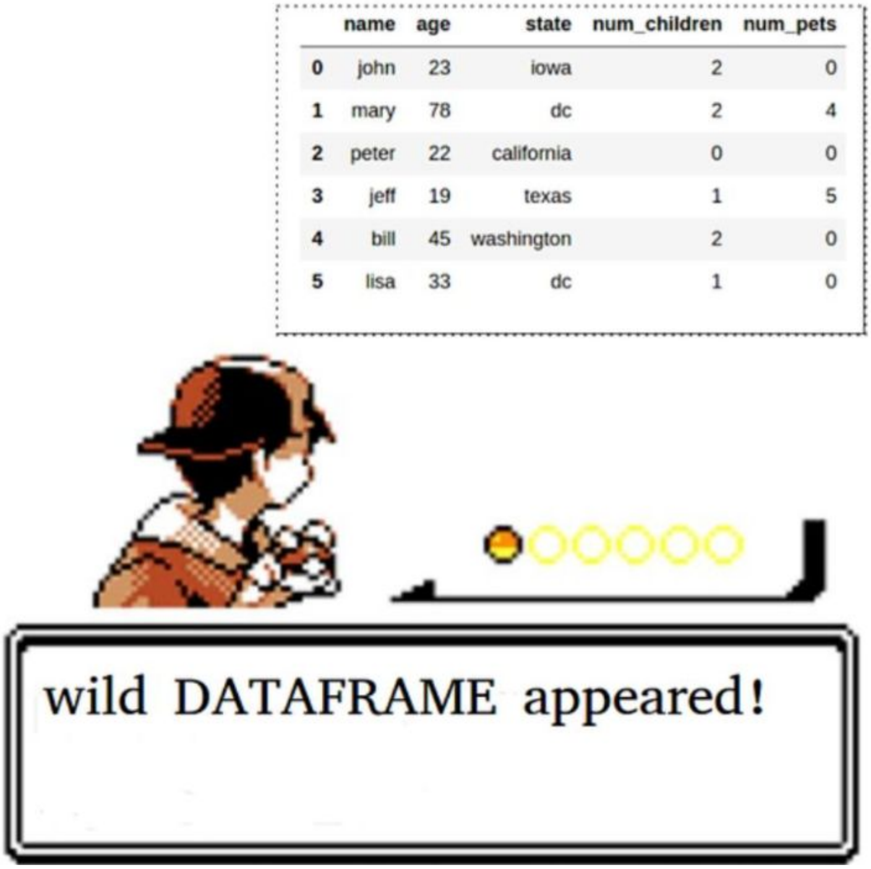
# Drug Consumption Risk based on Personality and Socio-Economic Status



| | name | age | state | num_children | num_pets |
|---|---|---|---|---|---|
| 0 | john | 23 | iowa | 2 | 0 |
| 1 | mary | 78 | dc | 2 | 4 |
| 2 | peter | 22 | california | 0 | 0 |
| 3 | jeff | 19 | texas | 1 | 5 |
| 4 | bill | 45 | washington | 2 | 0 |
| 5 | lisa | 33 | dc | 1 | 0 |

Emma Ritcey / 15er21@queensu.ca / 20007918
Rylen Sampson / 16ras6@queensu.ca / 20051918

## Background and Motivation

With the ability for individuals to obtain access to both legal and illegal drugs becoming easier in today’s age, the negative side effects of recreational use will become more apparent. One of these side effects which poses a dangerous threat to individuals is substance misuse. Previous attempts to tackle substance misuse come from a clinical background where the use of collected data is not highly emphasized or used appropriately to deliver insightful results.

The motivation for this project is to approach this issue using data analytics and machine learning models trained on data quantified from personality measurement tests conducted on numerous individuals and other descriptors pertaining to each individual. This would further help those studying and tackling the issue of substance abuse with their research.

## Description

The goal of this project is to determine the existence of indicators of drug consumption risk in the form of personality and socio-economic characteristics. The question to be answered is whether a specific type of person is more likely to participate in drug use, and whether there is a specific drug this person is more likely to consume.

## Direction of Solution

To determine whether there exist groups of people who are at a higher risk of becoming users of a specific type of drug, clustering will be used. Various different clustering algorithms will be experimented with, such as k-means and DBSCAN, to determine whether distinct groups exist. The data will also be separated and clustered by type of drug, to analyze if there are drugs that are easier to determine whether someone is at risk of using it.

Feature selection will also be used to determine whether subsets of the attributes carry more predictive power than others, as well as whether some of the attributes in the dataset hold very little predictive power and can be removed.

Further machine learning and data analytics methods may be used as we are inspired throughout the semester.

## Data Sources

Our primary source for data for this project is the “Drug Consumption (quantified) Data Set” found on University of California, Irvine’s Centre for Machine Learning and Intelligent Systems data repository website. Further details about the dataset are listed there as well as examples of the different attributes.

- https://archive.ics.uci.edu/ml/datasets/Drug+consumption+(quantified)#

As the project continues to move along, further data sources may be sought out for additional information related to the topic.

## Performance Metrics

Clustering performance metrics for our project include:

- Similarity between objects in the same cluster
- Accuracy to predict volatile substance abuse
- Mutual Information based scores
- Calinski-Harabasz Index
- The level of distinction between clusters

If regression methods are employed then root mean square error will be used to evaluate its accuracy. This may be necessary for predicting an individual's level of usage in a past time period or the models confidence they will use again in the future. The latter model measurement would be useful for researchers gauging if an intervention program was effective or not.