# 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 predictive modeling to assess the strength of this relationship.

Our findings suggest that while age-related trends in substance use are present, challenges such as class imbalance and small dataset size hinder the models' ability to accurately predict age based solely on drug use. Despite these challenges, the analysis highlights the potential for using substance use patterns as part of a broader predictive framework in public health and policy contexts.


# 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 if an individual is a Youth (under 21 years old) or Adult (21 and over) 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 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. |

| Variable Name | Description |
|---|---|
| hallucinogen_frequency | Median number of times hallucinogens were used in the past 12 months. |
| 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 and Results

### Read Data

We downloaded the dataset in `01-download-dataset.R` by retrieving it from a raw GitHub URL, and saving it locally in the data/raw directory.

### Data Cleaning & Transformation

The dataset underwent the following preprocessing steps in `02-data-clean-transform.R`:

1. Replaced missing values: Instances of `-` in character columns were converted to `NA` to handle missing data appropriately.

2. Converted character columns to numeric: All character columns (except age) were coerced into numeric format, introducing `NA`s where conversion was not possible.

3. Created a new column, `class`, to classify participants as `youth` or `adult` based on their age, where individuals aged 20 or younger were classified as `youth` and those older than 20 were classified as `adult`.

## EDA

Since we aim to predict age category using drug use patterns, we performed some exploratory data analysis (EDA) to gain a better understanding of our dataset in `03-eda.R`. The following shows the most valuable plots and insights we found.

We first want to take a look at the data to understand what we are dealing with. Let's view the first row of the data as an example:

```
  age    n alcohol.use alcohol.frequency marijuana.use marijuana.frequency
1  12 2798         3.9                 3           1.1                   4
  cocaine.use cocaine.frequency crack.use crack.frequency heroin.use
1         0.1                 5         0              NA        0.1
  heroin.frequency hallucinogen.use hallucinogen.frequency inhalant.use
1             35.5              0.2                      52          1.6
  inhalant.frequency pain.releiver.use pain.releiver.frequency oxycontin.use
1                 19                 2                      36           0.1
  oxycontin.frequency tranquilizer.use tranquilizer.frequency stimulant.use
1                24.5              0.2                      52           0.2
  stimulant.frequency meth.use meth.frequency sedative.use sedative.frequency
1                   2        0             NA          0.2                  13
  class
1 youth
```

We can also get a summary of each variable to note the ranges of data we have, and if we have any NA values present:

```
     age                  n          alcohol.use    alcohol.frequency
 Length:17          Min.   :2223   Min.   : 3.90   Min.   : 3.00
 Class :character   1st Qu.:2469   1st Qu.:40.10   1st Qu.:10.00
 Mode  :character   Median :2798   Median :64.60   Median :48.00
                    Mean   :3251   Mean   :55.43   Mean   :33.35
                    3rd Qu.:3058   3rd Qu.:77.50   3rd Qu.:52.00
                    Max.   :7391   Max.   :84.20   Max.   :52.00
```

```
marijuana.use   marijuana.frequency  cocaine.use    cocaine.frequency
Min.   : 1.10   Min.   : 4.00       Min.   :0.000   Min.   : 1.000
1st Qu.: 8.70   1st Qu.:30.00       1st Qu.:0.500   1st Qu.: 5.000
Median :20.80   Median :52.00       Median :2.000   Median : 5.250
Mean   :18.92   Mean   :42.94       Mean   :2.176   Mean   : 7.875
3rd Qu.:28.40   3rd Qu.:52.00       3rd Qu.:4.000   3rd Qu.: 7.250
Max.   :34.00   Max.   :72.00       Max.   :4.900   Max.   :36.000
                                                    NA's   :1
  crack.use      crack.frequency   heroin.use     heroin.frequency
Min.   :0.0000   Min.   : 1.00   Min.   :0.0000   Min.   :  1.00
1st Qu.:0.0000   1st Qu.: 5.00   1st Qu.:0.1000   1st Qu.: 39.62
Median :0.4000   Median : 7.75   Median :0.2000   Median : 53.75
Mean   :0.2941   Mean   :15.04   Mean   :0.3529   Mean   : 73.28
3rd Qu.:0.5000   3rd Qu.:16.50   3rd Qu.:0.6000   3rd Qu.: 71.88
Max.   :0.6000   Max.   :62.00   Max.   :1.1000   Max.   :280.00
                 NA's   :3                        NA's   :1
hallucinogen.use hallucinogen.frequency  inhalant.use   inhalant.frequency
Min.   :0.100    Min.   : 2.000         Min.   :0.000   Min.   : 2.000
1st Qu.:0.600    1st Qu.: 3.000         1st Qu.:0.600   1st Qu.: 3.375
Median :3.200    Median : 3.000         Median :1.400   Median : 4.000
Mean   :3.394    Mean   : 8.412         Mean   :1.388   Mean   : 6.156
3rd Qu.:5.200    3rd Qu.: 4.000         3rd Qu.:2.000   3rd Qu.: 6.625
Max.   :8.600    Max.   :52.000         Max.   :3.000   Max.   :19.000
                                                        NA's   :1
pain.releiver.use pain.releiver.frequency oxycontin.use    oxycontin.frequency
Min.   : 0.600    Min.   : 7.00           Min.   :0.0000   Min.   : 3.00
1st Qu.: 3.900    1st Qu.:12.00           1st Qu.:0.4000   1st Qu.: 5.75
Median : 6.200    Median :12.00           Median :1.1000   Median :12.00
Mean   : 6.271    Mean   :14.71           Mean   :0.9353   Mean   :14.81
3rd Qu.: 9.000    3rd Qu.:15.00           3rd Qu.:1.4000   3rd Qu.:18.12
Max.   :10.000    Max.   :36.00           Max.   :1.7000   Max.   :46.00
                                                           NA's   :1
tranquilizer.use tranquilizer.frequency stimulant.use   stimulant.frequency
Min.   :0.200    Min.   : 4.50          Min.   :0.000   Min.   :  2.00
1st Qu.:1.400    1st Qu.: 6.00          1st Qu.:0.600   1st Qu.:  7.00
Median :3.500    Median :10.00          Median :1.800   Median : 10.00
Mean   :2.806    Mean   :11.74          Mean   :1.918   Mean   : 31.15
3rd Qu.:4.200    3rd Qu.:11.00          3rd Qu.:3.000   3rd Qu.: 12.00
Max.   :5.400    Max.   :52.00          Max.   :4.100   Max.   :364.00

   meth.use      meth.frequency    sedative.use    sedative.frequency
Min.   :0.0000   Min.   :  2.00   Min.   :0.0000   Min.   :  3.00
1st Qu.:0.2000   1st Qu.: 12.00   1st Qu.:0.2000   1st Qu.:  6.50
```

```
Median :0.4000   Median : 30.00   Median :0.3000   Median : 10.00
Mean   :0.3824   Mean   : 35.97   Mean   :0.2824   Mean   : 19.38
3rd Qu.:0.6000   3rd Qu.: 47.00   3rd Qu.:0.4000   3rd Qu.: 17.50
Max.   :0.9000   Max.   :105.00   Max.   :0.5000   Max.   :104.00
                 NA's   :2
   class
Length:17
Class :character
Mode  :character
```

We do seem to have a few missing values, but not very many so it doesn't seem to be a big problem for us.

Since alcohol is one of the more common drugs on this list, we can begin by looking at the proportion of individuals in each age group who have consumed alcohol in the past 12 months:

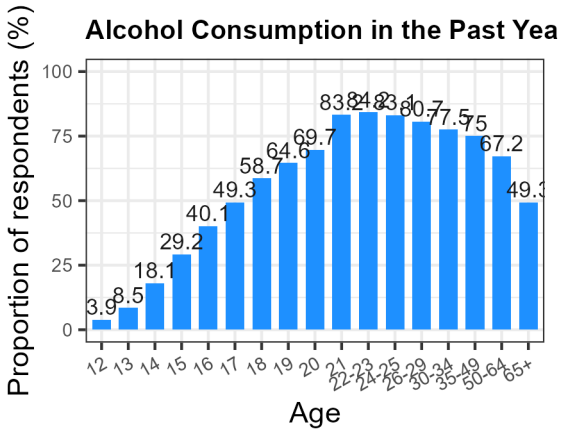


Figure 1: Proportion of individuals in each age group who have consumed alcohol in the past 12 months

We also applied the same approach to examine marijuana use over the same period:

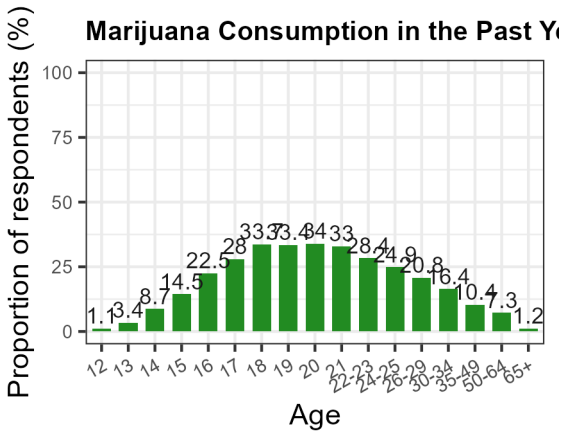


Figure 2: Proportion of individuals in each age group who have used marijuana in the past 12 months

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:

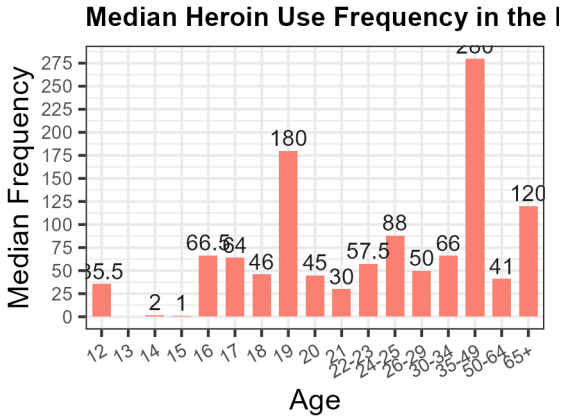


Figure 3: Proportion of individuals in each age group who have used heroin in the past 12 months

To explore the potential relationship between the frequency of heroin use and marijuana use

among individuals, we used a scatter plot of each age group's data points. We also added a linear regression line to help us determine whether there is any correlation between the two variables, and therefore whether heroin use is associated with marijuana use:

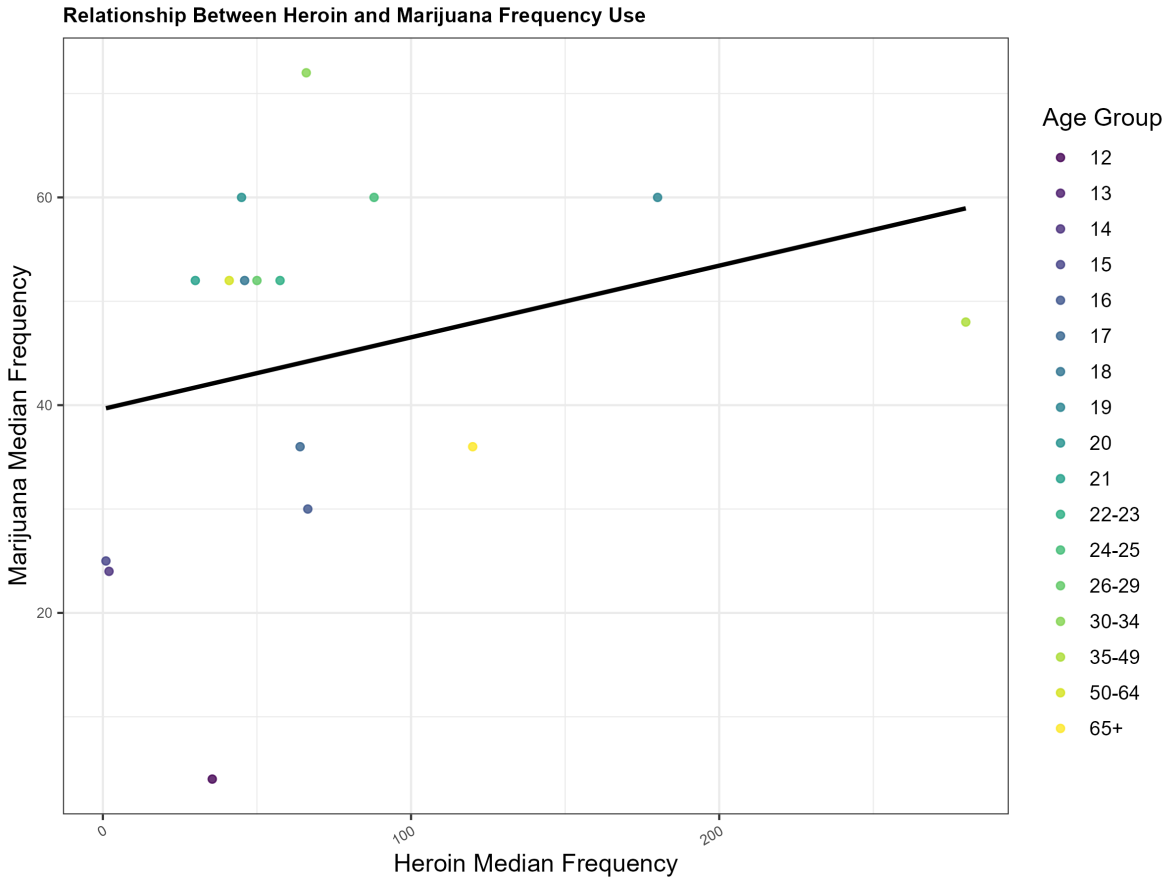


Figure 4: Relationship between frequency of heroin use vs. frequency of marijuana use

Finally, let's look at some EDA that compares the two main groups we are looking at: youth and adult.

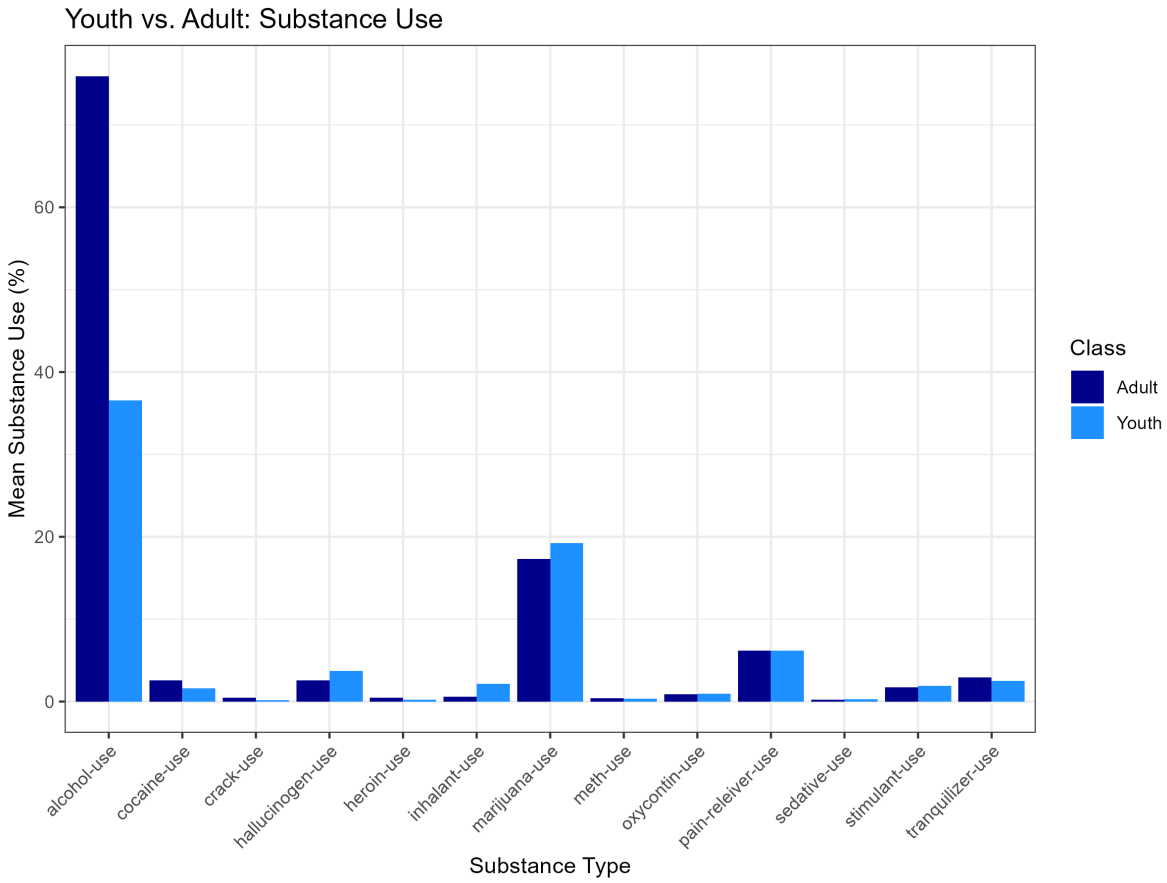


Figure 5: Comparing Substance Use between Adults and Youth

Some of these results are expected, such as the adults drinking more alcohol than the youth.

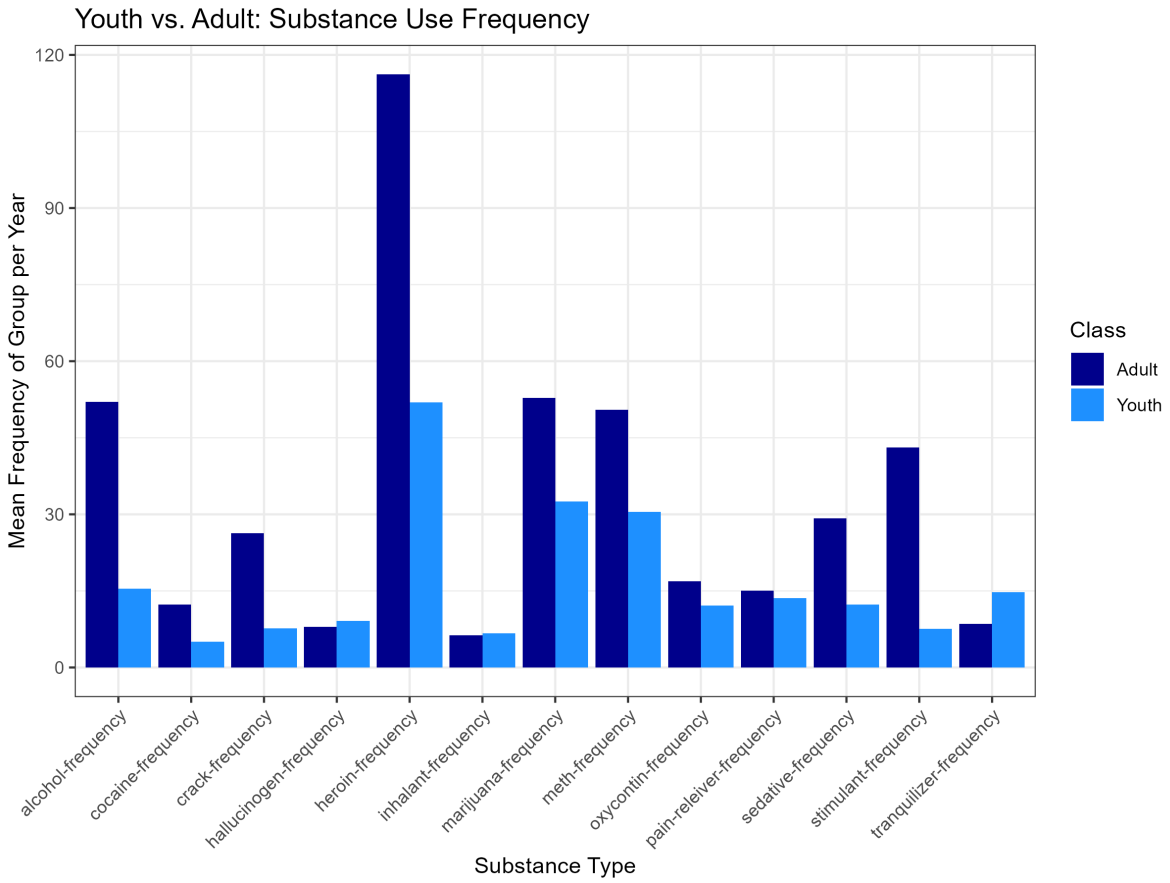


Figure 6: Comparing Frequency Use between Adults and Youth

In general, the adults seem to use more frequently throughout the year in comparison to the youth. There are some where this isn't the case, such as tranquilizer frequency, which is interesting to note.

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 an individual's age group (Under 21 or 21 & Over) based only on their reported substance use history. 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

First, since we are doing prediction, we did a train/test split on our data. This way, we can test and compare our models' ability to predict accurately.

We chose to run three different models on our data:

1. A Decision tree model
2. K-Nearest-Neighbors (KNN) Regression
3. Logistic Regression

This can be seen in the `scripts/04-analysis.R` file.

### Model Evaluation

After creating the models, we want to test their ability to predict the youth/adult class accurately. We do this by creating confusion matrices, where we can see a count of which model guesses the class correctly on our test data.

Below are the confusion matrices for each model we tested:

1. Decision Tree

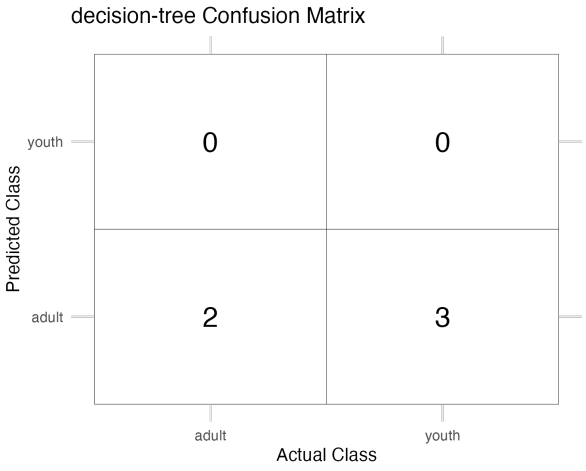


Figure 7: Confusion Matrix: Decision Tree

2. K Nearest Neighbours (KNN)

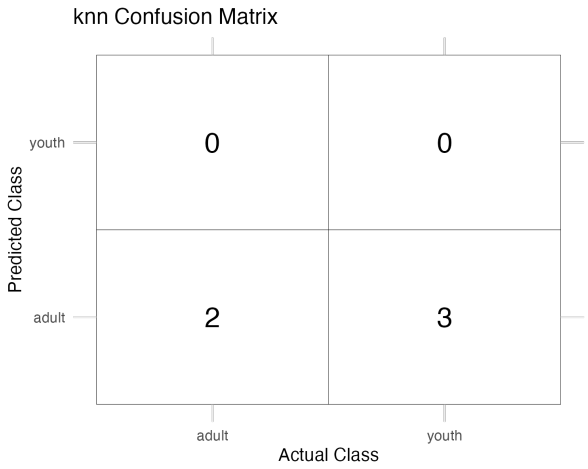


Figure 8: Confusion Matrix: KNN

3. Logistic Regression

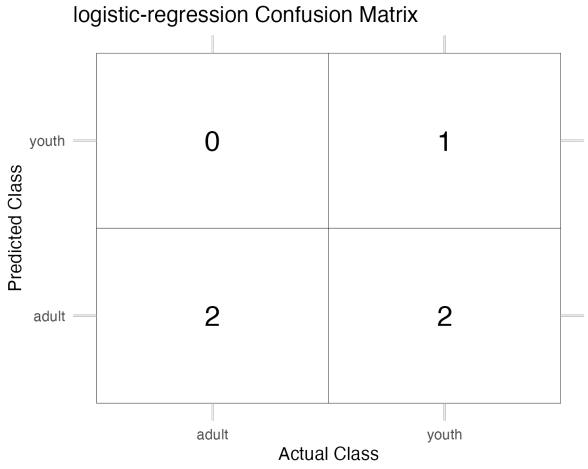


Figure 9: Confusion Matrix: Logistic Regression

Our models struggle to correctly classify youth vs. adult due to several key issues:

1. **Small Dataset Size**
   With only 17 rows, the models lack sufficient data to learn meaningful patterns. Machine learning models, like our decision tree, typically require more data to generalize well. Additionally, there are not many points for the regression models to work off of.

2. **Class Imbalance in Age Representation**
   The youth class consists of individual ages (e.g., 16, 17, 18), while the adult class consists of age ranges (e.g., 22-23, 25-49, 65+).
   Because of this, the models struggle to differentiate youth from adults, leading to false positives (misclassifying youth as adults).

3. **Lack of Feature Variation**
   If the key features (not including age) don't provide clear distinctions between youth and adult groups, the models may not have enough useful information to make accurate classifications. We do not have a wide variety of data, only information pertaining to drug use, and it is possible it may not be variable enough to accurately predict between the groups.

### Potential Solutions

1. **Increase Dataset Size**
   Collect more data to improve model performance. The Data is collected from a lot more individuals, and then aggregated to summarize between each age group. Using the original data would probably produce much better results.

2. **Feature Engineering**
   We could consider adding more/adjusting our features or restructuring age-related data.

3. **Resampling Techniques**
   Use a bootstrapping or synthetic sampling (such as SMOTE) technique to balance the class representation.

## Discussion

### Findings

Our models struggled with accurately classifying youth vs. adult due to several key issues identified in the analysis. The small dataset size likely hindered the models' ability to detect meaningful patterns. Additionally, the class imbalance between individual youth ages and adult age ranges made it difficult for the models to generalize and differentiate between these groups effectively. This was evident in the high rates of false positives, where youth were misclassified as adults. Lastly, the lack of feature variation beyond age likely contributed to the models' poor performance, as age alone may not be sufficient to distinguish between youth and adult categories.

We saw that our models were able to correctly predict the adult class, yet often made errors when predicting the youth class. However, since we have such few data points, we cannot be confident that the model is accurately predictign the adult class, or if it was more just luck.

We need more data points not only to train our model better, but also test our model and be confident it is a good fit. If we were to continue with this analysis in the future, we would likely need to find the dataset that this one aggregated the information from, that way we can build our models and understanding.

### Does This Align with Our Expectations?

Yes, these findings align with our expectations. Given the small dataset and the imbalance between the youth and adult classes, we anticipated that our models would struggle with classification accuracy. The confusion matrices for all three models-Decision Tree, KNN, and Logistic Regression- reflected the challenges we expected. The small dataset size likely exacerbated the issue, and the class imbalance was a known factor that we hypothesized would affect performance.

### Impact of Our Findings

These findings highlight the limitations of working with small datasets, particularly when dealing with imbalanced classes. In practical terms, the findings suggest that models built on such data may not be reliable for decision-making in contexts where accurate classification is crucial (e.g., youth-focused interventions or adult-targeted policies). The results also emphasize the importance of addressing class imbalance and ensuring that models are trained on sufficiently large and diverse datasets.

### Future Directions

This analysis raises several future questions:

1. **Expanding the Dataset**
   Revisiting the original "Drug Use by Age" dataset from FiveThirtyEight's repository, derived from the 2013 National Survey on Drug Use and Health, could provide deeper insights into substance use trends across age groups. With more time and resources, we could have leveraged the full dataset to increase sample size, address class imbalance, and engineer more features for improved model performance.
2. **Addressing Class Imbalance**
   To improve model performance, we could explore advanced resampling techniques like SMOTE or cost-sensitive learning, which would help balance the dataset and reduce the impact of class imbalance.

3. **Exploring Additional Features**
   Incorporating other relevant features, such as socio-economic factors, education level, or behavioral traits, could enhance classification accuracy and provide a better understanding of the distinctions between youth and adult groups.

## 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/.