**Project 2 - ETL-challenge**

Team 11 – Yu Zhou and Renee Sandidge

This project will evaluate a country’s happiness and relate it to its economic freedom. It will also evaluate the average temperature of the country.

* **Extract**

Data sets of interest. What makes a country happy? The data sets in this project investigate the “happiness” of a county, the economic freedom and the historical average weather temperature of the capital city.

Data was extracted from the following resources:

<https://www.kaggle.com/unsdsn/world-happiness>

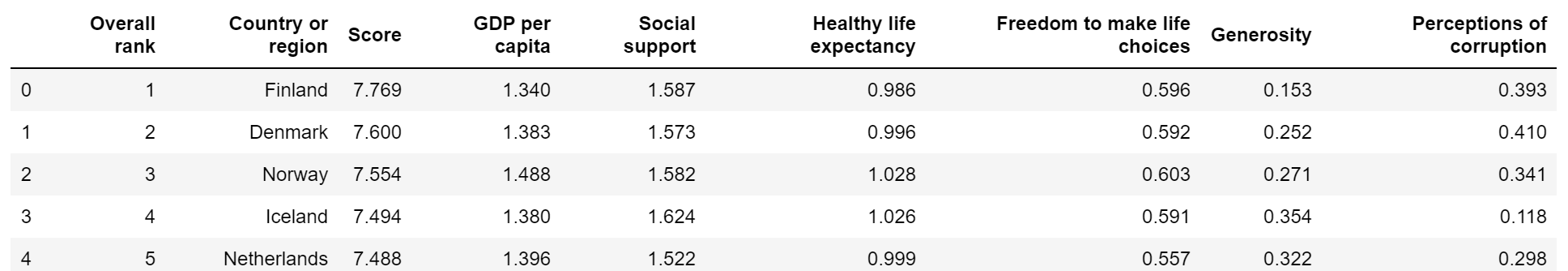
<https://www.heritage.org/index/about>

<https://geographyfieldwork.com/WorldCapitalCities.htm>

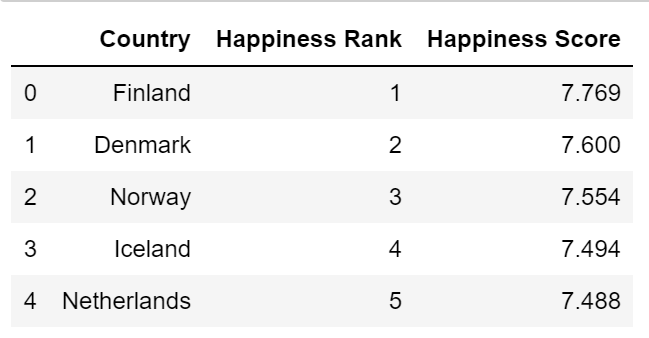
<https://openweathermap.org/api>

The websites provided data in a csv file types, html tables and API search query.

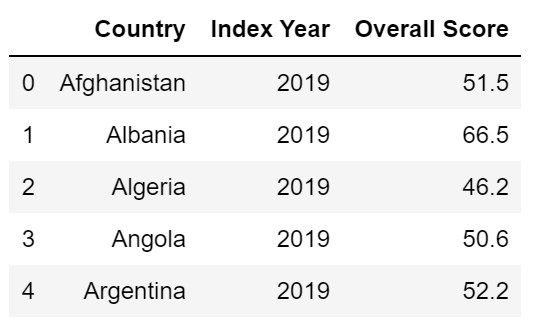
**Data Source 1:** The “Happiness Report” was extracted <https://www.kaggle.com/unsdsn/world-happiness> It consisted of a series of CSV files that were separated by year 2015 to 2020. It contained data for 156 countries. A few of the categories in happiness reports were GDP, healthy life expectancy, freedom, corruption, and generosity. We focused on overall score and rank.



Transformed data



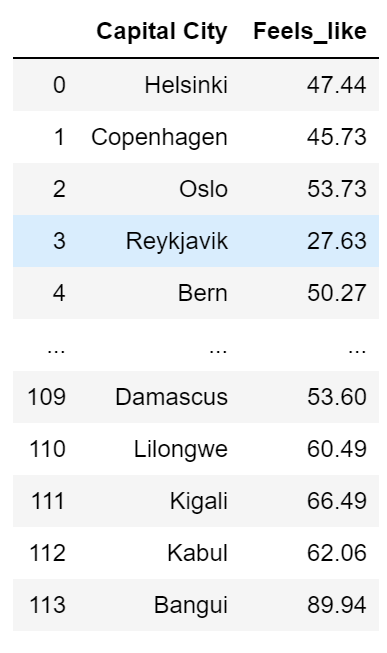
**Data Source 2:** The “Economic Freedom” report was pulled from <https://www.heritage.org/index/about?version=73> . This data was also in the form of csv file. The data contained over 187 countries and 5 years of data. Some of the categories in the data set were overall score, fiscal health, financial freedom and government spending. For this project we focused on year, country and overall score.



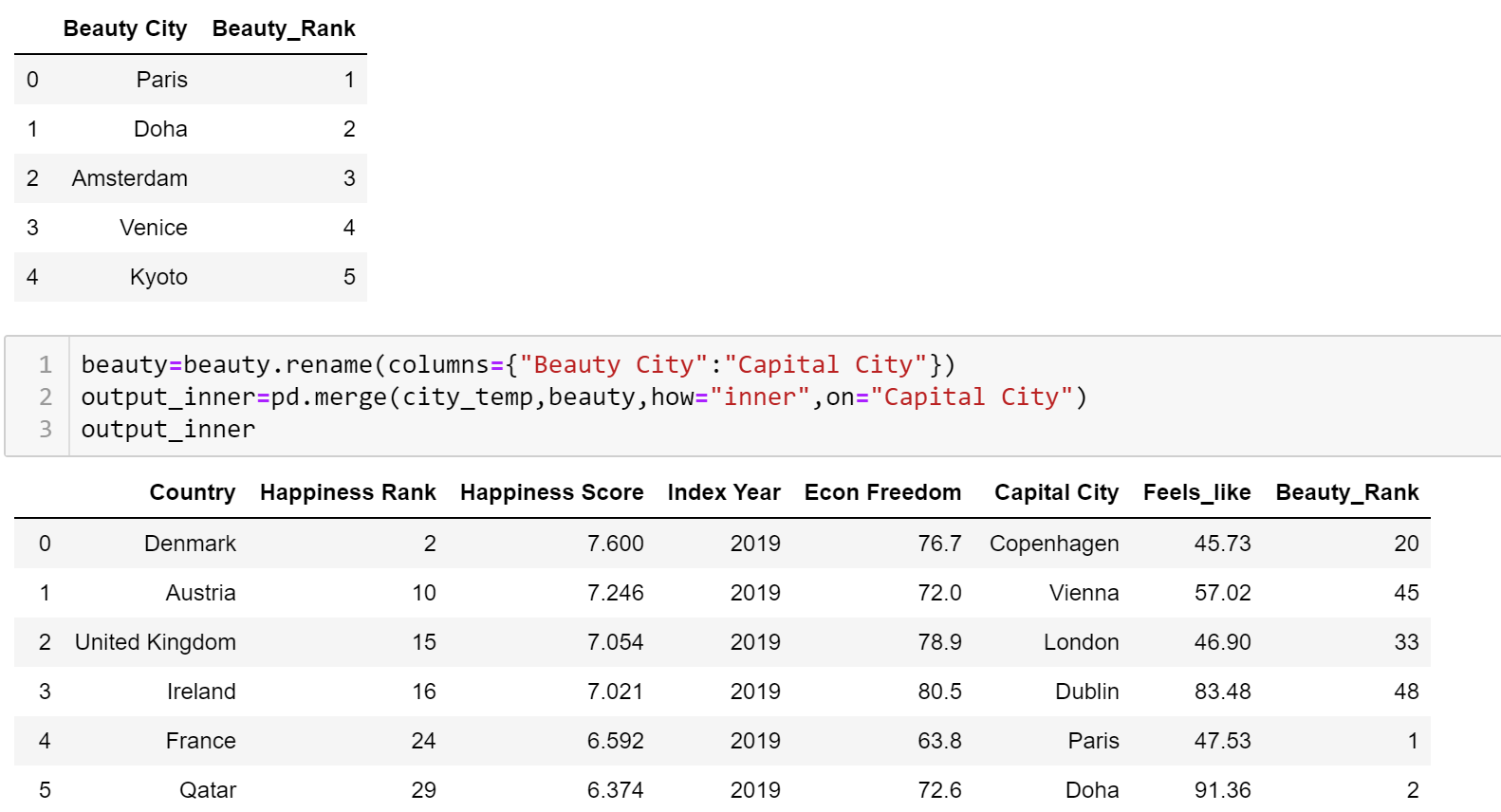
**Data Source 3:** Historical weather data was gathered from <https://openweathermap.org/api> using an api search.



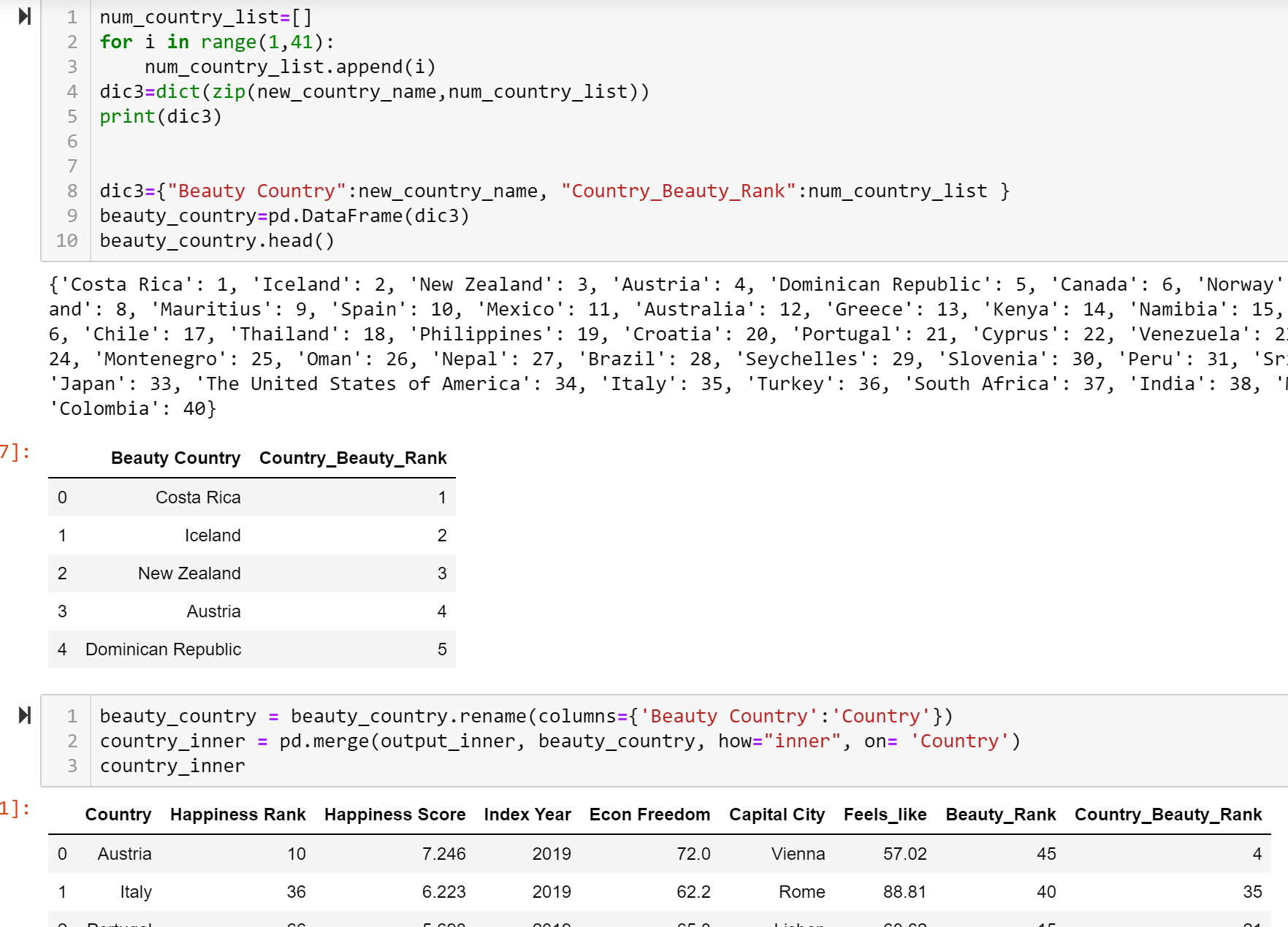
The search was conducted using city name. the capital city was needed for this search to work. A table was pulled in from <https://geographyfieldwork.com/WorldCapitalCities.htm> that provided the capital city of each country in the form of a html table.



**Data Source 4:** What is the most beautiful city in the world? This was found at <https://www.cntraveler.com/galleries/2016-01-08/the-50-most-beautiful-cities-in-the-world>. We used BeautifulSoup to capture and bring into Jupyter Notebook. A loop was used to obtain the city name. This was turned into a list and converted into a DataFrame. The DataFrame columns were changed to match the original DataFrame. We also assigned the ranking to the list. The two DataFrames were merged together using pd.merge, merging on “capital city”.



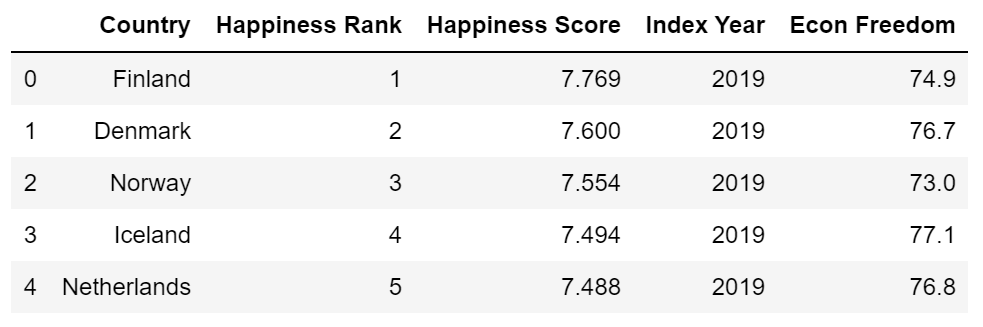
**Data Source 5:** What is the most beautiful country in the world? This was found at <https://www.cntraveler.com/gallery/most-beautiful-countries-in-the-world>. We used BeautifulSoup to capture and bring into Jupyter Notebook. A loop was used to obtain the country name. This was turned into a list and converted into a DataFrame. The DataFrame columns were changed to match the original DataFrame. The two DataFrames were merged together using pd.merge, merging on “country”.



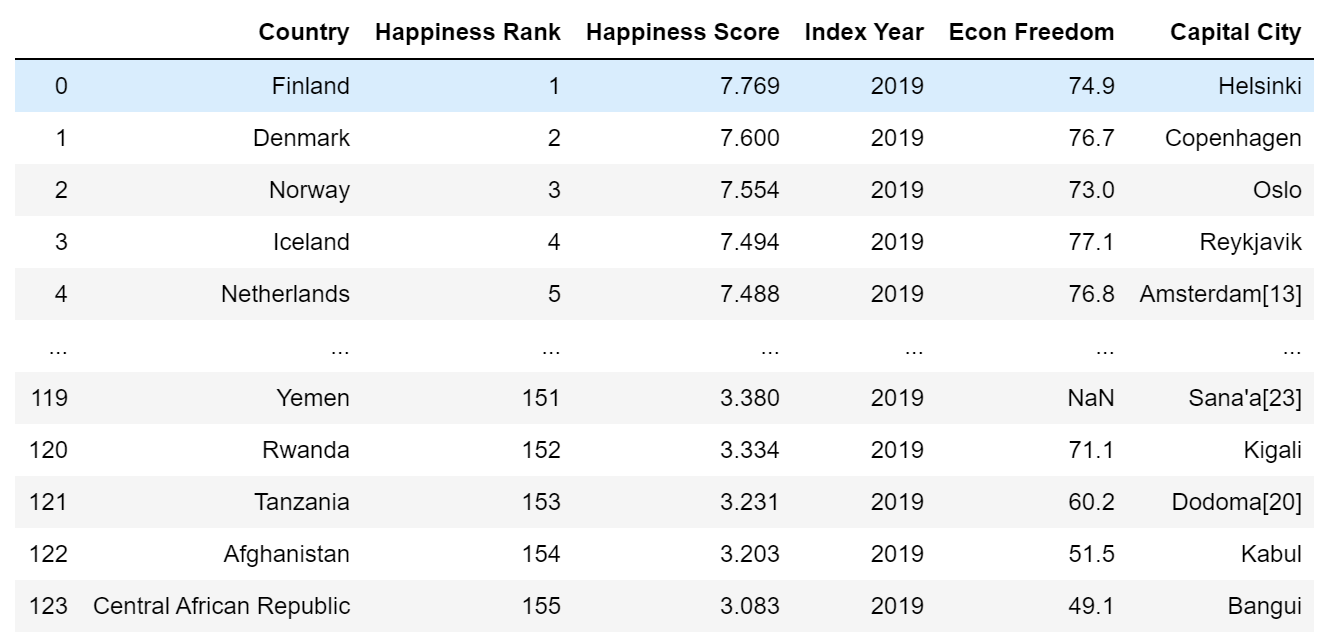
* **Transform**

Each csv file was imported into the Jupyter notebook as a separate file. Pandas pd.read\_csv was used to read each file. One learning from this import was the realization that columns names changed slightly with each year. The column names were updated to match. This made it possible to merge all the documents together into one data frame.

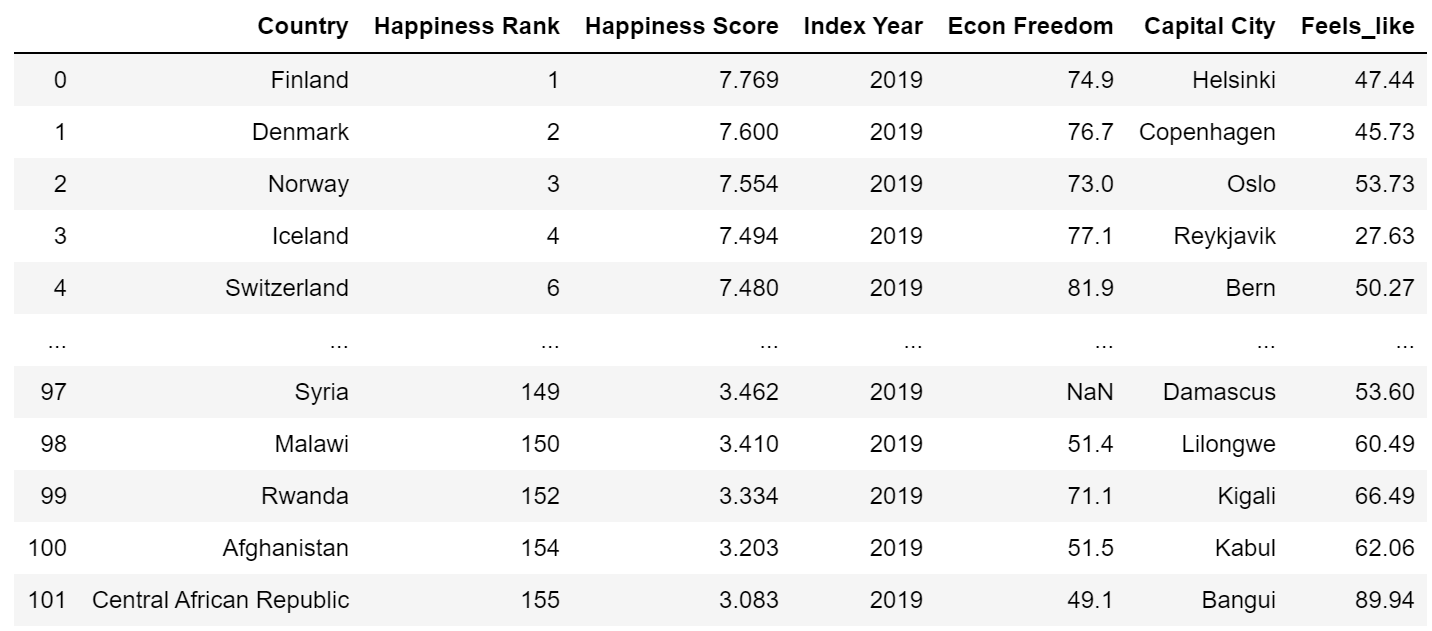
The economic freedom data was reduced by pulling in the columns that we found matched our happiness data. It was merged with the “Happiness Report” using pd.merge.



Another variable that was explored was how the weather of each of the countries in the “Happiness Report” affected the overall “Happiness Score” and the overall score from the economic freedom data. The first step was obtaining the capital city for each country. A table was pulled in from <https://geographyfieldwork.com/WorldCapitalCities.htm> that provided the capital city of each country in the form of a html table. The data was sorted by country then capital city.



Historical weather data was gathered from <https://openweathermap.org/api> using an api search. The data was merged into the Jupyter notebook.



The last two variables that were brought into the database were the most beautiful city and country in the world.

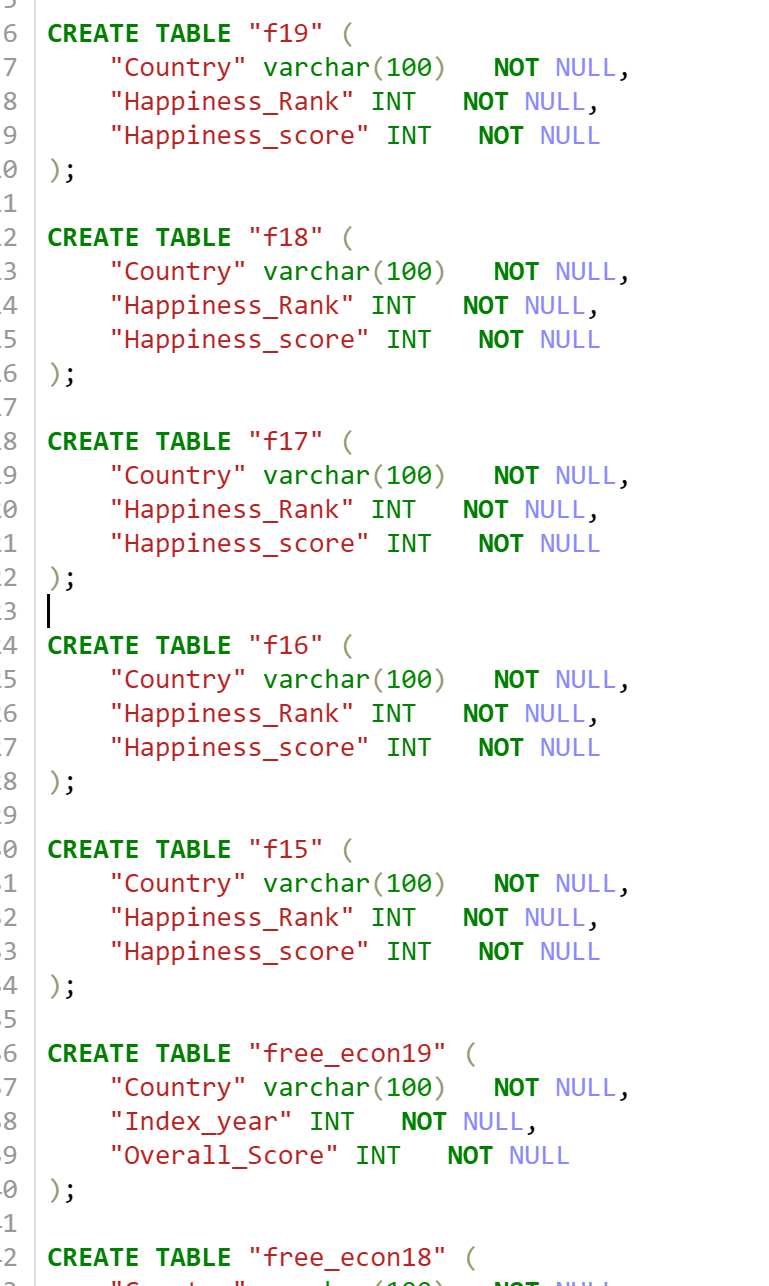
* **Load**

The last step was creating and loading everything into PGAdmin and creating a SQL database.

A screenshot of a computer

Description automatically generated

QuickDBD would only allow 10 tables.





**Finding Data**

Your project must use 2 or more sources of data. We recommend the following sites to use as sources of data:

* [data.world](https://data.world/)
* [Kaggle](https://www.kaggle.com/)

You can also use APIs or data scraped from the web. However, get approval from your instructor first. Again, there is only a week to complete this!

**Data Cleanup & Analysis**

Once you have identified your datasets, perform ETL on the data. Make sure to plan and document the following:

* The sources of data that you will extract from.
* The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc).
* The type of final production database to load the data into (relational or non-relational).
* The final tables or collections that will be used in the production database.

You will be required to submit a final technical report with the above information and steps required to reproduce your ETL process.

**Project Report**

At the end of the week, your team will submit a Final Report that describes the following:

* **E**xtract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).
* **T**ransform: what data cleaning or transformation was required.
* **L**oad: the final database, tables/collections, and why this was chosen.

Please upload the report to Github and submit a link to Bootcampspot.