**Project Report**

**Team Alcohol**

**Team Members:**

Johar Najmi

Will Seymour

Justin Thomas

**Extract:**

The team set out to take freely available datasets on the consumption of alcohol by country around the world along with additional data on alcohol use in countries such as drinking age and legal prohibitions on the use of alcohol and enable analysis of the health impacts of alcohol use by combining this data with statistics on alcohol related illnesses and other alcohol linked public health issues such as traffic incidents and alcohol related crime.

**Data Sources for Extraction:**

Drinking Age By Country Population. (2019-10-24). Retrieved 2019-12-14, from <http://worldpopulationreview.com/countries/drinking-age-by-country/>

* The data was obtained from this source in .csv format

FiveThirtyEight. “FiveThirtyEight Alcohol Consumption Dataset.” *Kaggle*, 26 Apr. 2019, <https://www.kaggle.com/fivethirtyeight/fivethirtyeight-alcohol-consumption-dataset>.

* The data was obtained from this source in .csv format

World Health Organization Global Health Indicators

<https://www.who.int/data/gho/data/indicators>

* The data was obtained from this source in .csv format

**Transform:**

All data was read in from csv files into Panda dataframes which worked well for this tabular data. For the Drinking age and FivethirtyEight datasets specific columns of interest were selected, column names were changed to be more descriptive and to ensure that the common/primary key column “country” was consistently named. The World Population Review and FiveThirtyEight alcohol consumption data frames were merged on the “country” column.

For the World Health Organization data multiple CSV Files were imported and merged The tables were joined on Location, "Dim1", and Year (Dim1) which is a column denoting if indicator is refers to males, females, or both sexes. The “location” column was also later renamed as “country” to match the alcohol demographic and consumption data tables during import to the database.

**Load:**

As the data was all tabular and initially relational based on related country columns in all data sets it was decided to use a local Postgres relational database to store the data for future analysis. An Alcohol database was crated by the Postgres Admin tool in the database. The sqlalchemy library was then used to insert two tables one called “alcohol” with the merged alcohol consumption and statistics data and another called “who\_data” with the joined World Health Organization data. The SqlAlchemy .to\_sql function was used to directly create the tables in the selected database.

**Overview of Github:**

**Root Folder**

* Project Report Word Doc
* World Health Organization Extract and Transform Notebook

**ELT Project Johars Final Copy**

* FiveThirtyEight and World Population Review Extract and Transform Notebook with some initial analysis and visualizations

**Export to Database**

* Jupyter Notebook with sqlalchemy code to load dataframes from all data sets and upload to local Postgres database Alchohol Database and related tables

**References for WHO Data (uses technical/health column names)**

**List of Columns and definition**

* Country
* Period - Year of observation
* Dim1 - Sex
* AttFracAllDeaths - Estimate of the proportion of all deaths attributal to alcohol for that country/year/sex
* AttFracCancerDeaths - Estimate of the proportion of all \*CANCER\* deaths attributable to alcohol for that country/year/sex
* AttFracLiverCirr - Estimate of the proportion of all cases of **liver cirrhosis** attributable to alcohol for that country/year/sex
* AttFracYLL - A score from 1 to 5, based on the percent of total YLL (Years of Life Lost) that can be attributable to alcohol, where 1 was the lowest percentage and 5 was the highest percentage. Total YLL is the estimate of the years of potential life lost due to premature deaths. YLL takes into account the age at which deaths occur, giving greater weight to deaths at a younger age and lower weight to deaths at older age.
* Dependence12MoPrev - Estimate of the percentage of adults who had alcohol dependence for that year (2016)
* DrinkersOnlyLitersAlc - Estimate of the equivalent amount of pure alcohol consumed in a year (2016) measured in liters, looking only at people who drink alcohol. (Will's note: a significant portion of any population will not drink any alcohol so this is useful to compare to the total per capita alcohol consumption stat)
* DrinkersOnlyCI - 95% confidence interval of DrinkersOnlyLitersAlc
* HarmfulUse12MoPrev - Estimate of the percentage of adults who had drank alcohol to the point of being harmful for that year (2016)
* UseDisorder12 - Estimate of the percentage of adults who had alcohol use disorder for that year (2016) (will's note: this is can be read as "the sum percentage of people with alcohol dependence or harmful use")
* TrafficCrash100k - Estimate of the proportion of traffic accidents per 100,000 population cause by alcohol
* TrafficCrashPct - Estimate of the percentage of all traffic accidents caused by alcohol
* GeometryCode - Country's code for mapping packages (GIS)
* CrimePct - Percent of crimes due caused by alcohol (e.g. DUI)