Logbook

Susan Reefman

19/11/2021

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

Meaning of the numeric values in the dataset can be seen in the codebook: https://www.datafiles.samhsa.gov/sites/default/files/field-uploads-protected/studies/TEDS-D-2019/TEDS-D-2019-datasets/TEDS-D-2019-DS0001/TEDS-D-2019-DS0001-info/TEDS-D-2019-DS0001-info-codebook_V1.pdf

```
# Load dataset
load("~/Bio-informatica/BFV3/Semester1/Themaopdracht09/tedsd_puf_2019.RData")

# get relevant data from data set
p <- "CASEID|AGE|EDUC|^EMPLOY|ALCDRUG|^SUB|^FREQ[1|2|3]|^ROUTE|^FRSTUSE|LOS|NOPRIOR|PSYPROB|DSMCRIT|REAS
data <- df[, grep(pattern=p, colnames(df))]


codebook <- read.csv("~/Bio-informatica/BFV3/Semester1/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

Variable	Type	Label	number.of.possible.values
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
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

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
data[data == -9] <- NA
```

Variation and distribution

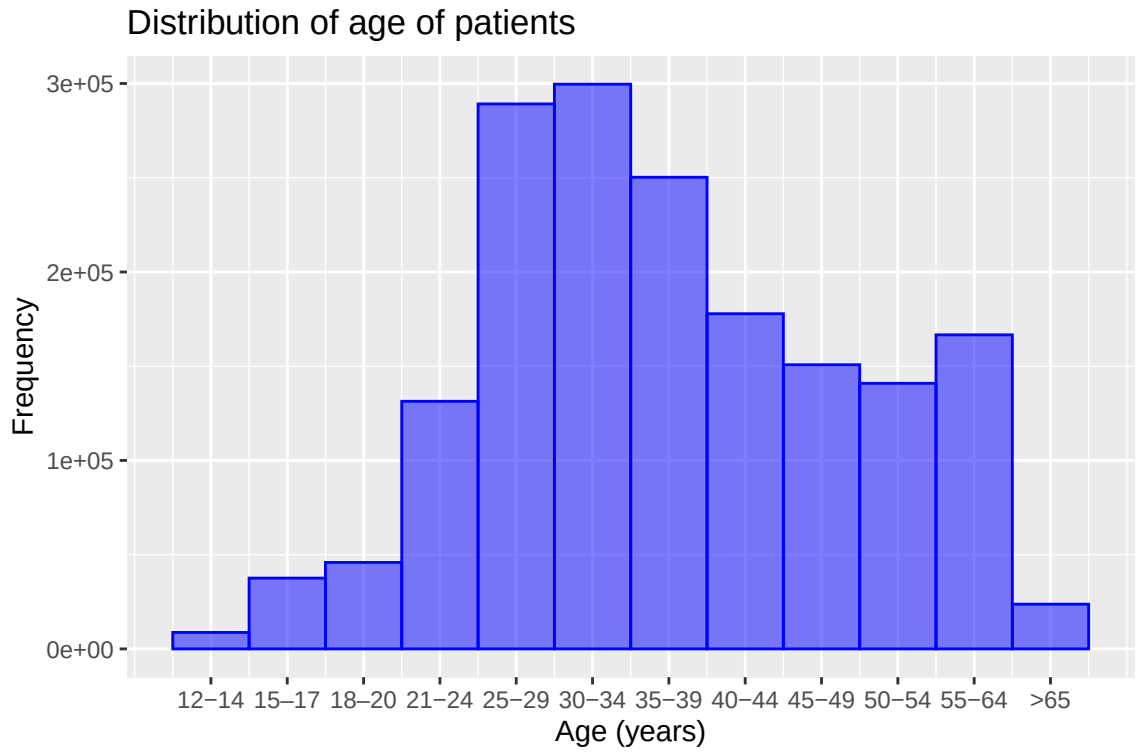
To get an idea of what the data includes an exploratory data analysis is done. The barplots display the amount of patients, in age groups, when they first used and their frequency of use. Also the amount of patients and their substance use and route of administration are displayed in a barplot. A boxplot is used to show the length of stay in treatment versus the type of substance use.

Results

With ggplot is a histogram made to depict the distribution of the age of patients in groups.

```
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)
```



The data is for used instances converted to strings instead of numeric values, this is done for convenient plotting purposes. Multiple stacked barplots are made to show the distribution of values in groups.

```
data$frstuse <- replace(data$FRSTUSE1, data$FRSTUSE1 == 1, "<11")
data$frstuse <- replace(data$frstuse, data$frstuse == 2, "12-14")
data$frstuse <- replace(data$frstuse, data$frstuse == 3, "15-17")
data$frstuse <- replace(data$frstuse, data$frstuse == 4, "18-20")
data$frstuse <- replace(data$frstuse, data$frstuse == 5, "21-24")
data$frstuse <- replace(data$frstuse, data$frstuse == 6, "25-29")
data$frstuse <- replace(data$frstuse, data$frstuse == 7, ">30")

df_frstuse <- count(data$frstuse)
df_frstuse$frstuse <- c("frstuse1")
colnames(df_frstuse) <- c("age", "freq", "frstuse")

data$frstuse2 <- replace(data$FRSTUSE2, data$FRSTUSE2 == 1, "<11")
data$frstuse2 <- replace(data$frstuse2, data$frstuse2 == 2, "12-14")
data$frstuse2 <- replace(data$frstuse2, data$frstuse2 == 3, "15-17")
data$frstuse2 <- replace(data$frstuse2, data$frstuse2 == 4, "18-20")
```

```
data$frstuse2 <- replace(data$frstuse2, data$frstuse2 == 5, "21-24")
data$frstuse2 <- replace(data$frstuse2, data$frstuse2 == 6, "25-29")
data$frstuse2 <- replace(data$frstuse2, data$frstuse2 == 7, ">30")

df2_frstuse <- count(data$frstuse)
df2_frstuse$frstuse <- c("frstuse2")
colnames(df2_frstuse) <- c("age", "freq", "frstuse")

data$frstuse3 <- replace(data$FRSTUSE3, data$FRSTUSE3 == 1, "<11")
data$frstuse3 <- replace(data$frstuse3, data$frstuse3 == 2, "12-14")
data$frstuse3 <- replace(data$frstuse3, data$frstuse3 == 3, "15-17")
data$frstuse3 <- replace(data$frstuse3, data$frstuse3 == 4, "18-20")
data$frstuse3 <- replace(data$frstuse3, data$frstuse3 == 5, "21-24")
data$frstuse3 <- replace(data$frstuse3, data$frstuse3 == 6, "25-29")
data$frstuse3 <- replace(data$frstuse3, data$frstuse3 == 7, ">30")

df3_frstuse <- count(data$frstuse)
df3_frstuse$frstuse <- c("frstuse3")
colnames(df3_frstuse) <- c("age", "freq", "frstuse")

total <- rbind(df_frstuse, df2_frstuse, df3_frstuse)

p1 <- ggplot(total, aes(fill=frstuse, y=freq, x=age)) +
            geom_bar(position="stack", stat="identity") +
            xlab("Age (years)") +
            ylab("Patients") +
            theme(axis.text.x = element_text(angle = 90))
```

```
data$freq <- replace(data$FREQ1, data$FREQ1 == 1, "No use in the past month")
data$freq <- replace(data$freq, data$freq == 2, "Some use")
data$freq<- replace(data$freq, data$freq == 3, "Daily use")

df_freq <- count(data$freq)
df_freq$freq_use <- c("freq1")
colnames(df_freq) <- c("frequency", "freq", "freq_use")

data$freq2 <- replace(data$FREQ2, data$FREQ2 == 1, "No use in the past month")
data$freq2 <- replace(data$freq2, data$freq2 == 2, "Some use")
data$freq2 <- replace(data$freq2, data$freq2 == 3, "Daily use")

df2_freq <- count(data$freq2)
df2_freq$freq_use <- c("freq2")
colnames(df2_freq) <- c("frequency", "freq", "freq_use")

data$freq3 <- replace(data$FREQ3, data$FREQ3 == 1, "No use in the past month")
data$freq3 <- replace(data$freq3, data$freq3 == 2, "Some use")
data$freq3 <- replace(data$freq3, data$freq3 == 3, "Daily use")

df3_freq <- count(data$freq3)
df3_freq$freq_use <- c("freq3")
colnames(df3_freq) <- c("frequency", "freq", "freq_use")

total1 <- rbind(df_freq, df2_freq, df3_freq)
```

```
p2 <- ggplot(total1, aes(fill=freq_use, y=freq, x=frequency)) +
            geom_bar(position="stack", stat="identity") +
            xlab("Frequency of use") +
            ylab("Patients") +
            theme(axis.text.x = element_text(angle = 90))
```

```
data$sub <- replace(data$SUB1, data$SUB1 == 1, "None")
data$sub <- replace(data$sub, data$sub == 2, "Alcohol")
data$sub <- replace(data$sub, data$sub == 3, "Cocaine/crack")
data$sub <- replace(data$sub, data$sub == 4, "Marijuana/hashish")
data$sub <- replace(data$sub, data$sub == 5, "Heroin")
data$sub <- replace(data$sub, data$sub == 6, "Non-prescription methadone")
data$sub <- replace(data$sub, data$sub == 7, "Other opiates and synthetics")
data$sub <- replace(data$sub, data$sub == 8, "PCP")
data$sub <- replace(data$sub, data$sub == 9, "Other hallucinogens")
data$sub <- replace(data$sub, data$sub == 10, "Methamphetamine")
data$sub <- replace(data$sub, data$sub == 11, "Other amphetamines")
data$sub <- replace(data$sub, data$sub == 12, "Other stimulants")
data$sub <- replace(data$sub, data$sub == 13, "Benzodiazepines")
data$sub <- replace(data$sub, data$sub == 14, "Other non-benzodiazepine tranquilizers")
data$sub <- replace(data$sub, data$sub == 15, "Barbiturates")
data$sub <- replace(data$sub, data$sub == 16, "Other non-barbiturate sedatives or hypnotics")
data$sub <- replace(data$sub, data$sub == 17, "Inhalants")
data$sub <- replace(data$sub, data$sub == 18, "Over-the-counter medications")
data$sub <- replace(data$sub, data$sub == 19, "Other")

df_sub <- count(data$sub)
df_sub$sub <- c("sub1")
colnames(df_sub) <- c("substance", "freq", "sub")

data$sub2 <- replace(data$SUB2, data$SUB2 == 1, "None")
data$sub2 <- replace(data$sub2, data$sub2 == 2, "Alcohol")
data$sub2 <- replace(data$sub2, data$sub2 == 3, "Cocaine/crack")
data$sub2 <- replace(data$sub2, data$sub2 == 4, "Marijuana/hashish")
data$sub2 <- replace(data$sub2, data$sub2 == 5, "Heroin")
data$sub2 <- replace(data$sub2, data$sub2 == 6, "Non-prescription methadone")
data$sub2 <- replace(data$sub2, data$sub2 == 7, "Other opiates and synthetics")
data$sub2 <- replace(data$sub2, data$sub2 == 8, "PCP")
data$sub2 <- replace(data$sub2, data$sub2 == 9, "Other hallucinogens")
data$sub2 <- replace(data$sub2, data$sub2 == 10, "Methamphetamine")
data$sub2 <- replace(data$sub2, data$sub2 == 11, "Other amphetamines")
data$sub2 <- replace(data$sub2, data$sub2 == 12, "Other stimulants")
data$sub2 <- replace(data$sub2, data$sub2 == 13, "Benzodiazepines")
data$sub2 <- replace(data$sub2, data$sub2 == 14, "Other non-benzodiazepine tranquilizers")
data$sub2 <- replace(data$sub2, data$sub2 == 15, "Barbiturates")
data$sub2 <- replace(data$sub2, data$sub2 == 16, "Other non-barbiturate sedatives or hypnotics")
data$sub2 <- replace(data$sub2, data$sub2 == 17, "Inhalants")
data$sub2 <- replace(data$sub2, data$sub2 == 18, "Over-the-counter medications")
data$sub2 <- replace(data$sub2, data$sub2 == 19, "Other")

df2_sub <- count(data$sub2)
df2_sub$sub <- c("sub2")
colnames(df2_sub) <- c("substance", "freq", "sub")
```

```
data$sub3 <- replace(data$SUB3, data$SUB3 == 1, "None")
data$sub3 <- replace(data$sub3, data$sub3 == 2, "Alcohol")
data$sub3 <- replace(data$sub3, data$sub3 == 3, "Cocaine/crack")
data$sub3 <- replace(data$sub3, data$sub3 == 4, "Marijuana/hashish")
data$sub3 <- replace(data$sub3, data$sub3 == 5, "Heroin")
data$sub3 <- replace(data$sub3, data$sub3 == 6, "Non-prescription methadone")
data$sub3 <- replace(data$sub3, data$sub3 == 7, "Other opiates and synthetics")
data$sub3 <- replace(data$sub3, data$sub3 == 8, "PCP")
data$sub3 <- replace(data$sub3, data$sub3 == 9, "Other hallucinogens")
data$sub3 <- replace(data$sub3, data$sub3 == 10, "Methamphetamine")
data$sub3 <- replace(data$sub3, data$sub3 == 11, "Other amphetamines")
data$sub3 <- replace(data$sub3, data$sub3 == 12, "Other stimulants")
data$sub3 <- replace(data$sub3, data$sub3 == 13, "Benzodiazepines")
data$sub3 <- replace(data$sub3, data$sub3 == 14, "Other non-benzodiazepine tranquilizers")
data$sub3 <- replace(data$sub3, data$sub3 == 15, "Barbiturates")
data$sub3 <- replace(data$sub3, data$sub3 == 16, "Other non-barbiturate sedatives or hypnotics")
data$sub3 <- replace(data$sub3, data$sub3 == 17, "Inhalants")
data$sub3 <- replace(data$sub3, data$sub3 == 18, "Over-the-counter medications")
data$sub3 <- replace(data$sub3, data$sub3 == 19, "Other")

df3_sub <- count(data$sub3)
df3_sub$sub <- c("sub3")
colnames(df3_sub) <- c("substance", "freq", "sub")

total2 <- rbind(df_sub, df2_sub, df3_sub)

p3 <- ggplot(total2, aes(fill=sub, y=freq, x=substance)) +
            geom_bar(position="stack", stat="identity")+
            xlab("Substance used") +
            ylab("Patients") +
            theme(axis.text.x = element_text(angle =90)) +
            ggtitle("Distribution of substance use of patients")
```

```
data$route <- replace(data$ROUTE1, data$ROUTE1 == 1, "Oral")
data$route <- replace(data$route, data$route == 2, "Smoking")
data$route <- replace(data$route, data$route == 3, "Inhalation")
data$route <- replace(data$route, data$route == 4, "Injection")
data$route <- replace(data$route, data$route == 5, "Other")

df_route <- count(data$route)
df_route$freq_route <- c("route1")
colnames(df_route) <- c("route", "freq", "freq_route")

data$route2 <- replace(data$ROUTE2, data$ROUTE2 == 1, "Oral")
data$route2 <- replace(data$route2, data$route2 == 2, "Smoking")
data$route2 <- replace(data$route2, data$route2 == 3, "Inhalation")
data$route2 <- replace(data$route2, data$route2 == 4, "Injection")
data$route2 <- replace(data$route2, data$route2 == 5, "Other")

df2_route <- count(data$route2)
df2_route$freq_route2 <- c("route2")
colnames(df2_route) <- c("route", "freq", "freq_route")
```

```r
data$route3 <- replace(data$ROUTE3, data$ROUTE3 == 1, "Oral")
data$route3 <- replace(data$route3, data$route3 == 2, "Smoking")
data$route3 <- replace(data$route3, data$route3 == 3, "Inhalation")
data$route3 <- replace(data$route3, data$route3 == 4, "Injection")
data$route3 <- replace(data$route3, data$route3 == 5, "Other")

df3_route <- count(data$route3)
df3_route$freq_route3 <- c("route3")
colnames(df3_route) <- c("route", "freq", "freq_route")

total3 <- rbind(df_route, df2_route, df3_route)

p4 <- ggplot(total3, aes(fill=freq_route, y=freq, x=route)) +
            geom_bar(position="stack", stat="identity") +
            xlab("Route of administration") +
            ylab("Patients") +
            theme(axis.text.x = element_text(angle = 90))
```

```r
grid.arrange(p1, p2, p4, top=textGrob("Distribution of use characteristics of patients"), ncol = 1)
```

```r
p3
```

```r
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", "Alcol
  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")
```

```r
# Length Of Stay data converted
data$los <- as.character(data$LOS)
data$los <- replace(data$los, data$los == 31, "31-45")
data$los <- replace(data$los, data$los == 32, "46-60")
data$los <- replace(data$los, data$los == 33, "61-90")
data$los <- replace(data$los, data$los == 34, "91-120")
data$los <- replace(data$los, data$los == 35, "121-180")
data$los <- replace(data$los, data$los == 36, "181-365")
data$los <- replace(data$los, data$los == 37, ">365")
```

Cleaning dataset

Because the dataset is very large and most of the data is not included in this project. Only the characteristics of the use of patients is necessary, which include: Type of use (alcohol, drugs or both), the substance of use,

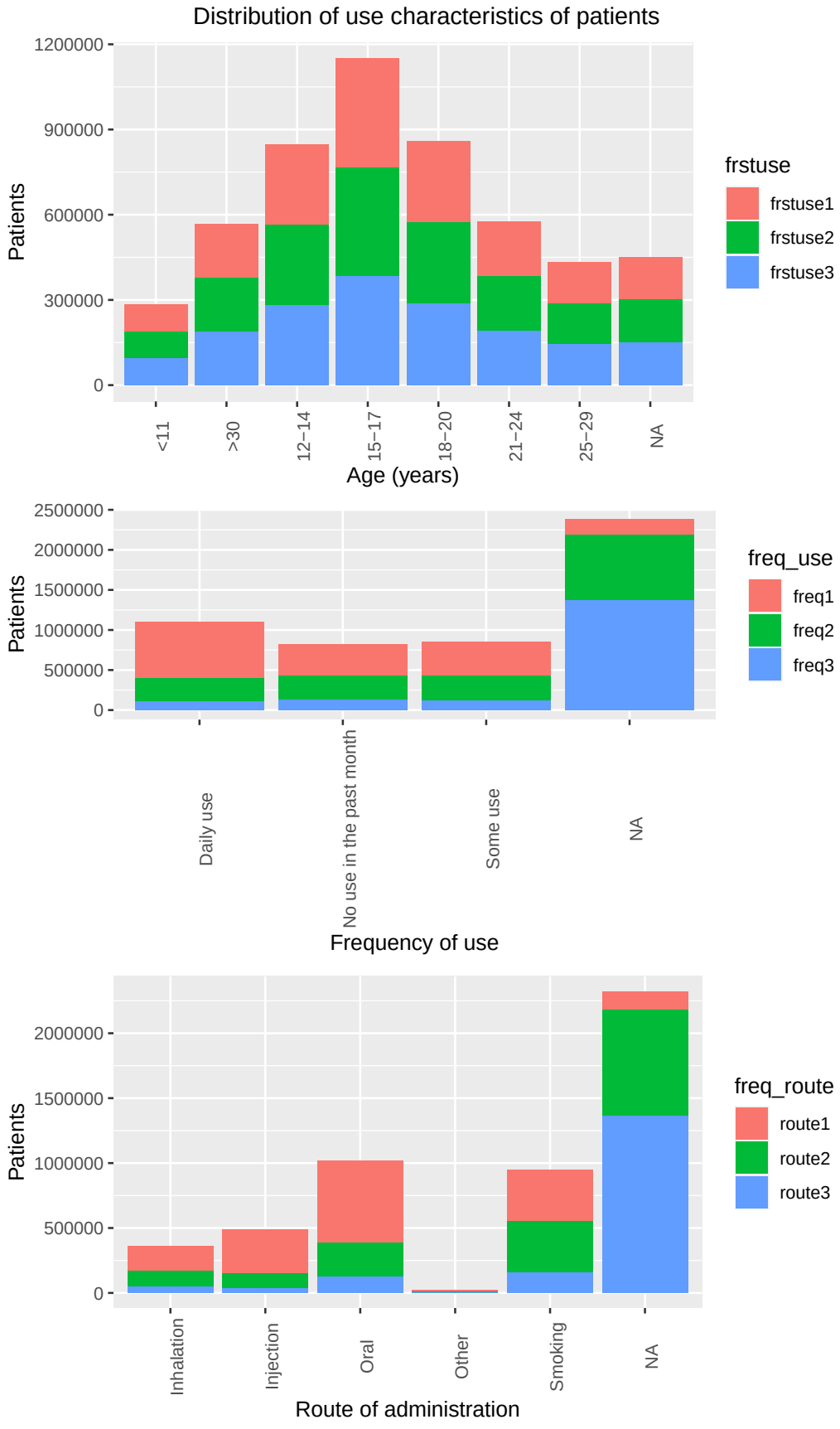


Figure 1: Figure 2: Stacked barplots of use characteristics of patients

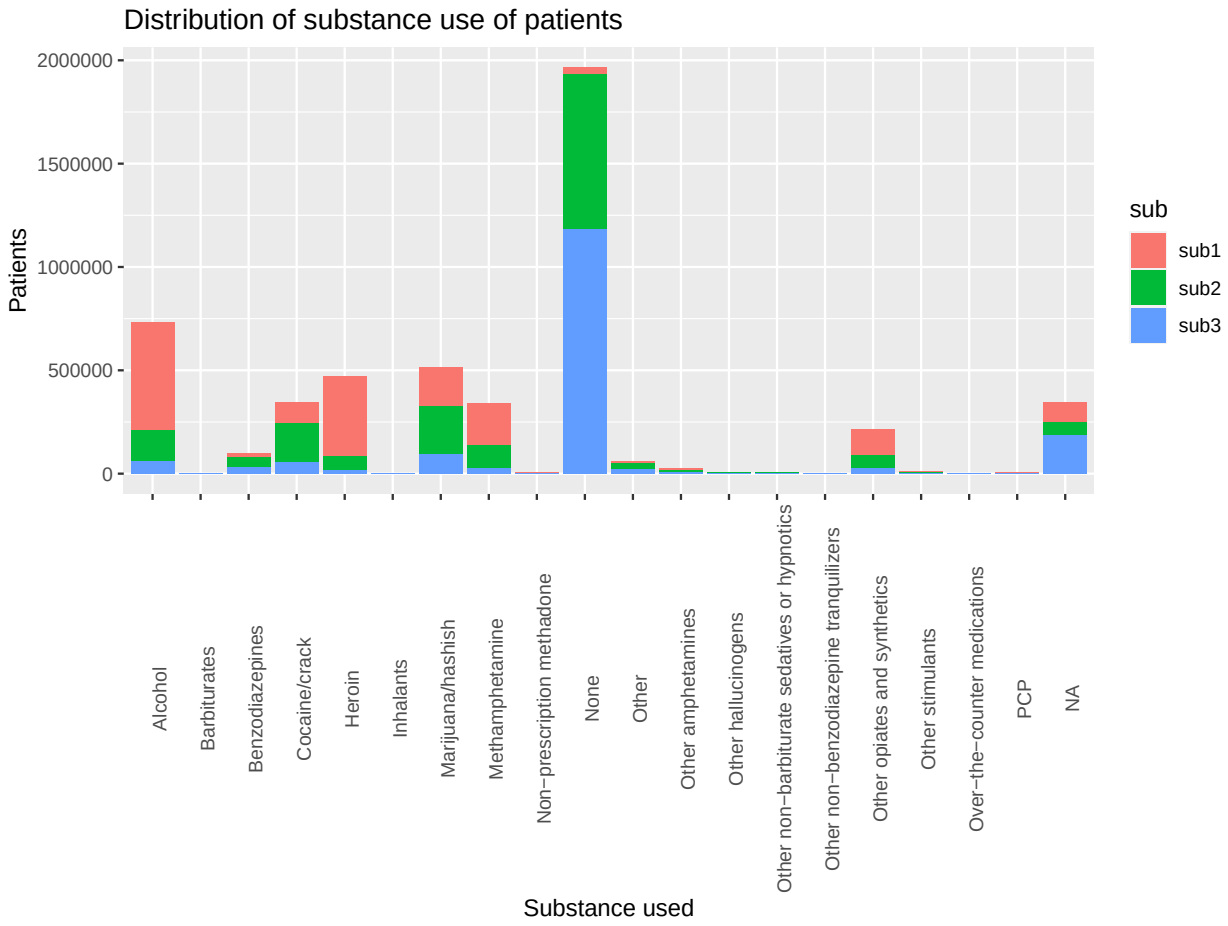


Figure 2: Figure 3: Stacked barplot of substance use of patients

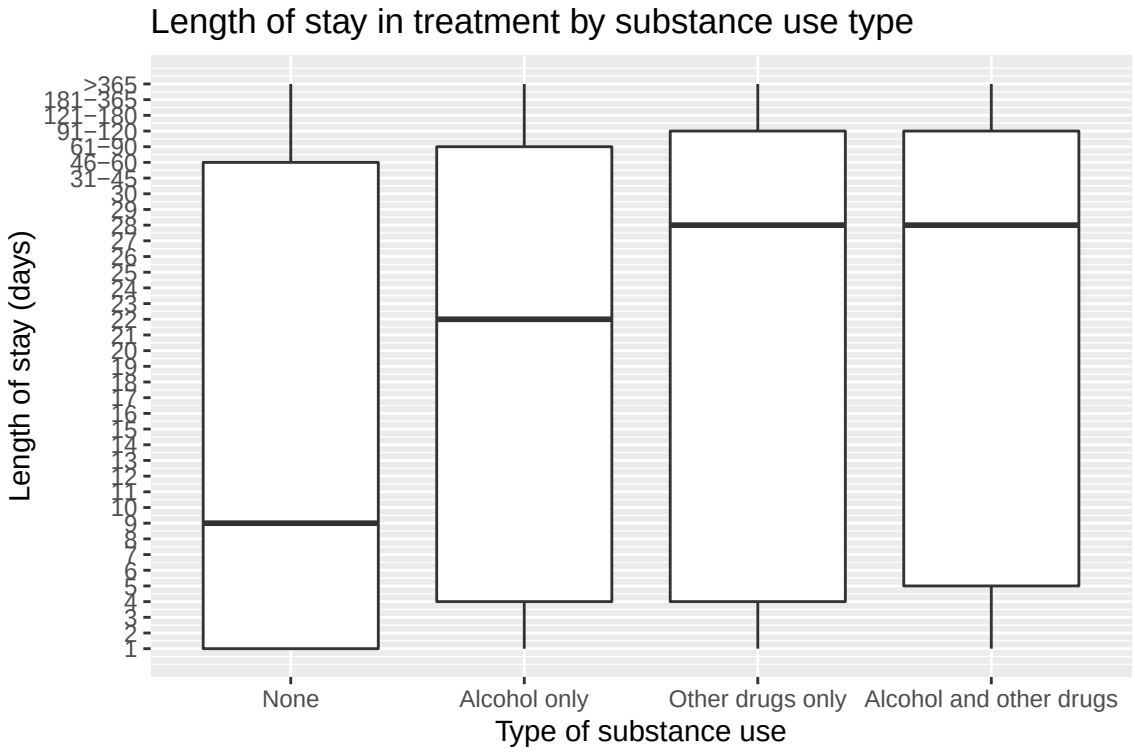


Figure 3: Figure 4: Boxplot length of stay in treatment in days by the substance use type of patients

the frequency of use and the age of first use. The length of stay is also necessary in this project. Because the dataset is this large, the patients which have information that is not available is removed from the dataset. This includes around 200.000 patients, the dataset still includes more than 1.5 million patients. Removing the missing values makes it also easier for the machine learning algorithm to classify instances. Thereafter a sample of the dataset is taken. Both the dataset and the sample are written to csv files, to be uploaded to Weka.

```
pat <- "type_drug|sub$|freq$|frstuse$|los"

# get relevant data from dataset
dataset <- data[, grep(pattern=pat, colnames(data))]

# remove rows with NA's
dataset <- na.omit(dataset)

dataset_sample <- sample_n(dataset, 1520681/10)
```

```
# Write datasets to csv files

write.csv(dataset, file = "dataset_thema09.csv")
write.csv(dataset_sample, file = "dataset_sample_thema09.csv")
```

Weka

Weka Explorer

The csv files are uploaded in the Weka Explorer. To validate the sample, both datasets were investigated with the ZeroR machine learning algorithm. In the table beneath, can be seen that the percentage of instances (in)correctly classified are not significant different. Taking the sample as the dataset to investigate further. The performance of standard machine learning algorithms was investigated. The classifier and attribute investigated were noted as well as the percentages (in)correctly classified instances.

```
weka <- read.table("~/Bio-informatica/BFV3/Semester1/Themaopdracht09/Weka_results/weka.csv",
                   sep = ',',
                   header = T)

knitr::kable(weka)
```

Classifier	Attribute	Percentage.Correctly.Classified.Instances	Percentage.Incorrect.Classified.Instances
ZeroR	los	13.6%	86.4%
ZeroR	sub	33.7%	67.3%
ZeroR	freq	45.5%	54.5%
ZeroR	frstuse	24.4%	75.6%
ZeroR	type_drug	53.9%	46.1%
OneR	los	16.9%	83.1%
OneR	sub	54.1%	45.9%
OneR	freq	53.7%	46.3%
OneR	frstuse	32.2%	67.8%
OneR	type_drug	71.4%	28.6%
J48	los	44.7%	55.3%
J48	sub	62.5%	37.5%
J48	freq	53.9%	46.1%
J48	frstuse	32.8%	67.2%
J48	type_drug	72.6%	27.4%
NaÃ¯ve Bayes	los	14.3%	85.7%
NaÃ¯ve Bayes	sub	58.6%	41.4%
NaÃ¯ve Bayes	freq	51.8%	48.2%
NaÃ¯ve Bayes	frstuse	26.9%	73.1%
NaÃ¯ve Bayes	type_drug	70.7%	29.3%
SimpleLogistics	los	14.6%	85.4%
SimpleLogistics	sub	58.9%	41.1%
SimpleLogistics	freq	52.3%	47.7%
SimpleLogistics	frstuse	27.6%	72.4%
SimpleLogistics	type_drug	72.5%	27.5%
NearestNeighbor	los	98%	2%
NearestNeighbor	sub	99%	1%
NearestNeighbor	freq	100%	0%
NearestNeighbor	frstuse	100%	0%

Classifier	Attribute	Percentage.Correctly.Classified.Instances	Percentage.Incorrect.Classified.Instances
NearestNeighbor	type_drug	100%	0%

Weka Experimenter

Using the Weka Experimenter machine learning algorithms with 10-fold cross-validations were performed. Results of each machine learning algorithm is separately saved in the Weka_results folder.

```
pat <- "Key_Run|Key_Fold|Percent_correct|Percent_incorrect|True_positive_rate|False_positive_rate|True_
coln <- c("Key run", "Key fold", "% correct", "% incorrect", "TPR", "FPR", "TNR", "FNR", "Time training

# Read table from csv file
zeror <- read.table("~/Bio-informatica/BFV3/Semester1/Themaopdracht09/Weka_results/ZeroR.csv",
                        sep = ',',
                        header = T)

# Get relevant information
zeror <- zeror[, grep(pattern=pat, colnames(zeror))]
colnames(zeror) <- coln

knitr::kable(head(zeror), digits = 3)
```

Key run	Key fold	% correct	% incorrect	TPR	FPR	TNR	FNR	Time training	Time testing
1	1	13.638	86.362	0	0	1	1	0.132	0.054
1	2	13.638	86.362	0	0	1	1	0.105	0.052
1	3	13.638	86.362	0	0	1	1	0.115	0.051
1	4	13.638	86.362	0	0	1	1	0.106	0.052
1	5	13.636	86.364	0	0	1	1	0.116	0.049
1	6	13.636	86.364	0	0	1	1	0.117	0.056

```
# Read table from csv file
oner <- read.table("~/Bio-informatica/BFV3/Semester1/Themaopdracht09/Weka_results/OneR.csv",
                        sep = ',',
                        header = T)

# Get relevant information
oner <- oner[, grep(pattern=pat, colnames(oner))]
colnames(oner) <- coln

knitr::kable(head(oner),  digits = 3)
```

Key run	Key fold	% correct	% incorrect	TPR	FPR	TNR	FNR	Time training	Time testing
1	1	10.071	89.929	0.005	0.005	0.995	0.995	0.922	0.212
1	2	10.126	89.874	0.004	0.005	0.995	0.996	0.734	0.178
1	3	10.237	89.763	0.006	0.006	0.994	0.994	0.389	0.165
1	4	10.347	89.653	0.004	0.005	0.995	0.996	0.423	0.164
1	5	9.949	90.051	0.003	0.006	0.994	0.997	0.793	0.226
1	6	10.215	89.785	0.004	0.005	0.995	0.996	0.468	0.256

```
# Read table from csv file
naivebayes <- read.table("~/Bio-informatica/BFV3/Semester1/Themaopdracht09/Weka_results/naivebayes.csv"
                         sep = ',',
                         header = T)

# Get relevant information
naivebayes <- naivebayes[, grep(pattern=pat, colnames(naivebayes))]
colnames(naivebayes) <- coln

knitr::kable(head(naivebayes),  digits = 3)
```

Key run	Key fold	% correct	% incorrect	TPR	FPR	TNR	FNR	Time training	Time testing
1	1	85.705	14.295	0.000	0	1	1.000	0.249	0.539
1	2	85.715	14.285	0.000	0	1	1.000	0.243	0.455
1	3	85.482	14.518	0.000	0	1	1.000	0.255	0.378
1	4	85.602	14.398	0.000	0	1	1.000	0.247	0.497
1	5	85.839	14.161	0.001	0	1	0.999	0.250	0.381
1	6	85.786	14.214	0.000	0	1	1.000	0.323	0.374

```
# Read table from csv file
Ikb <- read.table("~/Bio-informatica/BFV3/Semester1/Themaopdracht09/Weka_results/Ikb.csv",
                         sep = ',',
                         header = T)

# Get relevant information
Ikb <- Ikb[, grep(pattern=pat, colnames(Ikb))]
colnames(Ikb) <- coln

knitr::kable(head(Ikb),  digits = 3)
```

Key run	Key fold	% correct	% incorrect	TPR	FPR	TNR	FNR	Time training	Time testing
1	1	17.023	82.977	0.056	0.034	0.966	0.944	0.142	2704.603
1	2	17.053	82.947	0.058	0.034	0.966	0.942	0.138	2704.603
1	3	17.214	82.786	0.064	0.036	0.964	0.936	0.148	2704.603
1	4	17.061	82.939	0.058	0.035	0.965	0.942	0.137	2704.603
1	5	17.128	82.872	0.066	0.034	0.966	0.934	0.135	2704.603
1	6	17.319	82.681	0.060	0.036	0.964	0.940	0.136	2704.603

```
# Read table from csv file
SL <- read.table("~/Bio-informatica/BFV3/Semester1/Themaopdracht09/Weka_results/SimpleLogistics.csv",
                         sep = ',',
                         header = T)

# Get relevant information
SL <- SL[, grep(pattern=pat, colnames(SL))]
colnames(SL) <- coln

knitr::kable(head(SL),  digits = 3)
```

Key run	Key fold	% correct	% incorrect	TPR	FPR	TNR	FNR	Time training	Time testing
1	1	72.977	27.023	0.056	0.034	0.966	0.944	0.161	2891.072
1	2	72.947	27.053	0.058	0.034	0.966	0.942	0.184	2922.201
1	3	72.786	27.214	0.064	0.036	0.964	0.936	0.132	2638.446
1	4	72.939	27.061	0.058	0.035	0.965	0.942	0.133	2633.514
1	5	72.872	27.128	0.066	0.034	0.966	0.934	0.136	2634.736
1	6	72.681	27.319	0.060	0.036	0.964	0.940	0.132	2634.835

```
# Read table from csv file
smo <- read.table("~/Bio-informatica/BFV3/Semester1/Themaopdracht09/Weka_results/SMO.csv",
                        sep = ',',
                        header = T)

# Get relevant information
smo <- smo[, grep(pattern=pat, colnames(smo))]
colnames(smo) <- coln

knitr::kable(head(smo),  digits = 3)
```

Key run	Key fold	% correct	% incorrect	TPR	FPR	TNR	FNR	Time training	Time testing
1	1	11.519	88.481	0.082	0.048	0.952	0.918	2187.026	0.427
1	2	11.549	88.483	0.078	0.046	0.954	0.922	2089.946	0.589
1	3	11.319	88.343	0.090	0.047	0.953	0.910	2002.441	0.397
1	4	11.349	88.938	0.082	0.046	0.954	0.918	1764.868	0.458
1	5	11.519	88.481	0.090	0.045	0.955	0.910	1953.202	0.411
1	6	11.519	88.481	0.093	0.045	0.955	0.907	1807.116	0.407

Optimize algorithms

For further optimization Naive Bayes and Simple Logistics algorithms are chosen. First Naive Bayes is investigated. Changing to the `Train/Test Percentage Split (data randomized)` test option: With 10 runs and the split percentage ranging from 40 to 90%, the algorithm classifies between the 70.66% (40% split percentage) and 70.74% (66% split percentage). This is not better than the original Naive Bayes algorithm. Thereafter the k-fold cross-validation was tested. But also here, no optimization by changing the number of folds higher or lower.

Moving on to the Simple Logistics algorithm: Changing again the `Train/Test Percentage Split (data randomized)` test option: With 10 runs and the split percentage ranging from 40 to 90% the algorithm, the algorithm classified every time 72% of the instances correctly. This is not worse, but also no optimization. Optimization was achieved with increasing the number of folds with the k-fold cross-validation test-option. With 15-folds, 75% of the instances were correctly but the time performing the algorithm doubled, so in general would the increase not be sufficient.

```
meta <- c("Bagging", "Stacking", "Boosting") #rows
training_correct <- c("43.3%", "13.6%", "13.6%")
training_incorrect <- c("56.7%", "86.4%", "86.4%")

fold_correct <- c("9.6%", "13.6%", "13.6%")
fold_incorrect <- c("90.3%","86.4%", "86.4%")
```

```
training <- data.frame(training_correct, training_incorrect)
row.names(training) <- meta
colnames(training) <- c("Correctly classified", "Incorrectly classified")
knitr::kable(training)
```

	Correctly classified	Incorrectly classified
Bagging	43.3%	56.7%
Stacking	13.6%	86.4%
Boosting	13.6%	86.4%

```
fold <- data.frame(fold_correct, fold_incorrect)
row.names(fold) <- meta
colnames(fold) <- c("Correctly classified", "Incorrectly classified")
knitr::kable(fold)
```

	Correctly classified	Incorrectly classified
Bagging	9.6%	90.3%
Stacking	13.6%	86.4%
Boosting	13.6%	86.4%

Visualizing

Visualizing the results in a ROC curve, figure 5, can be done in the Explorer. By running the Naive Bayes algorithm with 10-fold cross-validation, being the best model, the ROC curve can be visualized. The area under the curve is 0.6664

Creating a learning curve can be done in the advanced mode of the Experimenter. For destination choose `CSVResultListener` and for `Result generator`, `RandomSplitResultProducer` with again 90.0 as `trainPercent`.The `splitEvaluator` is set to `CostSensitiveClassifier` which is set to the Naive Bayes classifier. After taking the data and error rate out of Weka a learning curve is made. As you can see in figure 6 the bottom plot the algorithm learns most when 90% of the data is used. The algorithm starts to learn slow when less than 20% is used.