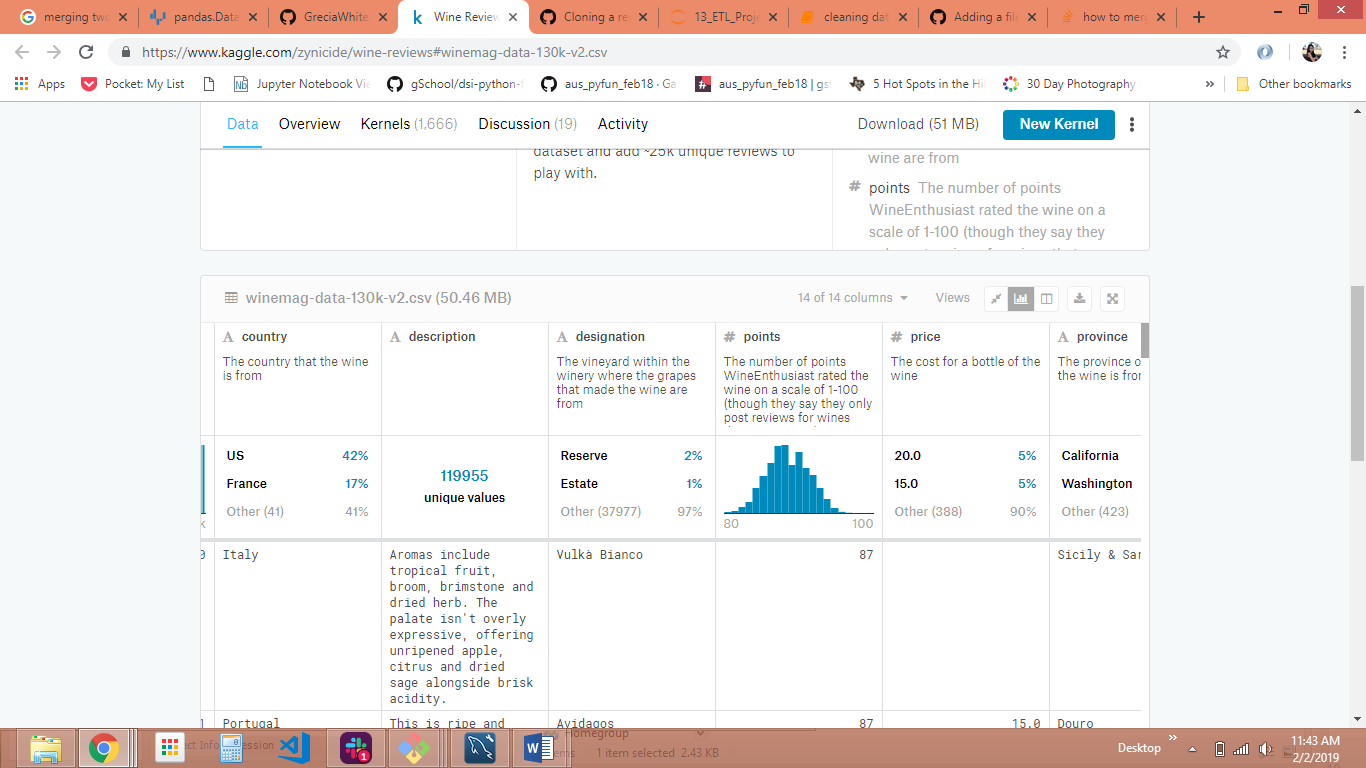
Wine Review Report

Sarah Marquardt, Alan Choate, Grecia White

Extracting Data

For this ETL project our group chose to focus on wine reviews and ratings. We extracted, transformed and loaded our data using Jupyter Notebook libraries and MySQL. We first downloaded two csv files from <https://www.kaggle.com/zynicide/wine-reviews#winemag-data-130k-v2.csv>, originally scraped from winemag.com. One file contained 130 columns x 14 rows and the second contained 151 columns x 11 rows of data.



Transforming Data

For the second part, both csv files were uploaded to Jupyter Notebook and manipulated with pandas library. In total, we imported pandas, pymysql, and from sqlalchemy we imported create\_engine. We converted the csv files into data frames and selected just the columns we wanted; description, points, country, province, variety, price. We only focused on the US so we created a new data frame containing only the US as a country. We also dropped all the empty rows, since some didn’t have a price. Finally, we concatenated our data frames into one and converted this to a single CSV file.

Loading Data

We used MySQL to create database with a single table with our chosen columns from our final csv file. We created an ID column as the primary key. Using the “to\_sql” command we loaded our database to MySQL. In our first attempt to load it we were given the error, “incorrect string value”. To fix this we had to convert the database and the table character set and collation to UTF-8 MB4.

Lessons learned

It’s helpful to have an order of operations beforehand to prevent duplicate efforts. Also, the version of MySQL matters because the older version doesn’t support the UTF-8 without the MB4 format. This could be due to the character count on one of our Description columns.