Classification

University of St. Gallen
School of Management, Economics, Law,
Social Sciences, International Affairs
and Computer Science

Assignment 4

Data Analytics I: Predictive Econometrics
Prof. Jana Mareckova

submitted by

Cyril Janak, 16-611-287
Jonas Husmann, 16-610-917
Niklas Kampe, 16-611-618
Robin Scherrer, 18-617-969

22.12.2021

Contents

Requirements

To solve the following tasks, the required libraries and the data sets are loaded first.

```
library(rpart)
library(rpart.plot)
library(dplyr)

load("GHA/drugs.RData")
```

Exercise 1

The share of males who consume soft drugs is ~29.18%

```
(m_s_drug <- (nrow(drugs[drugs$Gender=="male" & drugs$Soft_Drug==T,]) /
   nrow(drugs[drugs$Gender=="male",]) * 100) %>%
   round(., digits = 2) %>%
   paste0(., "%"))
```

```
## [1] "29.18%"
```

Exercise 2

The difference between the share of male and female hard drug consumers is ~2.74%

```
m_h_drug <- nrow(drugs[drugs$Gender=="male" & drugs$Hard_Drug==T,]) /
   nrow(drugs[drugs$Gender=="male",])

f_h_drug <- nrow(drugs[drugs$Gender=="female" & drugs$Hard_Drug==T,]) /
   nrow(drugs[drugs$Gender=="female",])

(diff_h_drug <- ((m_h_drug - f_h_drug) * 100) %>%
      round(., digits = 2) %>%
      paste0(., "%"))
```

```
## [1] "2.74%"
```

Exercise 3

From the shares of soft drug consumption for each age group, one can observe that only 16-17 year-olds consume soft drugs. Therefore, the consumption of soft drugs is decreasing in age, but not strictly as the groups of 18-19 and 20-24 year-olds are not consuming any soft drugs at all.

```
share_softdrugs_16_17 <- round((nrow(drugs[drugs$Age=="16-17 years" &
                                              drugs$Soft_Drug==T,]) /
   nrow(drugs[drugs$Age=="16-17 years",]))*100, digits = 2)
share_softdrugs_18_19 <- round((nrow(drugs[drugs$Age=="18-19 years" &
                                              drugs$Soft_Drug==T,]) /
   nrow(drugs[drugs$Age=="18-19 years",]))*100, digits = 2)
share_softdrugs_20_24 <- round((nrow(drugs[drugs$Age=="20-24 years" &
                                              drugs$Soft_Drug==T,]) /
   nrow(drugs[drugs$Age=="20-24 years",]))*100, digits = 2)

(shares_softdrugs <- data.frame(
   age = c("16-17 Years", "18-19 Years", "20-24 Years"),
   share = c(share_softdrugs_16_17, share_softdrugs_18_19, share_softdrugs_20_24)))
```

```
##           age share
## 1 16-17 Years  48.5
## 2 18-19 Years   0.0
## 3 20-24 Years   0.0
```

Exercise 4

The chi-squared test results in a X-squared statistic of 9.40 at a p-value of 0.025. Hence, the hypothesis of independence is rejected (0.025 < 0.05) and the earnings range and soft drug consumption are indeed dependent at a condifence interval of 5%.

```
drugs_table <- table(drugs$Earning, drugs$Soft_Drug)
chi_squared <- chisq.test(drugs_table)
(statistics <- chi_squared$statistic)
```

```
## X-squared
##  9.401385
```

```
(p_value <- chi_squared$p.value)
```

```
## [1] 0.02440394
```