Documentation of Decisions

We decided to convert the CSV files to Latin 1 because UTF-8 could not read certain symbols in the city names.

We selected only the columns we needed: the cities, the countries, and the yearly values from both; and the latitudes and longitudes, country codes, and populations from cities15000.csv.

We separated the Celsius and Fahrenheit temperatures, previously listed in one column, into separate columns to make them easier to read. We used the split function in Python to accomplish this. We then worked to remove the parentheses from the Fahrenheit column to make it look a bit better. We did this by deleting the current Fahrenheit column and replacing it with another which converted the values of the Celsius column to Fahrenheit. We did this because our attempts to use “str.replace” did not work.

We decided to get to work removing duplicates from the Average Temperature of Cities data frame. Once we tested for duplicates using the cities, country code, latitude, and longitude columns, we found that there were no duplicates. We repeated the process with the cities1500 data frame.

Cities1500 uses country codes instead of country names, and as we discovered several of these codes were missing. We sorted out all entries with a null country code value as well as other null values. We then went back to the Average Temperature of Cities data frame and searched for null values just to make sure that we didn’t miss anything. We decided to remove any entry that had null values for latitude, longitude, or population as we considered those values too important for what we wanted to accomplish with the project.

For the missing country codes, we attempted to replace the null values manually. We created a list out of the cities with null values that we named “null\_cities.” We very quickly ran into problems, however, and eventually decided that it would be better to drop those cities and move on. We used the “null\_cities” list to filter out and then drop the cities with null country code values.

We then returned to the “all\_df” data frame to ensure that all of the country codes were the proper, two-digit codes that we had decided to keep and not the three-digit codes for the sake of keeping our data uniform.

We then proceeded to load our data into a SQL database.

We belatedly realized that we needed to covert our column names to all lower case for them to work in SQL, so we proceeded to do that using the “.rename” function and the “str.lower” function.

In the process of preparing the panda data frames for SQL, we discovered a row where all of the information was off by one column, with no city name listed. We decided to drop that row.