**Report Group 7**

### **Data analysis on Drug Consumption and related factors**

This project’s goal is to analyse the factors that influence drug consumption between different age groups and countries.

The widespread concern about the use of illicit drugs has to take into consideration health, educational, individual's background and influence of other people.

The misuse of drugs can cause shifts in a person's behaviour and leads to the inability to control the use and the dosage, raising mental and physical problems.

As time passes the individual may need to increase the dosage and find it extremely difficult to go without the drug. What we did is an analysis on the drug consumption connected to parameters such as education, starting age of use and personality.

**DATASETS:**

We used 5 datasets for our project:

1. **Drug Consumption**: <https://github.com/deepak525/Drug-Consumption>

This Database contains records for 1885 respondents. For each respondent 12 attributes are known: Personality measurements which include neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness. It also contains general demographic factors such as level of education, age, gender, country of residence and ethnicity.

1. **The starting age of drug users in Europe:** <https://www.emcdda.europa.eu/data/stats2020/tdi_en>

The variables we dealt with are:

* LOCATION = European countries
* AGE = age groups representing the numbers of individuals that started to do drugs in that period
* MEAN = age average in each country

1. **Arrest data from 2010 to 2019:** <https://catalog.data.gov/dataset/arrest-data-from-2010-to-2019>

This dataset reflects arrest incidents in the City of Los Angeles from 2010 to 2019

* AGE = age of prisoners when they were sentenced
* ETHNICITY = traits that define the individual
* INJURY DESCRIPTION = the reason why they went to prison
* TIME = how much time they spent in prison

1. **Accidental drug related deaths from 2012 to 2018:** <https://catalog.data.gov/dataset/accidental-drug-related-deaths-2012-2018>

This dataset contains a listing of each accidental death associated with drug overdose in Connecticut from 2012 to 2018. A "Y" value under the different substance columns indicates that particular substance was detected.

* RACE = ethnicity of a population group
* AGE = age when they died
* INJURY DESCRIPTION = the way they died for OD (injection, inhalation, mix of drugs etc.)
* DRUG = which drug killed these people

1. **Education index:**

**http://hdr.undp.org/en/data**

* COUNTRY = country in which they collected data
* VALUE = number from 0 to 1 that shows the level of education in each country

**FIRST PART (Andrea Salusso)**

**RESEARCH QUESTIONS**

**1)** **Is Neuroticism connected to the misuse of drugs?**

* **Which kind of drugs people use if their level of Nscore is high/low?**
* **Is there a difference between legal and illegal drugs for what concerns Nscore?**
* **How often do people use different kinds of drugs?**
* **Is the use of drugs correlated to the education level?**
* **Does the use of alcohol matter when using drugs?**

We started working on the drug consumption dataset to answer these questions. We started the process of data cleaning and created new dictionaries containing the values of Oscore, Nscore, Ascore, Escore and Cscore in order to make them more readable. We realized a dictionary containing all the kinds of drugs we needed to compare data. Then we created a function that replaced the values of the use of drugs starting from 0, which means "Never Used", until 6, which means "Used in the last day".

In the same function we dropped all the data we didn't need, using just all the kinds of drugs and paying particular attention to the two values Oscorse and Nscore, which correspond to Openness to experience and Neuroticism. We replaced the values of the dataset corresponding to Oscore and Nscore with the dictionary we created earlier to make it easier to read.

We created a function to analyse the Nscore comparing legal and illegal drugs: the real difference between legal and illegal drugs is that legal ones tend to have a general Nscore that goes up and down without any strong jumps.

People who use or use drugs more often, tend to have higher Nscore, especially with heroin. We can see a big increase

in the group of people that use it everyday or almost everyday.

One of our questions was: how often do people use different kinds of drugs? To answer, we set up a function that showed the use, from 0, that means “Never used”, to 6, that means “used in the last day”.

We can see a pattern for what concerns stronger drugs: People that use Heroin, Cocaine, Ketamine and LSD have a tendency to use it rarely than lighter drugs that could be used everyday.

To satisfy the question about education, we focused on the mean of each drug based on sex. As we saw, males started doing drugs generally earlier, with a mean of -0.19060786850477202 which means between going to college and getting a diploma, while females at 0.183193991507431, which translates into the period between getting a professional certificate and getting a university degree.

At last we compared the use of alcohol to the use of drugs: From what we saw comparing a kind of drug with alcohol, we found that Hallucinogens generally tend to be used less at the same time as alcohol, because people probably tend to

prefer the use of cannabis or other hallucinogenic drugs to alcohol(with the exception of mushrooms). As we can see, the more frequently people use these kinds of drugs, the less they use alcohol. On the other hand, more addictive drugs such as cocaine, ketamine and Heroin have higher alcohol consumption that comes with them.

**Question 2: (Andrea Salusso)**

**Is there a correlation between people of different ethnicities in drug related arrests? How much time do they spend in prison on average?**

In order to answer this question, we used the drug arrest dataset. First of all we located the code representing drug related arrests and dropped the columns we didn’t need.

Since this dataset contained 1.3 million entries, we decided to take just 10,000. Of these 10,000, we located the people with drug related arrests. We changed the age range and added some values in order to make it clearer.

We created a function to get the mean of days people had to spend in prison based on each ethnicity. What we found out is that just in the city of Los Angeles, the average of days spent in prison between white and black people has a difference of 73 days, basically 2 months and a half, which is one of the reasons why, even though 5% of illicit drug users are African American, yet African Americans represent 29% of those arrested and 33% of those incarcerated for drug offenses.

**Question 3:(Andrea Salusso)**

**Is there a correlation between the level of education in Europe and the starting age of drug use?**

First we took the education dataset and created a new database with Country and Edu as parameters, in order to work with just these two values. Then we took all the education values and we did the mean of these values in order to find a unique value that states the average education level of a state.

Then we created a new dataset with the values of the previous dataset and we added the columns “Mean Age” from the starting age dataset and summed the values together to find a unique number.

As a conclusion, we noticed that the country with the overall lowest value is Luxembourg, which is expected because the mean age for the country is 16 years old. What is actually interesting is that Italy has a overall score higher than Germany, even though the education index is higher in Germany. This is due to the fact that the mean age in Italy is 21, while in Germany 18.

The highest scores we faced are Greece and UK, because they have the highest starting age and a general high education index.

**Question 4: (Nicolò Marcato, Giovanni Gatto)**

**Does education play a key role in drug use? Does dropping out of school consist of an increase in drug use?**

To answer to this question, we used the dataset drug\_consumption.csv

The dataset contains different information about age, education, frequency of drug use, divided per type of drug. Each information in the data frame is divided with numeric variables and each one corresponds to a specific information.

The dataset contained lots of columns of different drugs, so we had to choose (with which to work)

We were interested only about age, education, and frequency of heroin, cannabis and alcohol use, so we had to clean the dataset.

Data cleaning:

- Creating a dictionary, for the range of ages, in which every numeric variables are the keys, and its correspondent meaning are the values

- Creating another dictionary for the types of education, in which the numeric variables are the keys, and the meanings are values

- Removing from Education the information that are not relevant for the research question

Starting from the data cleaning of the dataset, drug\_consumption we analyze specific cases of education (people who left school, people who got a degree and people who did not get a degree) and drugs (Heroin, Cannabis and Alcohol).

First of all we analyzed the frequency with which heroin is consumed. We came up with discovering that people who left school and people who did not get a degree, on average, consume more heroin in comparison to people who get a degree.

Then we analyzed the use of alcohol in comparison with the three types of education that we took in consideration, and we saw, thanks to a pie chart, that people with a degree consume more alcohol in respect to people that left school and did not get a degree.

At the end we analyzed the use of cannabis; confronting the three pie charts we discovered that people that did not get a degree consume more cannabis than the other two categories of education.

So we can say that there is a relation between the drop of the studies (school or degree) and the use of drugs, in particular we discover that cannabis and heroin are related with the drop of the studies and alcohol, a drug more accessible, is used more bypeople with a degree.

**Question 5: (Nicolò Marcato, Giovanni Gatto)**

**Difference of starting age and overdose death for age between Netherlands and United Kingdom. In which of the two countries people start before assuming drugs? And in which country there’s the biggest number of death caused by overdose? Suppose why**

To answer to this questions, we use two datasets: starting\_age\_drug.csv and Overdose\_deaths.csv

Data cleaning:

* Removing the information that are not relevant from both datasets
* Merging the two datasets

We took in consideration the starting ages and overdose deaths of Netherlands and the United Kingdom and we created two plots, the first one shows the mean starting age and we discover that in the Netherlands the average age, 19 years, is before in comparison of the one in the UK that is over 20. From the second chart that represents the overdose deaths per inhabitant, we can see that in the Netherlands overdose’s deaths are much less than the ones in the UK.

**Question 6: (Giovanni Gatto, Nicolo Marcato)**

**Average age of arrests for drugs, alcohol and weapons? Average age in EU?**

To answer to this questions we used two datasets: Arrest\_Data\_from\_2010\_to\_2019.csv and Starting\_age\_drugs.csv.With the Arrest\_Data\_from\_2010\_to\_2019.csv we compute an average about drug crimes, weapon carry and drunkenness, we summarize it in a chart that shows the average age per crime; we discover that for weapon carry the age is, on average, less than the other two.

Then we add another element to analyze from the second dataset, starting using drugs, we compute a mean and thanks to another chart we can compare the average age for the four events.

**Question 7: (Matilde de March)**

**Which is the most common drug used which caused death by overdose?**

* **Is it more common for a genre to die?**
* **Are there any differences between males and females drug consumption?**

The dataset used to answer to this question is Accidental\_Drug\_Related\_Deaths\_2012-2018.csv

There are lots of columns in this dataframe which are not relevant for the analysis, most of them have a lot of Nans values, for example Residence Country or State, Death City, etc.

The major data used in this analysis are the drugs used by each person which caused overdose.

Data cleaning part:

* Removing columns which are not relevant for the analysis
* Filling Nans Values with 0
* Replacing string values with integer numbers
* Changing columns type from object to integer

Drug uses are marked with a 'Y', in order to answer to the research question these instances must be changed with ones, so that they can be summed up and made into other analyses. Other strings have been used to describe the type of drug used, so the same replacement has been done.

So answering the research question, yes there are differences between males and females on the drug used.

First of all males overdoses cases are much higher than females. So it is more common for males to die of overdoses than females. Then, as we can see, there is a difference between males and females on drug use that causes overdose: females use more Benzodiazepine, while males use more Cocaine, even though the most common drugs used by both genres are Heroin and Fentanyl.

Matilde De March 887371

Giovanni Gatto 884965

Nicolo’ Marcato 884926

Andrea Salusso 885032