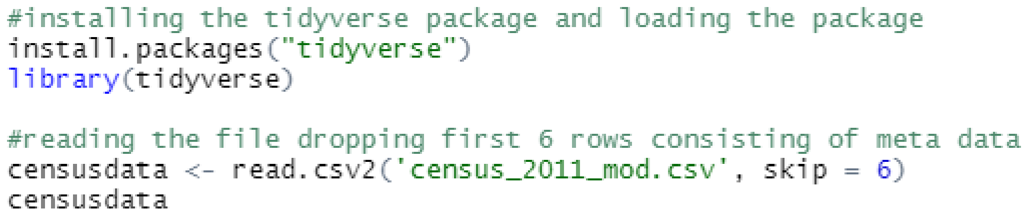
Data Mining Assignment 1

Submitted By:

Dileshwori Joshi (28385)

Importing the data

- ✓ Firstly, the objects in the local environment was removed and current working directory was checked.
- ✓ Tidyverse package was installed and loaded.
- ✓ Different packages provided by tidyverse like readr, dplyr, tidyr, and ggplot2 are used in the reading, manipulating, tidying and visualizing the data respectively.
- ✓ While reading the csv file, first six rows are skipped as it consisted of the meta data. Along with it last six rows are also removed.



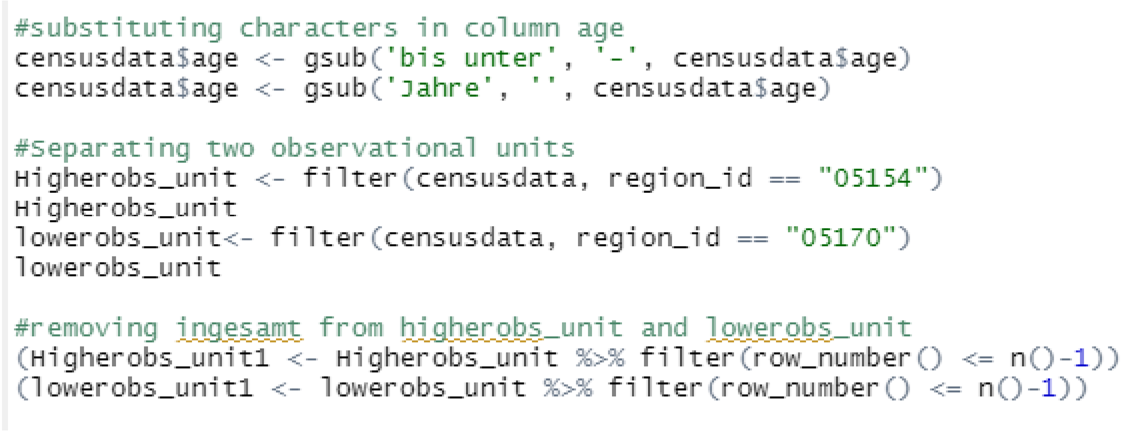
```
#installing the tidyverse package and loading the package
install.packages("tidyverse")
library(tidyverse)

#reading the file dropping first 6 rows consisting of meta data
censusdata <- read.csv2('census_2011_mod.csv', skip = 6)
censusdata
```

Fig 1: Code snippet for loading the library and importing the file

Cleaning the data

- ✓ Two observational units were separated, one with region_id =“05154” and other with region_id=“05170” .
- ✓ The row consisting of Ingesamt was removed to make visualization clearer.
- ✓ The observational unit with region_id =“05154” was pivoted in such a way that columns from employed_male and employed female were saved to column named category.
- ✓ The observational unit with region_id=“05170” was also arranged in such a way that it consists of three columns: age, category and cases



```
#substituting characters in column age
censusdata$age <- gsub('bis unter', '-', censusdata$age)
censusdata$age <- gsub('Jahre', '', censusdata$age)

#Separating two observational units
Higherobs_unit <- filter(censusdata, region_id == "05154")
Higherobs_unit
lowerobs_unit<- filter(censusdata, region_id == "05170")
lowerobs_unit

#removing ingesamt from higherobs_unit and lowerobs_unit
(Higherobs_unit1 <- Higherobs_unit %>% filter(row_number() <= n()-1))
(lowerobs_unit1 <- lowerobs_unit %>% filter(row_number() <= n()-1))
```

Fig 2: Code Snippet for cleaning data

Visualizing the data - I

- ✓ Observational units with region_id =“05154” was visualized using bar graph provided by ggplot2 package of tidyverse.

- ✓ The employed_male and employed_female population for different age groups are visualized using the bar graph as shown in Fig 3. The null rows values were removed before visualization.

- ✓ One of the finding from the bar graph is that the highest number of employed male and female are from age group 45-50.

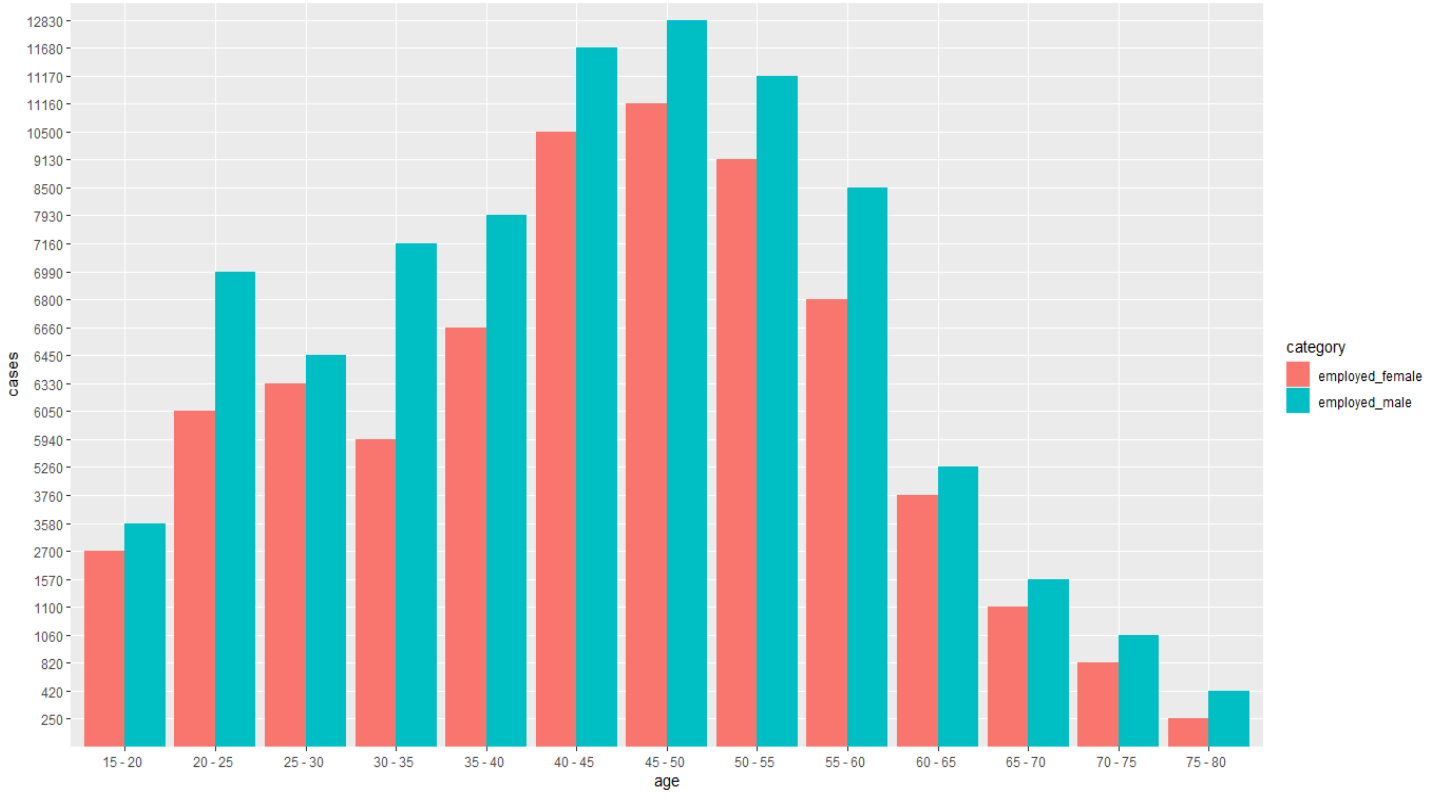


Fig 3: Visualizing age vs. cases with respect to category

Visualizing the data - II

- ✓ Another small observational unit is taken for region_id=05170.
- ✓ Four attributes employed_male, employed female, unemployed_male and unemployed_female according to the age distribution is plotted.

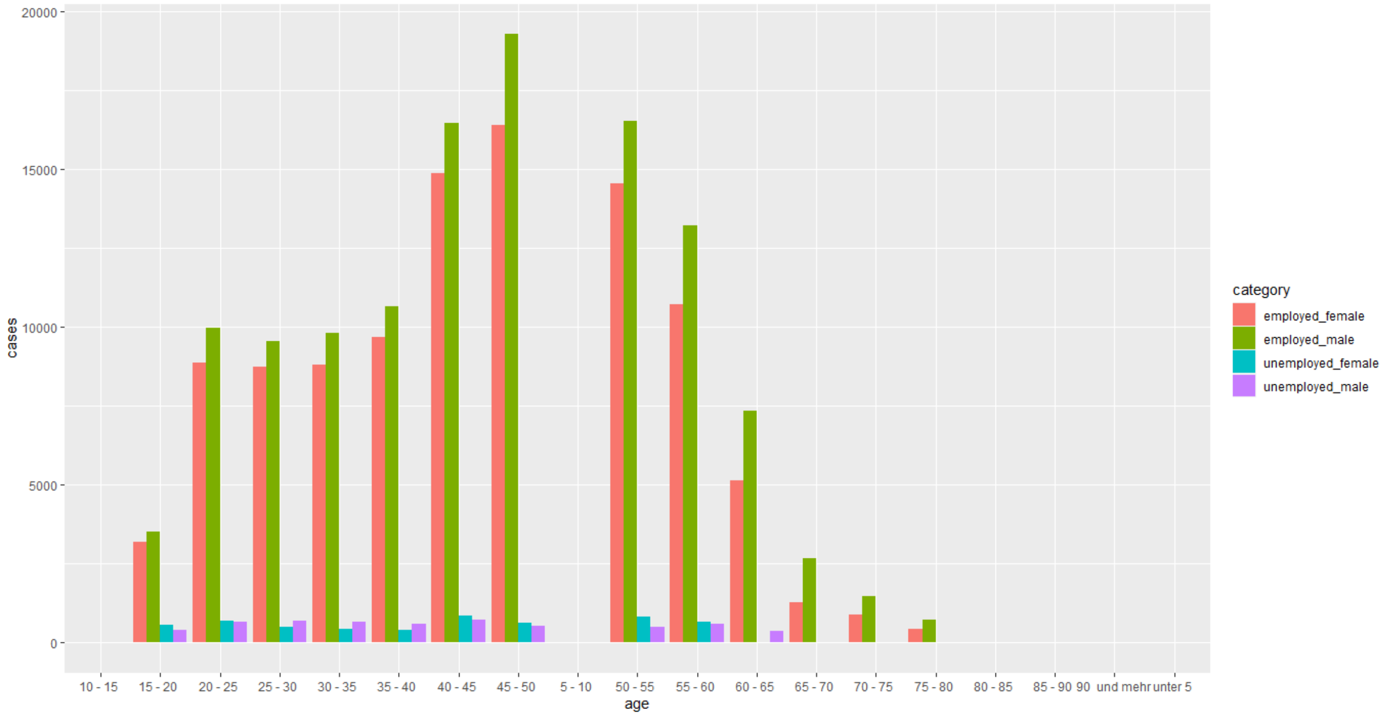


Fig 4: Visualizing employed_female vs. employed_male vs unemployed_male vs unemployed_female population in region_id="05170"

Thank You