Christian Matava: Craft Beers and Breweries

Extracting:

1. Used Kaggle to find all three of our csv files used in this project

Transformations for the breweries.csv; done in Breweries.ipynb:

1. First I examined the file. Then I printed the data in pandas to see what it looked like on there. I realized it had a header row that needed to be skipped in order to transform the data. So I performed a skiprows function.
2. Next I renamed the columns. The first column was named “Unnamed: 0”, which is a meaningless column name as what it actually refers to is the Brewery ID# which is what I changed it to. After that I renamed the “name” column to “brewery”.

Transformations for the beers.csv; done in Craft Beers.ipynb

1. First I examined the file. Then I printed the data in pandas to see what it looked like on there. Once again I had a title row that needed to be skipped in order to transform the data. In order to do that I performed a skiprows function.
2. Next I deleted columns that I did not need. The first one being the “ibu” column. While it is an interesting data set, many of the rows didn’t have data for this column. As a result I made the decision that it would be better to just get rid of the column entirely as opposed to having missing data.
   1. The next column I deleted was the “Unnamed: 0” column. This column was basically useless as it was the same as the index and didn’t offer any new information. Unlike in the Breweries.csv file, in which that “Unnamed: 0” column referred to the Brewery ID, the beers.csv file already had a Brewery ID column.
   2. Lastly, I deleted the “id” column for similar reasons. It didn’t add any useful or new information.
3. Renaming the Columns:
   1. I performed a “df.rename” function, and changed “name” to “craft\_beer” as it was more descriptive.
4. Reordering the Columns:
   1. Here I wanted to reorder the columns so that they were presented in a more meaningful way. For example the “abv” column was before the “craft\_beer” column. Why would you need to know the abv before the name of the beer?
   2. So I performed a reindex function and ordered the columns by: “brewery\_id”, “craft\_beer”, “style”, “abv”, and “ounces”.

Merging Breweries.ipynb and Craft Beers.ipynb

1. I wanted to merge these two together in order to get a combined data frame. Furthermore, I wanted the data frame to show all of the craft beers and what brewery they were from.
   1. In order to do this I merged the two together with a “pd.merge” function in which it was merged on the “brewery\_id” column with an inner merge. Both files had a Brewery ID that related the two together.
2. Once they were merged I wanted the combined columns in the data frame to be reordered in a way that would make the most sense, with the most important information first and the least important at the end.
   1. I performed a reindex function and ordered the columns as, “brewery\_id”, “brewery”, “craft\_beer”, “style”, “abv”, “ounces”, “state”, “city”
3. Lastly, I looked at what I had and realized that the dataframe was ordered by the index and had all of the different beers and breweries in a random and clustered order. But I wanted all of the craft beers from each brewery to be in order and next to each other. So I performed a “.sort\_values” function. Which ordered the data frame in numerical order based on the “brewery\_id”.

Loading:

1. Loaded it through SQLAlchemy onto AWS.
2. I created two tables, ‘breweries’ and ‘craft\_beer’.
3. While trying to load the ‘craft\_beer’ table, I found out there were duplicates in the data, so I had to perform a “drop\_duplicates” function so that it could be loaded to SQL. Additionally, I had to make two primary keys for this table in order for it to work. I made the “brewery\_id” and “craft\_beer” as the two primary keys.
4. While loading ‘breweries’ table into SQL, the column “state” caused an error because “state” is a reserved word in SQL. So I had to go back and change the column name to “state\_”, which fixed it.

Brian Holt: Wineries

**Transformation of winery data:**

1. I worked from one CSV file from Kaggle. First thing I did was load it into my jupyter notebook. I used Pandas DataFrame to clean the table.
2. I deleted unnecessary columns that didn’t pertain to the data I wanted to show, ie. Taster name and taster twitter handle names.
3. I next wanted to show only wineries from the US and not the other countries. Christian’s data was only for craft breweries in the US.
4. After I did that I wanted to check to see how many rows I was left with to make sure the data still had plenty of rows. It did.
5. I next renamed the province columns name to state since I am only dealing withUnited States.
6. I next reordered the columns to show Title first to see the wine list first.
7. Next I wanted to get rid of all rows that had blank values. I ended up deleting 17,596 rows in columns--region 1, region 2, and designation.
8. Next I organized the states by alphabetical order.