DVisualizers

Drug Abuse

Ghazal Ayobi, Khushnur Binte Jahangir

Visual Analytics

1. **Datasets Description**

Drug Consumption is a collection of 1885 responses containing 32 variables. The main goal of this dataset is the Five-Factor Model of personality and evaluation of drug consumption risks. Furthermore, this dataset has demographic information about each respondent. Such as age, gender, education, country, and ethnicity. As mentioned this dataset has personality attributes of drug users. These measures are as follows: neuroticism, extraversion, openness to experience, agreeableness, conscientiousness, impulsivity, and sensation seeking. Moreover, participants responded regarding the usage of 18 legal and illegal drugs. The link to the dataset can be found [here](https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29). The dataset is collected by Khushnur.

The second data is from the National Survey on Drug Use and Health (NSDUH) which is a leading source of statistical information on the use of illicit drugs, alcohol, and tobacco and mental health issues in the United States. This is a population survey of ages 12 and older. The dataset is available on the NSDUH website. The dataset for 2020 has 32983 observations and 2890 variables. The questions in the dataset include age at first use, as well as lifetime and past usage of alcohol, marijuana, cocaine (including crack), hallucinogens, heroin, inhalants, tobacco, pain relievers, tranquilizers, stimulants, and sedatives. Participants are also asked about their personal and family income, health care access, and coverage, illegal activities and arrest records, problems resulting from the use of drugs, and perceptions of risks. Demographic data include gender, race, age, ethnicity, educational level, employment status, income level, veteran status, household composition, and population density. The dataset is available in many downloadable formats. The dataset is collected by Ghazal. The link to the NSDUH 2020 dataset can be found [here](https://www.datafiles.samhsa.gov/dataset/national-survey-drug-use-and-health-2020-nsduh-2020-ds0001).

The third dataset is available in the Substance abuse and mental health archive (SAMHDA) which collects data through multiple sources and surveys. SAMHDA's goal is to provide access to public-use data files and documentation to support a better understanding of substance use disorders and mental health in the United States. SAMHDA contains data for all the US states from 2015 to 2020. This data archive has detailed information on alcohol, Tobacco, Marijuana, and other drugs. The benefit of this dataset is it provides a general overview of drug consumption across the United States. The datasets can be downloaded from SAMHDA archives. The dataset is collected by Ghazal. The link to the dataset can be found [here](https://pdas.samhsa.gov/saes/state)

The fourth dataset is smoking, alcohol consumption, and use of illicit drugs which is available on the website of *Our World in Data.* This is multi-table source provides information on the topic for all the world's countries. The first table covers data on the deaths from Tobacco, alcohol, and drug across all countries from 1990 to 2019. The second table contains a share of the population that has alcohol or illicit drug use disorders. The third table is the gender ratio with alcohol or drug use disorders and the last table is deaths from drugs or illicit drugs across different age groups in the world. The data tables from *Our World in Data* can be found [here](https://ourworldindata.org/drug-use). The dataset is found by Khushnur.

1. **Main Datasets**

The main dataset for this project is the Drug Consumption dataset which is found [here](https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29). The dataset contains records for 1885 respondents for 12 personality measures. The data was collected between March 2011 and March 2012 by Elaine Fehrman approved by the University of Leicester’s Forensic Psychology Ethical Advisory Group. During the data collection Revised NEO Five-Factor Inventory (NEO-FFI-R) questionnaire was employed. The five personality measures are as follows: Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness, Fehrman et al., 2017. The dataset is downloaded from the University of California, Irvine, UCI Machine learning repository. The donor of the database to the UCI is Evgeny M. Mirkes from the University of Leicester.

The second dataset we will use in this project is data tables available on the *Our World in Data* article on drug use which can be found [here](https://ourworldindata.org/drug-use). The data tables in this source provide general information on drug use across the world. The data is published in the Institute for Health Metrics and Evaluation, Global Burden of Disease (2019). *Global Burden of Disease* study identifies and provides two estimates for substance use deaths. The first is a direct death from a substance use disorder and the second is drugs can increase the risk of premature death. For our purpose, we will use deaths from tobacco, alcohol and other drugs across the world from 1990 to 2019. The second table we will use is the share of the population with alcohol and drug disorders across males and females and different age groups. The data tables can be downloaded from the *Our World in Data* website.

1. **Research Question**

The selected two datasets answer to our research questions. The dataset collected by Fehrman et al., 2017 introduces personality measurements for drug users. We aim to use the second dataset for introducing the general situation of drug use across the world over time. These are great sources to answer the following questions. Our first research question is: *What demographics are the biggest consumers of Alcohol? What is the correlation between education and drug consumption?* The data tables available in *Our World in Data have* sufficient data to answer the first part of the question. The reason is these data tables cover drug consumption data across many decades and countries. We do not change our first question as we have sufficient data to respond to this question.

The second research question is; *What is the relationship between alcohol consumption and other most consumed drugs among the* ***youth****, based on their personality measurement scores?* Based on the available data and research goal we would like to change the research question as follows: how drug consumption has changed over time, and how personality measurement scores are correlated with drug consumption? The reason we modified the research question is that based on the data, we can identify how drug usage has changed over the decades. In addition, drug users' personality score is a vital source to identify whether there are any correlated between specific personality scores and drug consumption.

The last research question is *Which drugs are consumed the most among* ***young people****? What is the relationship between social status and the most used drugs?* We plan to answer this question using data tables available on the Our World in Data website and the Drug consumption dataset by Fehrman et al., 2017. We keep this question the same. Both datasets have sufficient data to answer both parts of this question.

1. **Data Wrangling and Cleaning**