**Project Title: Alcohol & Mortality**

**Name of Group: Group 1**

Team Members:

Heather Mott

Rosali Gonzalez

Saleha Ahmed

Steven Rodriguez

Tinu Adepoju

Walter DePalma

**Team Name: Data Crushers**

**Project Description/Outline:**

Our group intends to analyze mortality rates by state and compare with the alcohol consumption rates in states for several different causes of death.

**Data sources:**

Mortality rate by year for county, state from various causes of death.

<https://www.kaggle.com/IHME/us-countylevel-mortality>

Alcohol consumption by state and type of beverage.

<https://pubs.niaaa.nih.gov/publications/surveillance117/pcyr1970-2019.txt>

**Story:**

Is there a correlation between consumption of ethanol and mortality rates ?

Available data:

Consumption of alcohol per capita in each of 50 states + DC from 1970 - 2019

Mortality rates by state and county for 21 categories in 8 different years from 1980  - 2014

Data Cleaning

:

:

:

Data Analysis

Trend in Alcoholic consumption:

* Map  of total consumption all years combined ?
* Box plot of consumption by year.

Outliers ?  Can we identify outlier states ?

* Overall correlations (All morts and all years) -- >weak!
  + Scatter plot / regression / correlation coefficient

Table of significant correlations (mort/year combinations with P < 0.05):

Positive correlation – 1) Cirrhosis and 2) Mental\_Substance abuse occupy 9 of the top ten list

~0.50 correlation coefficient over 1 decade for Cirrhosis.

Respiratory rate significant in 1980 only..

In 1980 maybe association between drinking and smoking which no longer holds (speculation).

Follow-up question ..if correlation corresponds with ban on smoking in restaurants, bars..->no time to explore.

Neg- correlations – 1) Diabetes and 2) Cardiovascular diseases seem to benefit from more drinking (!?!?)

No strong positive correlations over same years to balance out negative correlation (drinkers dying of cirrhosis first would drive down death by other causes )

* Scatter / regression of strongest correlations on single strongest year.

* Scatter regression over all years combined for same mort categories showing strongest correlation at least 1 year.
* Correlation for Cirrhosis holds over all years combined.
* Correlation does not hold for Mental\_Substance Abuse over all years combined.
* Correlation (negative) does not hold for Cardiovascular Disease for all years combined.

Changes in environment, other factors..

Volume of alcohol consumed - maybe correlation only holds above a certain volume ?

* Map of Strongest mortality category over map of alcohol consumption for year of strongest correlation.

Does location contribute ?

**Conclusions:**

(some ideas..)

* General correlation over all years is weaker than individual years
* Correlation is moderately strong and positive for several mortality categories
* Correlations between consumption of ethanol and mortality rates associated with Cirrhosis and Mental\_substance abuse is moderately strong over multiple years.
* Ethanol consumption in the US has changed over the 24 years analyized - maybe weakening correlations to mortality rates.
* Negative correlations also found with cardio and diabetes… ?
* Next question might be to find out if correlations are driven by lower or higher consumption levels. (?)
  + Separate heavy drinking from moderate drinking…( maybe next project..)

-------------------------------------------------------------------------------------------------------------------------

**Data cleaning steps:**

* + Clean txt file of instructions
  + Read csv into dataframe
  + Divide alcohol consumption per capita by 10000 to have average gallons of alcohol per person.
  + Remove spreadsheet title rows
  + Read excel sheets into individual dataframes
  + Create csv files for beverage codes, and FIPS code
  + Read FIPS location and beverage type into dataframes
  + Merge nih alcohol df with beverage type and FIPS location dfs
  + Filter (.loc) mortality df for statewide rates only (drop county data)
  + Split mortality rate by  ' '  to get mean mortality rate and not '( confidence limits)' for each year.

**Data Analysis steps:**

* Map of total ethanol consumption by state
  + \_\_\_\_\_\_
* List of top correlations mort category and year.
  + Create dictionary of mortality rate dataframes - originally 21 sheets
  + Create dictionary of dataframes for NIH alcohol consumption data from years of interest.
  + Loop through each year of alcohol consumption within a loop of each mortality df to calculate pearsons correlation for each combination (mortality category / year)
  + Append to new dataframe each statistically significant correlation (mort category and year) for P < 0.05
  + Sort dataframe of significant correlations by correlation coefficient.
  + Review table to find strongest positive and negative correlations.
* Regression for strongest pos and neg correlation examples
* Strongest years and also combined ?
* Trend in alcohol consumption country wide
* Map of mort for strongest correlation