1. 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.

2. 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.

3. 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 of the slide. 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.

4. 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.

ANNEX

```
#remove the list variables
ls()
rm(list = ls())
ls()

#get the working directory
getwd()

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

#reading the file dropping first 6 consisting of meta data
censusdata <- read.csv2('census_2011_mod.csv', skip = 6)
#dropping last 5 unnecessary rows with filter function
censusdata <- censusdata %>% filter(row_number() <= n()-5)
censusdata

#structure of the data
str(censusdata)

#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"))
(lowerobs_unit<- filter(censusdata, region_id == "05170"))

#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))

#tidying the data
#selecting two rows for visualization in Higherobs_unit1
Higherobs_unitselect <- select(Higherobs_unit1, age, employed_male, employed_female)
#removing rows with "x" values
Higherobs_unitfiltered <- filter(Higherobs_unitselect, employed_male != "X" & employed_female !="x")

#pivoting to form columns category and cases
first_observation <- Higherobs_unitfiltered %>%
  pivot_longer(employed_male:employed_female, names_to = "category", values_to = "cases")

#visualization of first_observation
ggplot(first_observation, aes(x = age, y=cases, fill=category)) +
  geom_col(position = "dodge")

#selecting only rows for visualization in lowerobs_unit1
lowerobs_unit1 <- select(lowerobs_unit1, age, employed_male, employed_female, unemployed_male, unemployed_female)
#pivoting
second_observation <- lowerobs_unit1 %>%
  pivot_longer(employed_male:unemployed_female, names_to = "category", values_to = "cases")
#replacing x values with numeric zero in column cases
second_observation$cases <- as.numeric(gsub("x", 0, second_observation$cases))

#visualization for tidydata in second_observation
ggplot(second_observation, aes(x = age, y=cases, fill=category)) +
  geom_col(position = "dodge")
```