```r
# Data structure: 3. Factor

# Factor a type of data structures used to store categorical variables. (e.g., gender)

# Gender vector
gender <- c("Male", "Female", "Female", "Male", "Male")
is.vector(gender)
class(gender)

# Convert gender_vector to a factor
factor_gender <- factor(gender)
factor_gender

# There are two different types of categorical variables:

# nominal categorical variable and ordinal categorical variable.
# 1. Norminal categorical variable: There is no implied order among categories.
# For example, Male and Female, and Cat, Dog, and Turtle
# 2. Ordinal categorical variable: There is a natural ordering.
# For example, "Low", "Medium" and "High", and "Primary", "Middle", "High"

# No ordering
pet <- c("Cat", "Dog", "Turtle", "Dog", "Cat", "Cat")
factor_pet <- factor(pet)
factor_pet
class(factor_pet)
factor_pet[1] > factor_pet[2]

# Natural ordering
income <- c("High", "High", "Low", "Midium", "Low")
factor_income <- factor(income, order = TRUE, levels = c("Low", "Midium", "High")) # You can specify the order (or level)
factor_income
class(factor_income)

income[3] > income[1]
factor_income[3] > factor_income[1]

# In survey, abbreviations are often used because it is convenient to record.
# But this can be confusing when you use survey data.
# You can recover the full words using levels function.
edu <- c("E", "E", "H", "M", "H", "C")
factor_edu <- factor(edu, order = TRUE, levels <- c("E", "M", "H", "C")) # If you don't specify the level, it will be alphabetical.
levels(factor_edu) <- c("Elemenary","Middle","High", "College")
factor_edu
as.numeric(factor_edu)

# summarize the factor
summary(factor_edu)
table(factor_edu)
summary(edu)

# Exercise 1. ################################################################################

set.seed(pi)
r <- rnorm(2754, 0, 1)
income <- exp(r)
hist(income, breaks=100)

# find 0.25, 0.5, 0.75, 0.95 quantiles of income. You can use quantile.
quant <- quantile(income, c(0.25, 0.5, 0.75, 0.95))

# Construct a vector, income.level, as follows:
# If income is <= 0.25 quantile, "VL"
# If income is > 0.25 quantile and <= 0.5 quantile, "L"
# If income is > 0.5 quantile and <= 0.75 quantile, "M"
# If income is > 0.75 quantile and <= 0.95 quantile, "H"
# If income is > 0.95 quantile, "VH"


# Make an ordered factor from income.level. Specify the levels as c("VL","L","M","H","VH")
factor_income.level <-

levels(factor_income.level) <- c("Very Low", "Low", "Middle", "High", "Very High")
summary(factor_income.level)      # You can see the summary of factor_income.level.

# Construct a subvector income.high that includes income that belongs to "High" and "Very High".
income.high <-
hist(income.high, breaks = 20)

# Calculate the average income of people who belong to "Middle" and "High".


# What is the difference between average income of "Very High" and average income of "High"



# Exercise 2. ################################################################################

industry <- sample(c("Manufacture", "Service", "IT"), 100, replace=TRUE, prob=c(0.3, 0.5, 0.2))
stock <- rep(NA,100)
stock[industry == "Manufacture"] <- rnorm(sum(industry=="Manufacture"), 3, 2)
stock[industry == "Service"] <- rnorm(sum(industry=="Service"), 2, 4)
stock[industry == "IT"] <- rnorm(sum(industry=="IT"), 8, 8)

factor.industry <- factor(industry)

# How many manufacturing, service, and IT companies?


# Compare the average stock prices and their standard deviations among these three industries.
mean(stock[factor.industry == "Manufacture"])
mean(stock[factor.industry == "Service"])
mean(stock[factor.industry == "IT"])
```