Chocolate Bar Ratings

08.30 Cohort- Brenda: Madeline Ye

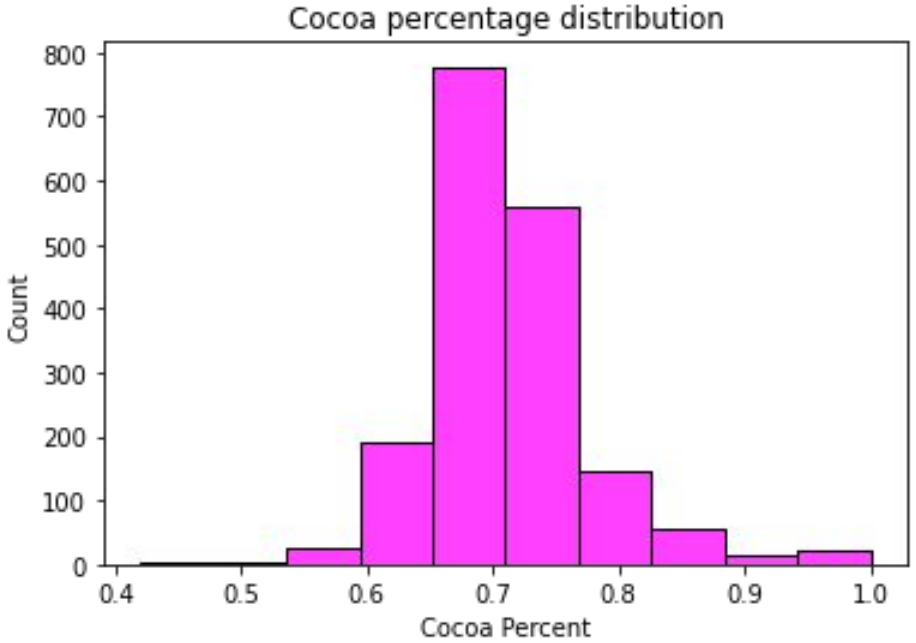
What's in the data set?

- 1,700 plain dark chocolate bars
- Company name
- Specific Geographic Region of origin for the bar
- Review Date
- Percentage of cocoa
- Company location
- Expert rating for the bar
- Variety(breed) of cacao bean used
- Broad geo region of origin for the cacao bean

These ratings were compiled by Brady Brelinski, Founding Member of the Manhattan Chocolate Society
http://flavorsofcacao.com/index.html
Kaggle:https://www.kaggle.com/rtatman/chocolate-bar-ratings

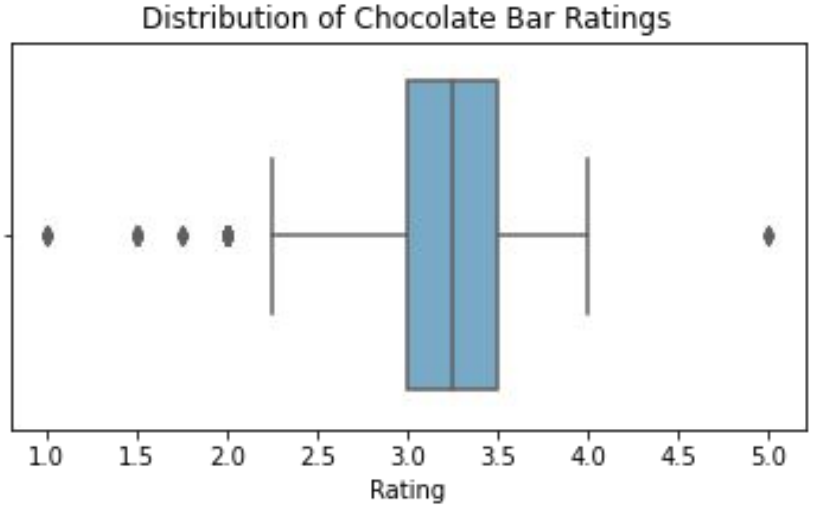
Exploratory Visuals: Cocoa percentage

- Most bars are in the high 60s, low 70s
- Normal distribution and some values at the tail ends



Exploratory Visuals: Rating

- Rated on:
  - Flavor
  - Texture
  - Aftermelt
  - Overall
- Investigated Outliers

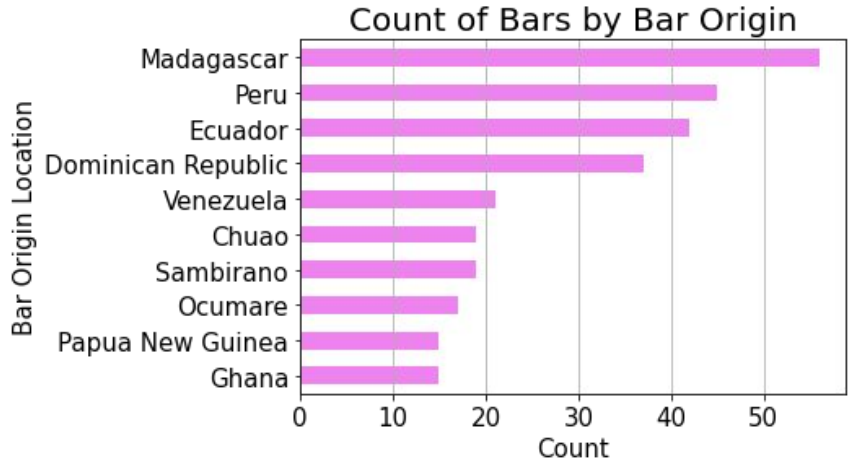


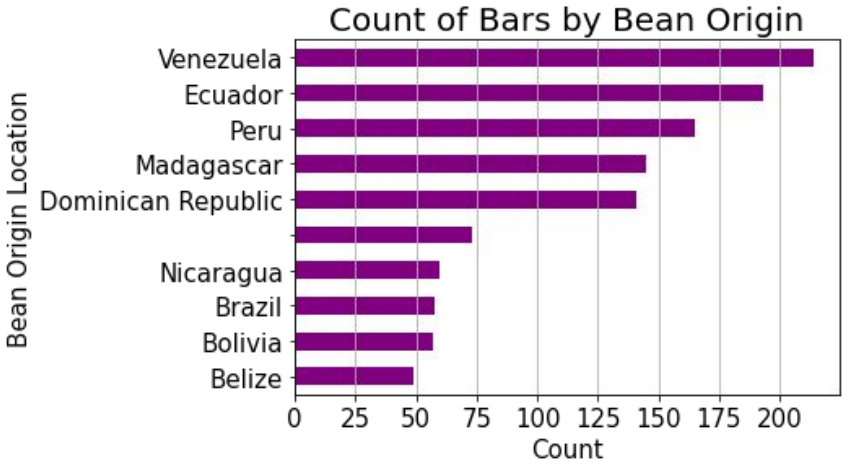
Flavors of Cacao Rating System:

- 5= Elite (Transcending beyond the ordinary limits)
- 4= Premium (Superior flavor development, character and style)
- 3= Satisfactory(3.0) to praiseworthy(3.75) (well made with special qualities)
- 2= Disappointing (Passable but contains at least one significant flaw)
- 1= Unpleasant (mostly unpalatable)

Exploratory Visual: Origins

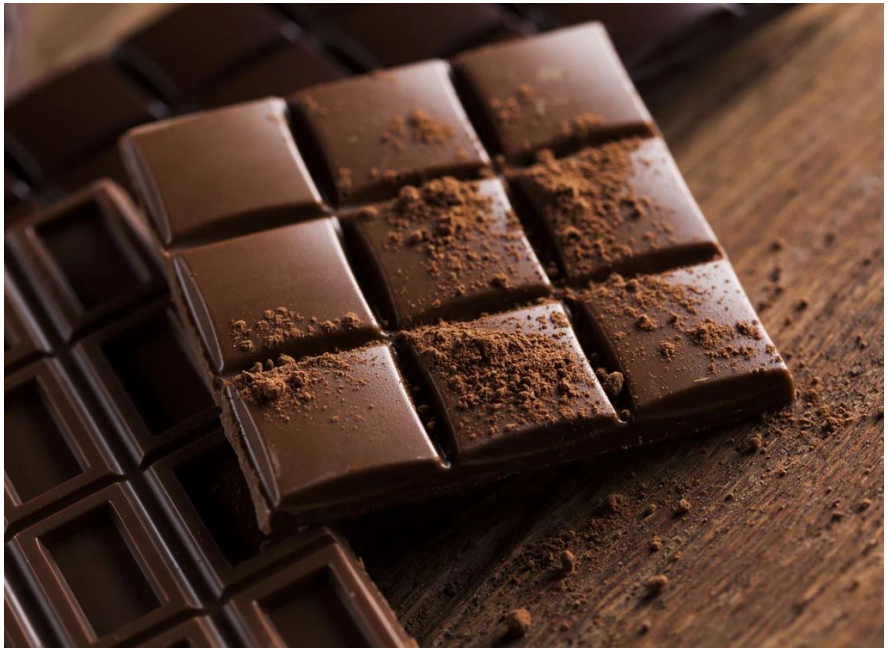
- Many more beans originate from Venezuela, Ecuador, Peru, etc… while Bar Origin is much more dispersed
- Bar Origin has 1038 unique values while Bean origin has 99





Modeling and Predicting Chocolate Bar Ratings

- Methods Used: Random Forest Regressor, KNN Regressor, Gradient Boosting Regressor
- Hypertuning done on all three models
- Best Method: K Nearest Neighbor Model(n = 5 neighbors)
  - Able to account for 79% of the variability of the rating score with the KNN model
  - Mean Average Error is about 0.08 off on ratings for chocolate bars.



Suggestions and Next Steps...

- Possible further hypertuning of the KNN model
  - Some overfitting occurred.
- Treating the target column as a categorical value therefore creating a classification problem instead
- Using Neural Networks since it is a larger dataset

Thank you!

Appendix

Cleaning Steps

- Column name formatting
- Two null values:
  - Bean origin and Bean type

```
#Checking the specific on these null values
df[df.isnull().any(axis=1)]
```

	Company(Maker)	Specific_Bean_Origin_Bar_Name	REF	Review_Date	Cocoa_percent	Company_location	Rating	Bean_type	Broad_Bean_Origin
1072	Mast Brothers	Madagascar	999	2012	72%	U.S.A.	2.5	Trinitario	NaN
1544	Soma	Three Amigos(Chuao, Wild Bolivia, D.R.)	676	2011	70%	Canada	4.0	NaN	Ven, Bolivia, D.R.

```
#Deciding to drop the null values as they are not a simple fix of imputing or backfilling and it is only 2 data points
df = df.dropna()
```

Cleaning Continued...

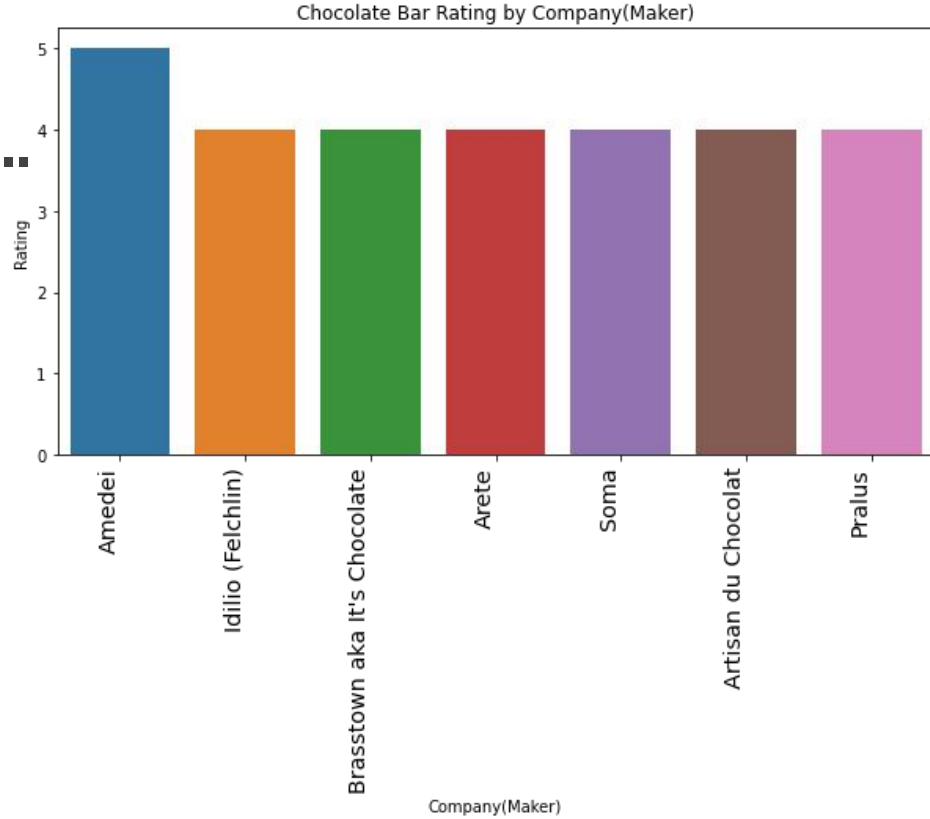
- Dropping Review Date Column, irrelevant
- Changing cocoa percentage to float
- Empty column value for Bean type

```
In [27]: df['Bean_type'].value_counts()
         #We see that there are a lot of blank values for bean type. Might be worth filling in this with "Unknown"

Out[27]:                          887
         Trinitario               418
         Criollo                  153
         Forastero                 87
         Forastero (Nacional)      52
         Blend                     41
         Criollo, Trinitario       39
         Forastero (Arriba)        37
         Criollo (Porcelana)       10
         Trinitario, Criollo        9
         Forastero (Parazinho)      8
         Forastero (Arriba) ASS     6
```

Exploratory Visuals Top Brands...

- Top 7 rated chocolate bars
- No other rating comes close to the top



Company(Maker)	Specific_Bean_Origin_Bar_Name	REF	Cocoa_percent	Company_location	Rating	Bean_type	Broad_Bean_Origin
Amedei	Toscano Black	40	0.7	Italy	5.0	Blend	
Amedei	Chuao	111	0.7	Italy	5.0	Trinitario	Venezuela