GROUP 3 PROPOSAL:  
Participants: Mabel Alamu, Arthur An, Dainty Cuevas, Sonny Maz

**DATA SETS:**

1. We’re using this 2020 alcohol substance abuse data from Kaggle: <https://www.kaggle.com/utkarshxy/who-worldhealth-statistics-2020-complete?select=alcoholSubstanceAbuse.csv>
2. And this Wiki list of countries by GDP per capita data from Kaggle: <https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)_per_capita>

**EXTRACT:**

* We are going to download the first file as CSV (Arthur, 10 minutes)  
  Then we will scrape the second data source and extract country and per capita GDP. (Group, 45 minutes)
* CHALLENGE: We'll use pandas (read\_csv) to import the first CSV file (Sonny, 20 minutes), and use MongoDB to scrape the second data source.

**TRANSFORM**

* Once imported, we will join the tables on country. (Dainty: 15 min)
* Rename all columns based on the indication from the first data source and drop the rows that we did not find meaningful (i.e. $US, nan).

**LOAD**

* We will do this as a group activity and load the cleansed dataframe into Postgres database. (Mabel and group, 45 minutes)

**Notes:**  
The activity took the span of the class on Wednesday 1/27/21; we plan to use extra time on

Saturday to explore the data further. Project to be complete Wednesday night.

**Web link:**

[**World Health Statistics 2020|Complete|Geo-Analysis**](https://www.kaggle.com/utkarshxy/who-worldhealth-statistics-2020-complete?select=alcoholSubstanceAbuse.csv)

Complete dataset with brief explanation.

[Wikipedia](https://en.wikipedia.org/)[**Wikipedia**](https://en.wikipedia.org/)

[**List of countries by GDP (nominal) per capita**](https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)_per_capita)