Deja Brew

Team Members: Ja-Rey Corcuera & Brianna Mendoza

Emails: u1156525@utah.edu & u1189033@utah.edu

UID's: u1156525 & u1189033

Project Description

In this project we will explore multiple coffee reviews, all given by professional coffee gurus, and explore how each rating compares, correlates, or differs from others. We aim to explore how these ratings might affect the bean producer's coffee sales, or how it might affect the coffee drinker's experience. We also want to see if we might be able to predict a price for a given roast given a price point, or a location, or even a roast level. Can the acidity of a roast determine the price? Or can the location? Is it important to consider the agtron level when picking a roast? Our aim is that we can acquire more useful information about a beverage that so many people consume on a daily basis that may enrich their experiences with it.

Data and Data Description

After scraping the relevant data from the website, we were left with 3010 reviews -- each had 14 columns of relevant data. The column data includes the location of the roaster, the origin of the coffee, roast level, estimated price, review date, aroma, body, flavor, aftertaste, agtron exterior, and agtron grind. Two extra rows were added based off of the data we already had to provide more comprehensive information. The first being a price in dollars per ounce -- this required that we drop the coffee that was not priced with US dollars, or the ones that were not priced by ounces. It was not many, so we concluded that it would be fairly insignificant compared to the amount of remaining data that we still had. Lastly, we quantified the roast level, assigning a value 0-5 based on the roast level. For instance, an extremely light roast would be represented as a 0, while an extremely dark roast would be represented by a 5. This would allow

us to use the qualitative data with other quantitative data more easily. The figure below shows three reviews in our dataframe and the attributes for each.

	Roaster	Name	Rating	Roaster Location	Coffee Origin	Roast Level	Est. Price	Dollars/Oz	Review Date	Aroma	Body	Flavor	Aftertaste	Agtron Ext	Agtron Gnd	Roast_Level_num
0	Paradise Roasters	Colombia Finca La Primavera Sidra	96	Minnesota	Colombia	Light	$48.00/12 ounces	4.00	March2022	9	9	10	9	64	82	0
1	SkyTop Coffee	Ethiopia Anaerobic Shantawene	94	New York	Ethiopia	Medium-Light	$20.00/12 ounces	1.67	March2022	9	9	9	8	58	80	1
2	Jaunt Coffee Roasters	Ethiopia Bensa Asefa Qonqana	94	California	Ethiopia	Light	$26.00/12 ounces	2.17	March2022	9	9	9	8	64	86	0

Ethics of Data

One major concern that we had was about where and how the beans are sourced. What we hoped was that every coffee is ethically sourced and that workers are not exploited, are paid fair wages, etc. Unfortunately, there was little information on whether that was considered when reviewing any given coffee. The only information about the source of the beans was the location. Sometimes there was an exact location, other times it was just the state, or even just the country. The unspecified sourcing leaves a great concern, as we do not wish to support any distributors or roasters that do not support their workers.

Another way that this may occur is if these reviews benefit the distributor or the roasters. If there is more traction for either the sourcer or the roaster after a review is posted and the coffee is not ethically sourced, this may contribute to the exploitation of the workers or even the land on which it was grown. Both are important to take into account.

Again, there is not much on this website as to whether the coffee is ethically sourced, but with further time with the data, we could manually check if each coffee is ethically sourced. However, if it is not, this does not do anything about the fact the review is already posted and may have already had any impact on the distributor or roaster, worker etc.

Methods

So far, we have been successful in finding the correlations between each of the columns and have represented them in multiple ways. See the figures below.

The following figure shows the correlations between each of the characteristics. Some initial interpretations that we can see are that the rating is most highly correlated with the flavor followed by the aroma. However, rating is negatively correlated with the roast level which seems to indicate that lighter roasted coffees tend to receive better ratings.

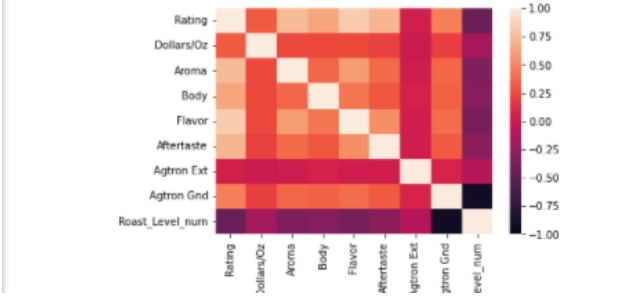
```
131]: 1 coffee_df.corr()
```

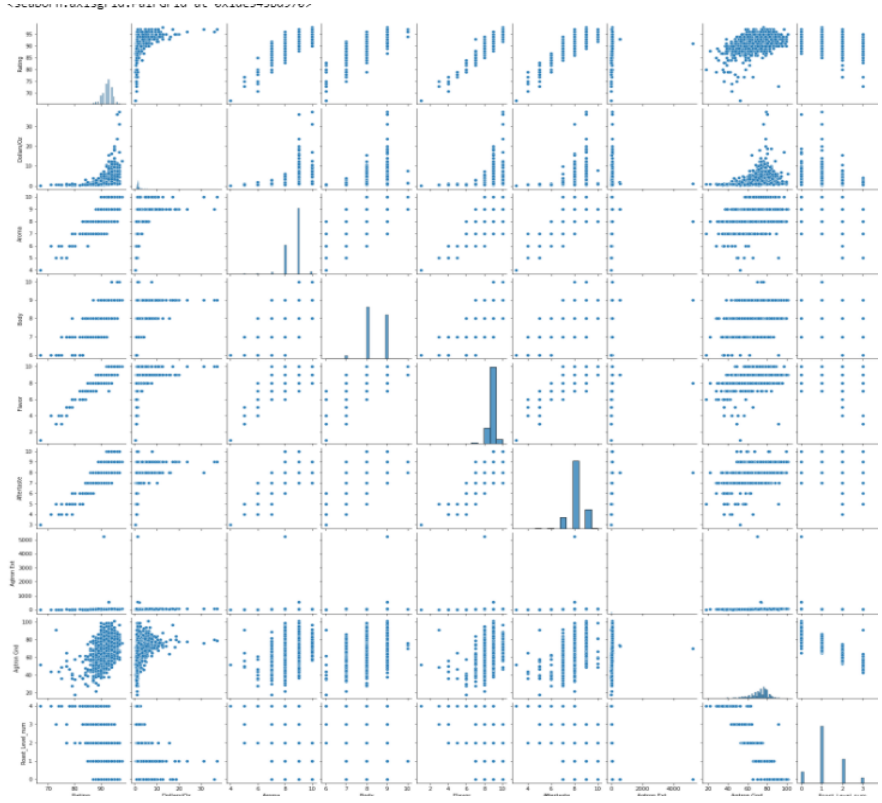
131]:

	Rating	Dollars/Oz	Aroma	Body	Flavor	Aftertaste	Agtron Ext	Agtron Gnd	Roast_Level_num
Rating	1.000000	0.311558	0.747029	0.643863	0.831778	0.721988	0.035427	0.467331	-0.462270
Dollars/Oz	0.311558	1.000000	0.236355	0.239126	0.235012	0.217382	0.005729	0.195762	-0.188394
Aroma	0.747029	0.236355	1.000000	0.360587	0.596813	0.381989	0.015439	0.363298	-0.360741
Body	0.643863	0.239126	0.360587	1.000000	0.427123	0.300708	0.050424	0.342259	-0.324906
Flavor	0.831778	0.235012	0.596813	0.427123	1.000000	0.534577	0.016960	0.388427	-0.391671
Aftertaste	0.721988	0.217382	0.381989	0.300708	0.534577	1.000000	0.028644	0.305447	-0.304751
Agtron Ext	0.035427	0.005729	0.015439	0.050424	0.016960	0.028644	1.000000	0.064191	-0.104185
Agtron Gnd	0.467331	0.195762	0.363298	0.342259	0.388427	0.305447	0.064191	1.000000	-0.908084
Roast_Level_num	-0.462270	-0.188394	-0.360741	-0.324906	-0.391671	-0.304751	-0.104185	-0.908084	1.000000

The figures below show the heatmap of the characteristics' correlations and also the scatter plot matrix.

```
In [132]: 1 sns.heatmap(data = coffee_df.corr(), vmin = -1, vmax = 1)
Out[132]: <AxesSubplot:>
```





We have completely scraped the website we are using for all the relevant data we need and have sufficiently cleaned it to the point of effortless usability. Moving forward, our goal is to be able to analyze these characteristics across coffees that come from different countries and continents of the world and be able to illustrate them graphically.

Preliminary Results

Through analyzing the above data, we can see that there are clear correlations between some of the characteristics, but close to none in others. We can see that roast level has a negative, fairly strong correlation with many of the other factors. Depending on the way the data is set up, that can mean a number of things, but for the sake of understanding, let us look at one of the correlations. There is a very strong positive correlation between aroma and rating. This tells us that the higher the aroma, the higher the rating. The exact correlation value is .747, which is fairly high. This correlation makes sense, as a beverage with a stronger aroma will likely have a

fuller taste and last longer in the drinker's mouth, creating a more lasting experience. For people who love coffee for the taste, this is likely a lovely experience, which explains the ratings.

We can explore individual correlation values in more depth as the project continues, but let us, lastly examine the pair plot to the left of the heatmap. We can observe that the correlations can be seen by the plotting of these values. The characteristics with the observed highest correlations in the heatmap have plottings indicating the correct correlations.

Each plot provides the same information but portrays it to us differently. The heat map provides possibly an easier-to-visualize representation of how tightly correlated each characteristic is to the next, but the pairplot lets us see more clearly how. And the values will numerically validate the other two plots.

There is still much to do in regards to our methods, but this is where we are now. We hope to look at locations next and see if there are any patterns in the coffee based on where it comes from. Is there a place that notoriously provides mediocre coffee, or is there one that consistently provides a high-quality brew? We hope to find something interesting within this data.

Peer Feedback

When meeting with the other group, they provided immense support for our project and offered insightful direction. One of the things they brought up was our stakeholder analysis. We mentioned to them that we had a hard time identifying who or what would hold stake with our data, and they offered the perspective of the roasters. Would this data, or the ratings themselves affect the roaster's or the distributor's sales or overall buyer interaction? We have since included it in our stakeholder analysis and ethical analysis.

We also discussed the biases in our data, as all of the reviews are done by the same group of 5-7 men. These ratings are centered around their own personal preferences. One can anticipate

that the evaluations were done as unbiased as possible, but there is no 100% guarantee. They may be conducting their own analysis' as objectively as they can but it may very well still be biased. To remedy this in a few of our computations, we are attempting to normalize the ratings, but there may still be skewed data.

Completed Milestones

Completed: Scraped all relevant data from the website; wrangled all necessary categorical data and transformed it into numerical data. Performed minimal regressions and correlations, started calculating value per 12oz cup of coffee.

Upcoming Milestones

Upcoming: create classifiers to recommend the perfect cup per user; find top 10 valued coffees; word maps, world countries csv file was obtained here; demonstration video.

Summary

After scraping our data, getting our proposal reviewed, and setting up a few regressions, we are confident that our data can effectively communicate the things we aim to communicate. We will be able to show correlations between price points and roast levels, and examine relationships between place of origin and the quality of the coffee. Our data is organized, converted, and ready to use for whatever we will need it for in the remainder of our project.

Assessment

We are on track to finish this project by the due date. Our data is fully accessible and usable, all we have to do is create our regressions, correlations, and maps. The most difficult part will likely be plotting the geolocation maps, but that may be something we can go without if need be.