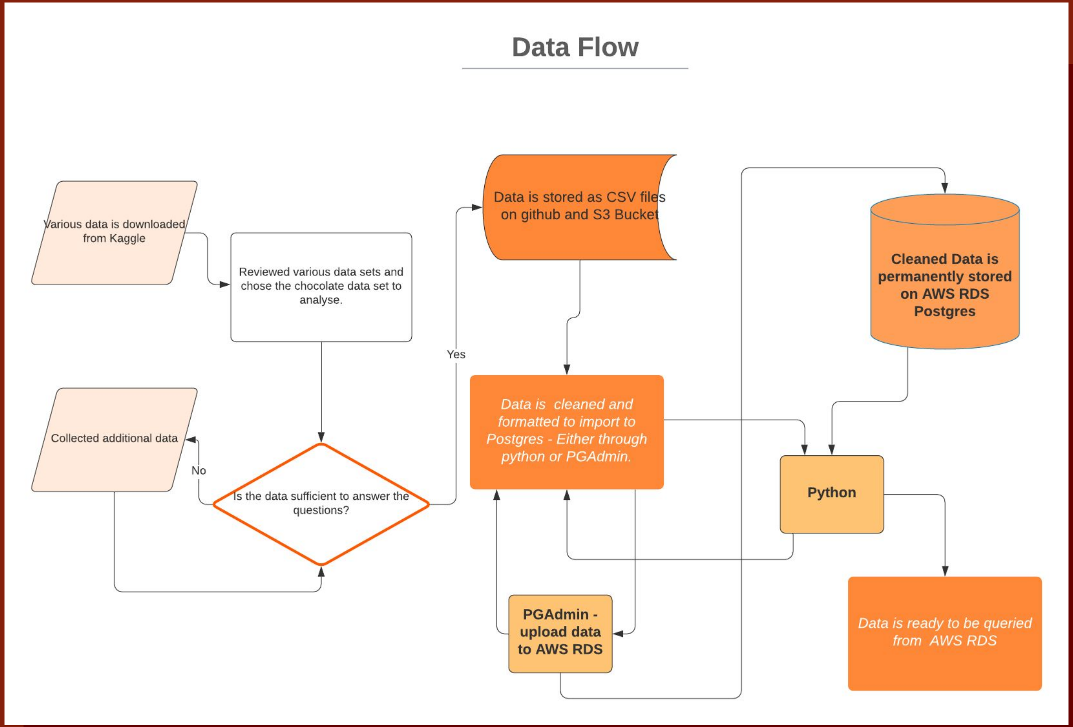
Binary Classifier	Multi Classifier
Random Forest/Balanced Random Forest	Random Forest
Logistic Regression	Extra Trees
SVM	Neural Net
Neural Net	

2. Review model weighting and confusion matrices desired to focus on recall and f1 score
3. Model selection preferred combination of high recall with explanation of results
4. Balanced RandomForest Classifier was selected as it had the best balance of recall and f1 score of the models.

Machine Learning

Model	Precision	Recall	F1 Score	Notes
Logistic Regression	0.44	0.40	0.42	Used Train_scaled, lbfgs
Random Forest	0.37	0.43	0.40	Used SMOTEENN resampling, entropy
Balanced Random Forest	0.30	0.80	0.44	Used Train_scaled, entropy
SVM	0.52	0.18	0.27	
Gradient Boost	0.55	0.18	0.28	
	Accuracy	**Loss**		
Neural Network	0.76	2.82		

Database Structure



Data Flow

Various data is downloaded from Kaggle

Reviewed various data sets and chose the chocolate data set to analyse.

Is the data sufficient to answer the questions?

No

Collected additional data

Yes

Data is stored as CSV files on github and S3 Bucket

Data is cleaned and formatted to import to Postgres - Either through python or PGAdmin.

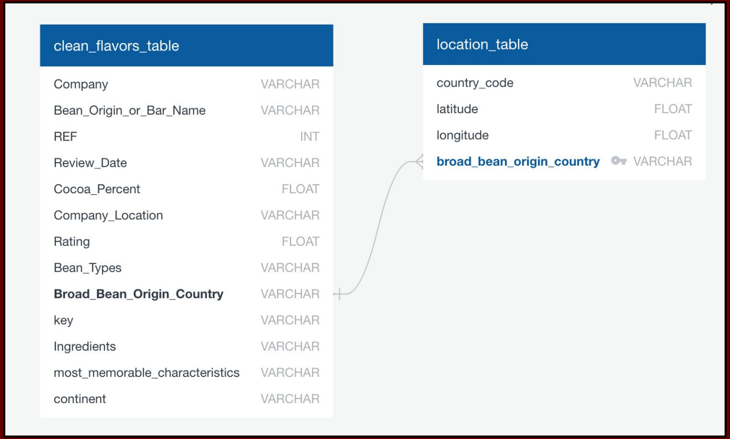
PGAdmin - upload data to AWS RDS

Cleaned Data is permanently stored on AWS RDS Postgres

Python

Data is ready to be queried from AWS RDS

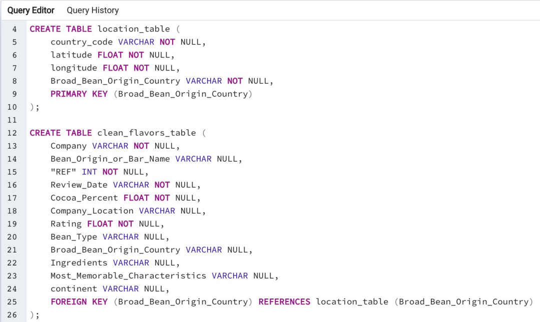
- Cleaned data to S3 Bucket
- PGAdmin to AWSRDS PostGresSql
- SQL Alchemy connection to Python for ML



clean_flavors_table	
Company	VARCHAR
Bean_Origin_or_Bar_Name	VARCHAR
REF	INT
Review_Date	VARCHAR
Cocoa_Percent	FLOAT
Company_Location	VARCHAR
Rating	FLOAT
Bean_Types	VARCHAR
Broad_Bean_Origin_Country	VARCHAR
key	VARCHAR
Ingredients	VARCHAR
most_memorable_characteristics	VARCHAR
continent	VARCHAR

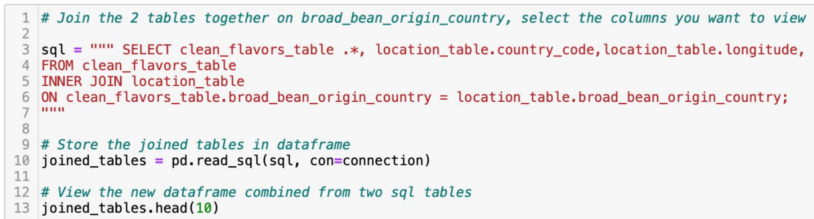
location_table	
country_code	VARCHAR
latitude	FLOAT
longitude	FLOAT
broad_bean_origin_country 🔑	VARCHAR

Database Structure



```sql
CREATE TABLE location_table (
    country_code VARCHAR NOT NULL,
    latitude FLOAT NOT NULL,
    longitude FLOAT NOT NULL,
    Broad_Bean_Origin_Country VARCHAR NOT NULL,
    PRIMARY KEY (Broad_Bean_Origin_Country)
);

CREATE TABLE clean_flavors_table (
    Company VARCHAR NOT NULL,
    Bean_Origin_or_Bar_Name VARCHAR NULL,
    "REF" INT NOT NULL,
    Review_Date VARCHAR NOT NULL,
    Cocoa_Percent FLOAT NOT NULL,
    Company_Location VARCHAR NULL,
    Rating FLOAT NOT NULL,
    Bean_Type VARCHAR NULL,
    Broad_Bean_Origin_Country VARCHAR NULL,
    Ingredients VARCHAR NULL,
    Most_Memorable_Characteristics VARCHAR NULL,
    continent VARCHAR NULL,
    FOREIGN KEY (Broad_Bean_Origin_Country) REFERENCES location_table (Broad_Bean_Origin_Country)
);
```

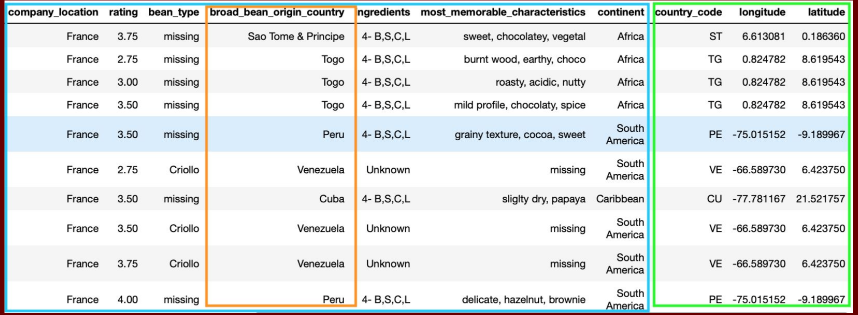
1. Table(s) creation



```python
# Join the 2 tables together on broad_bean_origin_country, select the columns you want to view

sql = """ SELECT clean_flavors_table .*, location_table.country_code,location_table.longitude,
FROM clean_flavors_table
INNER JOIN location_table
ON clean_flavors_table.broad_bean_origin_country = location_table.broad_bean_origin_country;
"""

# Store the joined tables in dataframe
joined_tables = pd.read_sql(sql, con=connection)

# View the new dataframe combined from two sql tables
joined_tables.head(10)
```

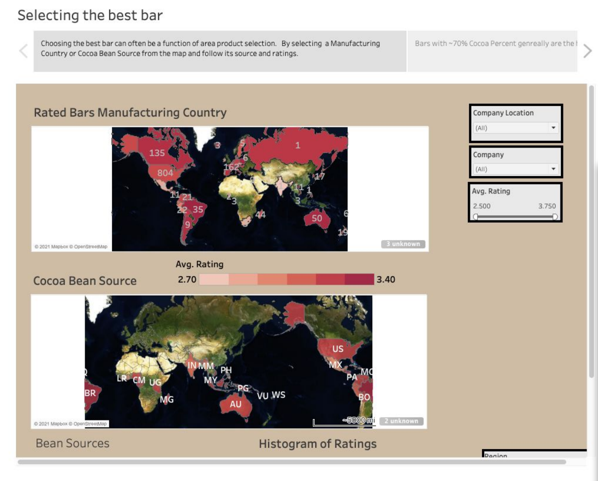
2. Create the join tables

3. Joined Master Table



company_location	rating	bean_type	broad_bean_origin_country	ngredients	most_memorable_characteristics	continent	country_code	longitude	latitude
France	3.75	missing	Sao Tome & Principe	4- B,S,C,L	sweet, chocolatey, vegetal	Africa	ST	6.613081	0.186360
France	2.75	missing	Togo	4- B,S,C,L	burnt wood, earthy, choco	Africa	TG	0.824782	8.619543
France	3.00	missing	Togo	4- B,S,C,L	roasty, acidic, nutty	Africa	TG	0.824782	8.619543
France	3.50	missing	Togo	4- B,S,C,L	mild profile, chocolaty, spice	Africa	TG	0.824782	8.619543
France	3.50	missing	Peru	4- B,S,C,L	grainy texture, cocoa, sweet	South America	PE	-75.015152	-9.189967
France	2.75	Criollo	Venezuela	Unknown	missing	South America	VE	-66.589730	6.423750
France	3.50	missing	Cuba	4- B,S,C,L	sligty dry, papaya	Caribbean	CU	-77.781167	21.521757
France	3.50	Criollo	Venezuela	Unknown	missing	South America	VE	-66.589730	6.423750
France	3.75	Criollo	Venezuela	Unknown	missing	South America	VE	-66.589730	6.423750
France	4.00	missing	Peru	4- B,S,C,L	delicate, hazelnut, brownie	South America	PE	-75.015152	-9.189967

Dashboard Tableau Dashboard - Story Board

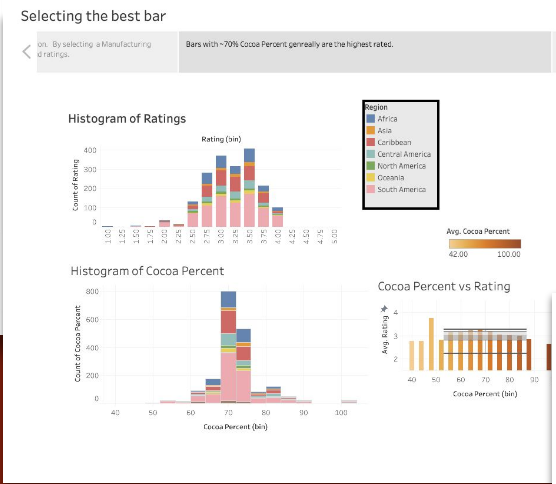


Selecting the best bar

Choosing the best bar can often be a function of area product selection. By selecting a Manufacturing Country or Cocoa Bean Source from the map and follow its source and ratings.

Bars with ~70% Cocoa Percent genreally are the h

- Select Bar Manufacturer / Cocoa Bean Country to see where the flow of beans to bar throughout the world. Filter on brand or rating



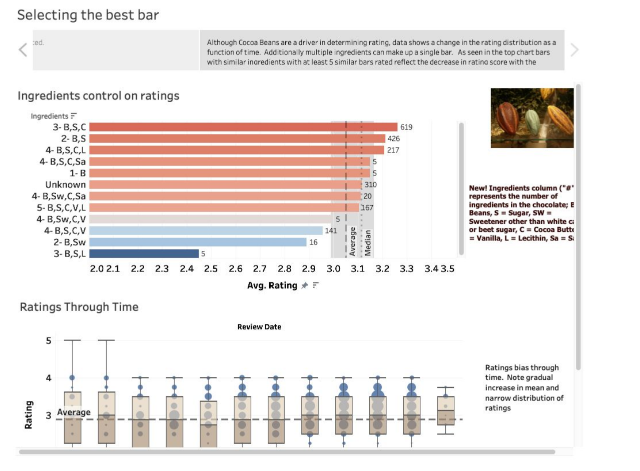
Selecting the best bar

on. By selecting a Manufacturing d ratings.

Bars with ~70% Cocoa Percent genreally are the highest rated.

Althou functi with s

- Examine Ratings relationship to Cocoa percentage….hint more is not necessarily better



Selecting the best bar

ted.

Although Cocoa Beans are a driver in determining rating, data shows a change in the rating distribution as a function of time. Additionally multiple ingredients can make up a single bar. As seen in the top chart bars with similar inaredients with at least 5 similar bars rated reflect the decrease in ratina score with the

New! Ingredients column ("#" represents the number of ingredients in the chocolate; B Beans, S = Sugar, SW = Sweetener other than white ca or beet sugar, C = Cocoa Butte = Vanilla, L = Lecithin, Sa = Sa

Ratings bias through time. Note gradual increase in mean and narrow distribution of ratings

- Review other controls on bar ratings, such as ingredients….the world loves sugar
- Chocolate critics narrow their voting over time and a general trend to increase the mean through time

Technology

Data Cleaning
Python-Pandas

Version Control
GitHub

Data Storage
Amazon S3

Database Admin
PGAdmin

EDA
Python and Seaborn

RDS- DataBase
AWS PostgresSQL

SQLAlchemy
Database Connector

Machine Learning
Pandas/Sci Kit Learn /Tensor Flow/ Imbalance Learn/Entropy

Dashboard
Tableau

Conclusions

- Cocoa percent is the biggest driver in choosing the best bar. 65-70% is the “sweet” spot

- Chocolate is complex and each chocolatier has their own preparation recipe

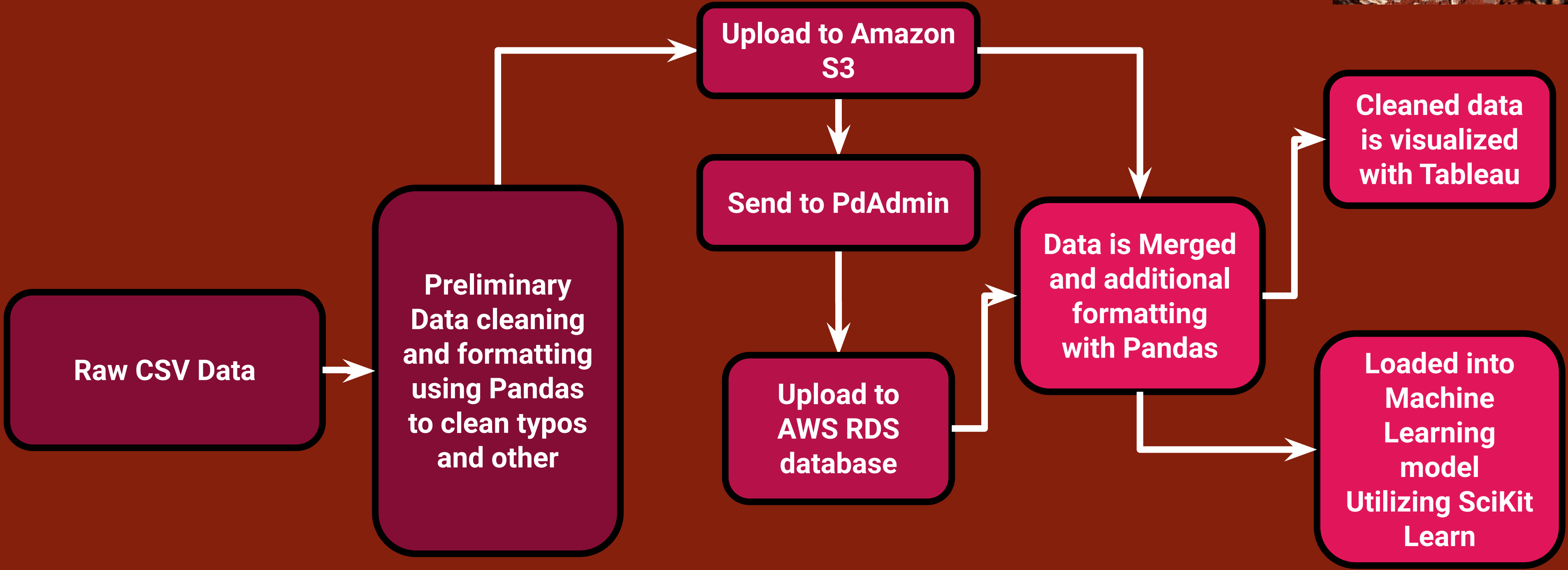
- Lots of hybridization of the beans means an ever growing chocolate portfolio

Appendix

Team Roles

- Square -(Bruce) Set up Repository and Branch structure
- Triangle - (Tahereh) Build additional ML models to see initial accuracy
- Circle - (Yan) Constructed EDR for PostGres SQL and begin building framework
- X- (Travis) Presentation and technology

Data Processing and Evaluation Workflow



Raw CSV Data

Preliminary Data cleaning and formatting using Pandas to clean typos and other

Upload to Amazon S3

Send to PdAdmin

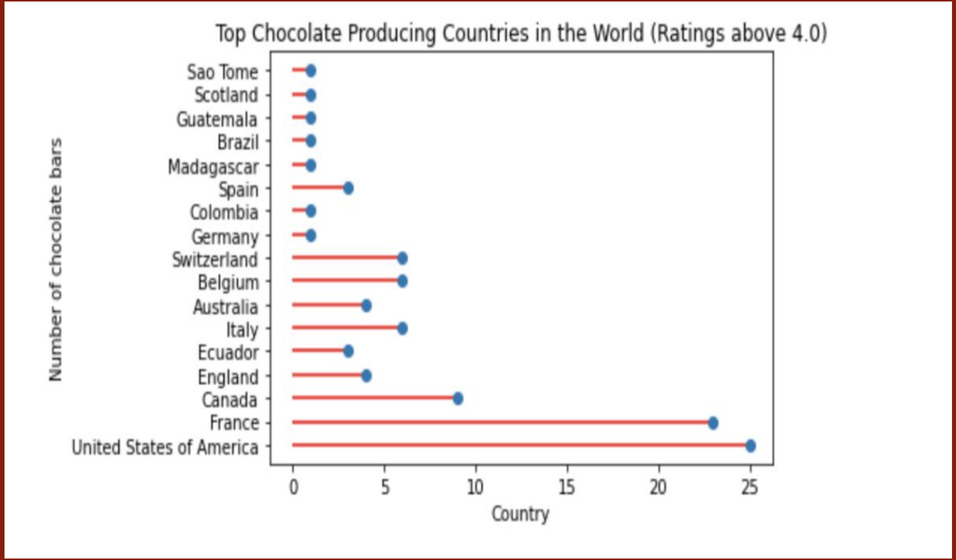
Upload to AWS RDS database

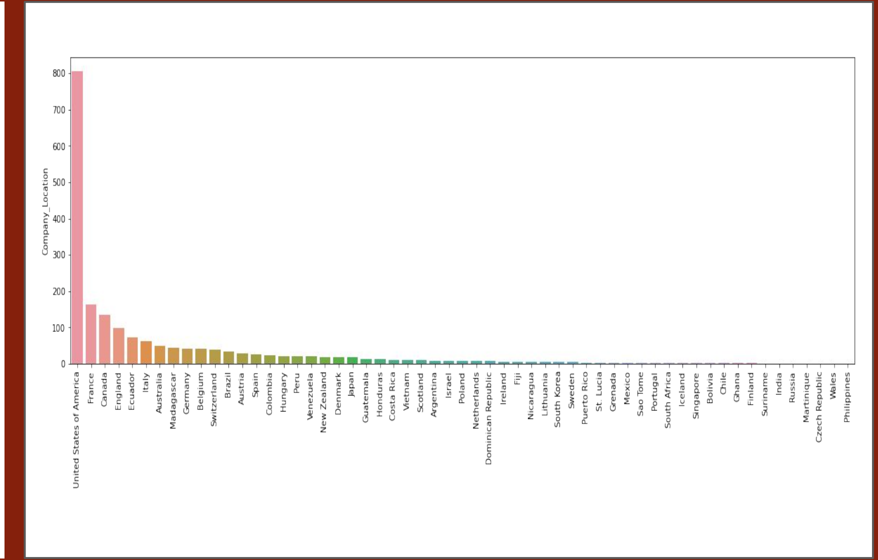
Data is Merged and additional formatting with Pandas

Cleaned data is visualized with Tableau

Loaded into Machine Learning model Utilizing SciKit Learn

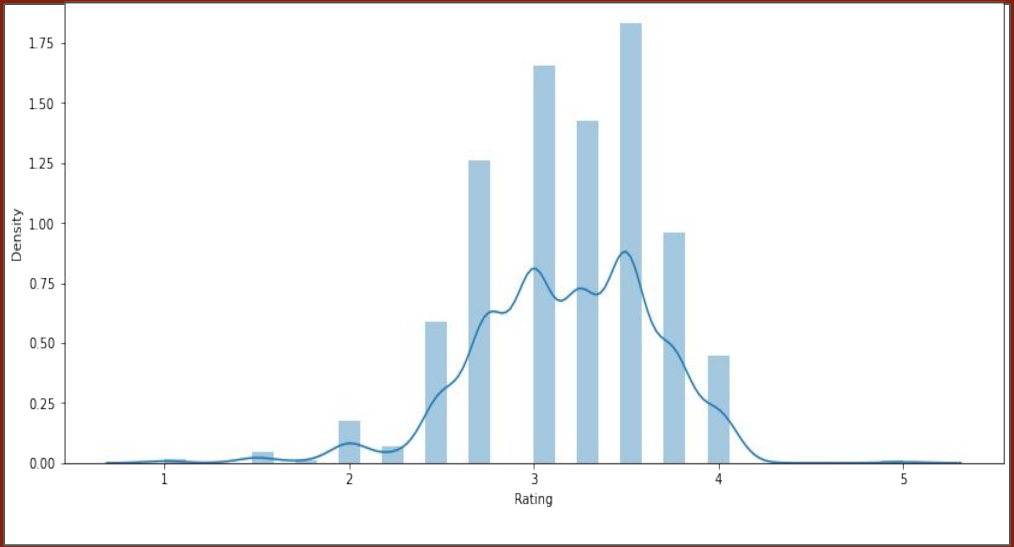
EDA / Analysis Example: Best Chocolate Crafting Country

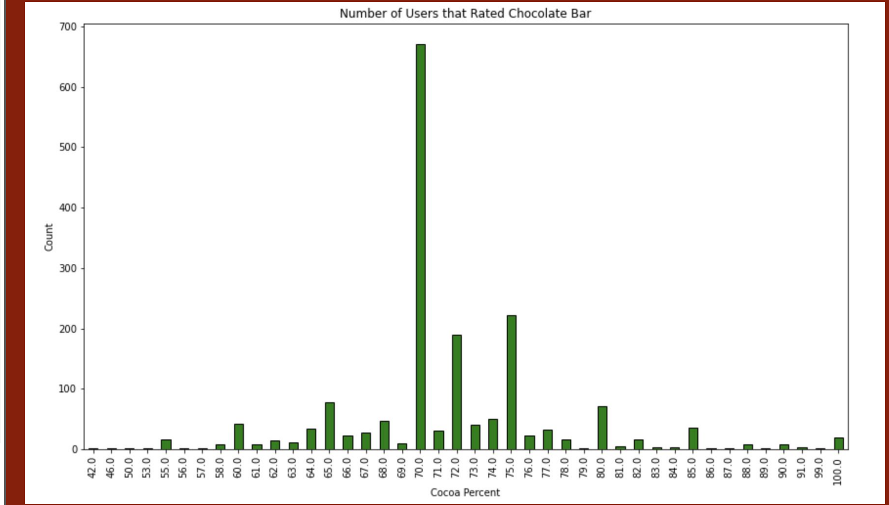




- European and North America nations dominate the list of artisan chocolatiers

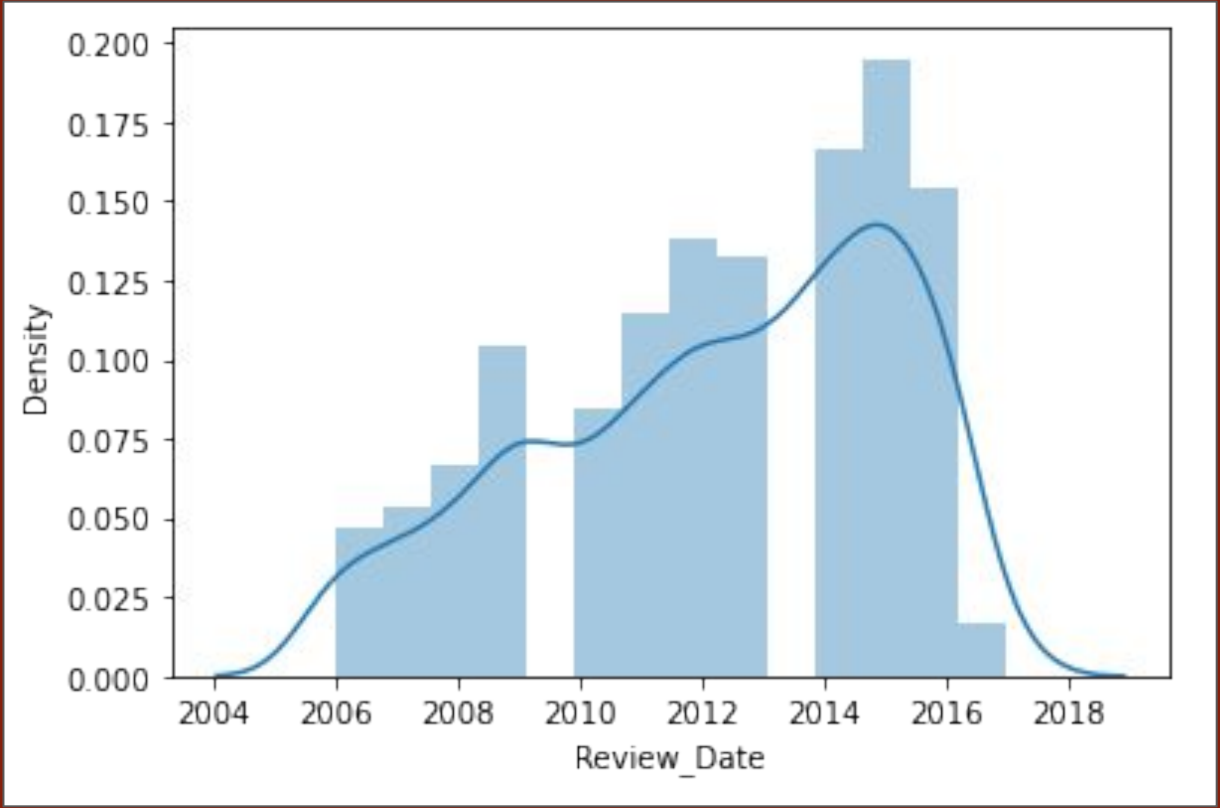
EDA / Analysis Example: Histogram of Bar Ratings





- **Slight left skew on the chocolate bar ratings**

Review Date density increase through time

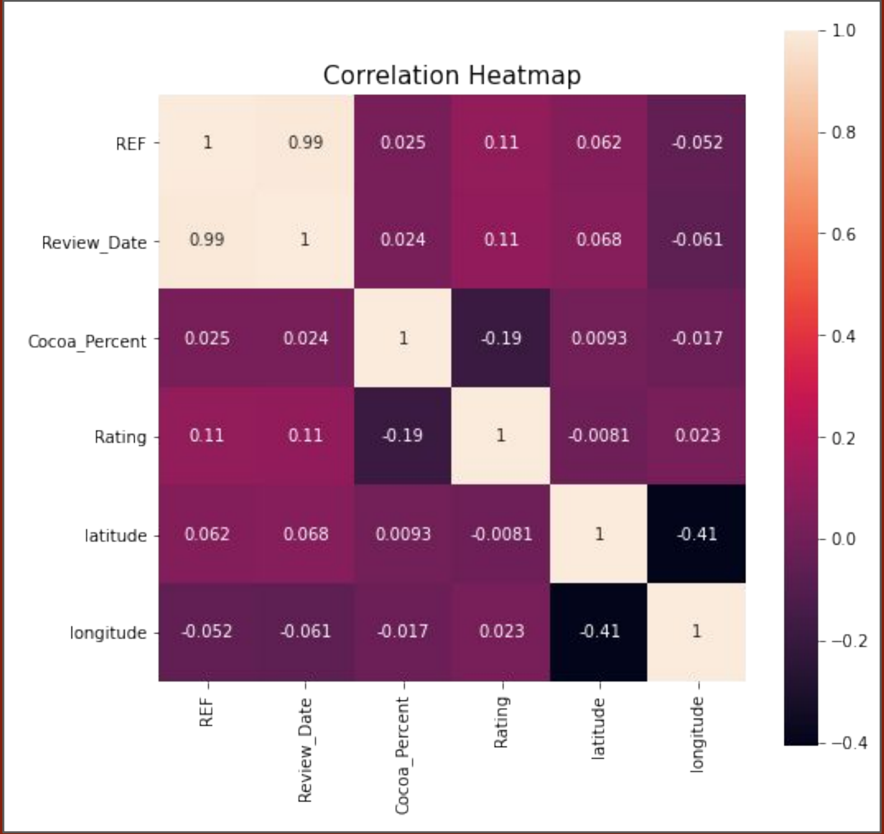


- Increase in number of rating through time
- Left skew

Feature correlation heatmap

- REF number is tied to date as stated in documentation

- Cocoa and Rating have the highest correlation of the features


Correlation Heatmap
REF
Review_Date
Cocoa_Percent
Rating
latitude
longitude
REF
Review_Date
Cocoa_Percent
Rating
latitude
longitude