# Project proposal

Team 20

```
library(knitr)
library(tidyverse)
library(broom)
library(htmltools)
```

## 1. Introduction

The research question we will focus on is the characteristics of an individual using hard drugs. In general, many people don't understand why or how other people become addicted to drugs. They may mistakenly think that those who use drugs lack moral principles or willpower and that they could stop their drug use simply by choosing to. In reality, drug addiction is a complex disease, and quitting usually takes more than good intentions or a strong will.

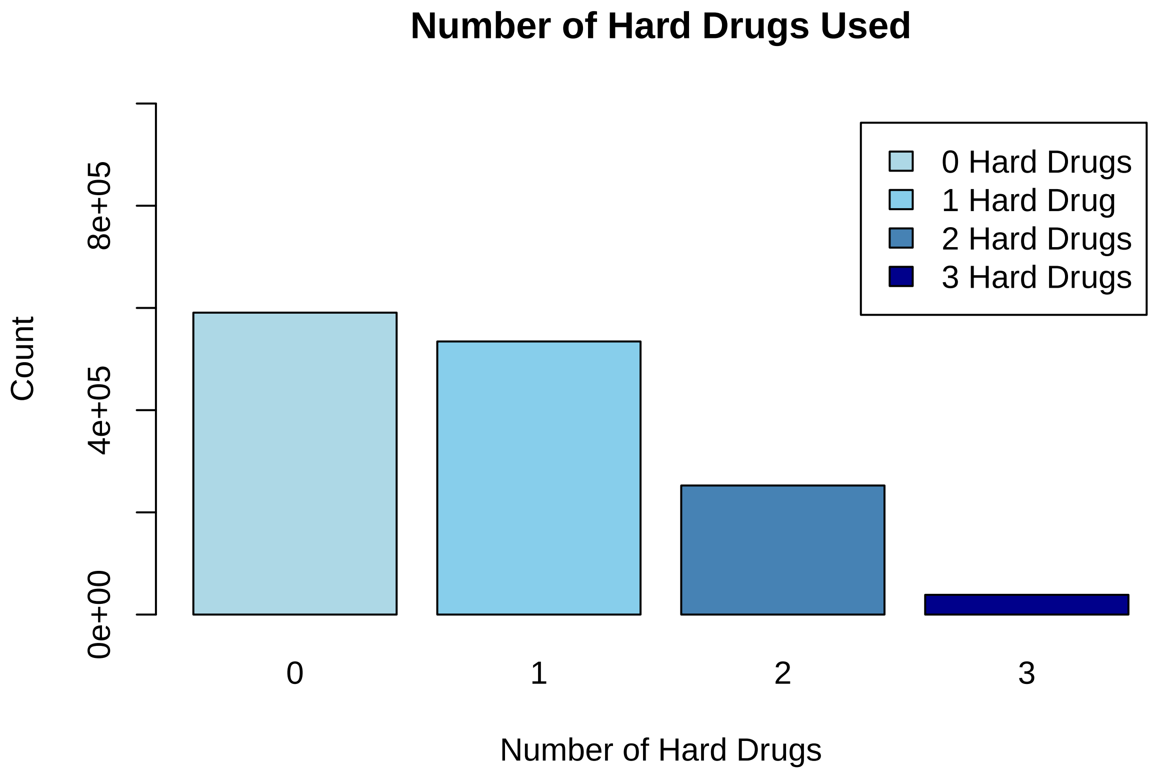
A significant portion of the global population uses hard drugs. From various research studies and articles we found online, it is possible to link the use of hard drugs to several reasons, including the individual's mental state, their physical condition, and the background from which they came. There are many dangers associated with the use of hard drugs, and over the years, accessibility to them has been increasing among various populations. The goal of our research is to understand the underlying factors contributing to addiction to hard drugs in order to construct a profile of individuals who use these substances. We aim to focus on different vulnerable populations to combat the phenomenon in a targeted and widespread manner. One of the reasons why understanding the characteristics of individuals using hard drugs can be challenging is the complex nature of drug addiction itself. Drug addiction is a multifaceted issue influenced by various biological, psychological, and social factors. These factors interact in unique ways for each individual, making it difficult to generalize or predict patterns of drug use and addiction.

Our approach is to find correlations between different variables and, based on that, build a logistic regression model that characterizes the addicted/user of hard drugs.

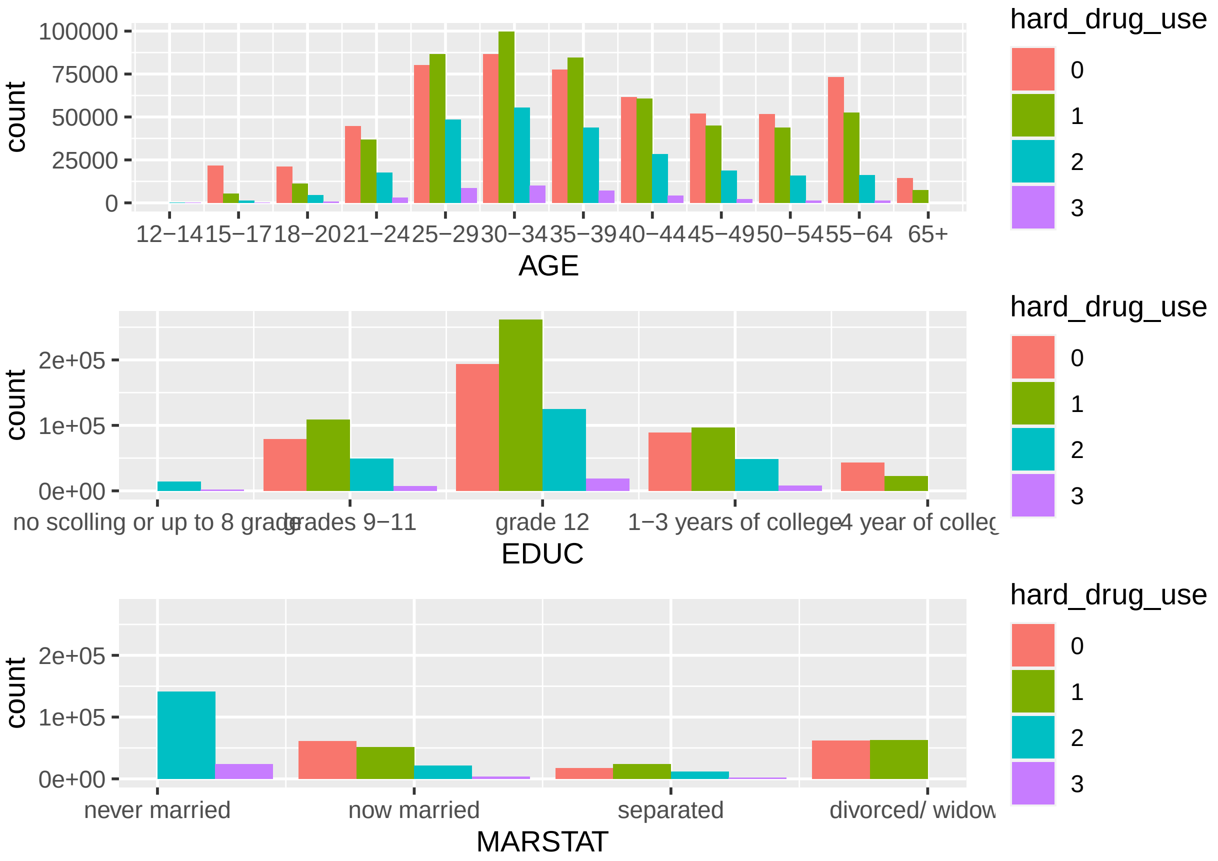
## 2. Data

The data set contains information on the admission of individuals aged 12 years or older, including their demographic characteristics such as age, gender, race, ethnicity, employment status, and substance use patterns. This data was collected in 2020 in the United States. Specifically, we will focus on variables related to the person's characteristics, such as their gender, age at admission, marital status, education, employment status, living arrangements, and source of income. We will also examine variables related to criminal justice involvement, including arrests in the past 30 days and criminal justice referrals. Additionally, we will focus on the primary, secondary, and tertiary substances used by the individual, as well as the route and frequency of use and age at first use. Furthermore, the dataset contains columns that explicitly indicate whether the individual declared the use of particular drugs upon admission.

## 3. Preliminary results



The graph generated illustrates the number of admissions categorized by the number of hard drugs used (0, 1, 2, or 3). Moving forward, we will conduct a thorough analysis to determine whether the use of a single hard drug is sufficient to classify an admission as a hard drug user. For the purposes of this study, hard drugs are defined as including cocaine/crack, heroin, other opiates and synthetics, PCP, hallucinogens, methamphetamine/speed, other stimulants, benzodiazepines, inhalants (e.g. gasoline, glue), and other drugs (e.g. ketamine). It is worth noting that the data may also include admissions for mental health-related concerns, which may involve the use of over-the-counter medication and calming pills. However, we did not consider these substances to be hard drugs for the purpose of this analysis.



The first graph portrays the distribution of hard drug use among different age groups, with the highest number of admissions observed in the 30-34 age group. This age group also exhibits the highest incidence of using 1, 2, or 3 hard drugs, while across all age groups, the most prevalent category is individuals who do not use any hard drugs, followed by those who use only one hard drug.

Moving on to the Education graph, the majority of the dataset comprises individuals who have completed 12 years of school. Additionally, individuals with the lowest education levels display a higher propensity to use hard drugs, while those with higher levels of education are less likely to use hard drugs or limit themselves to using only one type, by the information given at admission.

Finally, the last graph pertaining to marital status reveals that individuals who have never been married have the highest incidence of using hard drugs. Conversely, the groups that include individuals who are divorced or widowed tend to use only one hard drug or do not use any hard drugs at all.

```
## # A tibble: 4 x 3
## # Groups:   hard_drug_use [4]
##   hard_drug_use PRIMINC percentage
##   <fct>           <dbl>      <dbl>
## 1 0                   1       43.5
## 2 1                   5       42.2
## 3 2                   5       45.4
## 4 3                   5       47.2
```

The following table shows the percentages of the most common income groups among different subsets of hard drug admissions. It can be seen that the most common group that does not use hard drugs at all is the group of people with salaries, accounting for approximately 43.47% of this group. Additionally, it can be observed that admissions which have no source of income use hard drugs.

## 4. Data analysis plan

Week 1 -

1. Our top priority for the first week is to meticulously clean and refine the data set, as many states have submitted records that include multiple admissions for the same client. Additionally, we need to address the issue of missing values in the data set, as this can significantly impact our findings and conclusions. By the initial analysis we made, we can see that there are groups that act as a majority in the factors we have chosen. more over, we think that other variables in columns may have bias because of Simpson paradox, and we will need to normalize the data.
2. We will schedule a meeting with the primary researcher responsible for the data set to gain a more precise understanding of the primary research question and to ensure that our approach is aligned with the desired outcomes.
3. We will conduct thorough cross-referencing of our data set with external sources of information regarding hard drugs. This is necessary as the data set includes drugs that we are unsure of whether they are classified as hard drugs, necessitating further exploration and research.

Week 2 -

4. We will conduct correlation analyses to identify relationships between our variables. This process must be repeated multiple times for each variable, including each hard drug, age at start, status of living, and other relevant variables. We will compare and analyze the correlations between the variables and utilize them to build a reliable and effective logistic regression model.

Week 3 -

5. We will construct a detailed and comprehensive presentation and compose a thoroughly researched report outlining our findings and conclusions.

We believe that it is important to work collaboratively on this project since the data does not allow us to work independently and necessitates a team approach.

## Appendix

### Data README

```
# SISE2601 Project data description
================
Team name

This Markdown file describes the data folder structure and organization ...
```

### Source code

```
library(knitr)
library(tidyverse)
library(broom)
library(htmltools)
```

```r
opts_chunk$set(echo=FALSE) # hide source code in the document
knitr::opts_chunk$set(warning = FALSE, message = FALSE)

drugs <- read_csv("TEDSA_PUF_2020.csv")
# Create a vector of hard drug codes
hard_drug_codes <- c(3, 5, 7, 8, 9, 10, 12, 13,17,19)

# Define a function to count the number of hard drugs used by each individual
count_hard_drugs <- function(row) {
  count <- sum(row %in% hard_drug_codes)
  if (count > 3) count <- 3  # Limit the count to 3 (for 3 or more hard drugs)
  return(count)
}

# Apply the function to each row in the data
hard_drug_counts <- apply(drugs[, c("SUB1", "SUB2", "SUB3")], 1, count_hard_drugs)

# Count the occurrences of each count value
count_summary <- table(hard_drug_counts)

# Create a vector for the labels
labels <- c("0 Hard Drugs", "1 Hard Drug", "2 Hard Drugs", "3 Hard Drugs")

# Plot the stacked bar chart
barplot(count_summary, main = "Number of Hard Drugs Used",
        xlab = "Number of Hard Drugs", ylab = "Count", col = c("lightblue", "skyblue", "steelblue", "da
        legend = labels, beside = FALSE , ylim = c(0,1000000))
library(gridExtra)
drugs <- drugs %>%
  mutate(hard_drug_use = factor(hard_drug_counts, levels = 0:3))

age <- ggplot(drugs, aes(x = AGE, fill = hard_drug_use)) + geom_bar(position = "dodge") + scale_x_contin
education <- ggplot(drugs, aes(x = EDUC, fill = hard_drug_use)) + geom_bar(position = "dodge") + scale_
marital <- ggplot(drugs, aes(x = MARSTAT, fill = hard_drug_use)) + geom_bar(position = "dodge") + scale_
grid.arrange(age,education,marital,ncol = 1)
result <- drugs %>%
  filter(PRIMINC != -9) %>%
  group_by(hard_drug_use, PRIMINC) %>%
  summarise(count = n()) %>%
  group_by(hard_drug_use) %>%
  mutate(total_count = sum(count),
         percentage = count / total_count * 100) %>%
  filter(count == max(count)) %>%
  select(-count, -total_count)

print(result)
cat(readLines('../data/README.md'), sep = '\n')
#ADMYR - year of admission

#AGE - age at admission

#GENDER - this field identifies the client's biological sex
#RACE - this field identifies the client's race

#ETHNIC - this field identifies client's specific Hispanic or Latino origin, if applicable

#MARSTAT - this field describes the client's marital status. The following categories are compatible wi

#EDUC - This field specifies a) the highest school grade completed for adults or children not attending

#EMPLOY - this field identifies the client’s employment status.

#DETNLF - provides more detailed information about those clients who are coded as '04 Not in labor forc

#PREG - this field indicates whether a female client was pregnant at the time of admission.

#VET - this field indicates whether the client has served in the uniformed services.

#LIVARAG - Identifies whether the client is homeless, a dependent or living independently on his or her

#PRIMINC - this field identifies the client’s principal source of financial support.

#ARRESTS - indicates the number of arrests in the 30 days prior to the reference date.

#STFIPS - state FIPS codes consistent with those used by the U.S.

#REGION - geographic regions used are based on divisions used by the U.S. Census Bureau, with the addit

#DIVISION - census divisions are groupings of states that are subdivisions of the four Census regions.

#SERVICES - this field describes the type of treatment service or treatment setting in which the client

#METHUSE - this field identifies whether the use of opioid medications such as methadone, buprenorphine

#DAYWAIT - indicates the number of days from the first contact or request for a substance use treatment

#PSOURCE - this field describes the person or agency referring the client to treatment.

#DETCRIM - this field provides more detailed information about those clients who are coded as '07 Crimi

#NOPRIOR - indicates the number of previous treatment episodes the client has received in any substance

#SUB1 - This field identifies the client's primary substance use,.

#ROUTE1 - this field identifies the usual route of administration of the corresponding substance identi

#FREQ1 - Specifies the frequency of use of the corresponding substance identified in Substance Use.

#FRSTUSE1 - for alcohol use, this is the age of first intoxication. For substances other than alcohol,

#SUB2 - this field identifies the client's secondary substance use, same values as SUB1.

#ROUTE2 - this field identifies the usual route of administration of the corresponding substance identi

#FREQ2 - specifies the frequency of use of the corresponding substance identified in Substance Use, sam
```

```
#FRSTUSE2 - or alcohol use, this is the age of first intoxication. For substances other than alcohol, t

#SUB3 - this field identifies the client's secondary substance use, same values as SUB1.

#ROUTE3 - this field identifies the usual route of administration of the corresponding substance identi

#FREQ3 - specifies the frequency of use of the corresponding substance identified in Substance Use, sam

#FRSTUSE3 - or alcohol use, this is the age of first intoxication. For substances other than alcohol, t

#IDU - flag records if at least one valid primary, secondary, or tertiary substance was reported and if

#ALCFLG - flag records if alcohol was reported as the primary, secondary, or tertiary substance at the

#COKEFLG - flag records if cocaine or crack was reported as the primary, secondary, or tertiary substan

#MARFLG - flag records if marijuana or hashish were reported as the primary, secondary, or tertiary sub

#HERFLG - flag records if heroin was reported as the primary, secondary, or tertiary substance at the t

#METHFLG - flag records if non-prescription methadone was reported as the primary, secondary, or tertia

#OPSYNFLG - flag records if other opiates or synthetics were reported as the primary, secondary, or ter

#PCPFLG - flag records if PCP was reported as the primary, secondary, or tertiary substance at the time

#HALLFLG - flag records if hallucinogens were reported as the primary, secondary, or tertiary substance

#MTHAMFLG - flag records if methamphetamine/speed was reported as the primary, secondary, or tertiary s

#AMPHFLG - flag records if other amphetamines were reported as the primary, secondary, or tertiary subs

#STIMFLG - flag records if other stimulants were reported as the primary, secondary, or tertiary substa

#BENZFLG - flag records if benzodiazepines were reported as the primary, secondary, or tertiary substan

#TRNQFLG - flag records if other tranquilizers were reported as the primary, secondary, or tertiary sub

#BARBFLG - flag records if barbiturates were reported as the primary, secondary, or tertiary substance

#SEDHPFLG - flag records if other sedatives or hypnotics were reported as the primary, secondary, or te

#INHFLG - flag records if inhalants were reported as the primary, secondary, or tertiary substance at t

#OTCFLG - flag records if over-the-counter medications were reported as the primary, secondary, or tert

#OTHERFLG - Flag records if other substances were reported as the primary, secondary, or tertiary subst

#ALCDRUG - classifies client's substance use type as alcohol only, other drugs only, alcohol and other

#DSMCRIT - client's diagnosis is used to identify the substance use problem that provides the reason fo

#PSYPROB - indicates whether the client has co-occurring mental and substance use disorders.
```

```
#HLTHINS - this field specifies the client's health insurance at admission.

#PRIMPAY - this field identifies the primary source of payment for this treatment episode anticipated a

#FREQ_ATND_SELF_HELP - This field indicates the frequency of attendance at a substance use self-help gr
```