Chocolate Ratings and Weather

For our project, we found and were able to extract a data set in csv format from Kaggle which includes expert ratings of over 1,700 chocolate bars (datasets\_1919\_3310\_flavors\_of\_cacao file). Besides the ratings, the data set also includes data for other factors which may contribute to the rating of each chocolate bar, such as the company manufacturing the bar, the specific geo-region of origin for the bar, the manufacturer base country. We think the weather of the geo-regions where the cocoa beans were grown and harvested certainly have an impact on the chocolate flavor. Therefore, we decided to use the API to connect to and extract the weather data from <https://openweathermap.org/api> for the geo-regions given in the ‘datasets\_1919\_3310\_flavors\_of\_cacao’ file. The extracted weather data includes the latitude, longtitude, Max Temperature, Humidity, and Cloudiness of each regions.

To transform the data set into a nice and clean DataFrame, we dropped a few columns from the ‘datasets\_1919\_3310\_flavors\_of\_cacao’ file which includes data we think would not be valuable for the analysis, as well as renamed the column titles so it’s shorter and cleaner. We also removed the duplicated records. We transformed the extracted weather data into a nice DataFrame then saved it as a csv file.

Since the weather data set was extracted from API calls and we were unable to obtain the data for all the names listed in the “Geo\_Region\_for\_Bean” column of the ‘datasets\_1919\_3310\_flavors\_of\_cacao’ file due to region not found, we decided to use MongoDatabase to store our data sets for later analysis purpose. We named our database as “weather\_and\_chocolate” and stored our data sets in ‘chocolate\_rating’ and ‘weather’ collections.