**ETL Project**

**Matt Krenicki, Nina Anderson, Mir Ahmed, Dan Zanoria**

**Project Background and Need:**

* Pharma company is looking at commercial planning for new drug launch in antidepressant category. They want to better understand population behavior on country-by-country basis

**Objective:**

* Extract data from multiple sources (below), transform and load the data so it can be easily accessed / queried by the commercial planning organization

**Data Sources:**

* World Happiness Index: <https://www.kaggle.com/unsdsn/world-happiness>
* WHO Substance Abuse: <https://www.kaggle.com/utkarshxy/who-worldhealth-statistics-2020-complete?select=alcoholSubstanceAbuse.csv>
* List of Countries by Antidepressant Consumption: <https://www.kaggle.com/mathurinache/list-of-countries-by-antidepressant-consumption>

**Practical Question:**

* Currently, the company has disjointed data sets
* Company is asking for a common consolidated view of countries and their happiness, substance abuse, and antidepressant consumption data

**Approach:**

1. Locate source data
2. Extract data into files / DataFrame
3. Study/review the data
4. Clean data individually
5. Use common value (country) to merge the data sets
6. Upload the merged data into a database using at least one of the following:
   1. PostgresSQL
   2. Sqllite
   3. Mongo
7. Provide a toolset and instructions to be able to query the data loaded