**Notebook 22\_09\_04\_data\_wrangling:**

2) Data Wrangling

2.1) Load data and start explore

Data has extension .data; but this link is not very useful. But it says we can use Excel to open.

[DATA File Extension - What is .data and how to open? - ReviverSoft](https://www.reviversoft.com/file-extensions/data)

- Actually we can read from URL using .read\_csv()

* After reading data description from the website, it is not clear the features are continuos or categorical. Let import the data and look at the features more closely.
* Link to dataset is here <https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29>

2.1.1) Look at var2 (column1) - "age" to see if they have the same distribution as in the web.

2.1.2) Look at var14 (column13) - "Alcohol" and var15 to see if they have the same distribution as in the web.

2.2) Check missing values and duplicates

- No missing values.

- All column 0 (ID numbers) is unique.

- All people in the datasets are unique, i.e. no two people have all columns the sames (all features and outcome the same)

2.3) Take all features and two target var and further explore.

- Choose Amyl (var16) (because this is 'party drug') and Cannabis (var19) (because this is everywhere now)

- Rename columns so that the names make sense.

- Most of features in the dataset were categoretical originally. The dataset owners quantified them and we dataset we have has quantified features. While some quantified features make sense (for example personality traits), other quanfitied features don't (for example country and ethnicity). We will convert some features (age, gender, education, country, ethnicity) back to categoretical data type (we still keep their quantified values).

2.4) Create two target variables The two target variables Amyl and Cannabis have values from CL0 to CL7. Let transform to binary classification: "Never Used" and "Used over a Decade Ago" are considered "Non-user"; and all other classes form class "User"

- New dataframe drug\_1:

- We take only the needed columns and rename columns to make sense.

- rename column to make sense/

- New variable Age, Age\_level, Education, Education\_level, Country and Ethnicity

- New target variable Amyl\_binary (Non-user and User) and Amyl-user (values 0 or 1); the same for Cannabis.

EDA from here.

For EDA need

- Histogram of personality traits.

- Historgram for drug use as well => you will see one is rarely used and other is often used.

- Correlation

Summary (at the end of the notebook)