Project proposal

Team name: 7, El Puto's

1. Introduction

Research Question: Are there distinct differences between drug addicts in the US who are submitted to rehab multiple times compared to drug addicts who are submitted to rehab for the first time?

Our objective is to create a profile and identify similarities and differences between individuals who have participated in rehabilitation programs multiple times and those who are participating for the first time. Our research aims to provide a better understanding of these groups and match them with more appropriate treatments based on their group profile and tendencies.

This analysis could make a significant contribution to the mental health and drug addiction field. It can be challenging to match the right treatment or facility with different addicts. Our vision is that if we can create a distinct profile that affects the treatment given to the individual, funds can be used more accurately. This could lead to the creation of hospitals that are proficient with first-time patients and others that are proficient with returning drug addicts.

This problem is significant because many recovered drug addicts are constantly relapsing and are unable to leave the drug cycle. If we can help in improving the treatment they receive, we may be able to help them recover for good and live a better life.

Creating this profiling is difficult for many reasons, the first being the credibility of the data, as many addicts might withhold information on their admission to the rehab facility (for personal reasons, skepticism in the process and in their anonymity etc.) Moreover, the data is not a serial data, meaning we can't use prior admissions to rehab for returning addict, a data segment that could have been helpful to our cause.

Given that we have the admissions and discharges data from these facilities, we wish to create a tree model that can classify our rows based on the number of admissions to rehab a patient has, hopefully our tree will catch accurately as many groups as possible. Our groups will mainly focus on the number of times an addict was admitted to rehab, but might also take into consideration the patient's preferable drug type, and / or their mental health status upon admission. Than, we wish to point out the most significant indicators for each group (class) using the SHAP library

2. Data

Our data set holds records of admissions to rehab facilities over the US during the year of 2020. The data has both features from before admission to the rehab facility, from the day of admission and of the treatment offered to the patient. In our analysis we chose to focus on data from before and from the day of admission as personal features that might show differences in different groups in our research question. Some of the original features were too specific to our purpose, so we created 8 calculated features that use the original data and add to it more suitable information. All the different features used for our work are explained, both in meaning and in calculation method, and de-coded in the README file attached as an appendix. The original data set in encoded and all variables are numeric-labeled.

Quick explanation on the calculated features:

- CALC_RACE - Categorical field for the RACE feature.
- STATE_PARTY - The main party voted for in the patient's state presidential elections.
- SUBSTANCE - categorical value, the type of most meaningful substance a patient is consuming (method in README). - ADDICTIVE_LEVEL - The highest addictive level (relative) for a substance consumed by the patient.
  - FREQ - The highest frequency of drug consumption by the patient (self reported).
  - FRSTUSE_LEGAL , FRSTUSE_ILLEGAL - The age of the first use in legal or illegal drug
  - UNDER_INFLUENCE - Were drugs found in the patients body during admission?

After creating the calculated features and focus on the data features relevant to us (in the README file) the data has been filtered. The purpose of the filter is to ignore rows that might confuse our model. Our project focuses on creating a relative ID for different groups (and is not necessarily a prediction tool), so incomplete or ambiguous data is ignored. We focused on a few issues (index is for identification in the code):

1. (1) Addicts that are addicted to 2 or more illegal drug types (uppers/ downers, over the counter) might compromise our profile, that takes into consideration the substance used. They are currently marked as SUNDAY_FUNDAY in SUBSTANCE column.
2. (2) Null values (or unknown values) in original features: GENDER,MARSTAT, EDUC, EMPLOY, LIVARAG, PSOURCE, PSYPROB, NOPRIOR were removed.
3. (3) Null values in calculated features: SUBSTANCE, FREQ, and in both FRSTUSE_LEGAL and FRSTUSE_ILLEGAL were removed.

Our final data set has 23 features (without caseID) and 579254 rows.

3. Preliminary results

Exploratory data analysis, including some summary statistics and visualizations, along with some explanation on how they help you learn more about the problem. Obviously, you will add/implement more analysis as you work on your final project, but for the proposal stage, we want to see that project you're proposing is viable (and reasonable) to accomplish within the Class's time frame.

On this section, our purpose is to show a basic analysis of our final data set created above, as well as to show a few important relationships in the data. First let's show statistical information on the final features, and a glimpse into their first few values.

```
## [1] 579254     24
```

```
##      CASEID             AGE             GENDER        CALC_RACE
##  Min.   :      1   Min.   : 1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.: 306914   1st Qu.: 5.000   1st Qu.:1.000   1st Qu.:2.000
##  Median : 550939   Median : 7.000   Median :1.000   Median :2.000
##  Mean   : 578344   Mean   : 7.205   Mean   :1.314   Mean   :1.931
##  3rd Qu.: 820290   3rd Qu.: 9.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :1416638   Max.   :12.000   Max.   :2.000   Max.   :3.000
##
##     MARSTAT           EDUC       FRSTUSE_LEGAL    FRSTUSE_ILLEGAL
##  Min.   :1.000   Min.   :1.000   Min.   :-9.00    Min.   :-9.00
##  1st Qu.:1.000   1st Qu.:3.000   1st Qu.: 2.00    1st Qu.: 3.00
##  Median :1.000   Median :3.000   Median : 3.00    Median : 4.00
##  Mean   :1.715   Mean   :3.054   Mean   : 2.62    Mean   : 4.09
##  3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.: 4.00    3rd Qu.: 6.00
##  Max.   :4.000   Max.   :5.000   Max.   : 7.00    Max.   : 7.00
##                                  NA's   :200822   NA's   :213534
##      EMPLOY          PREG            VET         STATE_PARTY       LIVARAG
##  Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :0.000   Min.   :1.000
##  1st Qu.:2.00   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000
##  Median :3.00   Median :2.000   Median :2.000   Median :2.000   Median :3.000
##  Mean   :2.83   Mean   :1.992   Mean   :1.961   Mean   :1.398   Mean   :2.535
##  3rd Qu.:4.00   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:3.000
##  Max.   :4.00   Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :3.000
##
##     PRIMINC           ARRESTS           PSOURCE         DSMCRIT
##  Min.   :-9.0000   Min.   :0.00000   Min.   :1.000   Min.   :-9.000
##  1st Qu.: 1.0000   1st Qu.:0.00000   1st Qu.:1.000   1st Qu.: 4.000
##  Median : 2.0000   Median :0.00000   Median :2.000   Median : 5.000
##  Mean   : 0.7072   Mean   :0.06609   Mean   :3.391   Mean   : 4.733
##  3rd Qu.: 5.0000   3rd Qu.:0.00000   3rd Qu.:7.000   3rd Qu.: 8.000
##  Max.   : 5.0000   Max.   :2.00000   Max.   :7.000   Max.   :19.000
##
##  FREQ_ATND_SELF_HELP UNDER_INFLUENCE   SUBSTANCE     ADDICTIVE_LEVEL
##  Min.   :-9.0000     Min.   :0       Min.   :0.000   Min.   :1.000
##  1st Qu.: 1.0000     1st Qu.:1       1st Qu.:0.000   1st Qu.:3.000
##  Median : 1.0000     Median :1       Median :2.000   Median :3.000
##  Mean   : 0.9032     Mean   :1       Mean   :1.443   Mean   :3.184
##  3rd Qu.: 1.0000     3rd Qu.:1       3rd Qu.:3.000   3rd Qu.:4.000
##  Max.   : 5.0000     Max.   :1       Max.   :3.000   Max.   :4.000
##
##       FREQ           PSYPROB         NOPRIOR
```

```
##  Min.   :-9.000   Min.   :1.000   Min.   :0.000
##  1st Qu.: 2.000   1st Qu.:1.000   1st Qu.:0.000
##  Median : 2.000   Median :2.000   Median :1.000
##  Mean   : 2.222   Mean   :1.522   Mean   :1.821
##  3rd Qu.: 3.000   3rd Qu.:2.000   3rd Qu.:3.000
##  Max.   : 3.000   Max.   :2.000   Max.   :5.000
##
```

We are taking into consideration that our outcome might also include separate profiles based on the calculated SUBSTANCE feature and the original PSYPROB feature, as the patient's mental health and drug type might also have a major influence on his/her profile. In order to decide, correlation tests will be conducted between these 3 parameters. The goal will be to see if there is a notable correlation between SUBSTANCE <-> NOPRIOR and PSYPROB <-> NOPRIOR. If True, they will be used as explanatory variables for our model. if False, they will have to be represented in a different class and to be given with their own profile.

First relationship: NOPRIOR <-> PSYPROB

As in the README, here is the decoded data labels:

0 - No priors

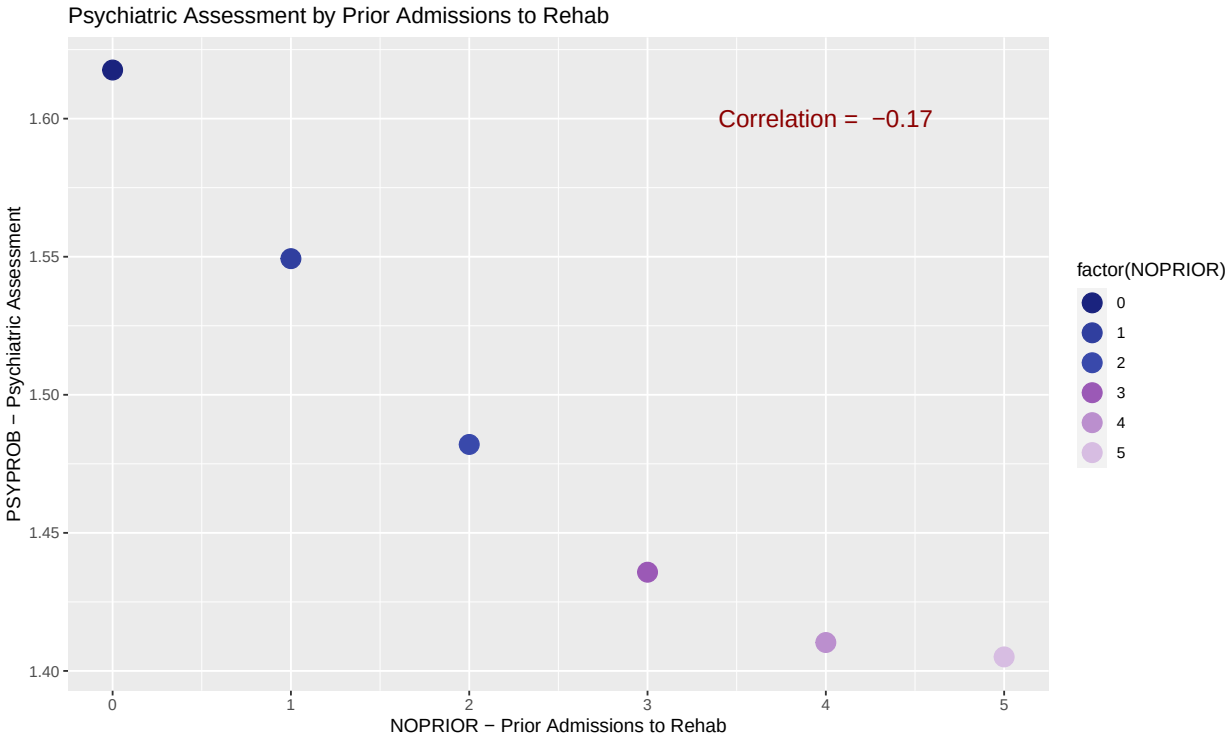
1 - 1 prior admission

2 - 2 prior admission

3 - 3 prior admission

4 - 4 prior admission

5 - 5 or more prior admission.



It appears that there is a medium negative correlation between NOPRIOR <-> PSYPROB so PSYPROB will be treated as an explanatory, X variable.

Second relationship: NOPRIOR <-> SUBSTANCE

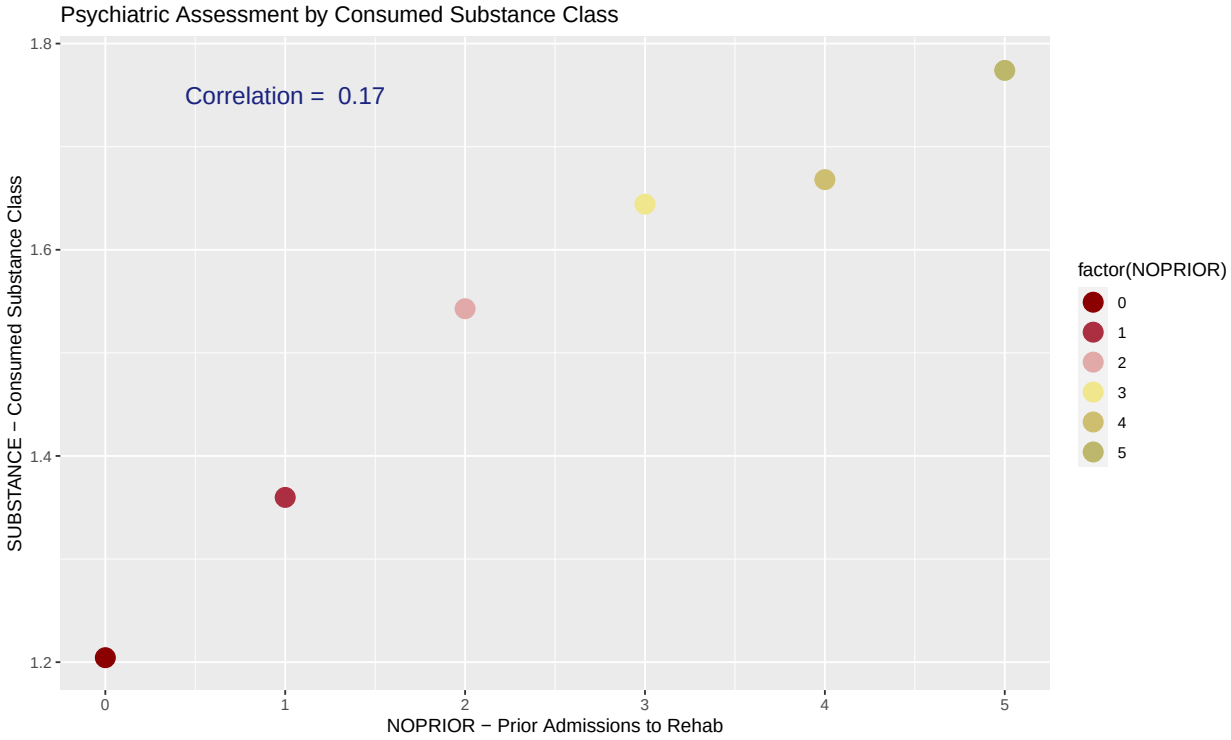
As in the README, here are the decoded data labels:

0 - legal in country

1 - over the counter prescriptions / other

2 - Stimulant

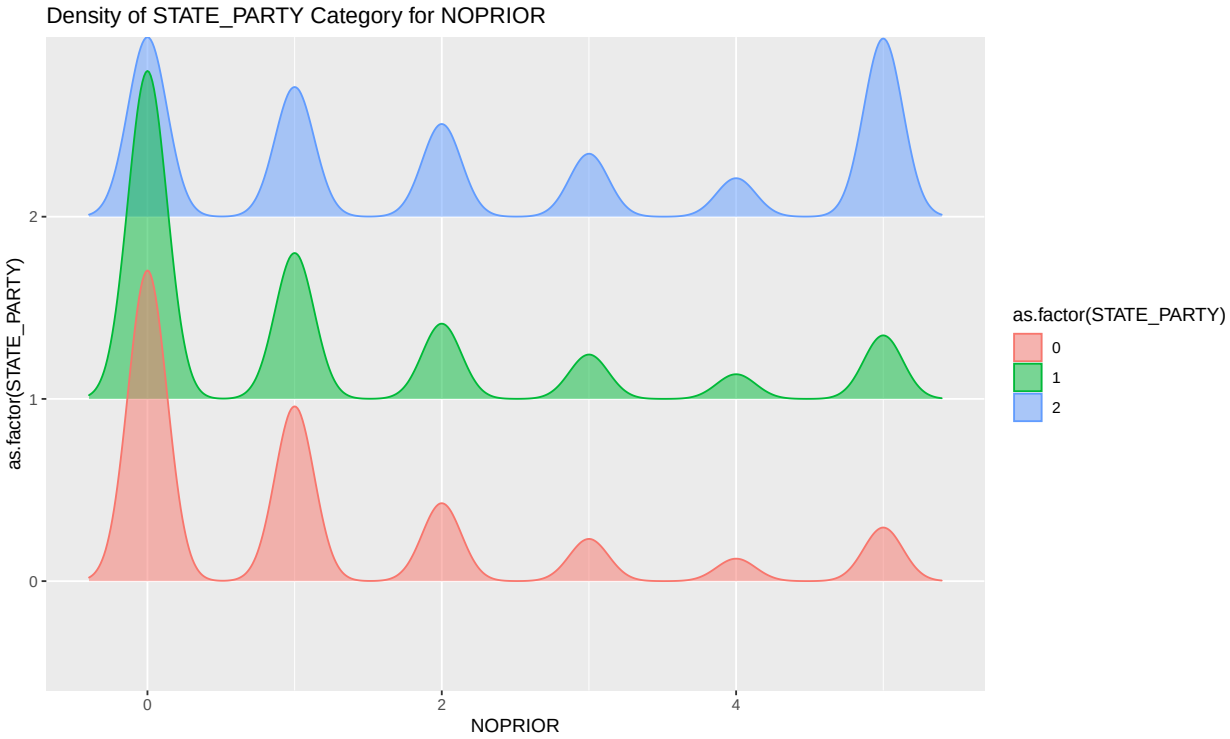
3 - Depressant

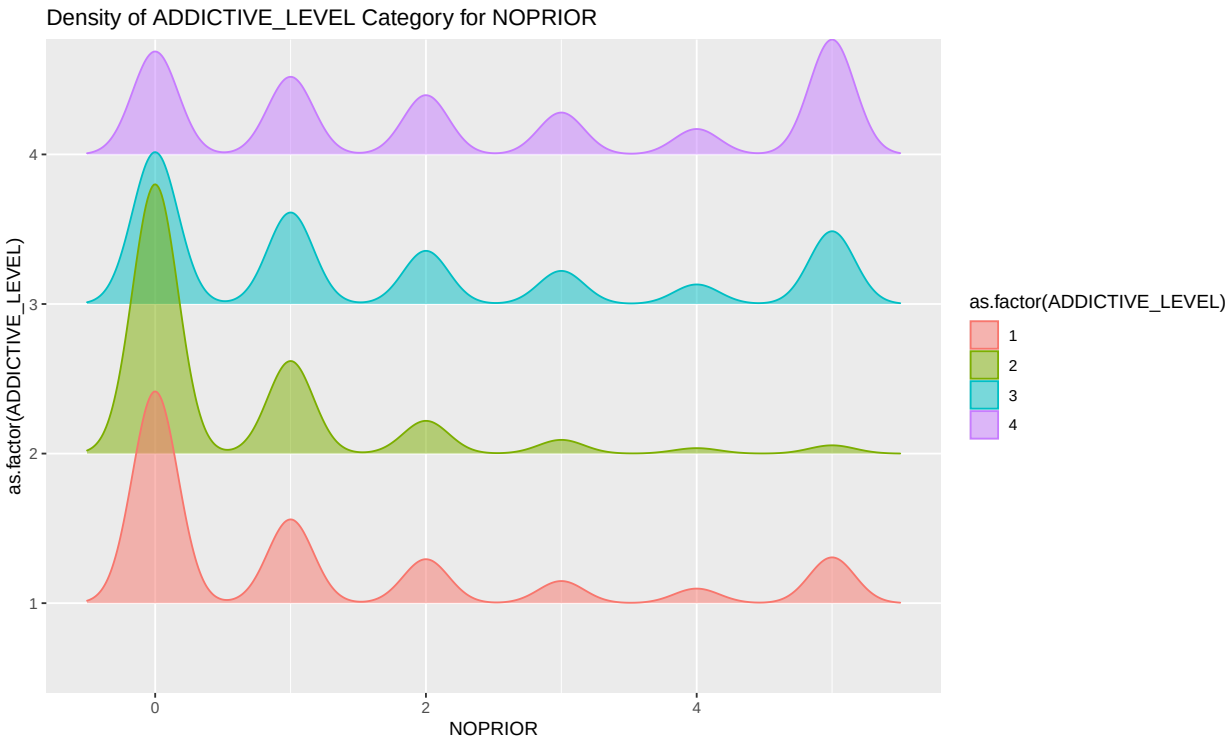


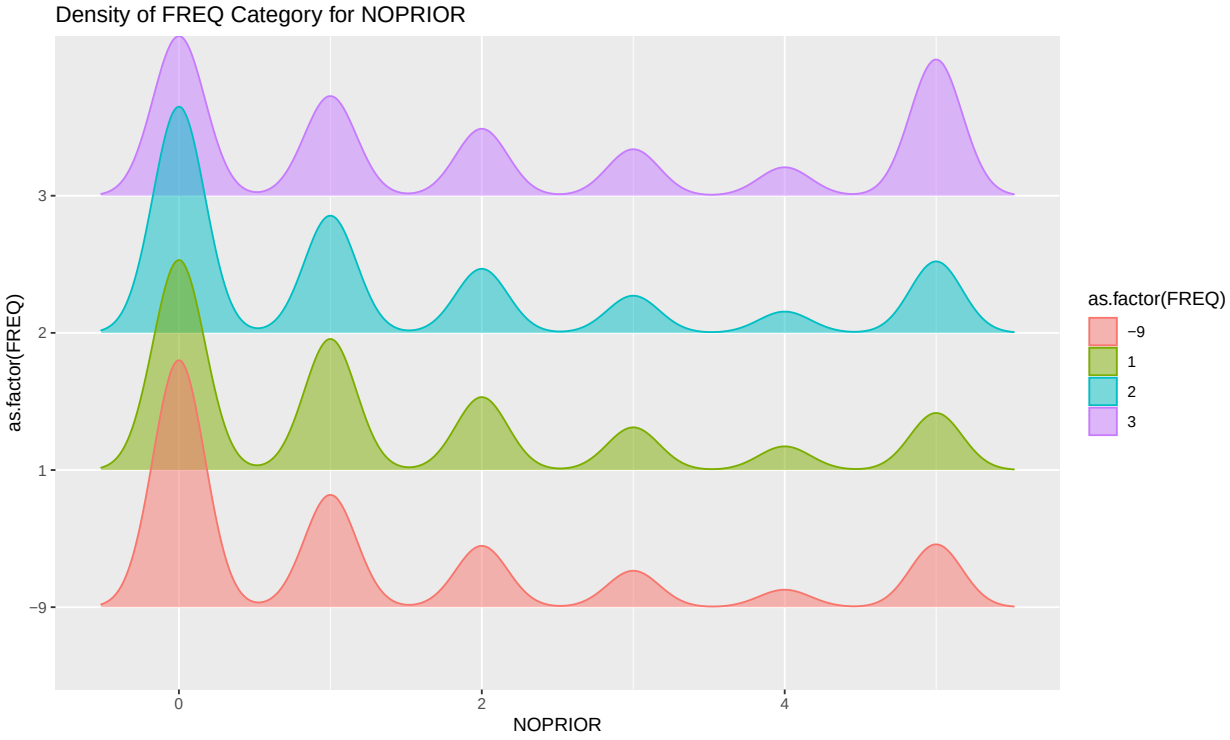
It appears that there is a medium positive correlation between NOPRIOR <-> SUBSTANCE so SUBSTANCE will be treated as an explanatory, X variable.

next, we'll show how the data disperses in a few of our chosen variables, to get a preliminary idea on the biggest most common features and their values in the set.

Categorical Features <-> NOPRIOR (Number of Prior Admissions to Rehab)

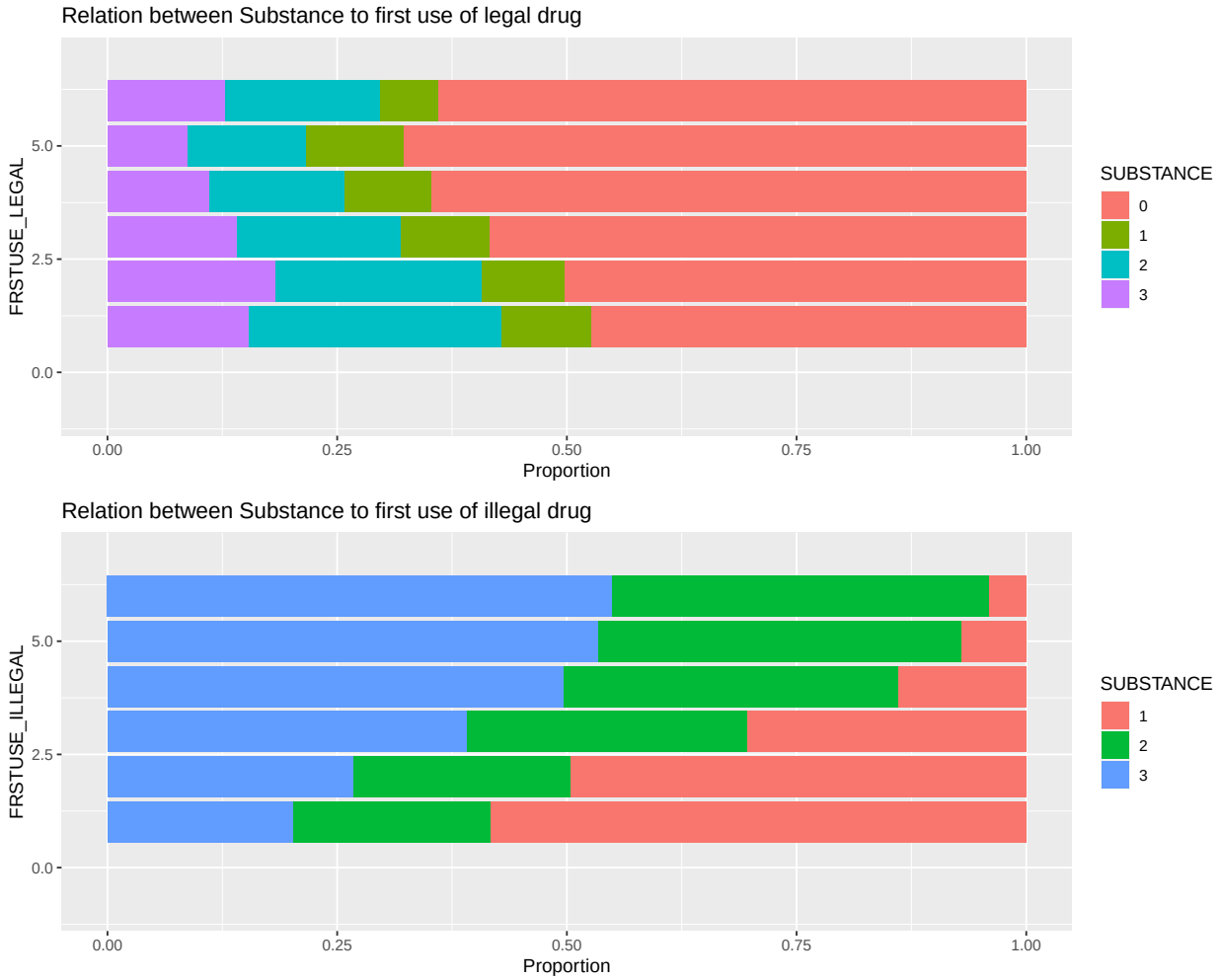






A few quick conclusions (remember, decoding of the Y axis's is in the README file): - We see that in democrat states (2) there are much more patients in 5th admission (and more), relative to the group size. One might think that citizens in republican states (0) "give up" after the first rehab. - We clearly see that the more addictive the consumed drug is, the greater is the number of admissions to 5th rehab (and more) - The frequency of use has a similar effect.

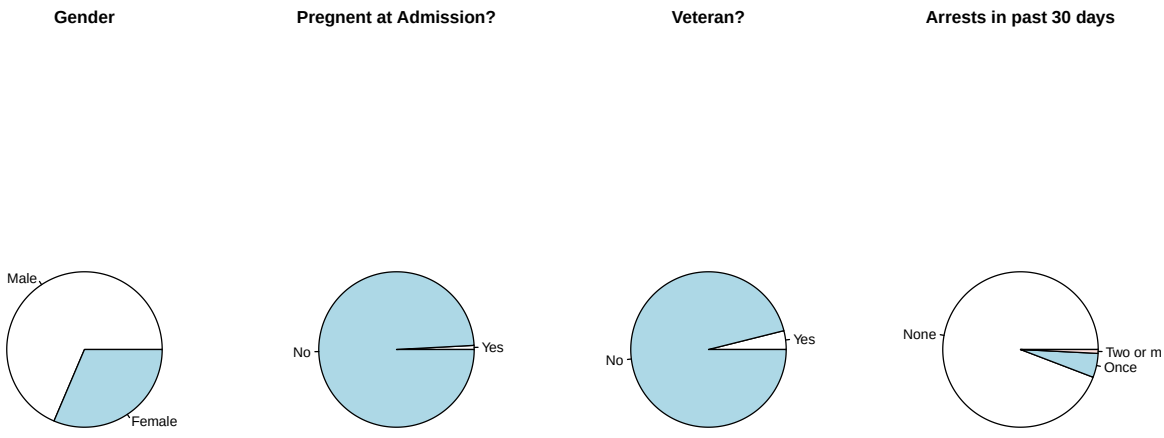
First Use <-> Substance



By carefully examining this visualization we can see that the sooner a person used a legal substance (age), the severity of the substance used on the day of admission to rehab is usually higher, which might indicate that the younger a person consume a legal drug for the first time, the higher are the chances of this drug to turn into a gate-way drug in the future (lead to use of worse drugs)

On the illegal drugs we see a clear and reversed conclusion: The older a person is when consuming their first illegal drug, the worse is the drug consumed on the day of admission to rehab.

Categories disperse A quick view on the relative quantities of each sub group out of the total:



4. Data analysis plan

What's next?

- Our outcome (response, Y) is the feature NOPRIOR as we wish to define profiles for returning rehab patients.
- Our predictor(explanatory, X) variables are, at least at the beginning (before proven not useful): AGE, GENDER, CALC_RACE, MARSTAT, EDUC, FRSTUSE_LEGAL, FRSTUSE_ILLEGAL, EMPLOY, PREG, VET, STATE_PARTY, LIVARAG, PRIMINC, ARRESTS, PSOURCE, DSMCRIT,FREQ_ATND_SELF_HELP, UNDER_INFLUENCE, ADDICTIVE_LEVEL, FREQ,PSYPROB, SUBSTANCE. These variables should compose a unique ID for each predicted class.
- Method: We believe that the best model for our purpose will be a tree classifier model, so that every class in the outcome will have unique branches reaching to it. After defining a good enough tree, we'll use the SHAP library in order to catch the most important feature for each of our target classes. This way we'll be able to get unique ID's for returning addicts in comparison to first timers, based on their most important features.
- To support our theory we'll need to prove that at least some of the classes we'll choose will have different characteristics (meaningful features in the decision tree). If we end up combining for example 2nd and 3rd timers, that's okay, but we do expect to find differences at least between 1st timers in rehab and the rest.
- Our work is divided to create load distribution, by each group member's availability. - Up until now one student was in charge of the calculated features, while the other 2 were in charge of the README file, method, proposal and learning the data. The exploratory data analysis was conducted by all 3 of us.
- For the rest of the project we believe that the correlation and finalizing of the research question's target classes will be one student's job, another will be in charge of building the model, and the third will be in charge of analyzing the results. That being said, we'll all be active at least for QA in every step of the way.

Stay tuned for more!

Data README

```
# \# SISE2601 Project data description

Team name 7 - Nir Rahav, Shahar Oded and Guy Zagorski.

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

CASEID (Case identification number) - Program generated case (record) identifier. A frequency distributi

AGE - Calculated from date of birth and date of admission and categorized.

| Value |       Label        |
|:-----:|:------------------:|
|   1   |    12-14 years     |
|   2   |    15-17 years     |
|   3   |    18-20 years     |
|   4   |    21-24 years     |
|   5   |    25-29 years     |
|   6   |    30-34 years     |
|   7   |    35-39 years     |
|   8   |    40-44 years     |
|   9   |    45-49 years     |
|  10   |    50-54 years     |
|  11   |    55-64 years     |
|  12   | 65 years and older |

GENDER - This field identifies the client's biological sex

| Value | Label  |
|:-----:|:------:|
|   1   |  Male  |
|   2   | Female |

CALC_RACE - This field identifies the client's race:

-   Black or African American: A person having origins in any of the black racial groups of Africa.

-   White: A person having origins in any of the original people of Europe, the Middle East, or North Af

-   Other - Use this category for instances in which the client is not identified in any category above

    -   Alaska Native (Aleut, Eskimo)

    -   American Indian

    -   Pacific Islander

    -   Asian

    -   Native Hawaiian

    -   Two or more races
```

Value	Label
1	Black or African American
2	White
3	Other race

MARSTAT (Marital status) - This field describes the client's marital status. The following categories a

- Never married: Includes clients who are single or whose only marriage was annulled.

- Now married: Includes married couples, those living together as married, living with partners, or c

- Separated: Includes those legally separated or otherwise absent from spouse because of marital disc

- Divorced, widowed

Value	Label
1	Never married
2	Now married
3	Separated
4	Divorced, widowed

EDUC (Education) - This field specifies:

- the highest school grade completed for adults or children not attending school

- current school grade for school-age children (3-17 years old) attending school.

Value	Label
1	Less than one school grade, no schooling, nursery school, or kindergarten to Grade 8
2	Grades 9 to 11
3	Grade 12 (or GED)
4	1-3 years of college, university, or vocational school
5	4 years of college, university, BA/BS, some postgraduate study, or more

FRSTUSE_LEGAL (Age at first use - legal material) - The field goes over FRSTUSE1, FRSTUSE2 and FRSTUSE3

Value	Label
1	11 years old and under
2	12-14 years
3	15-17 years
4	18-20 years
5	21-24 years
6	25-29 year
7	30 year and older
-9	Not use/Missing/unknown/invalid
NA	Not recorded (only illegal drug use records)

FRSTUSE_ILLEGAL (Age at first use - illegal material) - The field goes over FRSTUSE1, FRSTUSE2 and FRST

| Value | Label |

|:-----:|:---:|
1	11 years old and under
2	12-14 years
3	15-17 years
4	18-20 years
5	21-24 years
6	25-29 year
7	30 year and older
-9	Not use/Missing/unknown/invalid
NA	Not recorded (only legal drug use records)

EMPLOY (Employment status) - This field identifies the client's employment status.

- Full-time: Working 35 hours or more each week, including active duty members of the uniformed servi

- Part-time: Working fewer than 35 hours each week.

- Unemployed: Looking for work during the past 30 days or on layoff from a job.

- Not in labor force: Not looking for work during the past 30 days or a student, homemaker, disabled,

Value	Label
1	Full-time
2	Part-time
3	Unemploted
4	Not in labor force

PREG (Pregnant at admission) - This field indicates whether a female client was pregnant at the time of

Value	Label
1	Yes
2	No

VET (Veteran status) - This field indicates whether the client has served in the uniformed services (Ar

Value	Label
1	Yes
2	No

STATE_PARTY - This field calculate by the last 10 elections and classifies the country according to its

- Democrat - States where in seven of the last ten elections the majority voted for the Democratic Pa

- Republican - States where in 7 of the last ten elections the majority voted for the Republican Part

- Mitnadned - States where sometimes the majority votes for the Democratic Party and sometimes for th

Value	Label
0	Republican
1	Unstable / else

| 2 | Democratic |

LIVARAG (Living arrangements) - Identifies whether the client is homeless, a dependent (living with par

- Homeless: Clients with no fixed address; includes homeless shelters.

- Dependent living: Clients living in a supervised setting such as a residential institution, halfway

- Independent living: Clients living alone or with others in a private residence and capable of self-

Value	Label
1	Homeless
2	Dependent living
3	Independent

PRIMINIC (Source of income/support) - This field identifies the client's principal source of financial

Value	Label
1	Wages/salary
2	Public assistance
3	Retirement/pension, disability
4	Other
5	None
-9	Missing/unknown/not collected/invalid

ARRESTS (Arrests in past 30 days) - Indicates the number of arrests in the 30 days prior to the referen

Value	Label
0	None
1	Once
2	Two or more time

PSOURCE (Referral source) - This field describes the person or agency referring the client to treatment

- Individual (includes self-referral): Includes the client, a family member, friend, or any other ind

- Alcohol/drug use care provider: Any program, clinic, or other health care provider whose principal

- Other health care provider: A physician, psychiatrist, or other licensed health care professional;

- School (educational): A school principal, counselor, or teacher; or a student assistance program (S

- Employer/EAP: A supervisor or an employee counselor.

- Other community referral: Community or religious organization or any federal, state, or local agenc

- Court/criminal justice referral/DUI/DWI: Any police official, judge, prosecutor, probation officer

Value	Label
1	Individual (includes self-referral)

2	Alcohol/drug use care provider
3	Other health care provider
4	School (educational)
5	Employer/EAP
6	Other community referral
7	Court/criminal justice referral/DUI/DWI

DSMCRIT (DSM diagnosis) - Client's diagnosis is used to identify the substance use problem that provides

The discrete diagnosis codes have been recoded into categories related to use of and dependence on speci

Value	Label
1	Alcohol-induced disorder
2	Substance-induced disorder
3	Alcohol intoxication
4	Alcohol dependence
5	Opioid dependence
6	Cocaine dependece
7	Cannabis dependence
8	Other substance dependence
9	Alcohol abuse
10	Cannabis abuse
11	Other substance abuse
12	Opioid abuse
13	Cocaine abuse
14	Anxiety disorders
15	Depressive disorders
16	Schizophrenia/other psychotic disorders
17	Bipolar disorders
18	Attention deficit/disruptive behavior disorders
19	Other mental health condtion
-9	Missing/unknown/not collected/invalid/no or deferred diagnosis

FREQ_ATND_SELF_HELP (Attendance at substance use self-help groups in past 30 days) - This field indicat

Value	Label
1	Not attendance
2	1-3 times in the past month
3	4-7 times in the past month
4	8-30 times in the past month
5	Some attendance, frequency unknown
-9	Missing/unknown/not collected/invalid

UNDER_INFLUENCE - This field describe if the client arrive under any influence of drugs or alcohol to t

Value	Label
0	False
1	True

SUBSTANCE - This field describe the addiction of each person according to his statements in SUB1, SUB2 a

- Stimulant - Bitter drugs

- Depressant - depressant or "down" drugs

- legal in state - Legal drugs, according to the states.

- Over the counter / other - Medicines that are available in pharmacies and are consumed excessively

The values are ordered by severity.

Value	Label
0	legal in state
1	Over the counter / other
2	Stimulant
3	Depressant

: ADDICTIV_LEVEL - This field describe the patient's level of addiction. It's take the addictions and w

- Non addictive / Unknown - not addictive

- Low - low level addictive

- High - medium level addictive

- Very high - High level addictive

Value	Label
1	Non addictive / Unknown
2	Low
3	High
4	Very high

: FREQ (Frequency of use) - Specifies the frequency of use of the substances. The field uses a function

Value	Label
1	No use in the past month
2	Some use
3	Daily use
-9	Missing/unknown/not collected/invalid

: PSYPROB (Co-occurring mental and substance use disorders) - Indicates whether the client has co-occur

Value	Label
1	Yes
2	No

NOPRIOR (Previous substance use treatment episodes) - Indicates the number of previous treatment episod

Value	Label

```
|   0   |      No prior treatment episodes      |
|   1   |     One prior treatment episodes      |
|   2   |     Two prior treatment episodes      |
|   3   |    Three prior treatment episodes     |
|   4   |     Four prior treatment episodes     |
|   5   | Five or more prior treatment episodes |
```

Appendix

features engineering In this Section We'll place our features engineering as described in the data section.

Source code

```
library(knitr)
library(tidyverse)
library(broom)
library(htmltools)
library(ggplot2)
library(ggridges)
opts_chunk$set(echo=FALSE) # hide source code in the document
# Add original data import
admission_data <- read.csv("C:/Users/guy/Desktop/final_prop/data/TEDSA_PUF_2020.csv")
# /
#
# Feature engineering as described in the README and coded in the Appendix
#
# Filters and Selection to the engineered data
#
# Import new file
#/
admission_data <- read.csv("C:/Users/guy/Desktop/final_prop/data/admission_data_cooked.csv")

# Get an upper view on the data
dim(admission_data)
summary(admission_data)

# Add original data import
admission_data <- read.csv("C:/Users/guy/Desktop/final_prop/data/admission_data_cooked.csv")
admissions <- admission_data[!is.na(admission_data$PSYPROB) & !is.na(admission_data$NOPRIOR),]

ggplot(admissions, aes(x = NOPRIOR, y = PSYPROB, color = factor(NOPRIOR))) +
  geom_point(stat = "summary", fun.y = "mean", size = 5) +
  xlab("NOPRIOR - Prior Admissions to Rehab") +
  ylab("PSYPROB - Psychiatric Assessment") +
  ggtitle("Psychiatric Assessment by Prior Admissions to Rehab") +
  scale_color_manual(values = c("#1A237E", "#303F9F", "#3949AB", "#9B59B6", "#BB8FCE", "#D7BDE2", "#E8D
  annotate("text", x = 4, y = 1.6, label = paste("Correlation = ", round(cor(admissions$NOPRIOR, admiss

admissions <- admission_data[!is.na(admission_data$SUBSTANCE) & !is.na(admission_data$NOPRIOR),]

ggplot(admissions, aes(x = NOPRIOR, y = SUBSTANCE, color = factor(NOPRIOR))) +
```

```
  geom_point(stat = "summary", fun.y = "mean", size = 5) +
  xlab("NOPRIOR - Prior Admissions to Rehab") +
  ylab("SUBSTANCE - Consumed Substance Class") +
  ggtitle("Psychiatric Assessment by Consumed Substance Class") +
  scale_color_manual(values = c("#8B0000", "#AB2F40", "#E2A9A9", "#F0E68C", "#CDBE70", "#BDB76B")) +
  annotate("text", x = 1, y = 1.75, label = paste("Correlation = ", round(cor(admissions$NOPRIOR, admiss

ggplot(admission_data, aes(x = NOPRIOR, y = as.factor(STATE_PARTY), fill = as.factor(STATE_PARTY), colo
  geom_density_ridges(alpha = 0.5) +
  ggtitle("Density of STATE_PARTY Category for NOPRIOR")

ggplot(admission_data, aes(x = NOPRIOR, y = as.factor(ADDICTIVE_LEVEL), fill = as.factor(ADDICTIVE_LEVE
  geom_density_ridges(alpha = 0.5) +
  ggtitle("Density of ADDICTIVE_LEVEL Category for NOPRIOR")

ggplot(admission_data, aes(x = NOPRIOR, y = as.factor(FREQ), fill = as.factor(FREQ), color = as.factor(
  geom_density_ridges(alpha = 0.5) +
  ggtitle("Density of FREQ Category for NOPRIOR")


admission_data$GENDER <- as.numeric(admission_data$GENDER)
admission_data$PREG <- as.numeric(admission_data$PREG)
admission_data$VET <- as.numeric(admission_data$VET)
admission_data$ARRESTS <- as.numeric(admission_data$ARRESTS)
admission_data$SUBSTANCE <- as.character(admission_data$SUBSTANCE)

ggplot(admission_data, aes(y = FRSTUSE_LEGAL,
                  fill = SUBSTANCE)) +
  geom_bar(position = "fill") +
  labs(
    x = "Proportion",
    y = "FRSTUSE_LEGAL",
    fill = "SUBSTANCE",
    title = "Relation between Substance to first use of legal drug",
  ) +
  ylim(-1,7)

ggplot(admission_data, aes(y = FRSTUSE_ILLEGAL,
                  fill = SUBSTANCE)) +
  geom_bar(position = "fill") +
  labs(
    x = "Proportion",
    y = "FRSTUSE_ILLEGAL",
    fill = "SUBSTANCE",
    title = "Relation between Substance to first use of illegal drug",
  ) +
  ylim(-1,7)

admission_data$SUBSTANCE <- as.numeric(admission_data$SUBSTANCE)

# Set up a multi - paneled layout
par(mfrow = c(1,4))
```

```r
# Create the first pie chart
gender_table <- table(admission_data$GENDER)
gender_labels <- c("Male", "Female")
names(gender_table) <- gender_labels
pie(gender_table, main = "Gender")

# Create the second pie chart
preg_table <- table(admission_data$PREG)
preg_labels <- c("Yes", "No")
names(preg_table) <- preg_labels
pie(preg_table, main = "Pregnent at Admission?")

# Create the third pie chart
vet_table <- table(admission_data$VET)
vet_labels <- c("Yes", "No")
names(vet_table) <- vet_labels
pie(vet_table, main = "Veteran?")

# Create the forth pie chart
ARRESTS_table <- table(admission_data$ARRESTS)
ARRESTS_labels <- c("None", "Once","Two or more times")
names(ARRESTS_table) <- ARRESTS_labels
pie(ARRESTS_table, main = "Arrests in past 30 days")
cat(readLines('../data/README.md'), sep = '\n')
# Add original data import
admission_data <- read.csv("C:/Users/guy/Desktop/final_prop/data/TEDSA_PUF_2020.csv")

# Calculated features:

# First mutation of existing features (described in the README) -
admission_data$PREG[admission_data$PREG == -9] <- 2
admission_data$VET[admission_data$VET == -9] <- 2
admission_data$ARRESTS[admission_data$ARRESTS == -9] <- 0
admission_data$RACE[admission_data$RACE == -9] <- 7

# Create Party affiliation feature based on patients state
Republican <- c(1,2,5,18,20,21,22,28,29,30,31,38,40,45,46,47,48,49,54,56)
Democratic <- c(6,8,9,10,11,15,17,23,25,27,32,33,34,35,36,44,50)

# State party :
# Republican - 0
# Unstable / else - 1
# Democratic - 2

STATE_PARTY <- function(party){
    # If from a Democratic state
    if (party %in% Democratic) {
    return(2)
    # If If from a Republican state
  } else if (party %in% Republican) {
    return(0)
    # else - non-distinct-party affiliated state
    # If missing/unknown/not collected/invalid
```

```
  } else {
    return(1)
  }
}

tmp <- admission_data %>% select(STFIPS)

PARTY_v <- apply(tmp, 1, function(tmp) STATE_PARTY(tmp[1]))

# Activate calculated field

admission_data$STATE_PARTY <- PARTY_v


# Create new column FREQ containing the maximum value across FREQ1, FREQ2, and FREQ3
admission_data$FREQ <- apply(admission_data[, c("FREQ1", "FREQ2", "FREQ3")], 1, max, na.rm = TRUE)

# Create new feature UNDER_INFLUENCE on the day of admission
UNDER_INFLUENCE_v <- apply(admission_data[, c("ALCFLG", "COKEFLG", "MARFLG", "HERFLG", "METHFLG", "OPSY
admission_data$UNDER_INFLUENCE <- UNDER_INFLUENCE_v
# create a function SUB_ACCESS to assist in creating SUBSTANCE column
# legal -> TRUE , illegal -> FALSE
# vector of countries that marijuana is legal
weed_legal_states <- c(2,4,6,8,9,11,17,23,25,26,30,32,34,36,50,51)
SUB_ACCESS <- function(drug,state){
    # If alcohol
    if (drug == 2) {
    return(TRUE)
    # If weed and legal state
  } else if (drug == 4 & state %in% weed_legal_states) {
    return(TRUE)
  } else {
    return(FALSE)
  }
}

# Create a new function to indicate the type of the drug (by scale)
# legal in country - 0
# over the counter prescriptions / other - 1
# Stimulant - 2
# Depressant - 3
# Irrelevant - -9 (if value is 1 (NONE) or -9 (unknown))

SUB_CATEGORY <- function(drug,state){
    # If Stimulant
    if (drug %in% c(3, 10, 11, 12)) {
    return(2)
    # If Depressant
  } else if (drug %in% c(5, 6, 7, 13, 14, 15, 16)) {
    return(3)
    #if value is 1 (NONE) or -9 (unknown)
  } else if (drug %in% c(1,-9)) {
    return(-9)
```

```
    # If legal in state
  } else if (SUB_ACCESS(drug,state) == TRUE) {
    return(0)
  # Over the counter / other
  } else {
    return(1)
  }
}

SUBSTANCE_CLASS <- function(drug1,drug2,drug3,state){
    # Return 'SUNDAY_FUNDAY' if there is a combination of active drugs (in README)
    # else: Return max value for all 3 substances in use
  c1 = SUB_CATEGORY(drug1,state)
  c2 = SUB_CATEGORY(drug2,state)
  c3 = SUB_CATEGORY(drug3,state)

  c_range <- c(-9, 0, 1, 2, 3)
  c_vector <- c(c1, c2, c3)
  c_unique <- unique(c_vector) # keeps only unique values in vector, without values that appears twice
  c_range_check <- c_unique %in% c_range[3:5] # take all unique values in range [1,3], holds TRUE and F

  if (sum(c_range_check) >= 2) {
    return('SUNDAY_FUNDAY')
  } else {
    return(max(c_vector))
  }
}

tmp <- admission_data %>% select(SUB1,SUB2,SUB3,STFIPS)

SUBSTANCE_v <- apply(tmp, 1, function(tmp) SUBSTANCE_CLASS(tmp[1],tmp[2],tmp[3],tmp[4]))
# Activate calculated field

admission_data$SUBSTANCE <- SUBSTANCE_v


# Create vector for first use of legal drug
legal_drugs_firstuse = c()
illegal_drugs_firstuse = c()


FIRST_USE_LEGAL <- function(age1, drug1, age2, drug2, age3, drug3, state, legal_drugs_firstuse) {
  if (SUB_CATEGORY(drug1, state) == 0) {
    legal_drugs_firstuse <- c(legal_drugs_firstuse, age1)
  }
  if (SUB_CATEGORY(drug2, state) == 0) {
    legal_drugs_firstuse <- c(legal_drugs_firstuse, age2)
  }
  if (SUB_CATEGORY(drug3, state) == 0) {
    legal_drugs_firstuse <- c(legal_drugs_firstuse, age3)
  }
  if (length(legal_drugs_firstuse) == 0) {
    return (NA)
```

```
  } else {
    # Return min value for all 3 substances in use
    return (min(legal_drugs_firstuse))
  }
}

FIRST_USE_ILLEGAL <- function(age1,drug1,age2,drug2,age3,drug3,state,illegal_drugs_firstuse){
  if (SUB_CATEGORY(drug1,state) > 0) {
    illegal_drugs_firstuse <- c(illegal_drugs_firstuse, age1)
  }
  if (SUB_CATEGORY(drug2,state) > 0) {
    illegal_drugs_firstuse <- c(illegal_drugs_firstuse, age2)
  }
  if (SUB_CATEGORY(drug3,state) > 0) {
    illegal_drugs_firstuse <- c(illegal_drugs_firstuse, age3)
  }
  if (length(illegal_drugs_firstuse) == 0) {
    return (NA)
  } else {
    # Return min value for all 3 substances in use
  return (min(illegal_drugs_firstuse))
  }
}

tmp <- admission_data %>% select(FRSTUSE1,SUB1,FRSTUSE2,SUB2,FRSTUSE3,SUB3,STFIPS)

FRSTUSE_LEGAL_v <- apply(tmp, 1, function(tmp) FIRST_USE_LEGAL(tmp[1],tmp[2],tmp[3],tmp[4],tmp[5],tmp[6]
FRSTUSE_ILLEGAL_v <- apply(tmp, 1, function(tmp) FIRST_USE_ILLEGAL(tmp[1],tmp[2],tmp[3],tmp[4],tmp[5],tm
# Activate calculated field

admission_data$FRSTUSE_LEGAL <- FRSTUSE_LEGAL_v
admission_data$FRSTUSE_ILLEGAL <- FRSTUSE_ILLEGAL_v

# Define the severity of addiction potential for each drug category
# Non addictive / Unknown - 1
# Low - 2
# High - 3
# Very high - 4

addiction_levels <- c("1" = 1,
                      "2" = 3,
                      "3" = 3,
                      "4" = 2,
                      "5" = 4,
                      "6" = 3,
                      "7" = 4,
                      "8" = 3,
                      "9" = 2,
                      "10" = 3,
                      "11" = 3,
                      "12" = 3,
                      "13" = 3,
                      "14" = 3,
```

```
                      "15" = 4,
                      "16" = 3,
                      "17" = 2,
                      "18" = 2,
                      "19" = 1,
                      "-9" = 1)
ADDICTION_LEVEL <- function(SUB1, SUB2, SUB3) {
  a1 = addiction_levels[as.character(SUB1)]
  a2 = addiction_levels[as.character(SUB2)]
  a3 = addiction_levels[as.character(SUB3)]
  max_value <- max(a1, a2, a3)
  return(max_value)
}

# Assign addictive level to the most addictive drug the patient funds
# create a new column for addiction potential



tmp <- admission_data %>% select(SUB1,SUB2,SUB3)

ADDICTIVE_LEVEL_v <- apply(tmp, 1, function(tmp) ADDICTION_LEVEL(tmp[1],tmp[2],tmp[3]))
# Activate calculated field

admission_data$ADDICTIVE_LEVEL <- ADDICTIVE_LEVEL_v
CALC_RACE <- function(race){
    # If Black person
    if (race %in% c(4)) {
    return(1)
    # If White person
  } else if (race %in% c(5)) {
    return(2)
    # Other single race
  } else if (race %in% c(1,2,3,6,7,8,9)) {
    return(3)
    # If missing/unknown/not collected/invalid
  } else if (race %in% c(-9)) {
    return(-9)
  }
}

tmp <- admission_data %>% select(RACE)

RACE_v <- apply(tmp, 1, function(tmp) CALC_RACE(tmp[1]))
# Activate calculated field

admission_data$CALC_RACE <- RACE_v

# Take only relevant features for the final data set
admission_data <- admission_data %>%
  select(CASEID, AGE, GENDER, CALC_RACE, MARSTAT, EDUC, FRSTUSE_LEGAL, FRSTUSE_ILLEGAL, EMPLOY, PREG, VI
# Apply filters to create the final data set
```

```r
# FILTER (1)
admission_data <- admission_data[admission_data$SUBSTANCE != "SUNDAY_FUNDAY", ]
admission_data$SUBSTANCE <- as.integer(admission_data$SUBSTANCE)

#FILTER (2)
admission_data <- admission_data[admission_data$GENDER != -9 &
                                   admission_data$MARSTAT != -9 &
                                   admission_data$EDUC != -9 &
                                   admission_data$EMPLOY != -9 &
                                   admission_data$LIVARAG != -9 &
                                   admission_data$PSOURCE != -9 & admission_data$PSYPROB != -9 &
                                   admission_data$NOPRIOR != -9, ]

#FILTER (3)
admission_data <- admission_data[!(is.na(admission_data$SUBSTANCE) | is.na(admission_data$FREQ) |
                                     (is.na(admission_data$FRSTUSE_LEGAL) & is.na(admission_data$FRSTUSE

# Write out the data to the repository to save re-calculations in the future
write.csv(admission_data,"C:/Users/guy/Desktop/final_prop/data/admission_data_cooked.csv",
          row.names = FALSE)
```