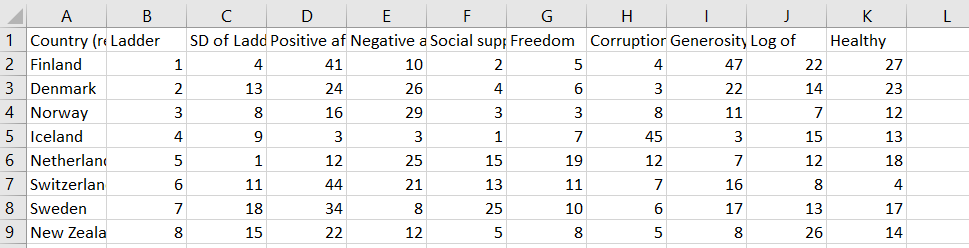
**ETL Project**

By: Marvin, Fahad, Lawson

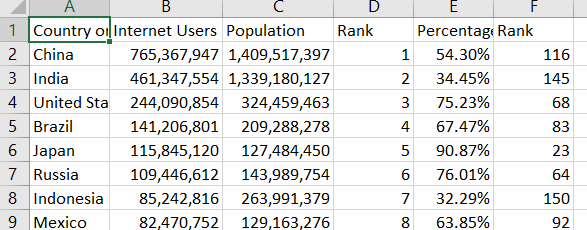
For this project we wanted to merge a happiness ranking with a broad-based technological advancement ranking to get a sense of which countries in the world were the happiest while also having being the most innovative technologically. Because we couldn’t find a direct ranking for technological advancement, we used internet usage per capita as an indirect proxy for technological advancement.

We started by **E**xtracting two data sets in the form of csv files. The first file (countries-by-internet-users.csv) which was obtained from Kaggle (<https://www.kaggle.com/tanuprabhu/list-of-countries-by-number-of-internet-users>) showed all of the countries in the world and had them sorted by the amount of people within that country that were internet users, while also showing the total population of each country and the percentage of the population of the country that were internet users. The second file (world-happiness-report-2019.csv) was obtained from data.world (<https://data.world/promptcloud/world-happiness-report-2019>) and it showed all of the countries in the world and sorted them by their Happiness Ranking, which was a combination of multiple other rankings including: Social Support, Freedom, Corruption, Generosity, GDP per Capita, and Health/Life Expectancy.

*Screen Clipping of Happiness Report csv*

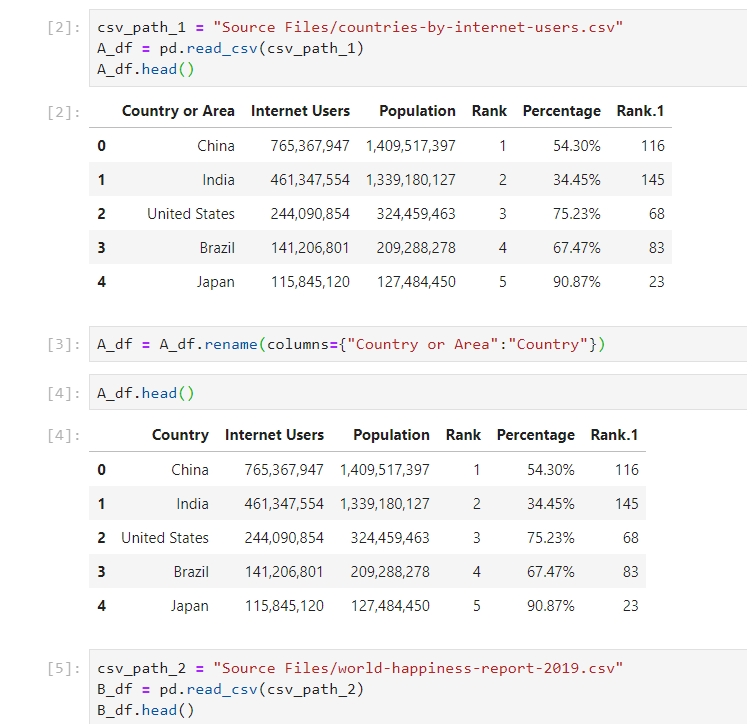


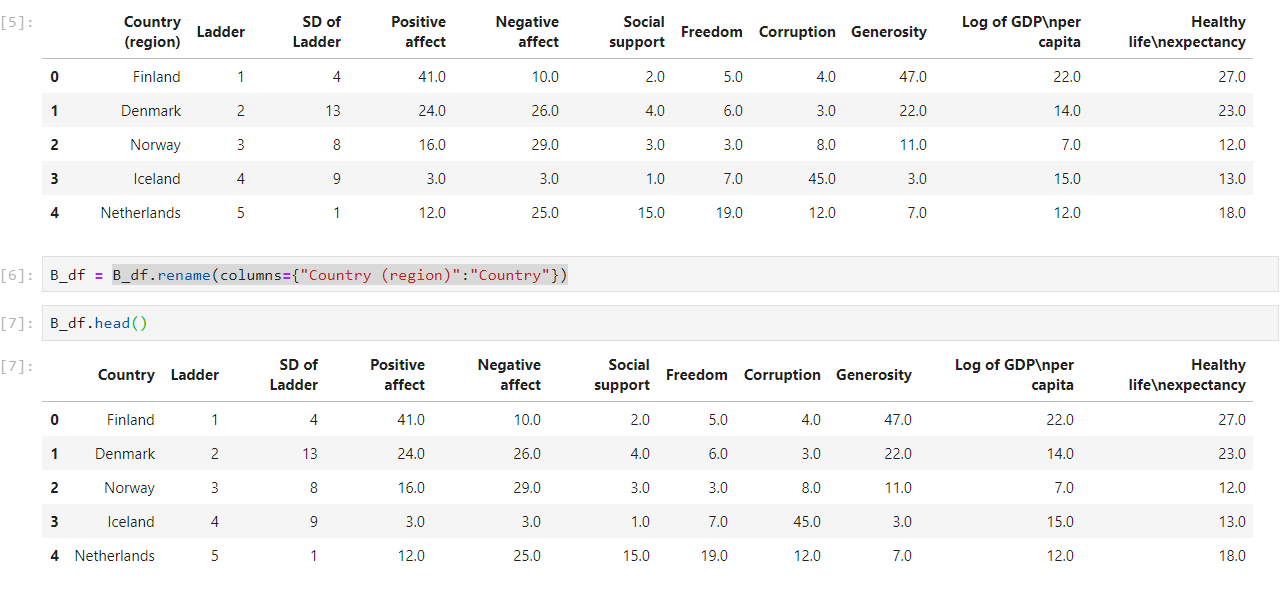
*Screen Clipping of Internet Usage csv*



After **E**xtracting the two datasets we stored them in our folders and began running code to **T**ransform the two datasets into something we could merge.

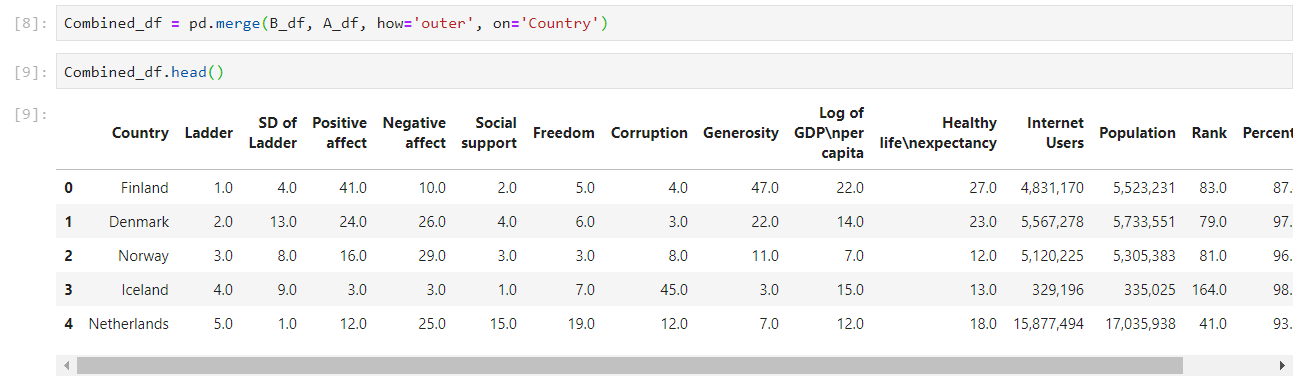
We first created file paths for the data sets so that it could be read by our Python code:



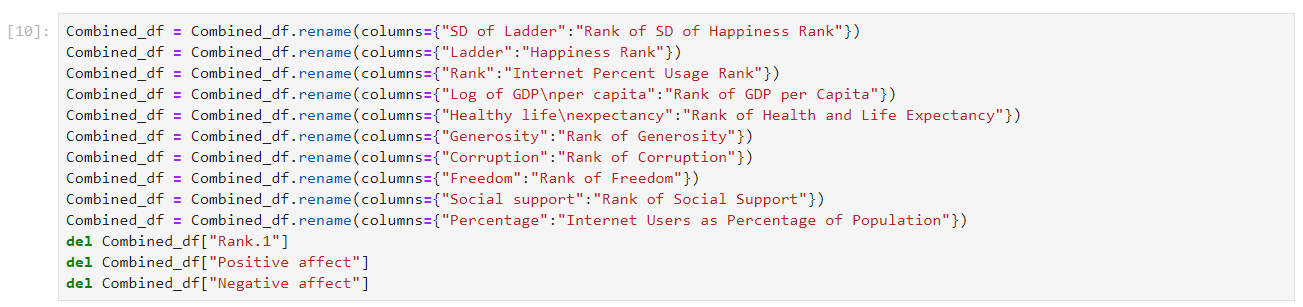


After create the file paths we changed the name of the index column for both files to Country as the Country names in each file would be the independent variables through which we would merge the files and all other dependent variables would be based off of (other columns in the files).

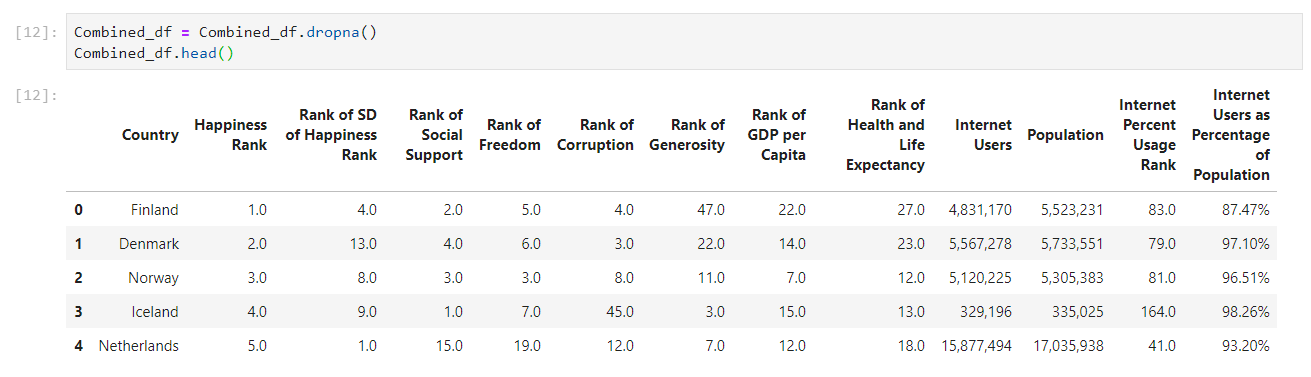
Once the index column title was made to match for both files, we merged the data sets into one Combined data frame.



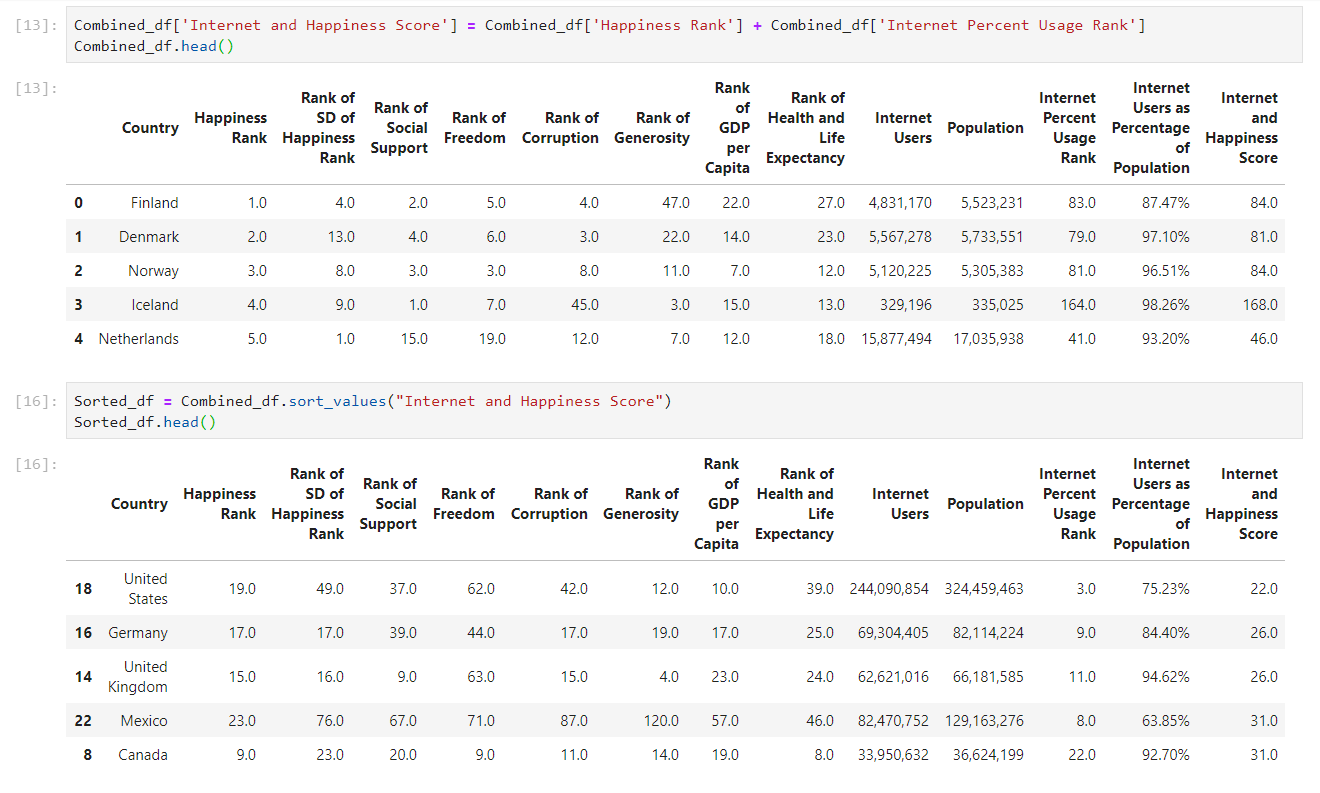
Once merged we undertook the painstaking task of renaming the columns of the Combined data frame into something cleaner and easier to understand, while also deleting columns that we deemed unnecessary.



After that we deleted all rows with multiple N/A values as it relates to happiness rankings as the two datasets had a different number of countries, resulting in the mismatched values not having a happiness rank.



Once that was complete, we wanted to create a new Column named Internet and Happiness Score, which was an equally weighted combination of the Happiness Rank and Internet Percent Usage Rank. This combined score would give us an indication of which countries were happy and broadly technologically advanced. When the score was created, we then sorted the entire dataset by the new metric.



In the last phase of our **T**ransformation process we reordered the columns to put our new score which sorted the combined data frame right next to the Country column, while also reordering the column in terms of importance to the data frame from left to right.



Once the **T**ransformation of our data was complete, we began the **L**oading process. We initially created a data base connection with PGAdmin and then we confirmed the tables and loaded the data frames into the data base we created.



We also ran a few querieson PGAdmin to get familiar with the application. Selecting parts of the Table that we deemed relevant and visualizing the combined data set in different ways (As described in the code below and its table outputs).





