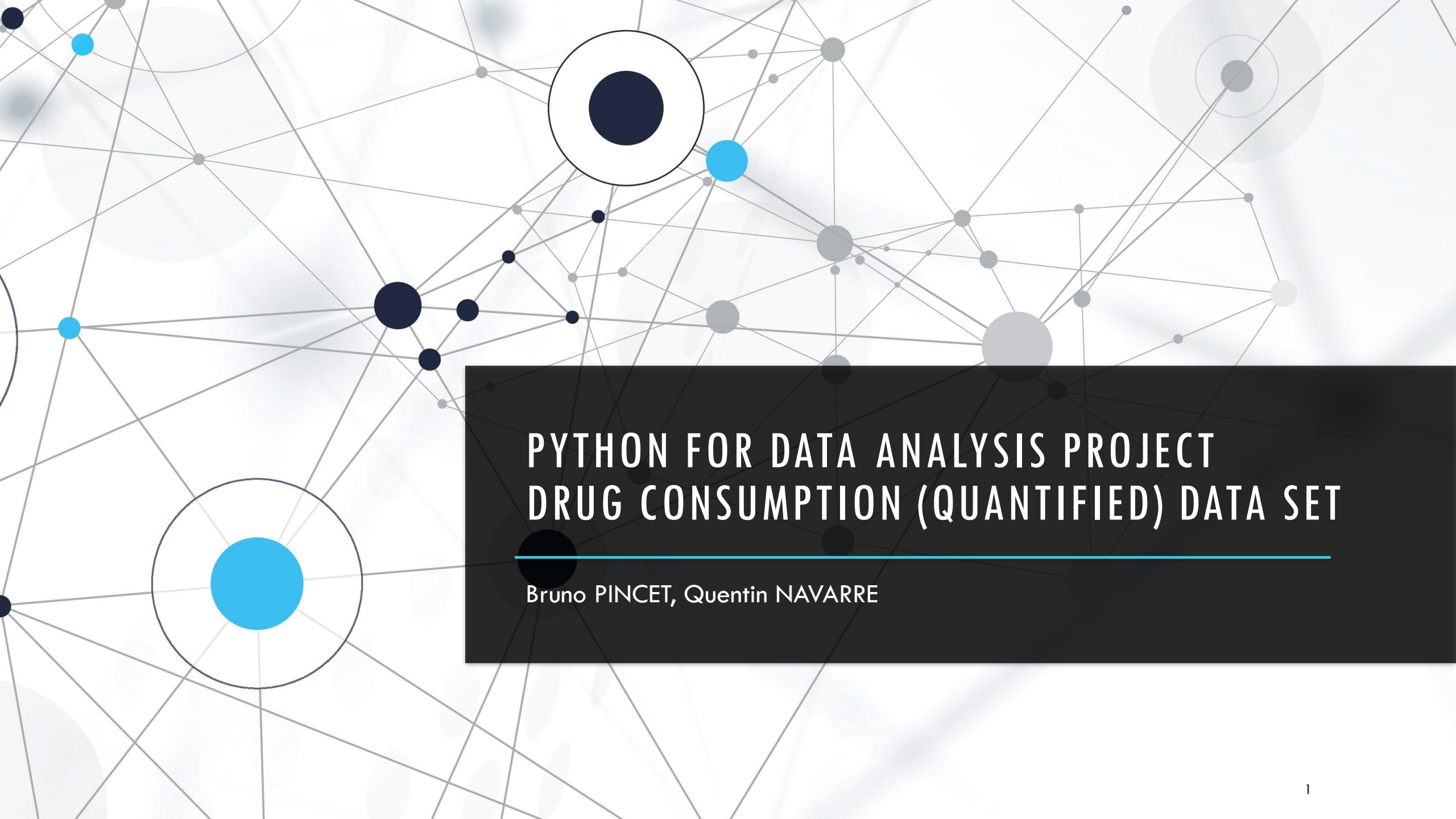


PYTHON FOR DATA ANALYSIS PROJECT
DRUG CONSUMPTION (QUANTIFIED) DATA SET

Bruno PINCET, Quentin NAVARRE

STEPS OF THE PROJECT

Data Cleaning

- Preparation of the data set

Data-visualization

- Data set presentation
- First link between the features and the target

Modeling

- Creation and optimization of various machine learning model in a classification problem

Flask API

- Web application to make prediction based on ML model

DATA CLEANING

BEFORE

	1	0.49788	0.48246	-0.05921	0.96082	0.12600	0.31287	-0.57545	-0.58331	-0.91699	...	CL0.4	CL0.5	CL0.6	CL0.7	CL0.8	CL0.9
0	2	-0.07854	-0.48246	1.98437	0.96082	-0.31685	-0.67825	1.93886	1.43533	0.76096	...	CL4	CL0	CL2	CL0	CL2	CL3
1	3	0.49788	-0.48246	-0.05921	0.96082	-0.31685	-0.46725	0.80523	-0.84732	-1.62090	...	CL0	CL0	CL0	CL0	CL0	CL0
2	4	-0.95197	0.48246	1.16365	0.96082	-0.31685	-0.14882	-0.80615	-0.01928	0.59042	...	CL0	CL0	CL2	CL0	CL0	CL0
3	5	0.49788	0.48246	1.98437	0.96082	-0.31685	0.73545	-1.63340	-0.45174	-0.30172	...	CL1	CL0	CL0	CL1	CL0	CL0
4	6	2.59171	0.48246	-1.22751	0.24923	-0.31685	-0.67825	-0.30033	-1.55521	2.03972	...	CL0	CL0	CL0	CL0	CL0	CL0

5 rows × 32 columns

Problems

- Column name instead of first row
- Unable to see features name
- Float data instead of clear categories

Solutions

- Addition of a row which value were columns name
- We renamed the columns
- We applied a dictionnary on each features to have clear data
- We deleted some outeliers

AFTER

id	age	gender	education	country	ethnicity	nscore	escore	oscore	ascore	cscore	...	ecstasy	heroin	ketamine	legalh	lsd	meth	mushrooms
2	25-34	Male	Doctorate degree	UK	White	29	52	55	48	41	...	Used in Last Month	Never Used	Used in Last Decade	Never Used	Used in Last Decade	Used in Last Year	Never Used
3	35-44	Male	Professional certificate/ diploma	UK	White	31	45	40	32	34	...	Never Used	Never Used	Never Used	Never Used	Never Used	Never Used	Used over a Decade Ago
4	18-24	Female	Masters degree	UK	White	34	34	46	47	46	...	Never Used	Never Used	Used in Last Decade	Never Used	Never Used	Never Used	Never Used

Problems

- Column name instead of first row
- Unable to see features name
- Float data instead of clear categories

Solutions

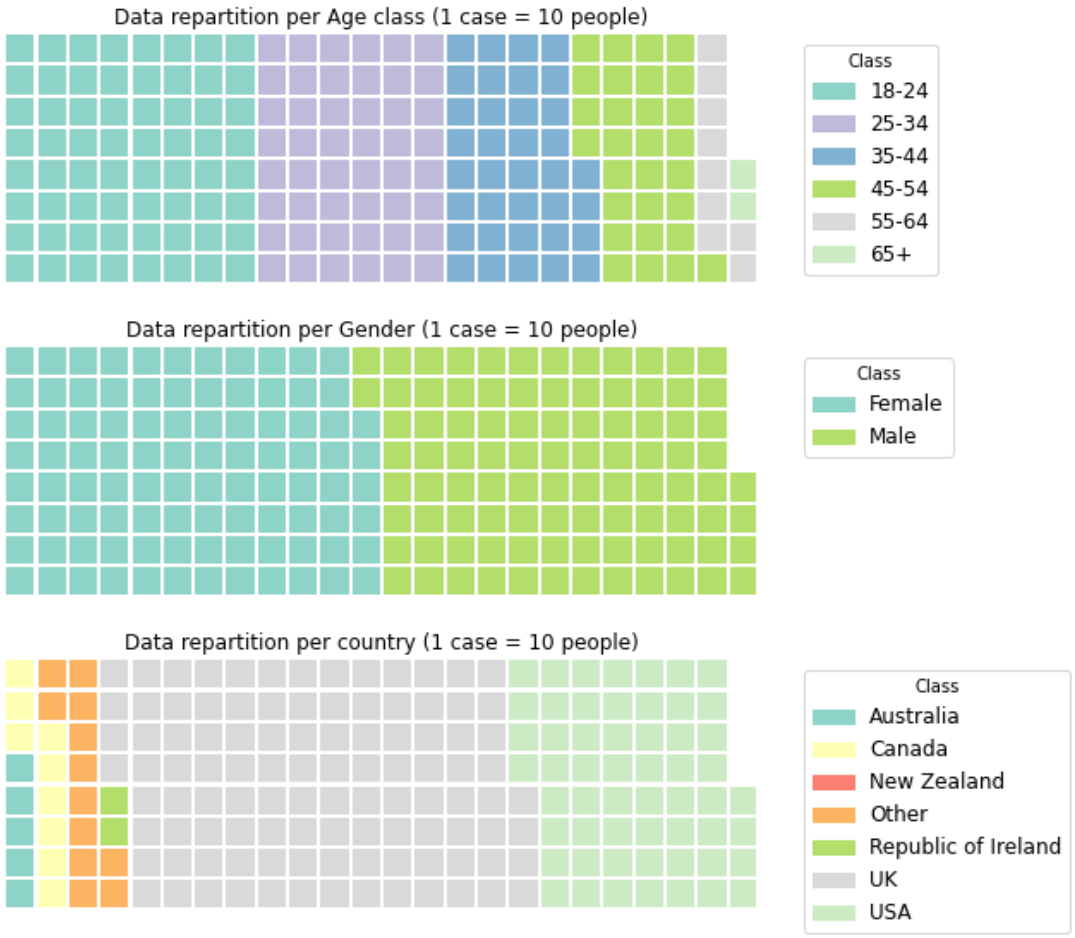
- Addition of a row which value were columns name
- We renamed the columns
- We applied a dictionnary on each features to have clear data
- We deleted some outeliers

DATASET INFORMATION

Database contains records for 1885 respondents. For each respondent 12 attributes are known

DATA SET INFORMATION

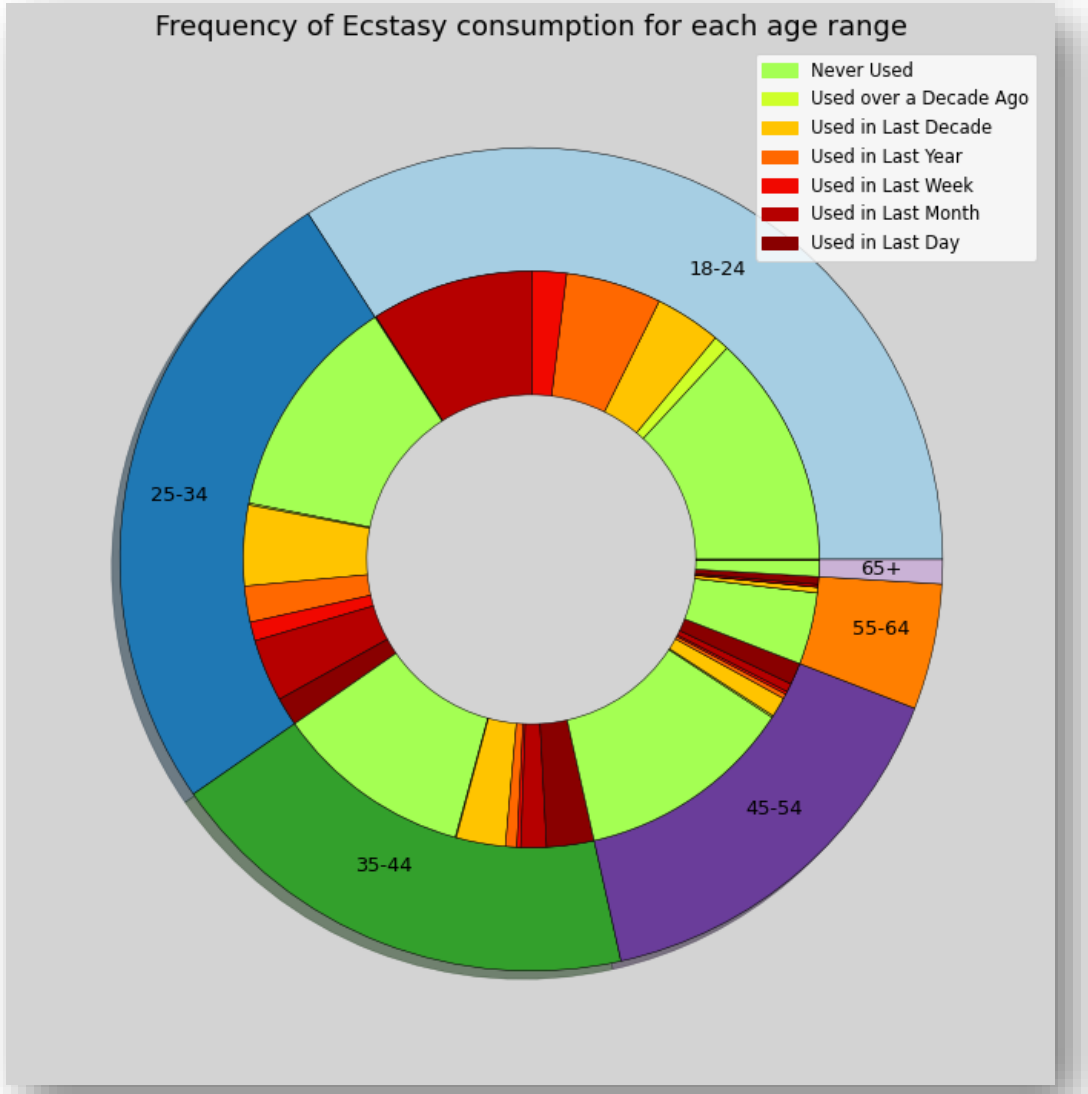
:



Personality measurements

- neuroticism
- extraversion
- openness to experience
- agreeableness
- conscientiousness
- impulsivity
- sensation seeking
- level of education
- age
- gender
- country of residence
- ethnicity

DATA SET INFORMATION



Questions concerning their use of 18 legal and illegal drugs

- Alcohol
- Amphetamines
- Amyl nitrite
- Benzodiazepine
- Cannabis
- Chocolate
- Cocaine
- Caffeine
- Crack
- Ecstasy
- Heroin
- Ketamine
- Legal highs
- LSD
- Methadone
- Mushrooms
- Nicotine
- Volatile substance
- Fictitious drug (Semeron)

DATA SET INFORMATION

Illegal drugs :

- Amphetamines
- Amyl nitrite
- Benzodiazepine
- Cannabis
- Cocaïne
- Crack
- Ecstasy
- Heroin
- Ketamine
- LSD
- Methadone
- Mushrooms
- Volatile substance

Legal drugs :

- Chocolate
- Caffeine
- Legal Highs
- Nicotine

Fictitious drug :

- Semeron

ASKED QUESTION ?

Can we find a relation between the personality measurements and the use of illegal drug ?

CREATION OF A NEW CATEGORICAL VARIABLE WITH 2 POSSIBLE OUTPUTS

Considered as a Drugged User

-Used in Last Decade

-Used in Last Year

-Used in Last Month

-Used in Last Week

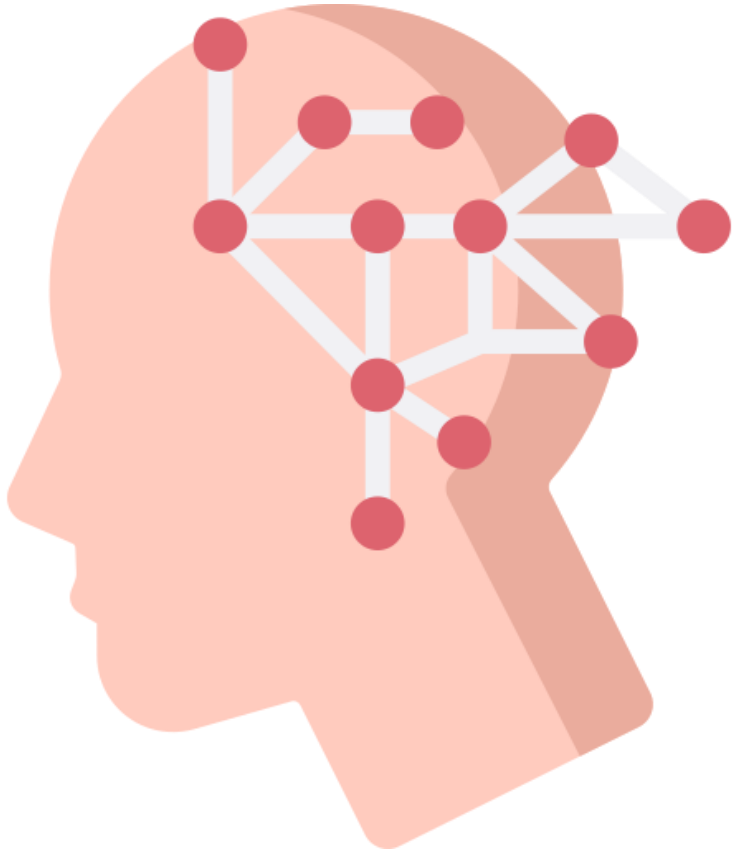
-Used in Last Day

-Never Used

-Used over a Decade Ago

Considered as a Clean User

FEATURES EXPLANATIONS



Features determined with the NEO PI-R, a personality inventory that assesses an individual on five dimensions of personality :

- Neurotocism
- Extraversion
- Openess to experience
- Agreeableness
- Conscientiouness

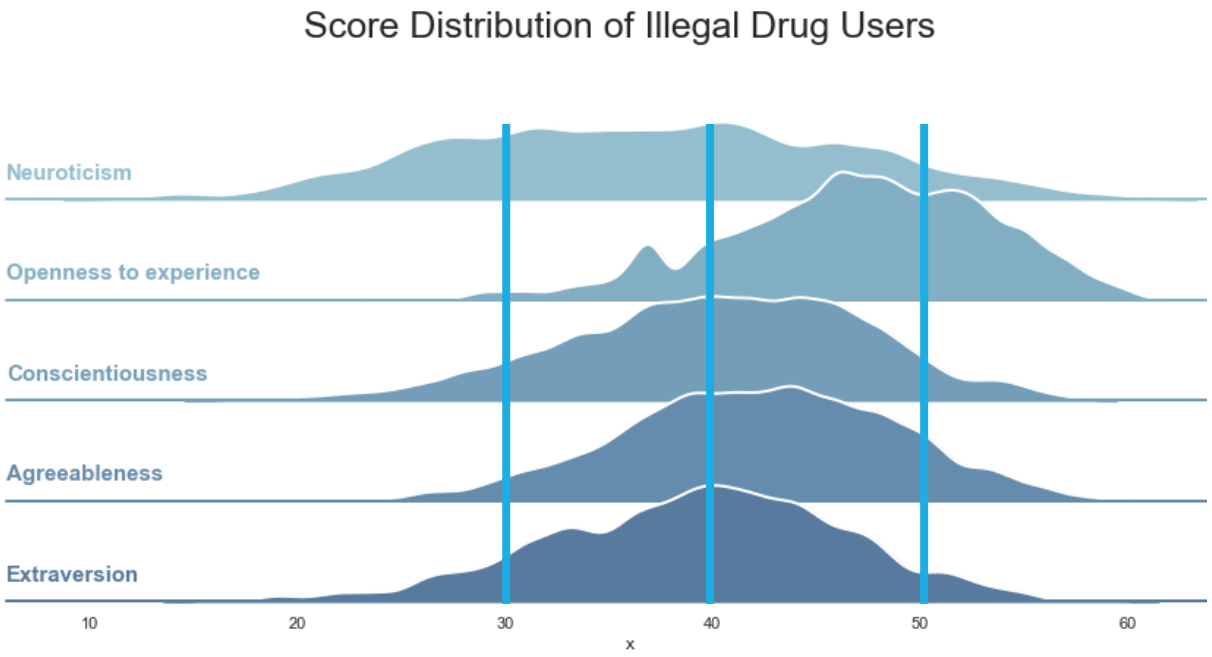
Each of these features is a score between 12 and 60, the higher the score, the more the characteristic correspond to the person.

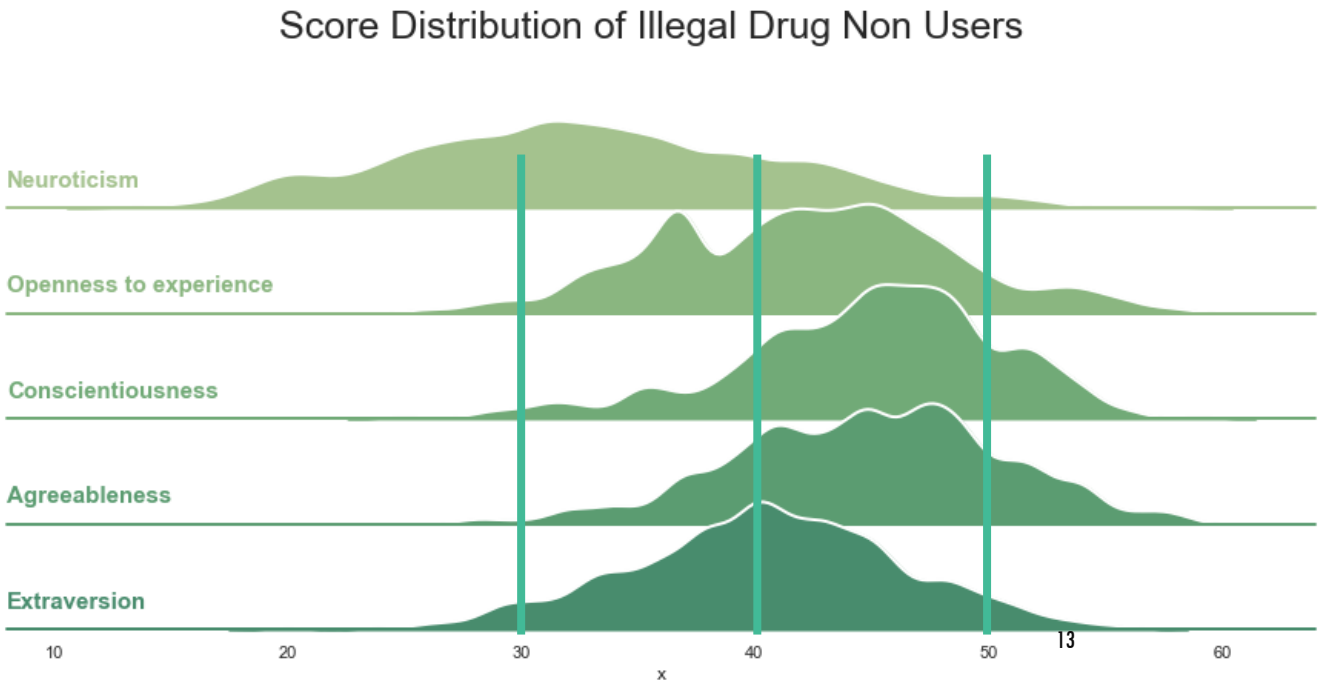
FIRST OBSERVATIONS

-Illegal drug users have slightly higher scores in neuroticism and openess to experience

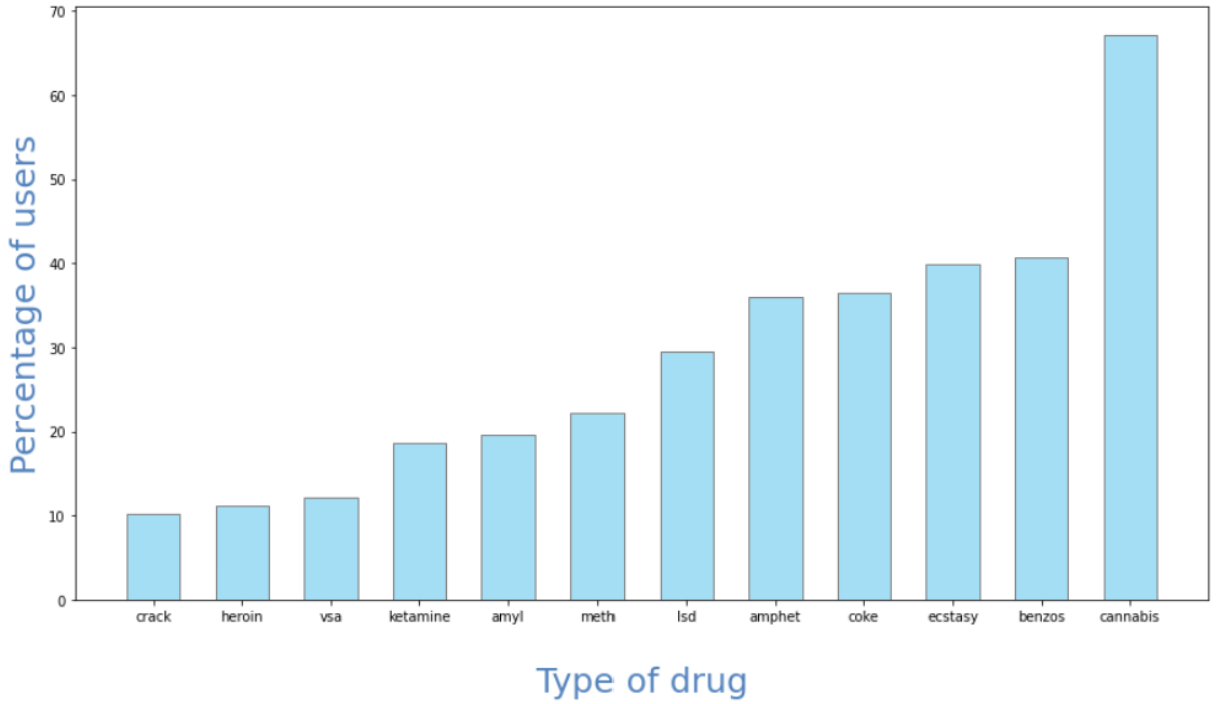
-Illegal drug users have slightly lower scores in conscientiousness and agreableness

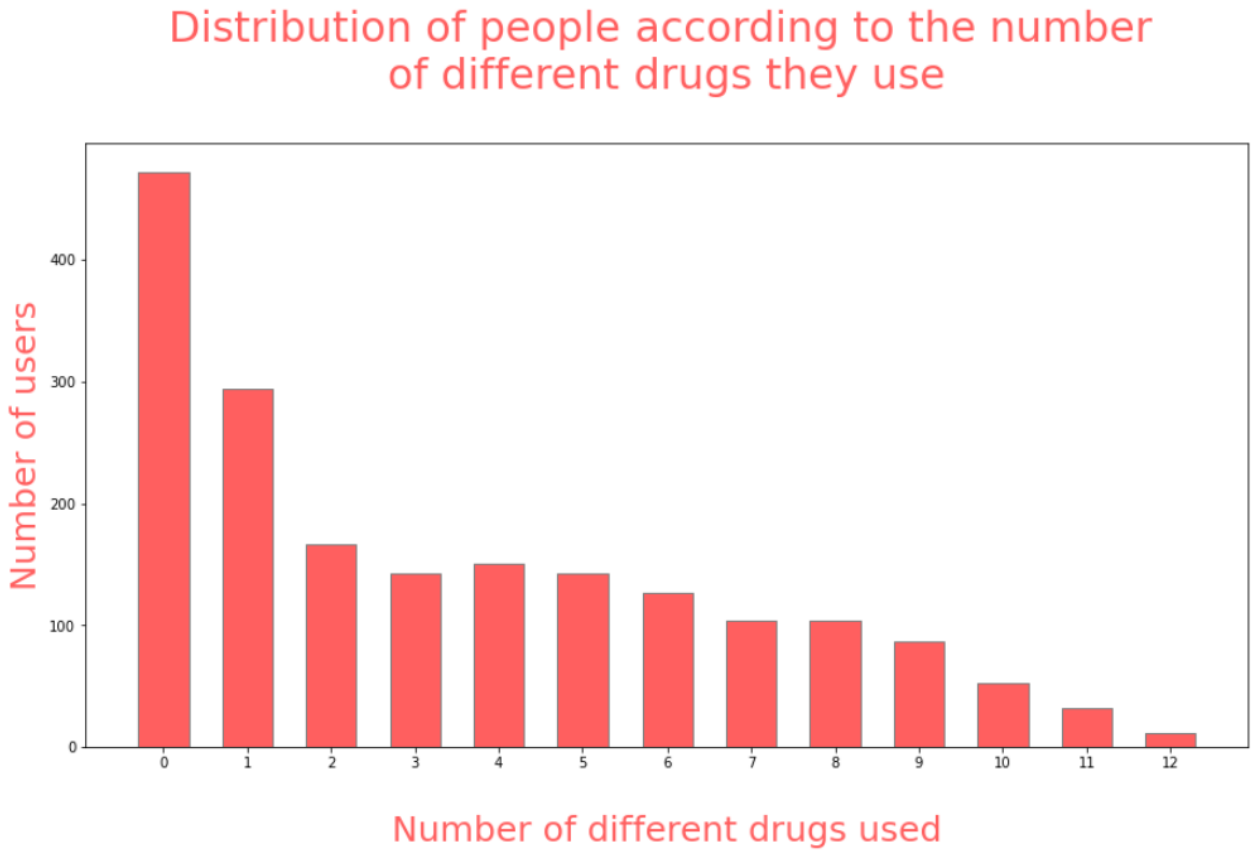
-The distribution of the score is much more compact with the clean people in comparison with the drug users





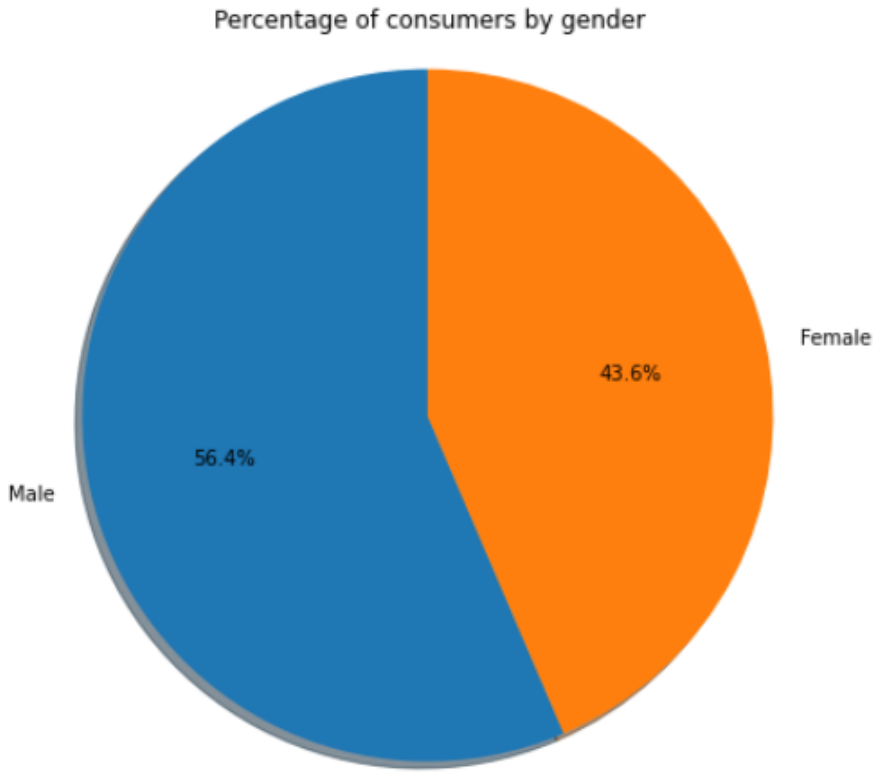
FIRST OBSERVATIONS

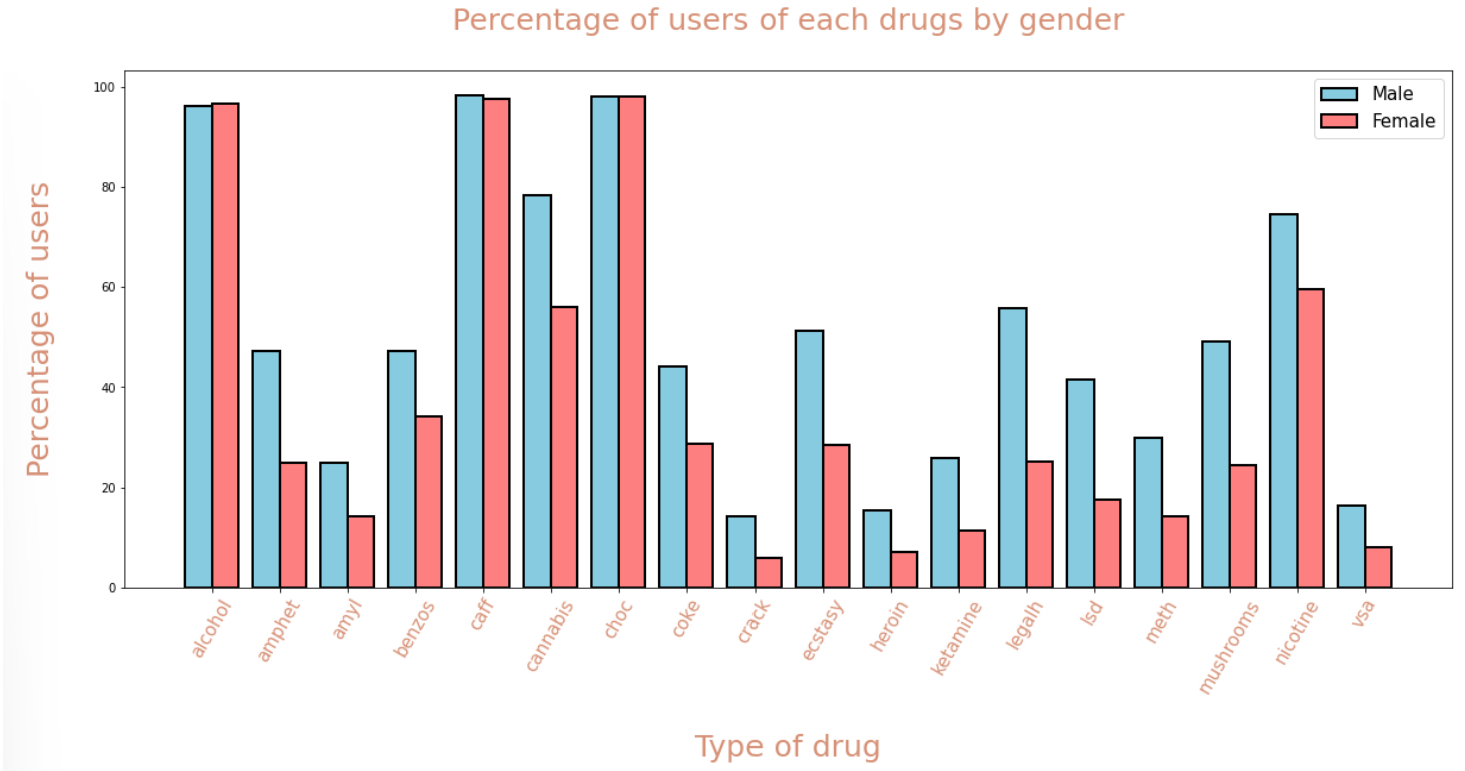




Not surprisingly the most common drug is cannabis and the more the number of different drugs consumed increases the fewer people there are in the category.

FIRST OBSERVATIONS

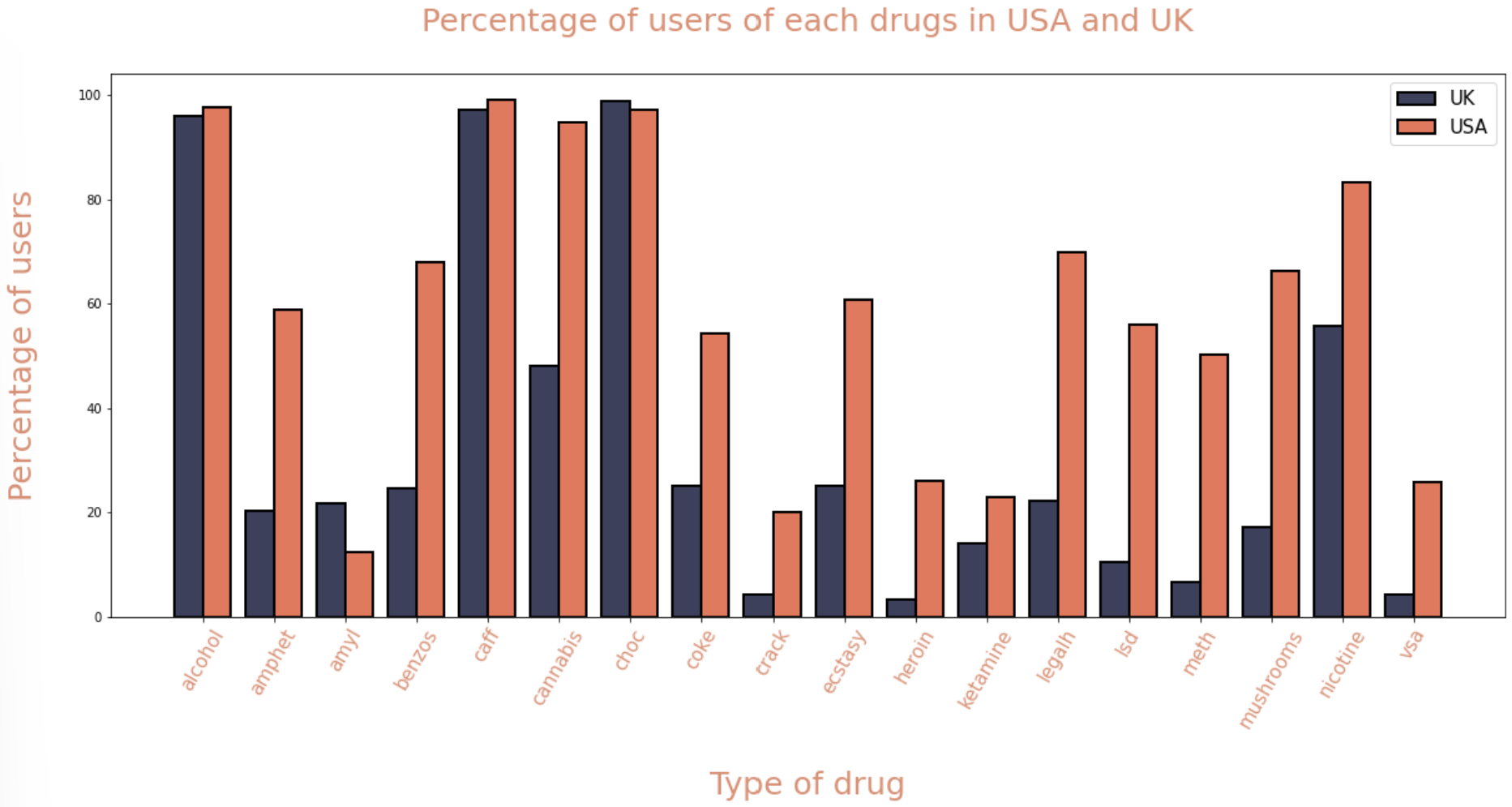




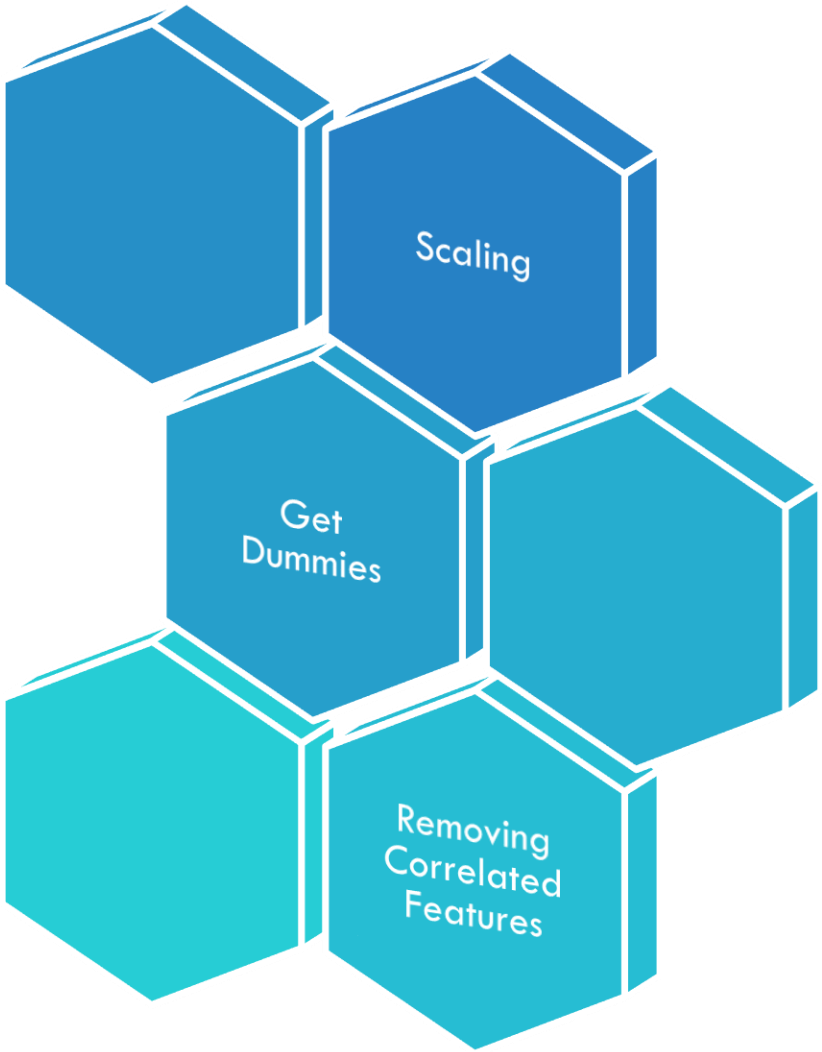
We can see that there are more men who use drugs than women, this is also shown on the graph on the right where we get more male users for each illegal drug while the results are the same for legal drugs.

FIRST OBSERVATIONS

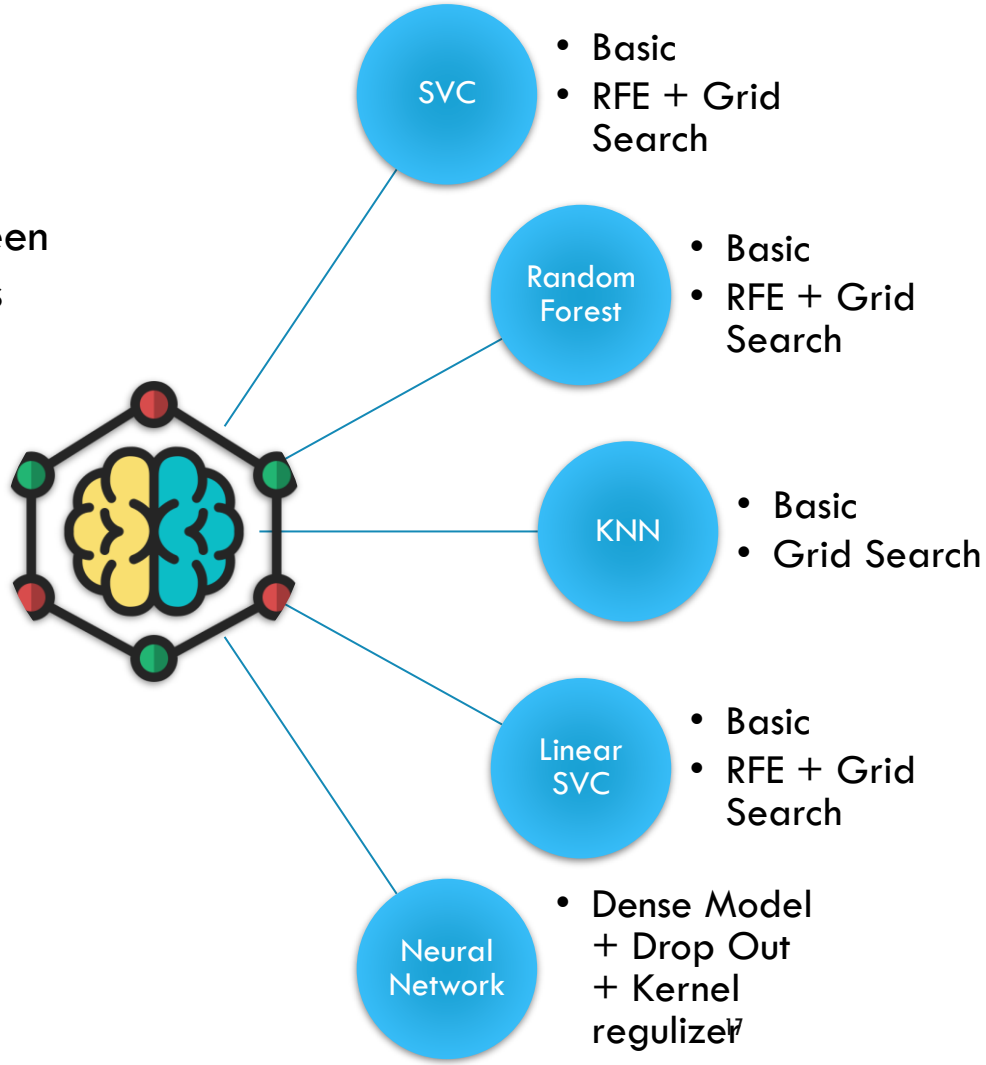
From this graphic we can observe 2 things:
Concerning the legal drugs, (alcohol, cafeine, chocolate) the percentage of users is almost the same if we compare various country.
However, concerning the hard drugs, the results are completely differents.



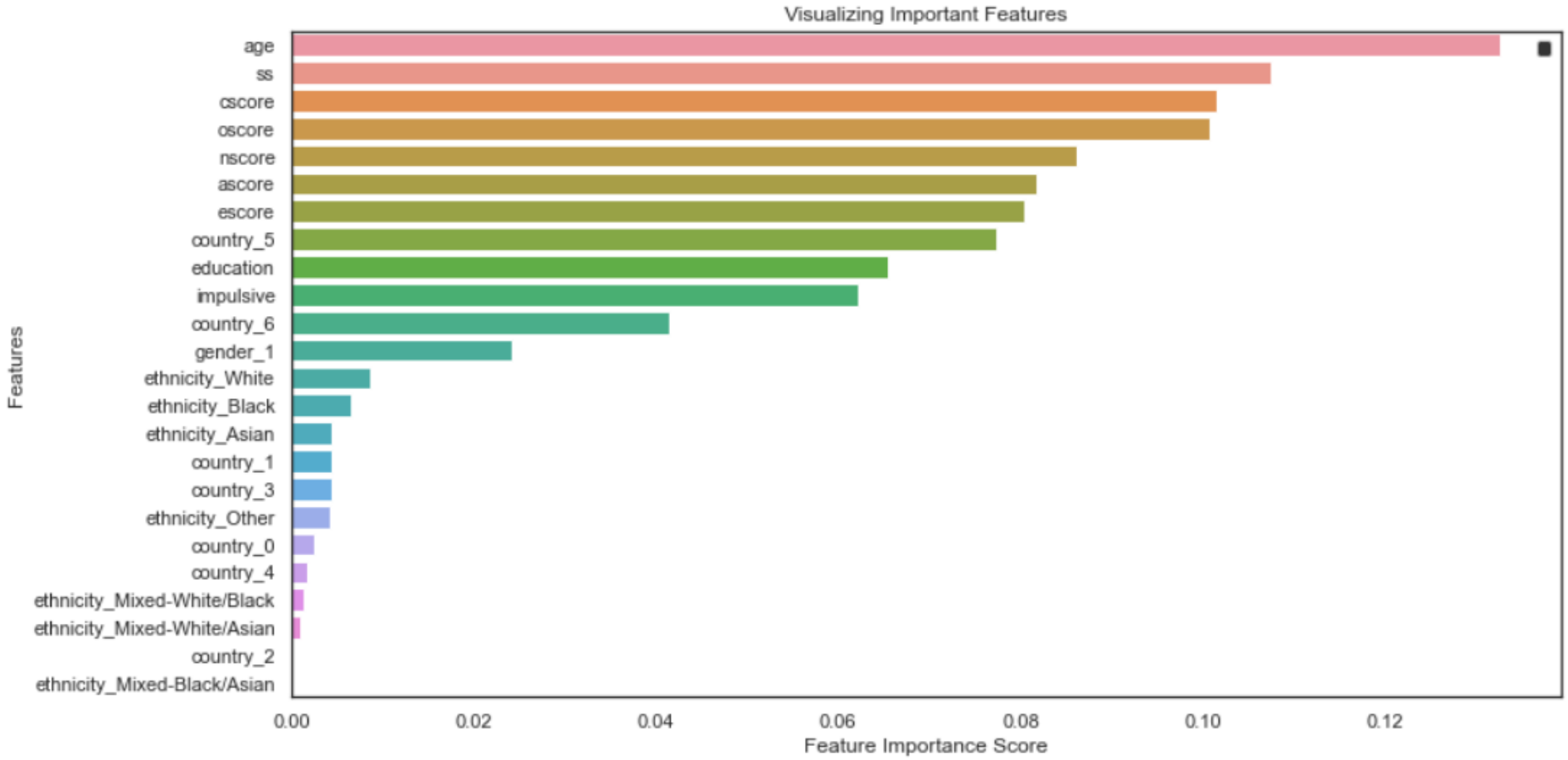
MACHINE LEARNING TO ANSWER THE QUESTION



Can we find a relation between the personality measurements and the use of illegal drug ?



MOST SIGNIFICANT CHARACTERISTICS FOR ILLEGAL DRUG USE



Apart from the age we can see that the most important features are mainly related to the personality of the individual.

FINAL RESULTS

The average accuracy of the model is close to 80%
We can thus predict if a person is probably drugged or clean based on various personality measurements.