ETL Project Scope March 11, 2021

## Extract: your original data sources and how the data was formatted

Our data is being pulled from one CSV file and one XLXS file that we found on Kaggle and Dataworld. We used Jupyter Notebooks to clean the data.

winemag-data\_first150k\_for\_project.csv

1. We were able to extract our data using sqlalchemy to create an engine.
2. Then we convert the CSV into a DataFrame

Wines.xlsx:

1. Since the file we were cleaning up was an .xlxs file and not a CSV file, we install openpyxl
2. Then we used it as our engine to read the excel file.
3. Next, we convert the file into a DataFrame

## Transform: what data cleaning or transformation was required.

Steps for cleaning the winemag-data\_first150k\_for\_project.csv:

1. We create new DataFrame with the selected columns we needed.
2. Then we identified the incomplete rows and eliminated them.
3. We then evaluated the data to determine that the data was the appropriate type.
4. We renamed Pinot Grigio to match Pinot Gris in the other dataset.
5. We then eliminated all wines that were not included in the top 6 favorite wines per <https://artwinepreserver.com/pages/types-of-wine>.
6. Finally, we reindexed the DataFrame with the remaining top 6 wines that we selected

Steps for cleaning the Wines.xlsx:

1. We eliminated the Vintage, County and Designation columns.
2. We identified the incomplete rows and then eliminated them.
3. Next, we evaluated the data to determine if the data was the appropriate type. We noted that the Price was an object rather than a float. We then converted it to a float.
4. We then eliminated all wines that were not include in the top 6 favorite wines per the Art Wine Preserve website.
5. Finally, we reindexed the DataFrame with the remaining top 6 wines that we selected.

## Load: the final database, tables/collections, and why this was chosen.

1. We connected to the database.
2. We then added tables.
3. Next, we used pandas to load the CSV converted DataFrame into the SQL database.
4. Then we used pgAdmin to pull our tables in and confirmed the data was added.
5. Why we chose to use SQL
   1. We used SQL because it is a good balance when you are working with numbers and text.
   2. SLQ was also chosen because it is relational database where Mongo does not relate as easily.
   3. We chose it so we could get more formular with SLQ due to the fact more companies use it and we wanted to stay completive.
   4. We also used this because Fred told us too. 😉

Bonus:

Create an Html with links to each DataFrames to display the final results.