Logbook

Susan Reefman

9/14/2021

Research topic and dataset

The research topic of this project is the treatment episodes of patients. The dataset used in this project is from the Treatment Episode Data Set (TEDS), an American data system of annual discharges from substance use treatment facilities. The used dataset is a TEDS-D dataset, which means it includes discharges from substance use treatment facilities. TEDS-D contains records on admissions of people of 12-years and older, and includes information of admission, substance use characteristics and discharges. The dataset includes information from patients who where at the facility in 2019. For this project some personal and substance use characteristics information are collected from the dataset. In the table the variables are listed with their meaning. The type of all variables are numeric and the number of possible values is different for each variable.

data source: https://www.datafiles.samhsa.gov/dataset/teds-d-2017-ds0001-teds-d-2017-ds0001

```
load("~/Documents/Themaopdracht09/tedsd_puf_2019.RData")

p <- "CASEID|AGE|EDUC|^EMPLOY|ALCDRUG|^SUB|^FREQ[1|2|3]|^ROUTE|^FRSTUSE|LOS|NOPRIOR|PSYPROB|DSMCRIT|REAS

# get relevant data from data set
data <- df[, grep(pattern=p, colnames(df))]

# replacing missing data with NA's
is.na(data) <- data == "-9"

codebook <- read.csv("~/Documents/Themaopdracht09/codebook.csv", header = TRUE,
                     sep = ",", na.strings = "N/A")

knitr::kable(codebook)
```

Variable	Type	Label	number.of.possible.values
AGE	Numeric	Age at admission	12
ALCDRUG	Numeric	Substance use type	4
CASEID	Numeric	Case identification number	NA
DSMCRIT	Numeric	DSM diagnosis (SuDS 4 or SuDS 19)	19
EDUC	Numeric	Education	5
EMPLOY	Numeric	Employment status at admission	4
EMPLOY_D	Numeric	Employment status at discharge	4
FREQ1	Numeric	Frequency of use at admission (primary)	3
FREQ2	Numeric	Frequency of use at admission (secondary)	3
FREQ3	Numeric	Frequency of use at admission (tertiary)	3
FREQ1_D	Numeric	Frequency of use at discharge (primary)	3
FREQ2_D	Numeric	Frequency of use at discharge (secondary)	3
FREQ3_D	Numeric	Frequency of use at discharge (tertiary)	3
FRSTUSE1	Numeric	Age at first use (primary)	7

Variable	Type	Label	number.of.possible.values
FRSTUSE2	Numeric	Age at first use (secondary)	7
FRSTUSE3	Numeric	Age at first use (tertiary)	7
LOS	Numeric	Length of stay in treatment (days)	37
NOPRIOR	Numeric	Number of previous substance use treatment episodes	2
PSYPROB	Numeric	Co-occurring mental and substance use disorders	2
REASON	Numeric	Reason for discharge	7
ROUTE1	Numeric	Route of administration (primary)	5
ROUTE2	Numeric	Route of administration (secondary)	5
ROUTE3	Numeric	Route of administration (tertiary)	5
SUB1	Numeric	Substance use at admission (primary)	19
SUB2	Numeric	Substance use at admission (secondary)	19
SUB3	Numeric	Substance use at admission (tertiary)	19
SUB1_D	Numeric	Substance use at discharge (primary)	19
SUB2_D	Numeric	Substance use at discharge (secondary)	19
SUB3_D	Numeric	Substance use at discharge (tertiary)	19

With this dataset a model can be made. In this project is researched whether the length of treatment can be predicted. The model is made by the following question: "Can the length of treatment be predicted based on the substance use, frequency of use and the age at first use by using machine learning?" To answer this question information is collected from the dataset. The personal information about the patients that is used is case id, age, education and employment status. Other information is about substance characteristics, for example; substance use type, substance use at admission and discharge and route of administration. The information about treatment episodes are also covered. This includes length of stay in treatment, number of previous substance treatment episodes and reason of discharge.

Exploratory data analysis

#missing data

To get an idea of what the data includes a couple histograms are made. In figure 1 is the age of patients shown. The age of patients is divided in groups: group 1 = 12–14 years, group 2 = 15–17 years, group 3 = 18–20 years, group 4 = 21–24 years, group 5 = 25–29 years, group 6 = 30–34 years, group 7 = 35–39 years, group 8 = 40–44 years, group 9 = 45–49 years, group 10 = 50–54 years, group 11 = 55–64 years, group 12 = 65 years and older.

The most patients fall under the groups 6, 7, 8 and are between 30 and 44 years old.

```
x <- 1:12
ggplot(data = data, aes(x = AGE)) +
  geom_histogram(bins = 12,
                 xlim = c(0,12),
                 ylim = c(0,1000),
                 fill = I("blue"),
                 col = I("blue"),
                 alpha = 0.5) +
  xlab("Age by groups") +
  ylab(NULL) +
  scale_x_continuous(labels=as.character(x),breaks=x) +
  scale_y_continuous(labels=NULL)
```

The age of first use of the patients is shown figure 2. The age of patients is divided in multiple groups: group 1 = 11 years and under, group 2 = 12–14 years, group 3 = 15–17 years, group 4 = 18–20 years, group 5

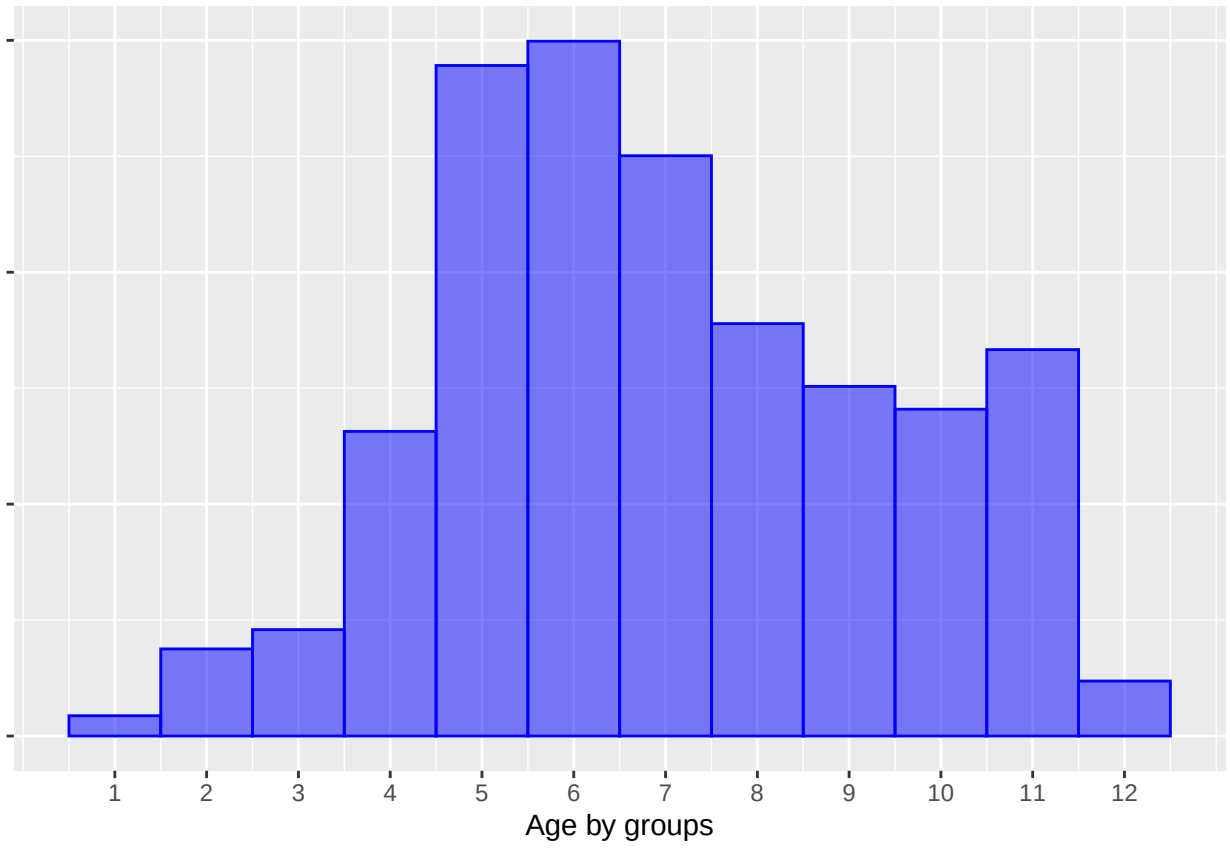

Figure 1: Histogram age of patients in groups

= 21-24 years, group 6 = 25-29 years, group 7 = 30 years and over. The frequency for the secondary and tertiary substance is lower, most patients have one an addiction for one substance.

```
par(mfrow = c(1, 2))

ggplot(data = data, aes(x = FRSTUSE1)) +
  geom_histogram(bins = 7,
                 col = I("blue"),
                 alpha = 0.2) +
  geom_histogram(data = data, aes(x = FRSTUSE2),
                 bins = 7,
                 col = I("green"),
                 alpha = 0.2) +
  geom_histogram(data = data, aes(x = FRSTUSE3),
                 bins = 7,
                 col = I("red"),
                 alpha = 0.2) +
  xlab("Age by groups") +
  ylab(NULL) +
  scale_x_continuous(labels=as.character(x),breaks=x) +
  scale_y_continuous(labels=NULL)
```

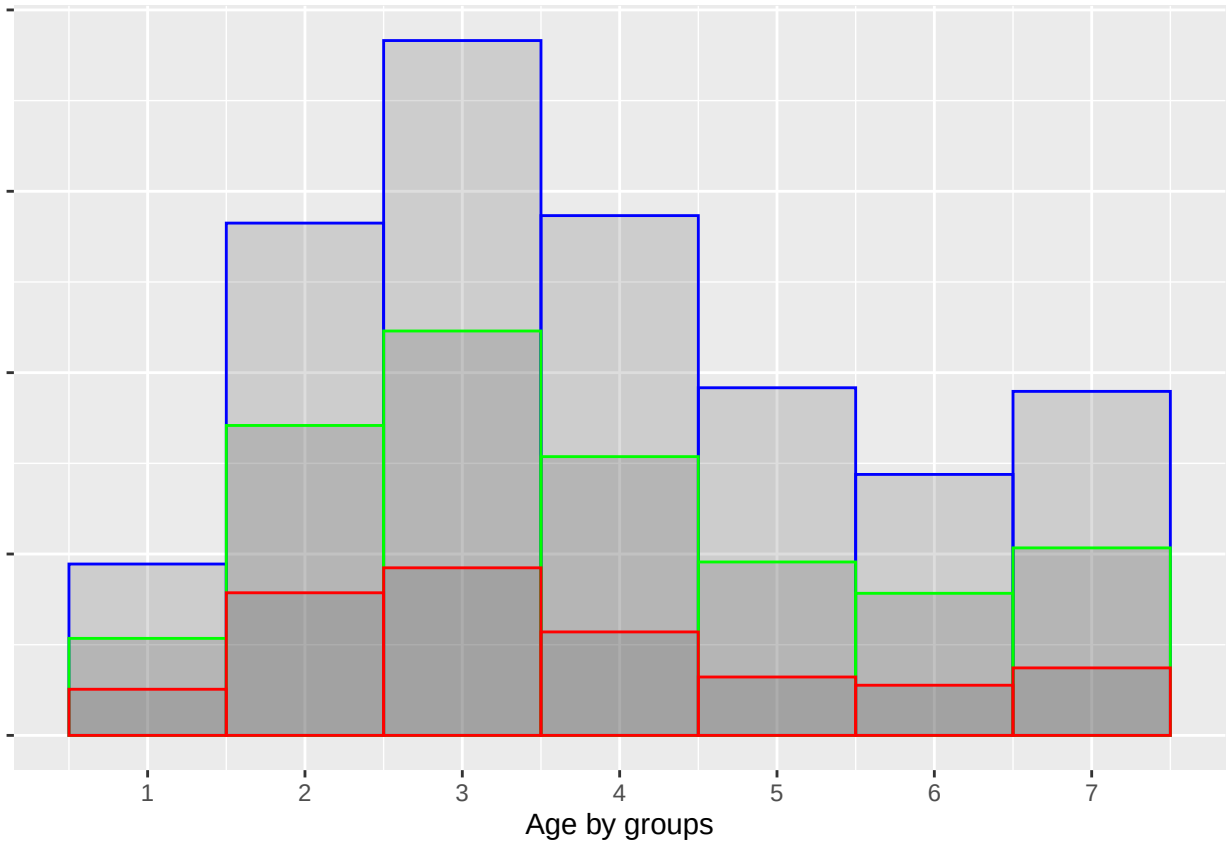

Figure 2: Histogram age of first use of patients in groups; blue = substance 1 - green = substance 2 - red = substance 3

```
ggplot(data = data, aes(x = FREQ1)) +
  geom_histogram(bins = 3,
```

```
                col = I("blue"),
                alpha = 0.2) +
geom_histogram(data = data, aes(x = FREQ2),
                bins = 3,
                col = I("green"),
                alpha = 0.2) +
geom_histogram(data = data, aes(x = FREQ3),
                bins = 3,
                col = I("red"),
                alpha = 0.2) +
xlab("Frequency of use in groups") +
ylab(NULL) +
scale_x_continuous(labels=c("No use in the past month","Some use","Daily use",4,5,6,7,8,9,10,11,12), l
scale_y_continuous(labels=NULL)
```

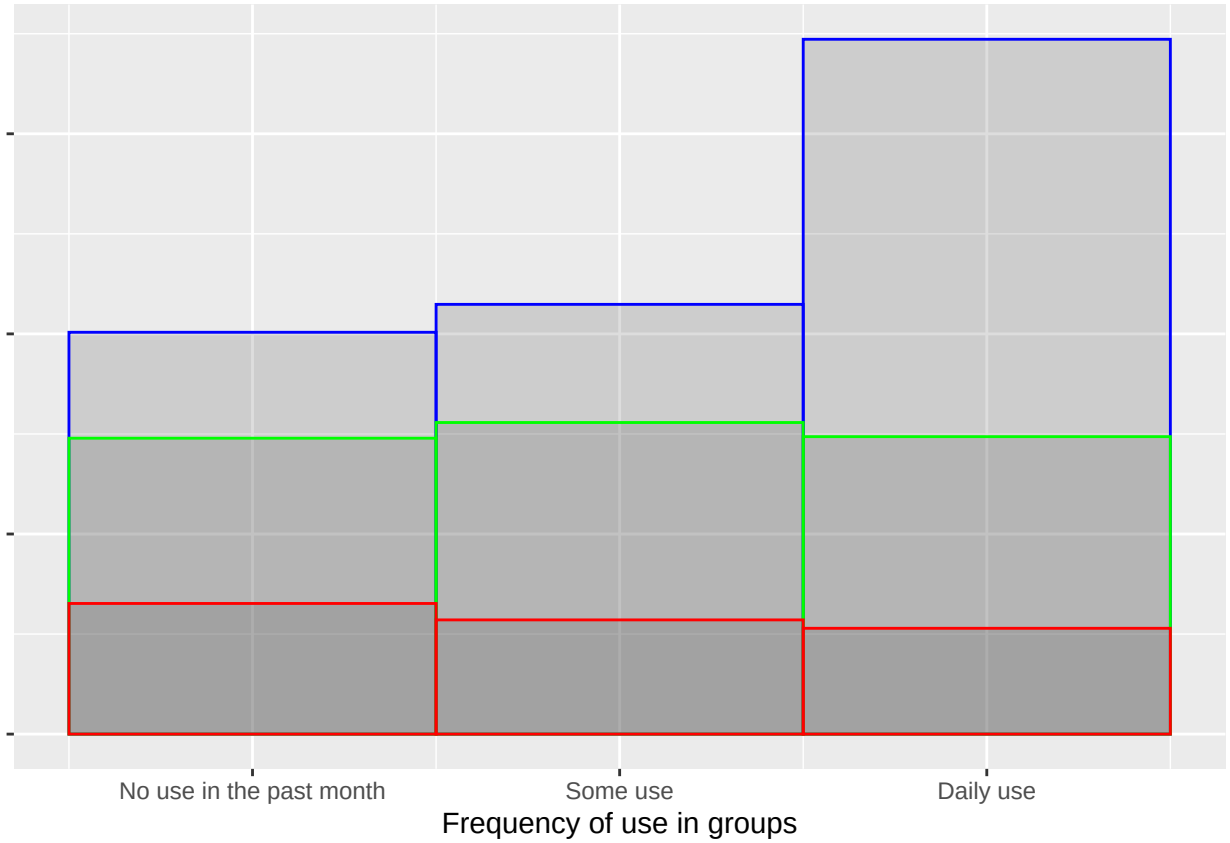

Figure 3: Histogram age of first use of patients in groups; blue = substance 1 - green = substance 2 - red = substance 3

Figure 3 shows the frequency of use of patients in groups. There are 3 groups; No use in the past month, some use and daily use. As can be seen most patients are in the group of daily use.