Sammi and David

**South American Coffee Production and Starbucks Locations**

Our group selected three CSV files from Kaggle, (CoffeeProductionConsumption.csv, StarbucksLocations.csv, and CountryCodes.csv), to use as the basis of a new and joined dataset. While we had a CSV file for producers and exporters of coffee that contained the names of countries, we only had a CSV file for the locations of Starbucks stores that contained Alpha2 or two-digit country codes. Our third CSV file contained columns with both sets of data for countries that we would use as our Rosetta Stone to merge the two disparate sets of data. Since we wanted a challenge, we ultimately chose not to use the third CSV, and selected a website with an imbedded HTML table. Our choice of website presented greater challenges. We were not able to parse our data with read\_html(), and we then needed to install the html5lib module or dependency. Our second choice was also unable to parse the data of the website. We had received a 403 Forbidden Error with use of both the modules. After consultation of Stack Overflow, we decided to use urllib as our module, and were successful with our read of the data. We then used Beautiful Soup with the html.parser to convert the data (in bytes) to a Beautiful Soup object. We were then able, with the help of the instructor, to convert the Beautiful Soup object into an HTML file (CountryCodes.html) that we were able to read as a Pandas dataframe. Our work for the previous section is contained in the notebook, CountryAlpha2Notebook.ipynb. Within Pandas, within the ETLTransformNotebook.ipynb, we were able to remove and rename various columns. We then created tables in pgAdmin4 (CoffeeDatabase.sql) and established a connection between Pandas and pgAdmin4. With the exchange of data between the two programs, through use of the to\_sql() function, we created a SQL query with two join statements to form our complete and corrected dataset. We encountered yet another problem: pgAdmin4 would only display 1000 rows of the total dataset of 26,000 rows drawn from the StarbucksLocations.csv file. We then used the read\_sql\_query function to import the completed and corrected dataset from pgAdmin4 into Pandas, where we then performed two filters for Brazil and Bolivia.