## ---------------------------------Finding Data-------------------------------

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

Kaggle:
https://www.kaggle.com/faressayah/data-visualization-seaborn-matplotlib-tutorial/comments

https://www.kaggle.com/mathurinache/world-happiness-report

## -------------------------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.

Project Proposal:

The question driving our ETL project will be…..Do countries with lower Happiness Index scores tend to consume more alcohol?

Our hypothesis is that less-happy countries will consume more alcohol, on average.

Extract:

- Our original two data sources will both be found on Kaggle. The first is a 2020 combined survey from the World Happiness Report. It is a csv that includes summary statistical data for each countries' population's perceived happiness, based on a 1-10 scale. This dataset also includes social data such as; Logged GDP per capita, social support, health life expectancy, and perceived corruption. The second dataset shows what we believe to be the number of alcoholic beverages consumed in a year on average by residents of each country who choose to consume alcohol.

Transform:

- Both datasets will need to be cleaned before they can be combined into a final database. While both datasets can be joined by country name, they have different regional/continental indicators that will need to be matched accordingly. The world happiness dataset also has a large number of superfluous columns that must be dropped. Some of the columns in this dataset should also be renamed to provide a clearer understanding of what they look to portray. Thankfully, the Drinks by Country dataset is more concise and requires no manipulation at this time. These functions will be performed via Pandas/Python on a Jupyter Notebook.

Load:

- Once the datasets are cleaned to our liking the data from the clean csvs will be pushed into pre-generated SQL tables within the same SQL database. We feel that Object Relational Mapping is the database format that best suits the type of data we are transforming. This is because the ORM tool automatically generates the data access code we need to write. Therefore, if later on we were to create an application that interacts with this data model, the amount of code we would need to write would be drastically reduced. Finally, the two tables will be joined on their primary key, which in this case will be 'country_name', producing our final product. A two sample t-test may be run on the data to determine whether or not our hypothesis was correct. A box-and-whisker plot based on continent for both Happiness Index and Alcohol Consumption may also be interesting to visualize.