**Project 2 Group 5: ETL Project**

**World Health & Happiness Reports**

**Data Sources:**

* World Happiness Report "2019\_happiness.csv" <https://www.kaggle.com/unsdsn/world-happiness>
* Alcohol & Substance Abuse "alcoholSubstanceAbuse.csv" <https://www.kaggle.com/utkarshxy/who-worldhealth-statistics-2020-complete>
* Tobacco Use for Ages 15+ "tobaccoAge15.csv" <https://www.kaggle.com/utkarshxy/who-worldhealth-statistics-2020-complete>

**Proposal:**

We chose the the World Happiness Report as our inital data file. The fields of interest are the country, happiness score, and GDP of the country. In the alcohol substance abuse report and the tobacco use report, the fields are gender and total consumption. In the data extraction portion, we will load in the csv files using Pandas. For data transformation, we will extract the columns of interest to our ETL project. From the transformed DataFrames, we will create a relational database connecting the different data sources using countries as the key. For data loading, we will create a connection to the database and append dataframes to tables. This database could provide insight to how statistics such as alcohol, tobacco, and susbtance abuse affect the different outcomes of the happiness report.

**Members**

* Rebecca Kwon: database setup, load tables
* Mary Olaitan: happiness csv extraction + transformation
* Neil Tipton: alcohol&substance csv extraction + transformation
* Daniel Garza: tobacco15+ csv extraction + transformation
* Kristen Wilder: technical report