Chocolate Bar Ratings Predictor

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

Topic Selection

Predicting chocolate bar ratings from variable such as:

- **Chocolateur**
- **Bean type**
- **Bean Source**
- **Cocoa Percent**

Reason for Topic Selection

- Examined several food and beverage topics
- 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.

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 sources

- Kaggle- Chocolate Bar ratings
  - https://www.kaggle.com/rtatman/chocolate-bar-ratings
- Kaggle - Countries and States Lat Long
  - https://www.kaggle.com/paultimothymooney/latitude-and-longitude-for-every-country-and-state
- Flavors of Cacao website
  - http://flavorsofcacao.com/chocolate_database.html

Four questions to answer

- Can we predict if a bar will be rated in the top 15%(>=3.75)?
  - 1 SD above the mean
- 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 and Matplot lib 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

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

Data Preprocessing and Cleaning

- Pandas will be the primary avenue for cleaning and preparing the data
- Basic data editing of the csv will initially be performed
- Cleaned Data will be uploaded to S3
- Python will read from AWS PostGres SQL and format for use in Machine Learning
- Tableau Dashboard uses cleaned merged csv exports from Python/Pandas

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

Data Preprocessing and Cleaning

- Corrected Typos and inconsistent country names from Broad Bean Origin and Company Locations
- Formatted numerical values consistently removing string characters
- Add Continent/Region column and merged in Ingredients and Most Memorable Characteristics from webscrape data using a unique concatenated key
- Handled outliers in Review Data and Cocoa Percent
- Dropped or replace null values

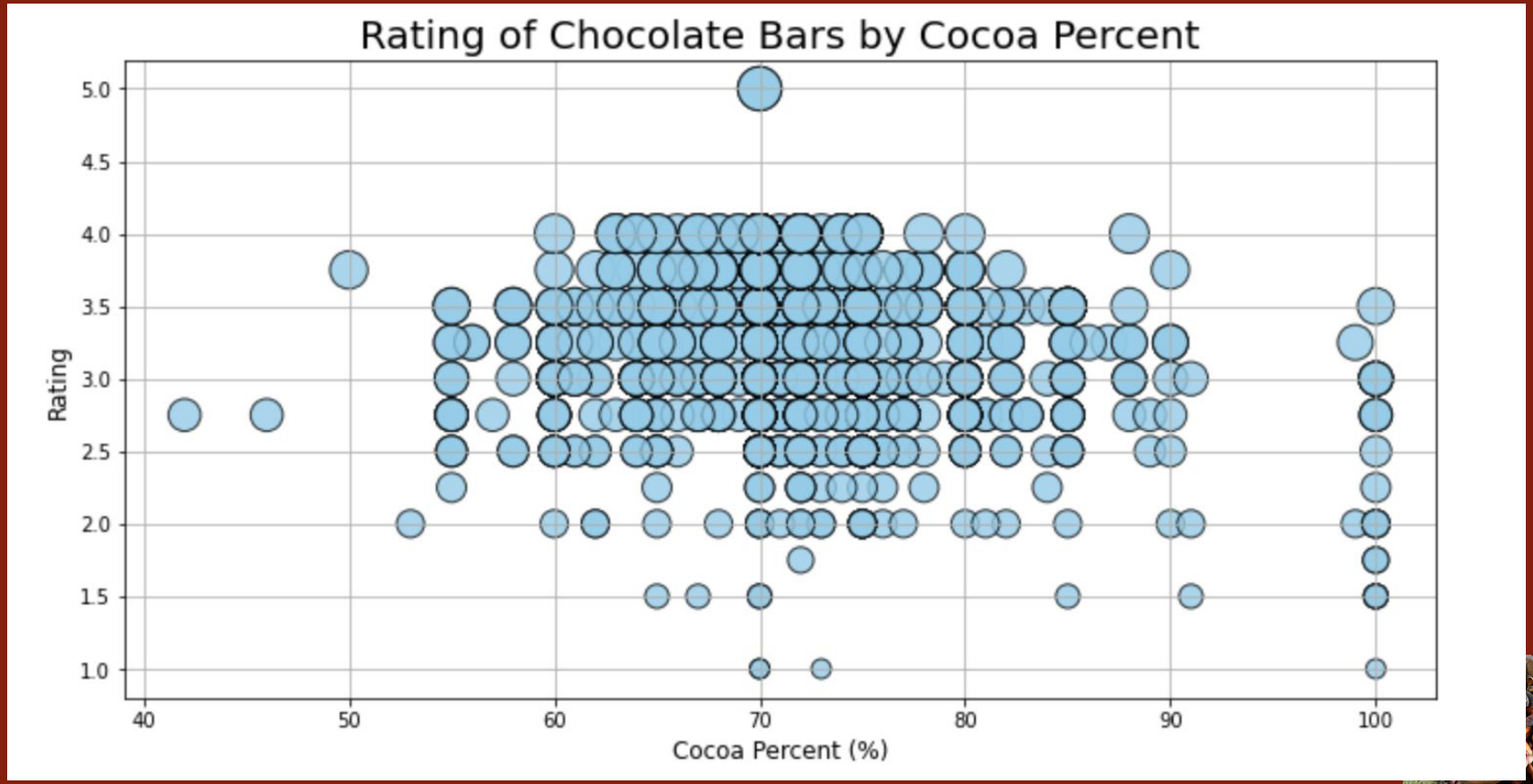
See README for details:

https://github.com/jilek/DataBootcampFinalProject/tree/main/Segment2_Deliverable

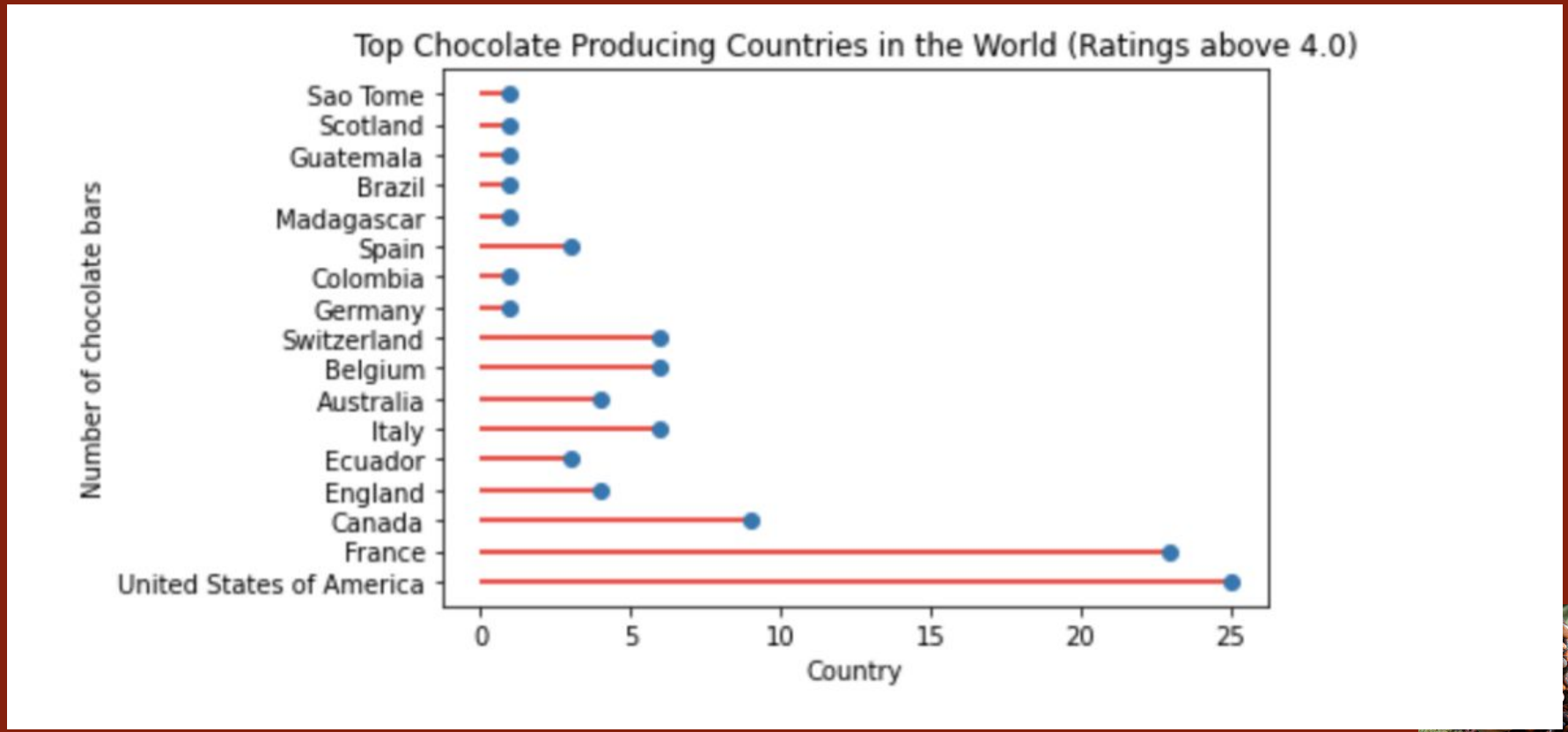
EDA / Analysis Example: Where are the best beans grown?

	Broad_Bean_Origin	REF	Review_Date	Cocoa_Percent	Rating
43	Solomon Islands	1811.000000	2016.250000	74.000000	3.437500
17	Haiti	1354.444444	2014.000000	71.333333	3.388889
19	Honduras	1478.666667	2014.533333	73.933333	3.350000
16	Guatemala	1352.758621	2013.896552	71.758621	3.344828
39	Republic of Congo	1091.600000	2012.600000	70.500000	3.325000

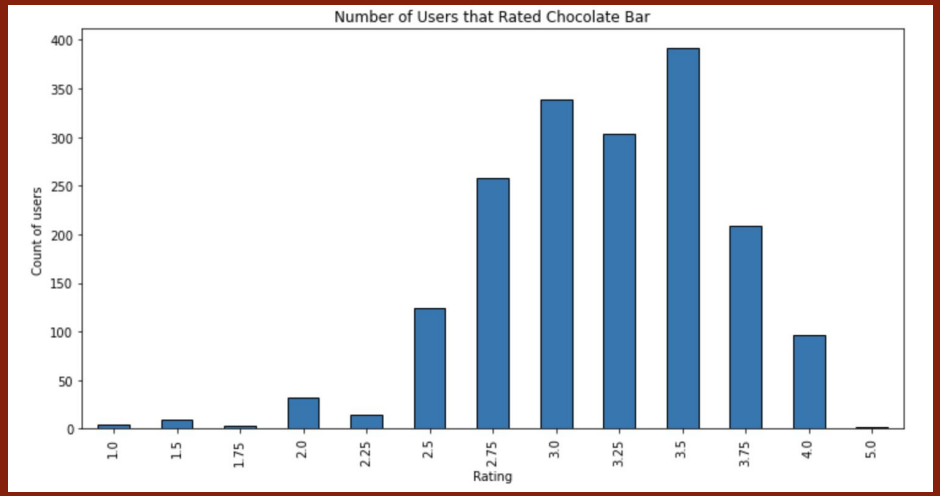
EDA / Analysis Example:Does Cocoa % lead to better bar rating

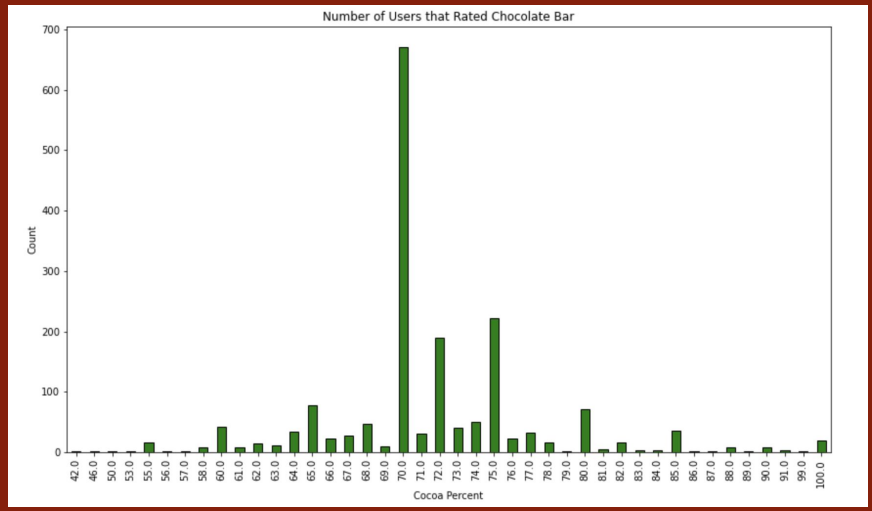


EDA / Analysis Example: Best Chocolate Crafting Country



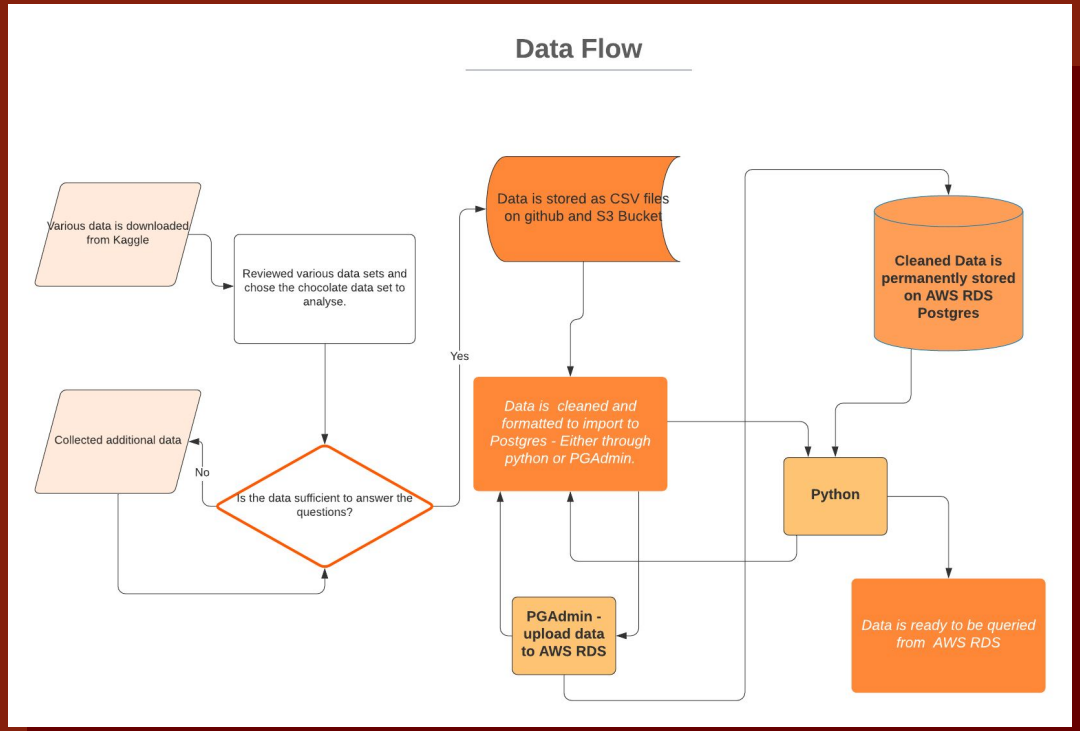
EDA / Analysis Example: Histogram of Bar Ratings





Slight left skew on the chocolate bar ratings

Database Structure



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

Database Structure

Query Editor Query History

```
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)
);
```

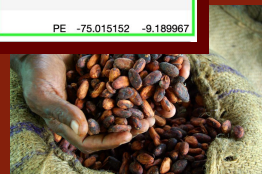
1. Table(s) creation

```
# 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	sligtly 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

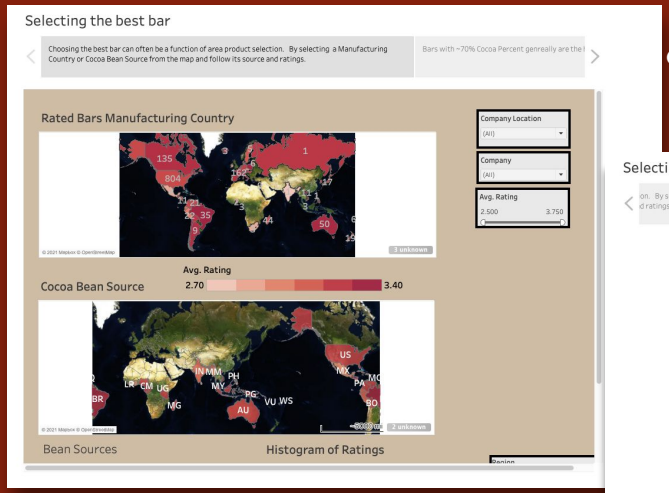


Machine Learning Model

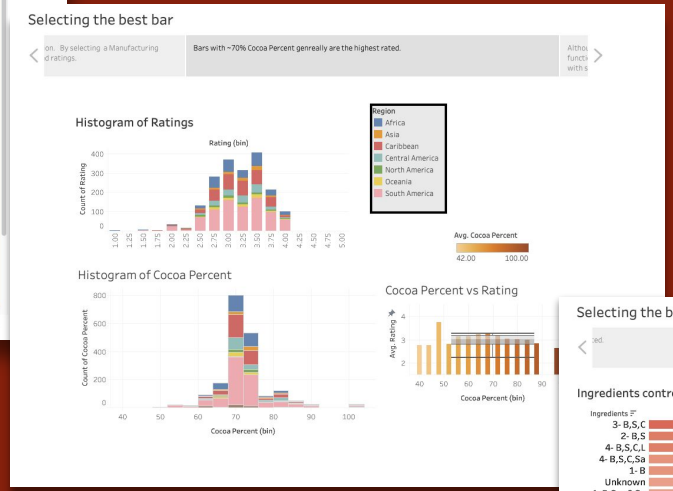
- Review five ML classifiers
  - Random Forest (Binary and Multi)
  - Extra Trees Classifier (Multi)
  - SVM
  - Neural Net
- Review model weighting and confusion matrices desired to focus on recall
- Model selection preferred combination of high recall with explanation of results
- RandomForest Binary classifier was selected as it multiclass results were subpar (below a coin toss) and Neural Net was not useful in giving insight (black box) to how a user might want to know the drivers for choosing a better chocolate bar

Dashboard

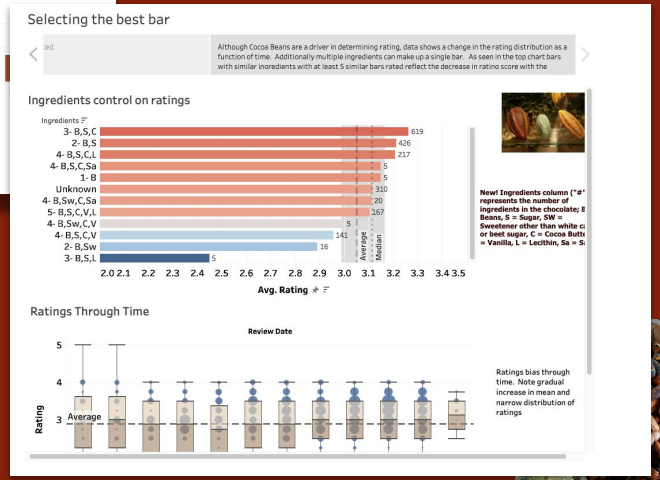
Tableau Dashboard - Story Board



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



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



- 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