DSCI 310 Project Group 7

Jade Chen, Jessica Luo, Heidi Lantz, and Nazia Edroos

Outline

Predicting Age from Drug Use Patterns: A Statistical Analysis

Summary

This study explores the relationship between drug use and age, aiming to determine whether a person's reported substance use can serve as a reliable predictor of their age. Using the "Drug Use by Age" dataset from FiveThirtyEight, which compiles data from the National Survey on Drug Use and Health (Services 2013), we employ statistical techniques to analyze the relationship between age and substance use. The dataset contains information about a given age group, and summary info about the percentage of people who have done a drug and the median frequency it is used at. There is information on a wide variety of drugs, including alcohol, marijuana, various illicit drugs, and more. We apply feature selection and predictive modeling to assess the strength of this relationship. Our findings contribute to understanding how drug use behaviors vary across different age groups and the potential for age prediction in public health and policy contexts.

(add more about the findings at the end)

Introduction:

Background and Research Question

Age can greatly influence many aspects of how we behave and what decisions we make, especially regarding substance use. Patterns of drug consumption change over time in a person's life, reflecting their social, biological, and environmental influences. Research indicates that substance use disorders, including those related to alcohol, tobacco, cannabis, and opioids, generally decrease with age (Traci Green 2010). Younger individuals often engage in higher rates of substance experimentation, while older adults tend to exhibit more stable or declining usage patterns. Understanding these age-related trends is essential for developing effective public health strategies, addiction prevention programs, and targeted interventions.

In this study, we investigate whether a person's reported drug use can serve as a predictor of their age (Services 2013). Previous research has highlighted the progression and predictors of substance use across the lifespan. For instance, studies have identified that early initiation of substance use is associated with an increased risk of developing substance use disorders later in life (Irma Arteaga 2010). Furthermore, patterns observed during adolescence are often associated with continued use or potential abuse in adulthood (Joseph Allen 2021). This tells us that finding out which age groups are using which drugs can be useful information for addressing and preventing issues like this. Moreover, life-course patterns of substance abuse reveal that older adults with substance misuse issues often reflect on their usage patterns,

providing insights into the development and persistence of these behaviors over time (Caitlin Foster 2019). Recognizing these patterns across different age groups can inform the creation of age-specific prevention and treatment programs. By assessing the strength of the relationship between age and drug use behaviors, our analysis contributes to discussions on behavioral health and its implications.

This leads us to our Research Question:

- *Can we accurately predict an individual's age based only on their reported patterns of substance use?*

About the Data

To address our research question, we utilize the "Drug Use by Age" dataset sourced from FiveThirtyEight's repository, which is derived from from the 2013 National Survey on Drug Use and Health. This study was conducted by the United States Department of Health and Human Services, Substance Abuse and Mental Health Services Administration (SAMHSA), Center for Behavioral Health Statistics and Quality.

The dataset covers self-reported drug use trends across 17 age groups in the United States, examining 13 different substances. The dataset includes both legal substances, such as alcohol and marijuana, and illicit drugs, offering a broad perspective on substance use trends for each age group.

Below is a summary of the variables included in the dataset:

Variable Name	Description
`age`	Age group of respondents (e.g., 12, '22-23', '35-49', '65+' etc.).
`n`	Number of people surveyed in each age group.
`alcohol_use`	Percentage of respondents who used alcohol in the past 12 months.
`alcohol_frequency`	Median number of times alcohol was used in the past 12 months.
`marijuana_use`	Percentage who used marijuana in the past 12 months.
`marijuana_frequency`	Median number of times marijuana was used in the past 12 months.
`cocaine_use`	Percentage who used cocaine in the past 12 months.
`cocaine_frequency`	Median number of times cocaine was used in the past 12 months.
`crack_use`	Percentage who used crack in the past 12 months.
`crack_frequency`	Median number of times crack was used in the past 12 months.
`heroin_use`	Percentage who used heroin in the past 12 months.
`heroin_frequency`	Median number of times heroin was used in the past 12 months.
`hallucinogen_use`	Percentage who used hallucinogens in the past 12 months.
`hallucinogen_frequency`	Median number of times hallucinogens were used in the past 12 months.

Variable Name	Description
`inhalant_use`	Percentage who used inhalants in the past 12 months.
`inhalant_frequency`	Median number of times inhalants were used in the past 12 months.
`pain_reliever_use`	Percentage who used pain relievers in the past 12 months.
`pain_reliever_frequency`	Median number of times pain relievers were used in the past 12 months.
`oxycontin_use`	Percentage who used OxyContin in the past 12 months.
`oxycontin_frequency`	Median number of times OxyContin was used in the past 12 months.
`tranquilizer_use`	Percentage who used tranquilizers in the past 12 months.
`tranquilizer_frequency`	Median number of times tranquilizers were used in the past 12 months.
`stimulant_use`	Percentage who used stimulants in the past 12 months.
`stimulant_frequency`	Median number of times stimulants were used in the past 12 months.
`meth_use`	Percentage who used methamphetamine in the past 12 months.
`meth_frequency`	Median number of times methamphetamine was used in the past 12 months.
`sedative_use`	Percentage who used sedatives in the past 12 months.
`sedative_frequency`	Median number of times sedatives were used in the past 12 months.

Methods & Results:

(will be done by Jessica and Jade)

- describe in written english the methods you used to perform your analysis from beginning to end that narrates the code the does the analysis.
- your report should include code which:
  - loads data from the original source on the web
  - wrangles and cleans the data from it's original (downloaded) format to the format necessary for the planned classification or clustering analysis
  - performs a summary of the data set that is relevant for exploratory data analysis related to the planned classification analysis
  - creates a visualization of the dataset that is relevant for exploratory data analysis related to the planned classification analysis
  - performs classification or regression analysis

  - creates a visualization of the result of the analysis

- note: all tables and figure should have a figure/table number and a legend

Data Cleaning

1. Handle missing values by removing rows with excessive missing data or imputing reasonable values.
2. Check format of variable and transform as needed.
3. Check for outliers in age-related variables.
4. Standardize or normalize predictor variables as needed.

age	n	alcohol.use	alcohol.frequency	marijuana.use	marijuana.frequency
12	2798	3.9	3	1.1	4
13	2757	8.5	6	3.4	15
14	2792	18.1	5	8.7	24
15	2956	29.2	6	14.5	25
16	3058	40.1	10	22.5	30
17	3038	49.3	13	28.0	36
18	2469	58.7	24	33.7	52
19	2223	64.6	36	33.4	60
20	2271	69.7	48	34.0	60
21	2354	83.2	52	33.0	52
22-23	4707	84.2	52	28.4	52
24-25	4591	83.1	52	24.9	60
26-29	2628	80.7	52	20.8	52
30-34	2864	77.5	52	16.4	72
35-49	7391	75.0	52	10.4	48
50-64	3923	67.2	52	7.3	52
65+	2448	49.3	52	1.2	36

EDA

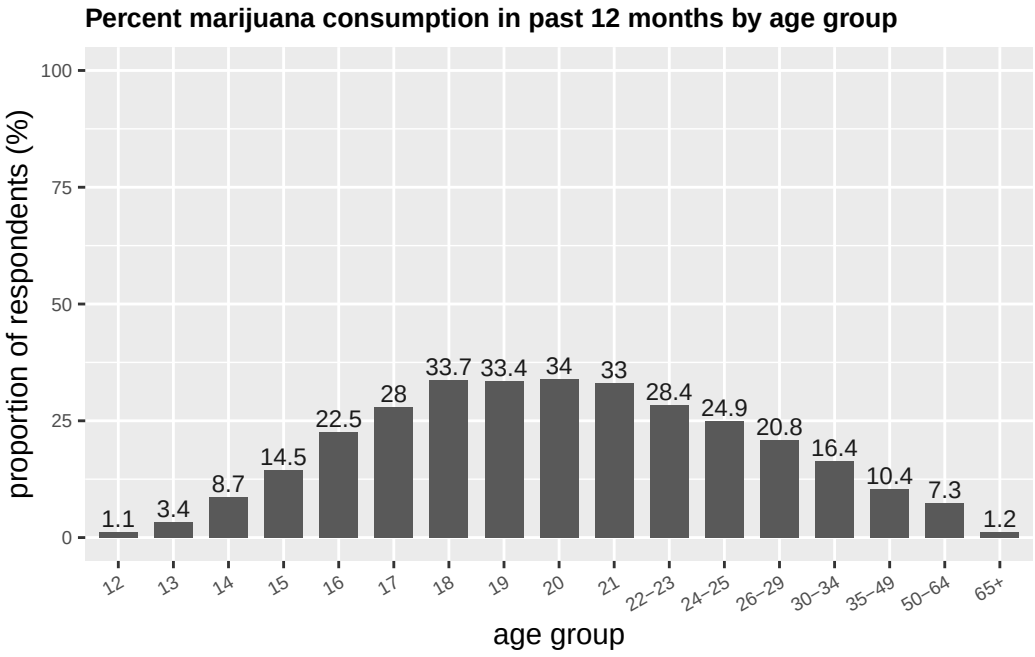
```
glimpse(data)
```

```
Rows: 17
Columns: 28
$ age                   <chr> "12", "13", "14", "15", "16", "17", "18", "19"~
$ n                     <int> 2798, 2757, 2792, 2956, 3058, 3038, 2469, 2223~
$ alcohol.use           <dbl> 3.9, 8.5, 18.1, 29.2, 40.1, 49.3, 58.7, 64.6, ~
```

```
$ alcohol.frequency       <dbl> 3, 6, 5, 6, 10, 13, 24, 36, 48, 52, 52, 52, 52~
$ marijuana.use           <dbl> 1.1, 3.4, 8.7, 14.5, 22.5, 28.0, 33.7, 33.4, 3~
$ marijuana.frequency     <dbl> 4, 15, 24, 25, 30, 36, 52, 60, 60, 52, 52, 60,~
$ cocaine.use             <dbl> 0.1, 0.1, 0.1, 0.5, 1.0, 2.0, 3.2, 4.1, 4.9, 4~
$ cocaine.frequency       <chr> "5.0", "1.0", "5.5", "4.0", "7.0", "5.0", "5.0~
$ crack.use               <dbl> 0.0, 0.0, 0.0, 0.1, 0.0, 0.1, 0.4, 0.5, 0.6, 0~
$ crack.frequency         <chr> "-", "3.0", "-", "9.5", "1.0", "21.0", "10.0",~
$ heroin.use              <dbl> 0.1, 0.0, 0.1, 0.2, 0.1, 0.1, 0.4, 0.5, 0.9, 0~
$ heroin.frequency        <chr> "35.5", "-", "2.0", "1.0", "66.5", "64.0", "46~
$ hallucinogen.use        <dbl> 0.2, 0.6, 1.6, 2.1, 3.4, 4.8, 7.0, 8.6, 7.4, 6~
$ hallucinogen.frequency  <dbl> 52, 6, 3, 4, 3, 3, 4, 3, 2, 4, 3, 2, 3, 2, 3, ~
$ inhalant.use            <dbl> 1.6, 2.5, 2.6, 2.5, 3.0, 2.0, 1.8, 1.4, 1.5, 1~
$ inhalant.frequency      <chr> "19.0", "12.0", "5.0", "5.5", "3.0", "4.0", "4~
$ pain.releiver.use       <dbl> 2.0, 2.4, 3.9, 5.5, 6.2, 8.5, 9.2, 9.4, 10.0, ~
$ pain.releiver.frequency <dbl> 36, 14, 12, 10, 7, 9, 12, 12, 10, 15, 15, 15, ~
$ oxycontin.use           <dbl> 0.1, 0.1, 0.4, 0.8, 1.1, 1.4, 1.7, 1.5, 1.7, 1~
$ oxycontin.frequency     <chr> "24.5", "41.0", "4.5", "3.0", "4.0", "6.0", "7~
$ tranquilizer.use        <dbl> 0.2, 0.3, 0.9, 2.0, 2.4, 3.5, 4.9, 4.2, 5.4, 3~
$ tranquilizer.frequency  <dbl> 52.0, 25.5, 5.0, 4.5, 11.0, 7.0, 12.0, 4.5, 10~
$ stimulant.use           <dbl> 0.2, 0.3, 0.8, 1.5, 1.8, 2.8, 3.0, 3.3, 4.0, 4~
$ stimulant.frequency     <dbl> 2.0, 4.0, 12.0, 6.0, 9.5, 9.0, 8.0, 6.0, 12.0,~
$ meth.use                <dbl> 0.0, 0.1, 0.1, 0.3, 0.3, 0.6, 0.5, 0.4, 0.9, 0~
$ meth.frequency          <chr> "-", "5.0", "24.0", "10.5", "36.0", "48.0", "1~
$ sedative.use            <dbl> 0.2, 0.1, 0.2, 0.4, 0.2, 0.5, 0.4, 0.3, 0.5, 0~
$ sedative.frequency      <dbl> 13.0, 19.0, 16.5, 30.0, 3.0, 6.5, 10.0, 6.0, 4~
```

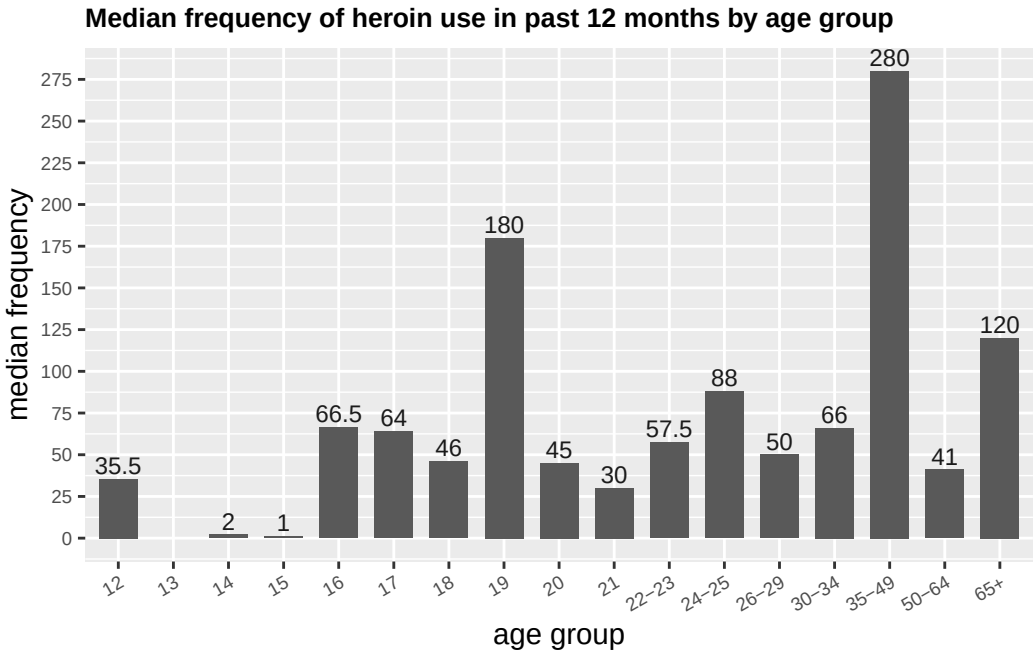
Since we aim to classify whether substance use during adolescence can predict substance use patterns later in life, we begin by visualizing the proportion of individuals in each age group who have consumed alcohol in the past 12 months:

We apply the same approach to examine marijuana use over the same period:



Both graphs reveal that these substances are commonly used during adolescence. However, while alcohol consumption increases significantly in adulthood, marijuana use tends to plateau

in early adulthood. To explore the use of hard drugs, such as heroin, crack and hallucinogens, we shift our focus to the median frequency of use rather than the proportion of users. This is because the proportion of individuals using harder substances is relatively small, making median frequency a more informative metric for understanding usage patterns. For instance, we can plot the distribution of heroin use across different age groups:



The motivation behind our research stems from the significant variation in substance use distribution based on the type of substance. Harder substances, in particular, exhibit distinct usage patterns compared to more commonly used substances like alcohol and marijuana. By analyzing these patterns, we aim to determine whether an individual's history of substance use during adolescence can predict their substance use behavior later in life. The insights gained from this research could inform prevention programs, helping them design targeted strategies to effectively advocate against drug abuse. This is especially critical for younger individuals, who are more susceptible to peer pressure and may benefit from early intervention.

Analysis

Split into test/training set.

Use linear regression, blah blah blah. Perform correlation analysis to identify significant predictors before modeling. Evaluate model performance using accuracy or RMSE.

Discussion:

(will be done by Nazia)

What we found:

summarize what you found

Does this align with our expectations?

discuss whether this is what you expected to find?

Impact of our findings

discuss what impact could such findings have?

Future directions

discuss what future questions could this lead to?

References

Caitlin Foster, Julie A. Gorenko, Candace Konnert. 2019. "Exploring Life-Course Patterns of Substance Abuse: A Qualitative Study." *Aging & Mental Health*, 378–85. https://www.tandfonline.com/doi/full/10.1080/13607863.2019.1693966.

Irma Arteaga, Arthur Reynolds, Chin-Chih Chen. 2010. "Childhood Predictors of Adult Substance Abuse." *National Library of Medicine*, 1108–20. https://pubmed.ncbi.nlm.nih.gov/27867242/.

Joseph Allen, Rachel Narr, Emily Loeb. 2021. "Different Factors Predict Adolescent Substance Use Vs. Adult Substance Abuse." *National Library of Medicine*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7755752/.

Services, US Department Of Health And Human. 2013. "FiveThirtyEight Drug Use by Age Dataset." https://github.com/fivethirtyeight/data/blob/master/drug-use-by-age/drug-use-by-age.csv.

Traci Green, et al. 2010. "Patterns of Drug Use and Abuse Among Aging Adults with and Without HIV: A Latent Class Analysis of a US Veteran Cohort." *National Library of Medicene*, 208–20. https://pubmed.ncbi.nlm.nih.gov/20395074/.