**ETL Project -Wines overview.**

As a new group we started talking and discovered we had in common the love for wine, our idea was to discover wine regions and their variety to be able to plan a wine trip.

We used [two Kaggle’s datasets](https://www.kaggle.com/zynicide/wine-reviews) to analyze wine regions, wineries, vineyards, wine varieties and their correspondent reviews around the world.

**The columns of interest:**

* Country
* Province
* Region
* Winery
* Vineyard
* Variety
* Title
* Points
* Price
* Review

**Files used:**

winemag-data150.csv

winemag-data130.csv

**Data Transformation**

In order to perform the desired analysis, we applied some data cleaning and transformation as follows:

1. Used Pandas to load csv files and store them as Data Frames
2. Created filtered Data Frames, removing columns such as region\_2, taste\_name, taster\_twitter\_handle that were not relevant for our analysis and rename columns for a better understanding
3. Merged both datasets, removing duplicated and blank values for region, price and vineyard columns.
4. Created catalogues with the correspondent id for country, province, region, winery, variety and vineyard.

**Data Load**

Finally, we transferred our cleaned and transformed dataset into a database. Creating a database with its respective tables to match the catalogues from Pandas’ Data Frames using Postgres and PgAdmin. Reconnected to database and created three-degree normalized tables, we chose this level of normalization to have more efficient data analysis.

**Summary**

Cleaning the data and transforming it into a way it was more suitable for our analysis was a real challenge, however; we could manage to clean the raw data and it was easier to normalize tables in SQL and have a good analysis of wine regions over the world and different types of wine.

Interfaz de usuario gráfica, Aplicación, Chat o mensaje de texto

Descripción generada automáticamente