Team: Allyson McInnis, Kayli Aguilera, Annie Donnelly, Liberty Heise

Coffee Quality Analysis

Tools Utilized:

- Pandas with Google Colab
- HTML/CSS/Bootstrap
- SQL for Loading into Tableau
- Tableau for tables

Things to do:

- ☐ Start an .ipynb with our dataset to ETL (Allyson)
  Subtract harvest year from grading year?
- ☐ Machine Learning Prediction of Coffee Quality (Annie)
- ☐ Load data to SQL, and upload data into Tableau for chart making (Liberty)
- ☐ Slides (Kayli)

Monday: notebook with ETL,
Wednesday:
Thursday: Begin SQL and Tableau to finalize

Monday: Finish ReadMe, GitHub and Colab to practice presentation

Requirements

Data Model Implementation (25 points)

- A Python script initializes, trains, and evaluates a model (10 points)
- The data is cleaned, normalized, and standardized prior to modeling (5 points)
- The model utilizes data retrieved from SQL or Spark (5 points)
- The model demonstrates meaningful predictive power at least 75% classification accuracy or 0.80 R-squared. (5 points)

Data Model Optimization (25 points)

- The model optimization and evaluation process showing iterative changes made to the model and the resulting changes in model performance is documented in either a CSV/Excel table or in the Python script itself (15 points)
- Overall model performance is printed or displayed at the end of the script (10 points)

GitHub Documentation (25 points)

- GitHub repository is free of unnecessary files and folders and has an appropriate .gitignore in use (10 points)
- The README is customized as a polished presentation of the content of the project (15 points)

Group Presentation (25 points)

- All group members speak during the presentation. (5 points)
- Content, transitions, and conclusions flow smoothly within any time restrictions. (5 points)

- The content is relevant to the project. (10 points)
- The presentation maintains audience interest. (5 points)

Grading

This project will be evaluated against the requirements and assigned a grade according to the following table:

Machine Learning

For the machine learning portion of the project, the csvs that were created during the ETL process were used. To be able to apply them to machine learning the data need to be converted into a usable form. The data points that were strings needed to be changed into integers. Once this was done the dataframes could be implemented to multiple models or.

The first model that was used was to see if the data could be clustered. It was found that the data could be clustered. By utilizing the elbow curve we found that it could be clustered into two groups and possibly three. This backed the thinking of using two groups, flavor and demographics. To simplify the machine learning, both groups were used together in one dataframe for most of the models.

To ensure that the dataset could be used for predictions a set of regressors were applied to it. We used regressors because our data is on a continuum rather than a binary scale. We found from the regressors that there appeared to be a good predictive data. As shown below.

When the regressors showed that the coffee dataset could possibly be predictive it was applied to a deep neural network. The deep neural network was used to see what predictions, or if we could make predictions based upon the data that we had. Many issues appeared, or more like one big issue, and that was our accuracy and losses were way out of proportion to the regressors. To sum it up our accuracy never got over 0. Despite changing the number of nodes, the number of layers, the activation, and the compile model loss the accuracy remained 0 and our losses were exponential. Even when breaking down our main data set into the clusters of flavor and demographic accuracy and loss remain the same. What was found after taking a deeper look at the data was that the data points were not independent of each other. In other words, the column that was being used for prediction was dependent on the sum of flavor columns.

ETL

Extract

First thing we did was download the database we found called “coffee_ratings.csv” which had the data qualities that we were searching for. We Extracted our data from the csv and imported it as a dataframe into jupyter notebook.

Transform

After that we made multiple data frames so that we can see the accuracy based on flavor profile and demographics specifically. We dropped all the nulls from both data frames so our data would be more concise. We then focused on demographic df and changed the integer that was in place for the dates used in the “grading_date” column to a datetime format; the column “grading_date” was also changed to a datetime year.

It was at this point when we were trying to narrow down some of our columns, did we see that there were some serious differences between the data in each column. The year sometimes had varying symbols in them that made it ununiform, as well as the elevation measurement types varied. The data was extracted after being cleaned for nulls and a quick hand-cleaning of data for anomalies in the columns was done. Other objects which would not have read into the dataframes properly along with columns that were deemed unnecessary were removed from the set. Finally we were able to upload the best version of the csv called “New_ETL.csv”.

In a last push to make our data more uniform, we made a “year_diff” column that subtracts the year that the coffee was graded (or tasted) from the year that coffee was actually harvested. We then made our new dataframes (demographic_df and flavor_profie_df) into new CSV’s so they would be ready to load into Postgresql.

Load

In our final steps, we uploaded our newly squeaky clean dataframes and concatinated them into one called coffee_df. We then pushed all of this to SQL and were able to see our new tables made within postgres.

Tableau + Visualization of Data

In order to understand more thoroughly what we were looking at in the numbers, the cleaned data was loaded into Tableau. Initially, characteristics like altitude and country of origin were graphed to see if there were any correlations but almost immediately it became apparent that something was wrong with our data. In this map, the sum of

coffee cupping incidents show strong representation in Central and South America. According to *World Population Review*, the countries with the highest coffee production are: Brazil, Vietnam, Colombia, India and Uganda. As one can see, most of these countries are well under-represented in this data set. We next look at the average cup point given, regardless origin. All coffee, regardless of Country of Origin received a rating higher than 80. The results finalized a suspicion that something was off with our data.

We created a graph that compared each of the flavor profiles against the total cup score and immediately saw that the flavor profiles mirrored the total cup score on each entry. This linear correlation confirmed our knowledge that the profiles were not independent from the total cup score. This was problematic.

Analysis and take away

- Our data was skewed towards collecting data on high ranked coffee from a variety of countries.
- Our data was linear, as Annie mentioned during the findings in the Machine Learning which means our data could have simply shown it's cards during multiple linear regressions.
- Our data was not complex enough to benefit from neural networks and machine learning.
- Linear Regression Analysis should always come first in order to determine if the data needs further Machine Learning analysis. Our scores for LR were very high, versus our scores in the ML, which were extremely low.
- A deeper look into the data is always important to gather general information and determine how the data interacts.
- Using different modes of analysis help inform and improve intuition and understanding of data.
- Next time, we would test for linear correlations first to ensure balanced data in this case, more variation in the cupping scores and more scores per country of origin.