Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**(ФГБОУ ВО «КубГУ»)**

**Факультет компьютерных технологий и прикладной математики**

**Кафедра вычислительных технологий**

**Отчёт лабораторная №2**

**Дисциплина: ОБРАБОТКА БОЛЬШЫХ ДАННЫХ**

Работу выполнил \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Осипов В.Р.

Направление подготовки 02.03.02 Фундаментальная информатика и

информационные технологии

Преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Яхонтов А.А.

Краснодар

2025

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

**Задания к лабораторной работе**

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):

Китай, Россия, США, Канада, Япония

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) : вариант брать по формуле No пп % No ссылки + 1.

1) <https://kudago.com/spb/list/33-luchshih-muzeya-peterburga/>

**Ход работы**

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.

Подключаем библиотеку rvest и создаём подключение к ссылкам внутри функции, которая будет парсить наши HTML:

library(rvest)

extract\_year\_data <- function(year) {

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

Читаем HTML:

webpage <- read\_html(url)

Смотрим на сайте HTML файл, чтобы найти, где данные располагаются в DOM-дереве. Можно заметить, что все данные находятся в теге tbody, а для каждой страны в теге tr внутри tbody. Начинаем парсинг:

rows <- html\_nodes(webpage, "tbody tr")

data <- lapply(rows, function(row) {

cells <- html\_nodes(row, "td")

list(

Country = html\_text(cells[2]),

Quality\_of\_Life\_Index = html\_text(cells[3]),

Purchasing\_Power\_Index = html\_text(cells[4]),

Safety\_Index = html\_text(cells[5]),

Health\_Care\_Index = html\_text(cells[6]),

Cost\_of\_Living\_Index = html\_text(cells[7]),

Property\_Price\_to\_Income\_Ratio = html\_text(cells[8]),

Traffic\_Commute\_Time\_Index = html\_text(cells[9]),

Pollution\_Index = html\_text(cells[10]),

Climate\_Index = html\_text(cells[11]),

Year = year

)

})

Filter(Negate(is.null), data)

}

Создаём датасет, где для 2014-2021 года собрана статистика:

years <- 2014:2021

all\_data <- do.call(rbind, lapply(years, function(year) {

year\_data <- extract\_year\_data(year)

do.call(rbind, lapply(year\_data, as.data.frame, stringsAsFactors = FALSE))

}))

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

Китай, Россия, США, Канада, Япония

Вводим страны в отдельный список и создаём сабдатасет для нужных нам стран:

countries <- c("China", "Russia", "United States", "Canada", "Japan")

filtered\_data <- subset(all\_data, Country %in% countries)

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.

Создаём функцию для постройки графиков по каждому качеству и выводим линии на графике для нужных нам стран:

indicators <- c(

"Quality\_of\_Life\_Index", "Purchasing\_Power\_Index", "Safety\_Index",

"Health\_Care\_Index", "Cost\_of\_Living\_Index", "Property\_Price\_to\_Income\_Ratio",

"Traffic\_Commute\_Time\_Index", "Pollution\_Index", "Climate\_Index"

)

plot\_indicator <- function(data, indicator, title) {

plot(NULL,

xlim = range(data$Year, na.rm = TRUE),

ylim = c(0, 200),

xlab = "Year", ylab = indicator, main = title)

colors <- c("red", "blue", "green", "orange", "purple")

countries <- unique(data$Country)

for (i in seq\_along(countries)) {

country\_data <- subset(data, Country == countries[i])

lines(country\_data$Year, country\_data[[indicator]],

type = "o", col = colors[i], pch = 16, lwd = 2)

}

legend("topright", legend = countries, col = colors, pch = 16, cex = 0.8)

}

for (ind in indicators) {

plot\_indicator(filtered\_data, ind, paste("Тенденция показателя:", ind))

}

Словесный анализ:

Япония, США и Канада лидируют по качеству жизни, причём Япония на первом месте. Они сильно отрываются от Китая и России.

Япония лидирует во всех категориях, кроме загрязнения. В загрязнении лидирует Китай.

Все страны, кроме как в качестве жизни, не сильно отстают друг от друга.

В 2019 году уровень безопасности в Китае сильно понизился из-за COVID-19.

В России самое долгое время путешествия из-за дорог и водителей.

4. С одной из страниц (по варианту): вариант брать по формуле No пп % No ссылки + 1.

1) <https://kudago.com/spb/list/33-luchshih-muzeya-peterburga/>

Логика задания абсолютно та же, что и в первом задании:

1. Создаём подключение к сайту.
2. Читаем HTML файл.
3. Смотрим DOM-дерево чтобы найти закономерность, где именно хранятся данные.
4. Парсим HTML по анализу DOM-дерева и собираем датафрейм.

Код для задания 4:

library(rvest)

extract\_museum\_data <- function() {

url <- "https://kudago.com/spb/list/33-luchshih-muzeya-peterburga/"

webpage <- read\_html(url, encoding = "UTF-8")

big\_item <- html\_node(webpage, "div.post-list-big")

items <- html\_nodes(big\_item, "article.post-list-item")

data <- lapply(items, function(item) {

name <- html\_text(html\_node(item, "h3.post-list-item-title a span"), trim = TRUE)

Encoding(name) <- "UTF-8"

address <- html\_text(html\_node(item, "address.post-list-item-info"), trim = TRUE)

if (is.na(address))

address <- html\_text(html\_node(item, "span.post-list-item-info--event-place"), trim = TRUE)

Encoding(address) <- "UTF-8"

description <- html\_text(html\_node(item, "div.post-list-item-description div p"), trim = TRUE)

Encoding(description) <- "UTF-8"

link <- html\_attr(html\_node(item, "a.post-list-item-title-link"), "href")

age\_restriction <- html\_text(html\_node(item, "span.list-age-restriction"), trim = TRUE)

Encoding(age\_restriction) <- "UTF-8"

list(

Название = name,

Адрес = address,

Описание = description,

Ссылка = link,

Возрастное\_ограничение = age\_restriction

)

})

Filter(Negate(is.null), data)

}

museums\_data <- extract\_museum\_data()

museums\_df <- do.call(rbind, lapply(museums\_data, as.data.frame, stringsAsFactors = FALSE))