ETL Project report for UWA Data Analytics Bootcamp Project 2 group 5

Participants:

1. **Solomon Dias**
2. **Victoria Giles**
3. **Vincent Gai**

Overview:

The project is about the consumption of coffee in countries which import and process the bean. The original data sources are as follows:

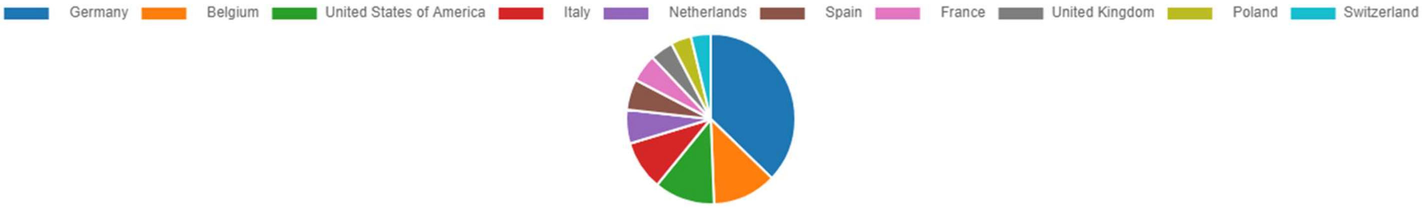
1. https://www.ico.org/new_historical.asp
2. https://www.kaggle.com/datasets/michals22/coffee-dataset?resource=download&select=Coffee_production.csv

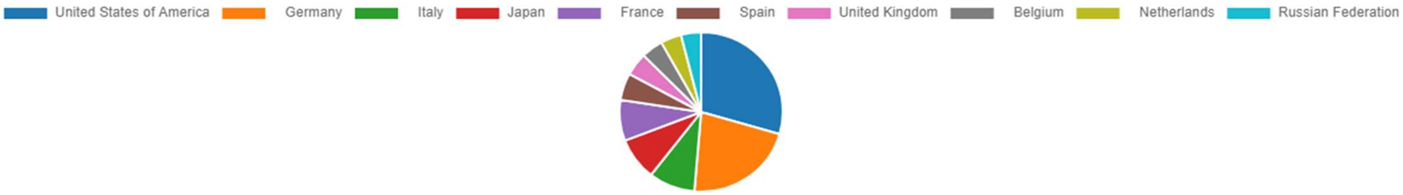
We used the Coffee_imports and coffee_re-export datasets to showcase the ETL process. We formatted the data using pandas and jupyter notebook. We used an inner join on the datasets and only selected the columns we were interested in. The datasets contained data on the imports and re- exports for the past 30 years from 1990 -2020. It also contained a column of the Total exports and the total imports.

Extracting and Transformation:

1. For the extraction of the data, we downloaded the datasets onto our local machines and then imported them into a Pandas DataFrame.
2. We then had to clean the dataset as there was a lot of unnecessary information for the past 30 years which was not needed for the analysis we wanted to perform.
3. After doing an inner join on the datasets using the 'Country' as a column we selected to columns we were interested in.
4. After the columns were selected, we used the 'set_index' function to set the index to country as this would make it easier to plot graphs and not confuse the reader.
5. A new column was then created with the total import's vs the re-export. This gave us the consumption for the past 30 years.
6. Using this new data, we calculated the percentage of consumption and created a new column.
7. The new DataFrame was then exported to .csv file ready to be uploaded to a database for further analysis.
8. The .csv file was then loaded into a SQL database. We chose was because the data was very structured in the .csv format.
9. An ERD was created displaying the relationship between each dataset coffee import and coffee export.
10. SQL PG-admin was used to create tables and load the original datasets.

Some images created using SQL:





Further Analysis:

1. We could have used another dataset and got information like the per-capita consumption vs export of coffee.
2. Also, the datasets could have contained more countries and data for further analysis. The ones provided here are very limited.
3. We could group the countries and create a world map to provide an interactive analysis of coffee consumption.