Logbook

Susan Reefman

05/10/2021

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))]

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
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

Variable	Type	Label	number.of.possible.values
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

In the dataset missing data is coded for '-9'. In this research those values are replaced with NA.

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

Variation and distribution

To get an idea of what the data includes a couple histograms are made (see results). The histograms display the amount of patients, in age groups, when they first used and their frequency of use. Also the amount of patients and their substance use are displayed in a histogram. A boxplot is used to show the length of stay in treatment versus the type of substance use.

Results

In figure 1 is the age of patients shown. The age of patients is divided in groups of around 4 years. Most patients are between 25 and 39 years old. The age of patients seems like a normal distrubution.

```
x <- 1:12
label <- c("12-14","15-17", "18-20", "21-24", "25-29", "30-34", "35-39", "40-44", "45-49", "50-54", "55-
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) +
  ggtitle("Distribution of age of patients") +
  xlab("Age (years)") +
  ylab("Frequency") +
  scale_x_continuous(labels=label,breaks=x)
```

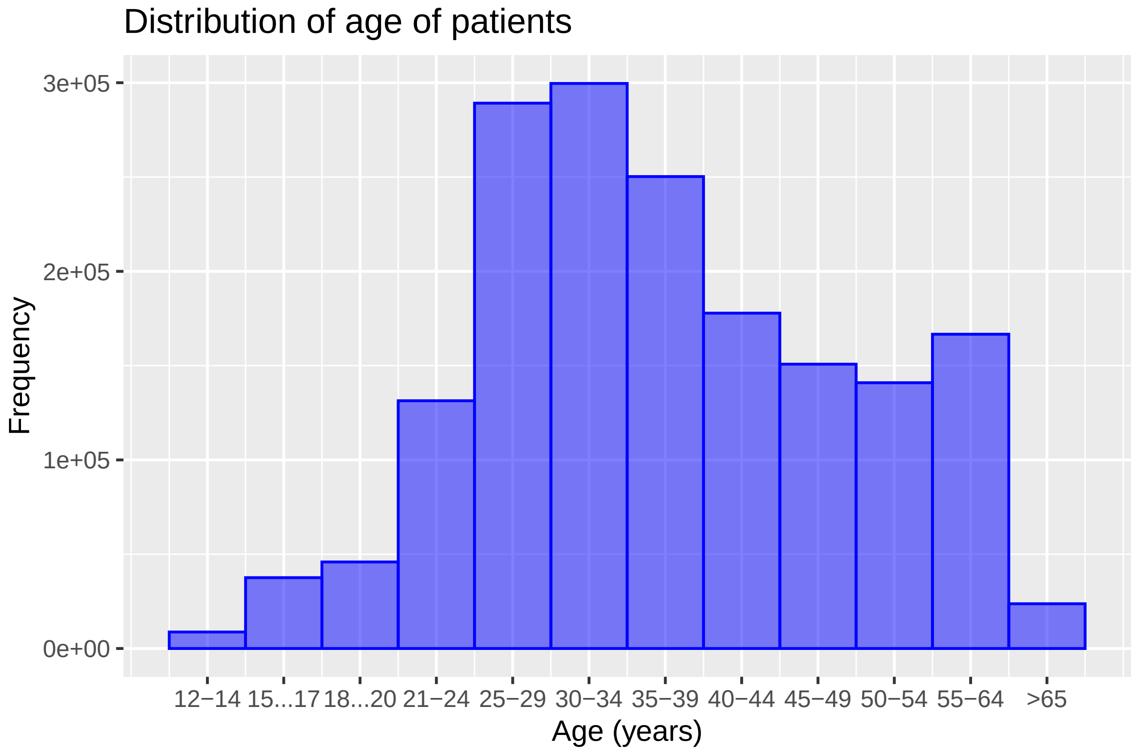

Figure 1: Histogram age of patients in groups

The age of first use of the patients and the frequency of use is shown figure 2. The age of patients is divided in multiple groups, in which the first is 12 years of younger and the last group is 29 years of older. The frequency for the secondary and tertiary substance is lower, most patients have one an addiction for one substance. For the frequency of substance use the primary substance is mostly use daily. The frequency of use of secondary and tertiary substance are very similar, the patients are almost evenly spread over the groups of frequency of use. The age of first use of patients seems normally distributed.

```
# histogram age of first use
x <- 1:7
label <- c("< 12", "12-14", "15-17", "18-20", "21-24", "25-29","> 29")
p1 <- 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 (years)") +
  ylab("Patients") +
  ggtitle("Frequency of patients in groups by age of first use")+
  scale_x_continuous(labels=label,breaks=x)

# histogram frequency of use
x <- 1:3
p2 <- 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") +
  ylab("Patients") +
  ggtitle("Frequency of patients in groups by frequency of use") +
  scale_x_continuous(labels=c("No use in the past month","Some use","Daily use"), breaks=x)

grid.arrange(p1, p2)
```

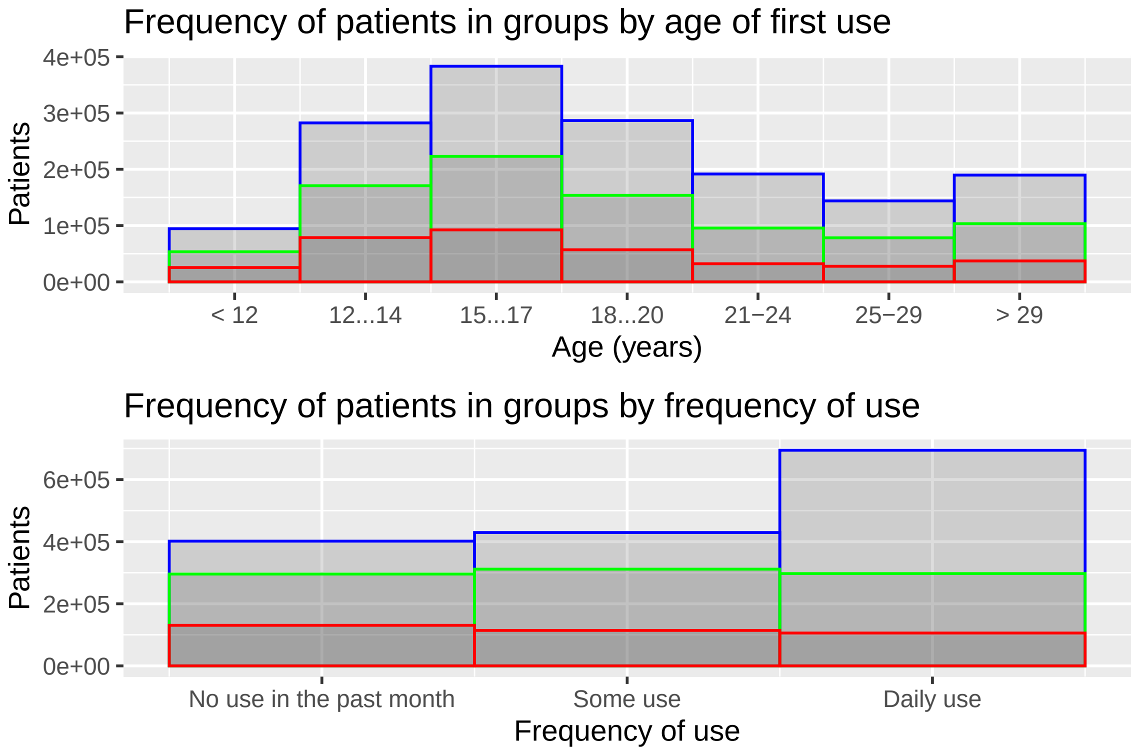

Figure 2: Top: Histogram of age of first use of patients in groups; Bottom: Histogram of frequency of use by patients in groups

Figure 3 displays the number of patients and their substance of use. Not surprisingly is the red (tertiary substance) very high for the None of substance of use. Not all patients have multiple substances they are addicited to. However, there are a few patients that filled in that the have none as primary substance of use (blue). Alcohol, Heroin and Methamphetamine are the most substances that are used by patients.

```
x <- 1:19
label <- c("None", "Alcohol", "Cocaine/crack", "Marijuana/hashish", "Heroin", "Non-prescription methadon
ggplot(data = data, aes(x = SUB1)) +
  geom_histogram(bins = 19,
                 col = I("blue"),
                 alpha = 1) +
  geom_histogram(data = data, aes(x = SUB2),
                 bins = 19,
                 col = I("green"),
                 alpha = 0.5) +
  geom_histogram(data = data, aes(x = SUB3),
                 bins = 19,
                 col = I("red"),
                 alpha = 0.2) +
  xlab("Substance used") +
  ylab("Patients") +
  scale_x_continuous(labels=label, breaks=x) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  ggtitle("Frequency of patients by substance use")
```

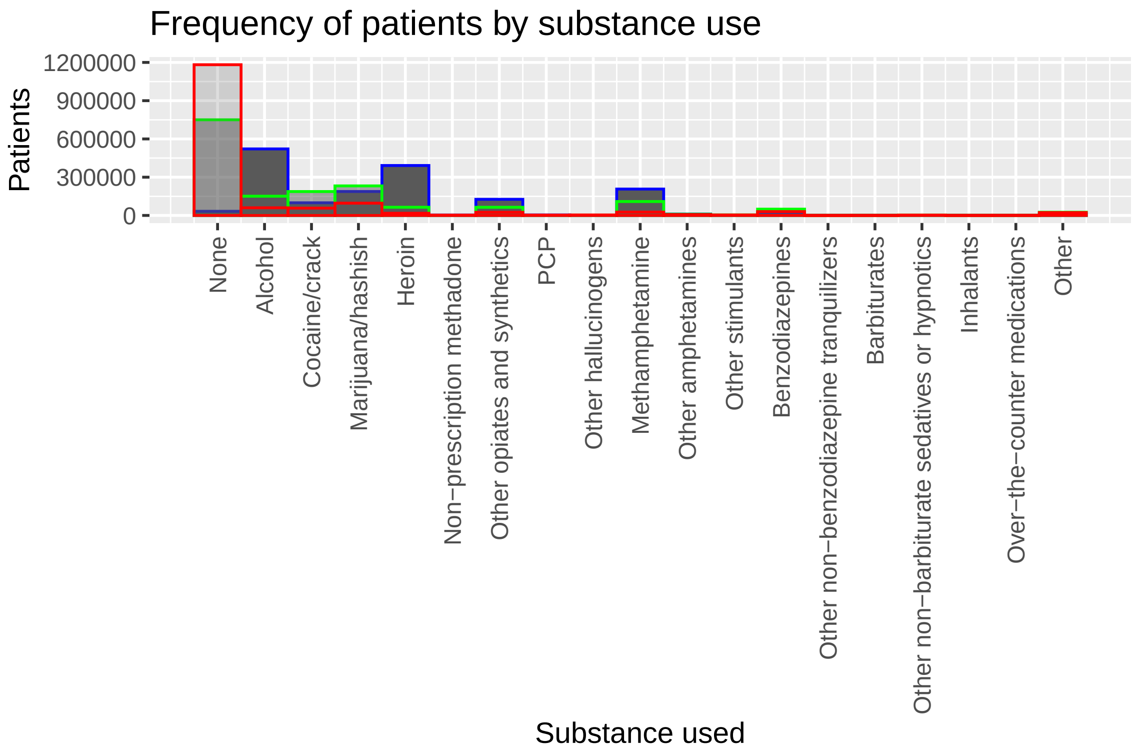

Figure 3: Histogram of substance use of patients

Figure 4 shows the usual route of administration of patients. Most patients use the oral route of administration of their substance use. Most patients that have a secondary substance use, use the smoking route of administration of that substance.

```
x <- 1:5
label <- c("Oral", "Smoking", "Inhalation", "Injection", "Other")
ggplot(data = data, aes(x = ROUTE1)) +
  geom_histogram(bins = 5,
                 col = I("blue"),
                 alpha = 0.1) +
  geom_histogram(data = data, aes(x = ROUTE2),
                 bins = 5,
                 col = I("green"),
                 alpha = 0.5) +
  geom_histogram(data = data, aes(x = ROUTE3),
                 bins = 5,
                 col = I("red"),
                 alpha = 1) +
  xlab("Route of substance") +
  ylab("Patients") +
  scale_x_continuous(labels=label, breaks=x) +
  ggtitle("Route of substance use")
```

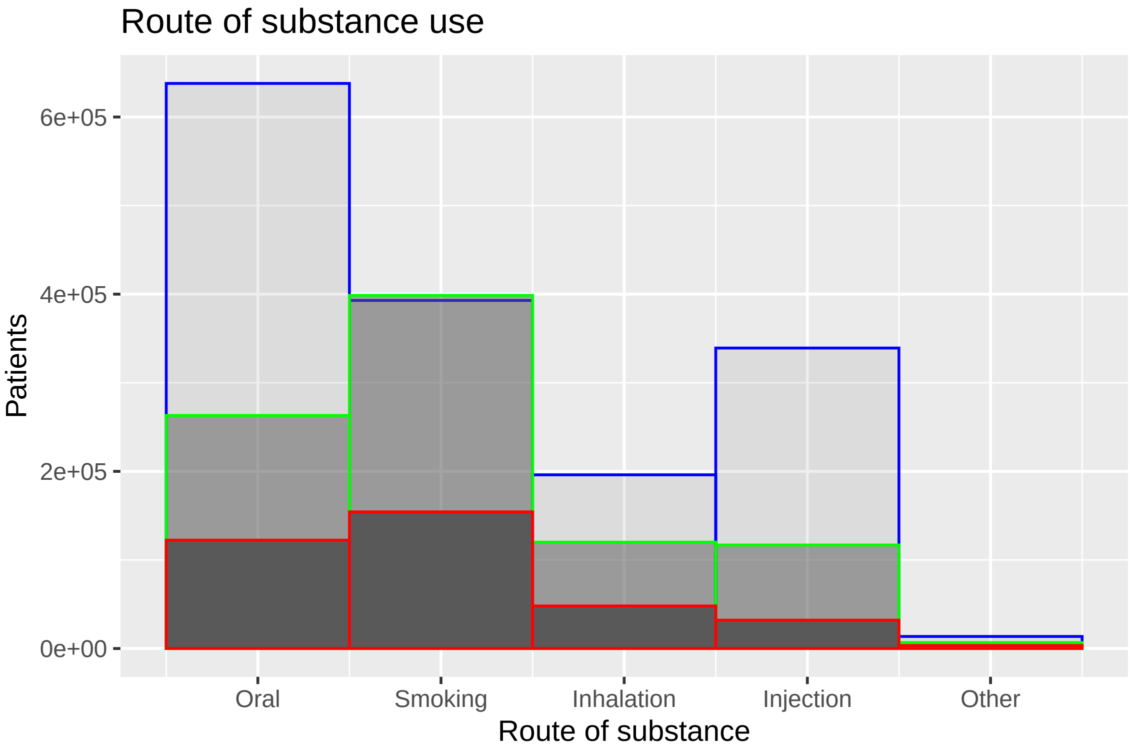

Figure 4: Histogram of route of substance administration of patients

The boxplot in figure 5 shows the distribution of patients length of stay in treatment in days versus the substance use type which they get treated for. The mean of length of stay when they indicated “none” as substance use type is 9 days. For alcohol and other drugs the treatment is much longer. For alcohol as substance use type the average length of stay is 22 days. For the patients with a substance use type of only drugs or alcohol and other drugs the average length of stay is 28 days. The length of stay has a large difference in all substance use type groups.

```
data$type_drug <- replace(data$ALCDRUG, data$ALCDRUG == 1, "Alcohol only")
data$type_drug <- replace(data$type_drug, data$type_drug == 2, "Other drugs only")
data$type_drug <- replace(data$type_drug, data$type_drug == 3, "Alcohol and other drugs")
data$type_drug <- replace(data$type_drug, data$type_drug == 0, "None")



y <- 1:37

ggplot(data = data, aes(factor(x=type_drug, level = c("None", "Alcohol only", "Other drugs only", "Alco
  geom_boxplot() +
  scale_y_continuous(labels = c(1:30,"31-45","46-60", "61-90", "91-120", "121-180", "181-365", ">365"),
                     breaks=y) +
  ylab("Length of stay (days)") +
  xlab("Type of substance use") +
  ggtitle("Length of stay in treatment by substance use type")
```

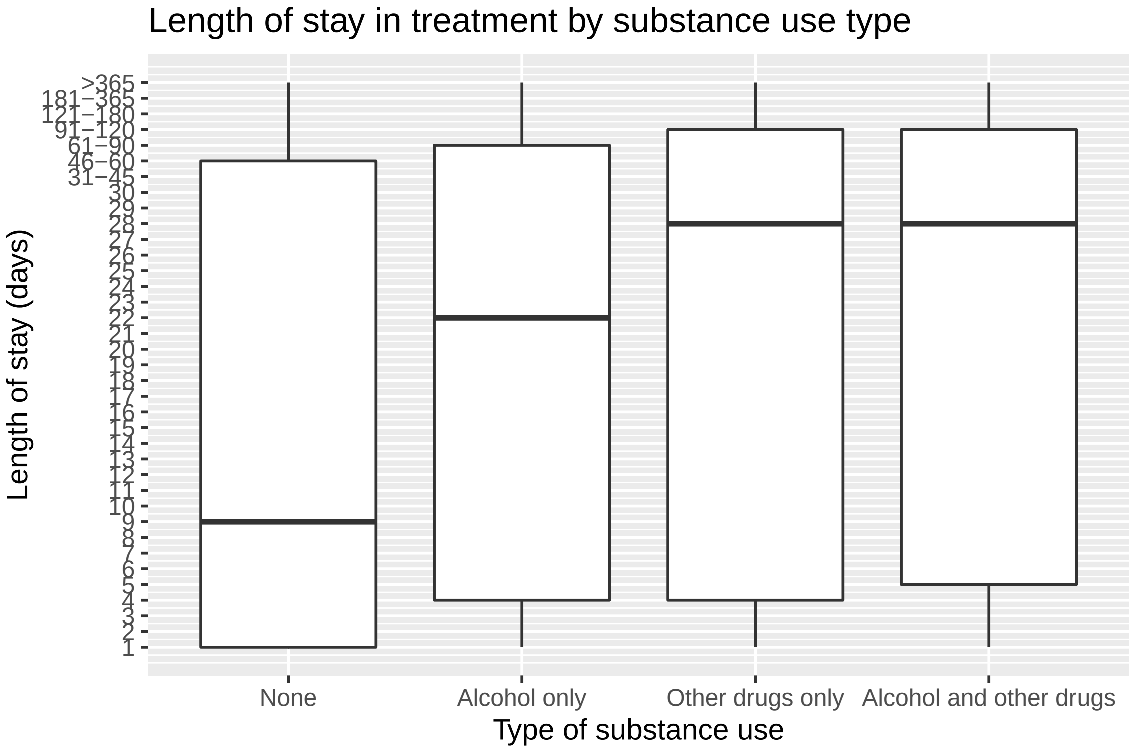

Figure 5: Boxplot length of stay in treatment in days by the substance use type of patients

Conclusion & Discussion

Overall, as can be seen in the figures (see Results), the data is widely spread. The age of patients and the age of first use of patients seem normally distributed. The attributes of variables can be discussed. In some cases the attributes are vague and can be wide interpreted. This is in the case of frequency of use, there are three groups in which the data is divided: No use in the past month, Some use and Daily use. As the first group is very clear, the second group is to vague. What is meant by some use? Also the third group, daily use, can be up for discussion; how much is the substance used daily.