Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования «Кубанский государственный университет»

Кафедра вычислительных технологий

**ОТЧЕТ**

о выполнении лабораторной работы №4

по дисциплине «Обработка больших данных»

Выполнил: ст. гр. 39

Пшеничнов А.А.

Проверил: преподаватель

Яхонтов А.А.

Краснодар

2025

Тема: Извлечение данных с WEB-страниц. Пакет rvest.

Цель: Научиться работать извлекать информацию сWEB-страниц с помощью инструментов языка R.

Задание:

1. В ходе лабораторной работы, необходимо собрать информацию об уровне жизни стран мира из таблиц сайта https://www.numbeo.com/quality-of-life/rankings\_by\_country.jsp?title=2021 с 2014 по 2021гг:

2. Каждый студент должен взять 5 стран (по варианту) .

3. Составить data.frame (возможно для каждой страны) так, чтобы иметь возможность проанализировать с помощью графиков изменение рейтингов для всех 10 показателей для всех своих 5-ти стран, прокомментировать в отчете результат.

Необходимо нарисовать на одном и том же графике рейтинг всех 5 стран, проанализировать результат, анализ словесно отразить в отчете. Проанализировать изменение во времени всех показателей указанных стран, подобрать наилучший (с вашей точки зрения) способ визуализации.

4. С одной из страниц (по варианту) :

7) https://www.culture.ru/museums/institutes/location-krasnodarskiy-kray-krasnodar

собрать информацию в data.frame, которя содержала бы: Название музея, его адрес и ссылку для перехода при клике на фото / ссылке на музей.

Ход работы:

1)

library(rvest)

html\_list <- list()

nodes\_list <- list()

df\_list <- list()

for (year in 2021:2014) {

url <- paste0('https://www.numbeo.com/quality-of-life/rankings\_by\_country.jsp?title=', year)

html\_page <- read\_html(url)

html\_list[[as.character(year)]] <- html\_page

tables <- html\_nodes(html\_page, 'table')

nodes\_list[[as.character(year)]] <- tables

df <- html\_table(tables[[2]], fill = TRUE) %>% as.data.frame()

rownames(df) <- df[, 2]

df <- df[, 3:11]

df\_list[[as.character(year)]] <- df

}2)



country <- c("Turkey", "Spain", "Romania", "Slovenia", "Germany")

3)

colors <- c('blue', 'green', 'red', 'purple', 'gold')

plot\_index\_over\_years <- function(index\_name, df\_list, years, countries, colors, title, ylabel, legend\_pos = 'topright') {

index\_data <- do.call(rbind, lapply(as.character(years), function(y) df\_list[[y]][countries, index\_name]))

index\_df <- as.data.frame(index\_data, row.names = years)

colnames(index\_df) <- countries

mn <- min(index\_df, na.rm = TRUE)

mx <- max(index\_df, na.rm = TRUE)

plot(years, index\_df[[countries[1]]], type = 'b', col = colors[1], lwd = 2, pch = 1, lty = 1,

ylim = c(mn - 13, mx + 13), xlab = 'Года', ylab = ylabel, main = title)

for (i in 2:length(countries)) {

lines(years, index\_df[[countries[i]]], type = 'b', col = colors[i], lwd = 2, pch = 1, lty = 1)

}

legend(legend\_pos, cex = 0.6, legend = countries, fill = colors)

return(index\_df)

}

years\_full <- 2014:2021

years\_climate <- 2016:2021 # для Climate Index, так как 2014 и 2015 отсутствуют

# Построения графиков по индексам

QLI <- plot\_index\_over\_years('Quality of Life Index', df\_list, years\_full, country, colors,

'Оценка индекса качества жизни', 'Индекс качества жизни', 'bottomright')

PPI <- plot\_index\_over\_years('Purchasing Power Index', df\_list, years\_full, country, colors,

'Оценка индекса покупательной способности', 'Индекс покупательной способности')

SI <- plot\_index\_over\_years('Safety Index', df\_list, years\_full, country, colors,

'Оценка индекса безопасности', 'Индекс безопасности')

HCI <- plot\_index\_over\_years('Health Care Index', df\_list, years\_full, country, colors,

'Оценка индекс медицинского обслуживания', 'Индекс медицинского обслуживания', 'bottomright')

CLI <- plot\_index\_over\_years('Cost of Living Index', df\_list, years\_full, country, colors,

'Оценка индекса прожиточного минимума', 'Индекс прожиточного минимума')

PPIR <- plot\_index\_over\_years('Property Price to Income Ratio', df\_list, years\_full, country, colors,

'Оценка отношения цены на жилье к доходу', 'Отношение цены на жилье к доходу')

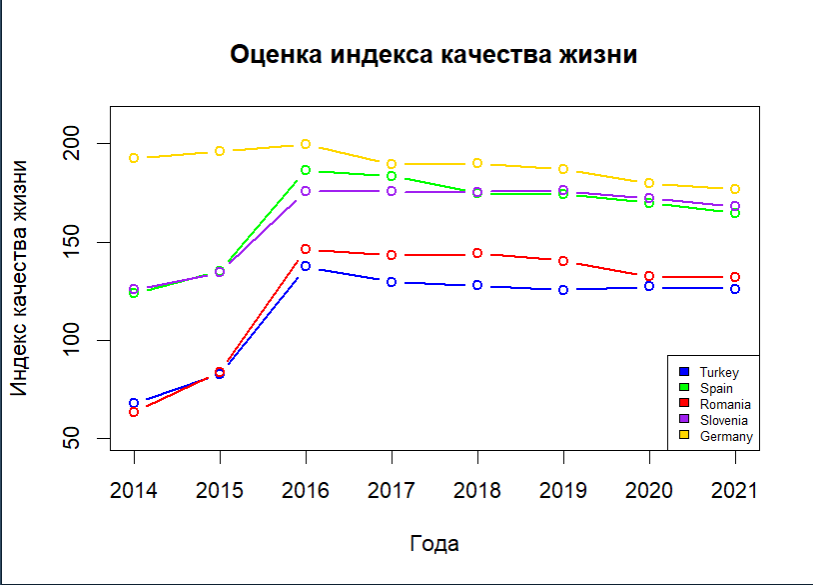
TCTI <- plot\_index\_over\_years('Traffic Commute Time Index', df\_list, years\_full, country, colors,

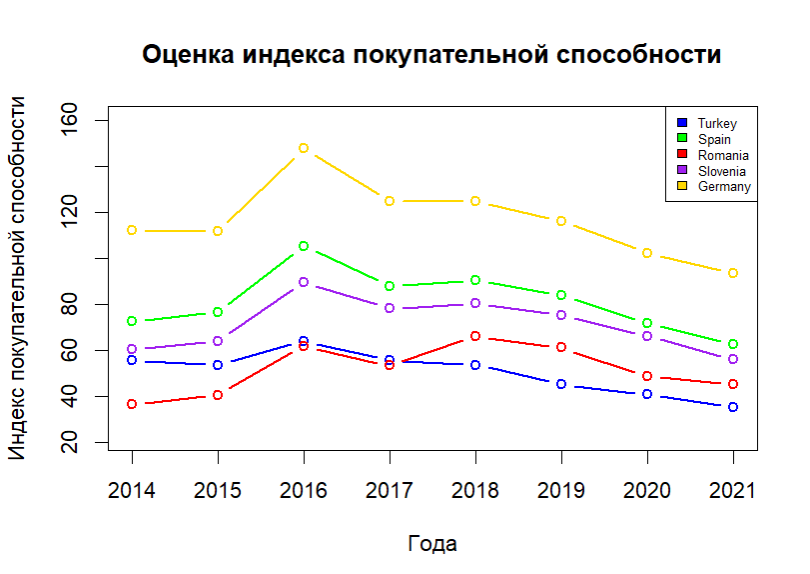
'Оценка индекса времени движения на дороге', 'Индекс времени движения на дороге')

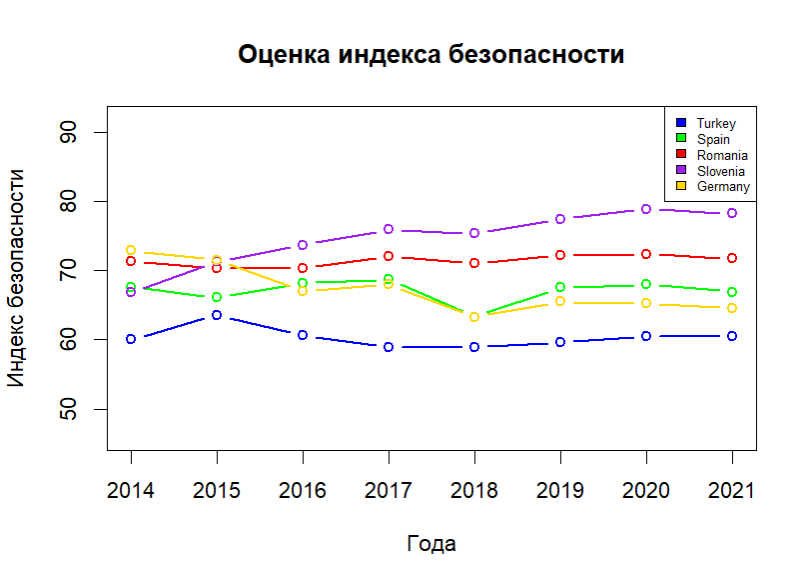
PI <- plot\_index\_over\_years('Pollution Index', df\_list, years\_full, country, colors,

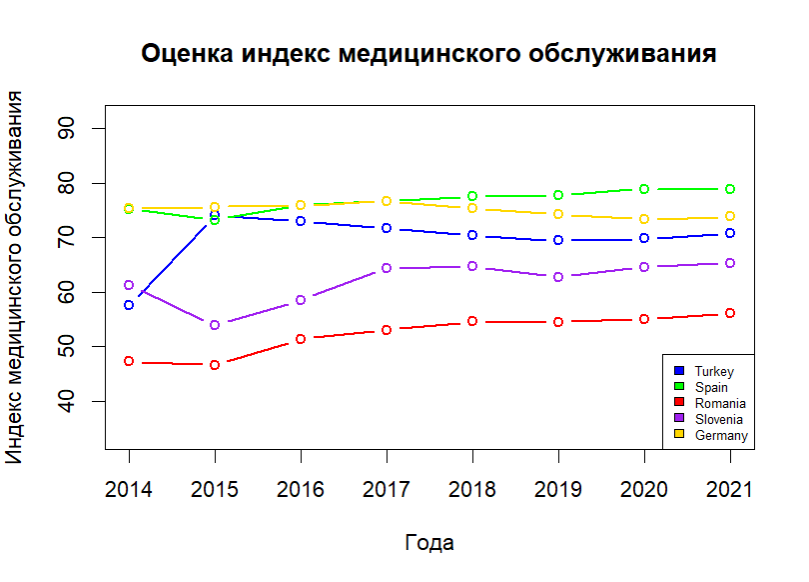
'Оценка индекса загрязнения', 'Индекс загрязнения')

CI <- plot\_index\_over\_years('Climate Index', df\_list, years\_climate, country, colors,

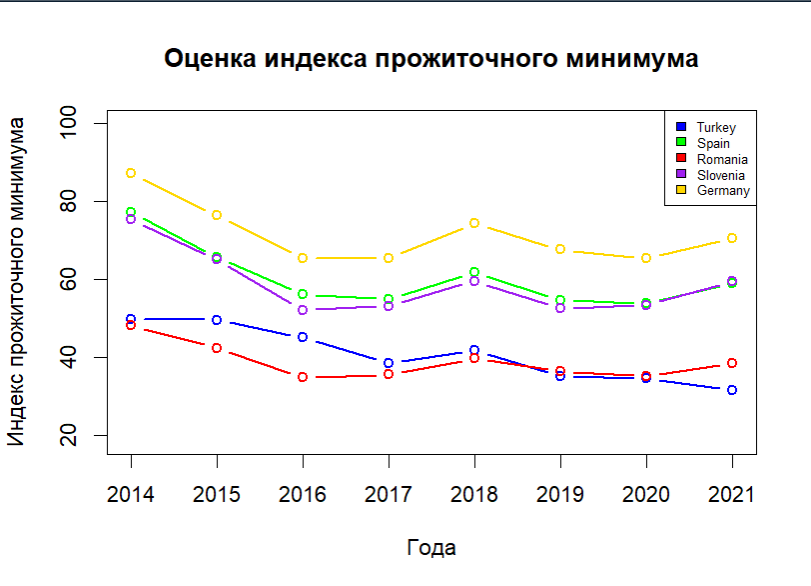
'Оценка климатического индекса', 'Климатический индекс', 'bottomright') 

Германия среди всех выбранных стран лидирует в качестве жизни на протяжении всех лет наблюдений. 

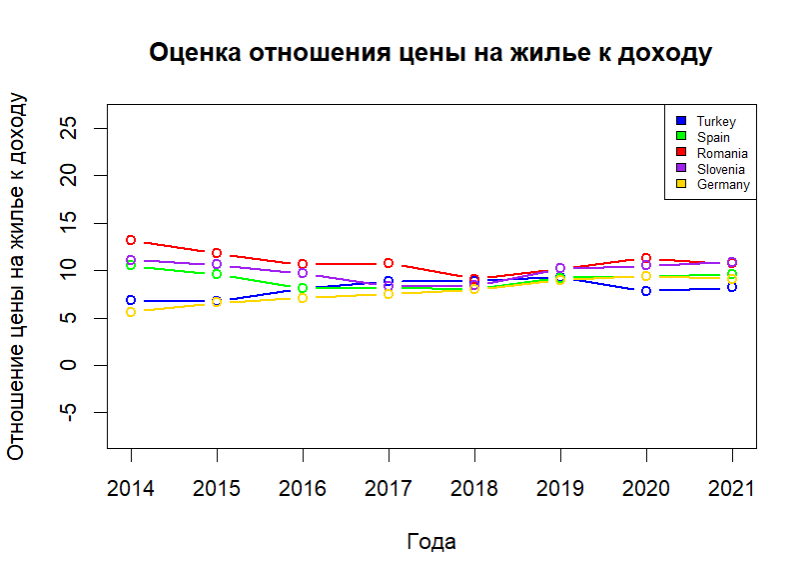
Ситуация повторяется с качеством жизни

Германия имеет средний уровень безопасности за всё время наблюдения. Словакия и Румыния показывают отличные результаты. 

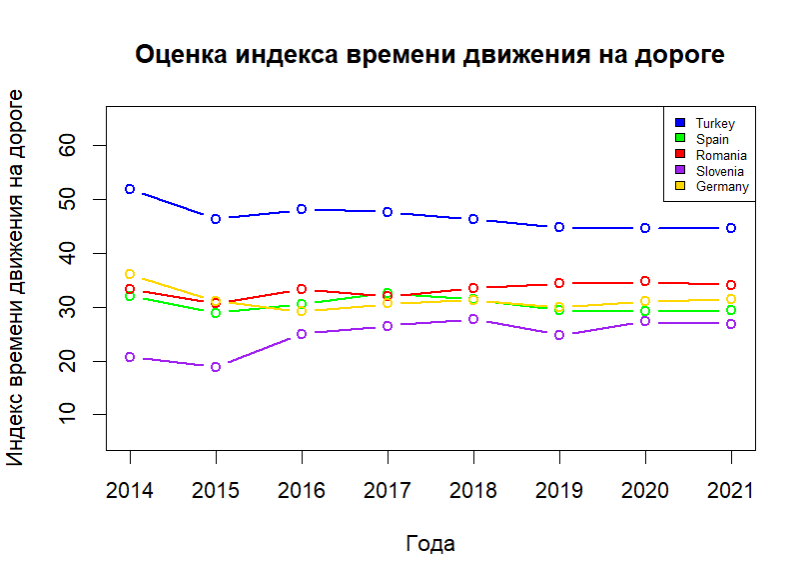
Забота о здоровье в Германии и Испании примерно на одном уровне, у Турции наблюдается огромный скачек в 2015 году. С каждым годом Забота о здоровье улучшается в исследуемых странах.



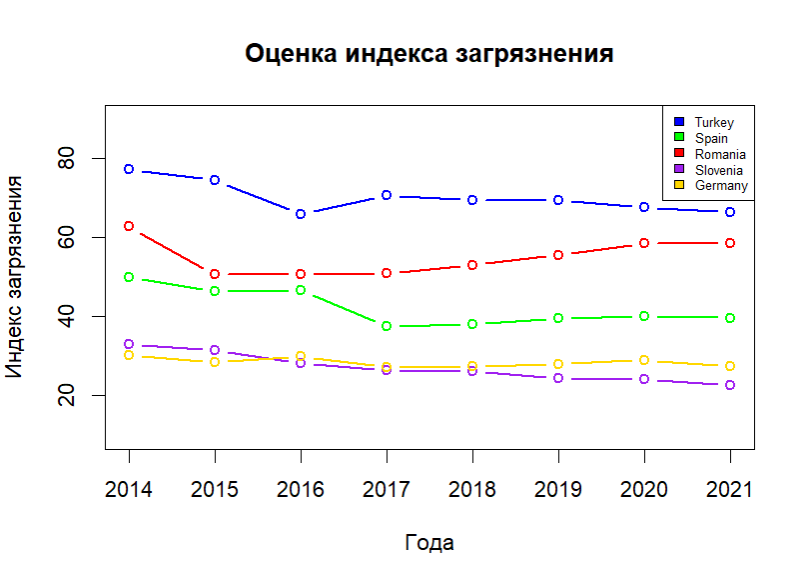
Стоимость жизни в Германии больше всего во все годы. В основном индекс прожиточного минимума уменьшается с каждым годом.



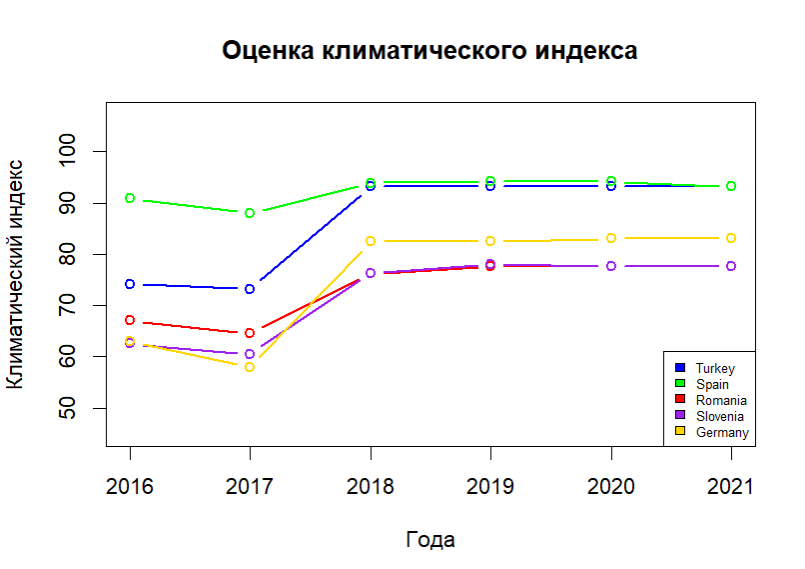
Цены на жилье в исследуемых странах с каждым годом стремятся к одинаковому значению.



В Турции Время движения на дороге самое большое.



Турция загрязнена больше всего.



Лучший климат в Испании.

4)

url <- "https://ru.wikipedia.org/wiki/Список\_музеев\_Ростовской\_области"

page <- read\_html(url)

# Извлекаем таблицы с классом wikitable

tables <- html\_nodes(page, "table.wikitable")

# Преобразуем таблицы в список data frame'ов

museum\_list <- lapply(tables, html\_table, fill = TRUE)

museum\_data <- bind\_rows(museum\_list)

# Приводим названия столбцов к нижнему регистру и заменяем пробелы на подчеркивания

names(museum\_data) <- tolower(names(museum\_data))

names(museum\_data) <- gsub("\\s+", "\_", names(museum\_data))

# Переименовываем столбцы для удобства

museum\_data <- museum\_data %>%

rename(

Name = starts\_with("назв"),

Location = matches("город|насел")

)

# Извлекаем ссылки из второго столбца каждой таблицы

links <- lapply(tables, function(table) {

# Извлекаем все строки таблицы

rows <- html\_nodes(table, "tr")

# Для каждой строки извлекаем второй <td> и тег <a> внутри него

sapply(rows[-1], function(row) { # Пропускаем первую строку (заголовок)

second\_td <- html\_node(row, "td:nth-child(2)")

if (!is.na(second\_td)) {

link <- html\_node(second\_td, "a")

if (!is.na(link)) {

html\_attr(link, "href")

} else {

NA

}

} else {

NA

}

})

})

# Объединяем ссылки из всех таблиц

all\_links <- unlist(links)

# Преобразуем относительные ссылки в полные

full\_links <- paste0("https://ru.wikipedia.org", all\_links)

# Добавляем столбец Link и заполняем его ссылками

museum\_data$Link <- NA

if (length(full\_links) >= nrow(museum\_data)) {

museum\_data$Link <- full\_links[1:nrow(museum\_data)]

} else {

museum\_data$Link[1:length(full\_links)] <- full\_links

}

# Просмотр результата

View(museum\_data) 