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 №4**

**Дисциплина: ОБРАБОТКА БОЛЬШЫХ ДАННЫХ**

Работу выполнил \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Дука В.А.

Направление подготовки 02.03.02 Фундаментальная информатика и

информационные технологии

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.

**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:

Это оценка общего качества жизни с использованием

эмпирической формулы, которая учитывает:

— индекс покупательной способности (чем выше, тем лучше),

— индекс загрязнения (чем ниже, тем лучше),

— отношение цены на жилье к доходу (ниже). лучше),

— индекс прожиточного минимума (чем ниже, тем лучше),

— индекс безопасности (чем выше, тем лучше),

— индекс медицинского обслуживания (чем выше, тем лучше),

— индекс времени движения на дороге (чем ниже, тем лучше)

— климатический индекс (чем выше, тем лучше).

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) : 8 вариант - <https://www.afisha.ru/krasnodar/museum/> - собрать информацию в data.frame, которая содержала бы: Название музея, его адрес, описание музея(если есть) и ссылку для перехода на музей.

**Ход работы:**

1. Для начала собрал данные с сайта, используя библиотеку rvest:

get\_numbeo\_data <- function(year) {

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

webpage <- read\_html(url)

tables <- html\_nodes(webpage, "table")

if (length(tables) < 2) {

stop("Не удалось найти нужную таблицу на странице: ", url)

}

data <- html\_table(tables[2], fill = TRUE)[[1]]

data$Year <- year

if ("Rank" %in% colnames(data)) {

data <- data[, !(names(data) %in% c("Rank"))]

}

return(data)

}

years <- 2014:2021

numbeo\_data <- do.call(rbind, lapply(years, get\_numbeo\_data))

if ("Rank" %in% colnames(numbeo\_data)) {

numbeo\_data <- numbeo\_data[, !(names(numbeo\_data) %in% c("Rank"))]

}

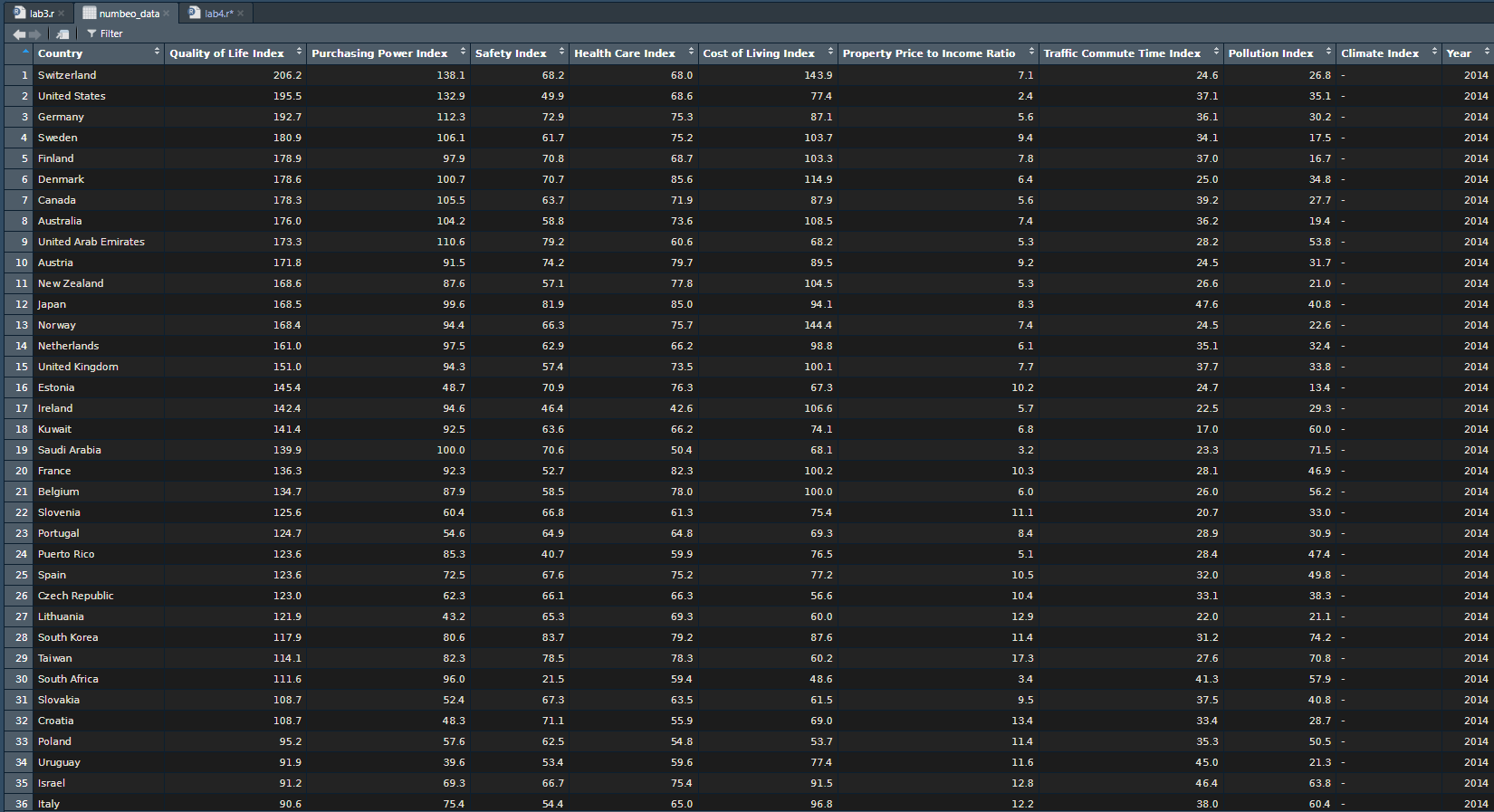


Рисунок 1 – получившийся дата-сет из таблицы сайта numbeo с показателями уровней жизни стран с 2014 по 2021 год

1. После этого отфильтровал дата-сет, выбрав только нужные страны своего варианта с 2014 по 2021 год:

countries <- c("United States", "Canada", "Austria", "United Kingdom", "Denmark")

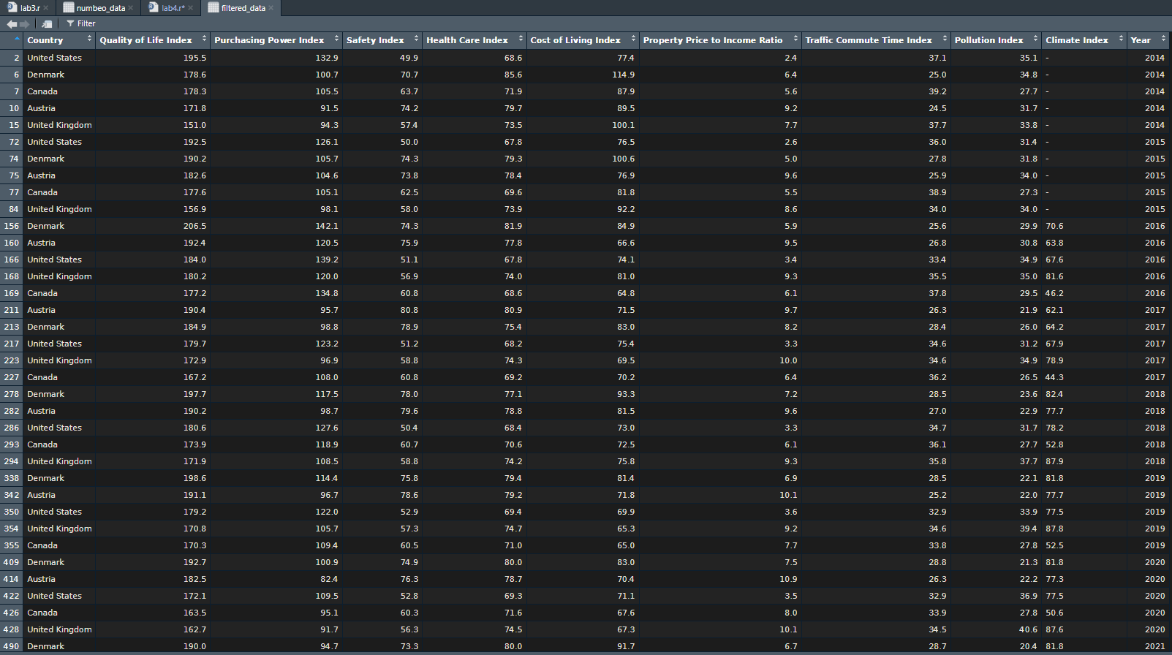
filtered\_data <- numbeo\_data[numbeo\_data$Country %in% countries, ]

Рисунок 2 – получившийся отфильтрованный по нужным странам дата-сет уровня жизни

1. *1) Сохраняем отдельно для каждой страны csv файл:*

output\_dir <- "C:/Users/twink/Desktop/vse/ycheba)))/6 СЕМ/Обработка БД Rstudio (Яхонтов - Приходько)/lab4/"

if (!dir.exists(output\_dir)) {

dir.create(output\_dir, recursive = TRUE)

}

if ("Rank" %in% colnames(filtered\_data)) {

filtered\_data <- filtered\_data[, !(names(filtered\_data) %in% c("Rank"))]

}

# Сохраняем данные для каждой страны в отдельный CSV-файл

for (country in countries) {

country\_data <- filtered\_data[filtered\_data$Country == country, ]

if (nrow(country\_data) > 0) {

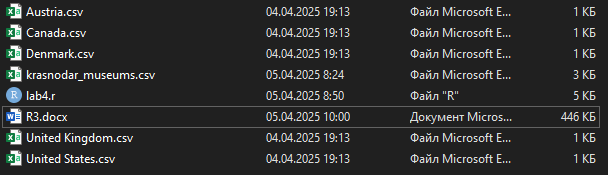
file\_path <- file.path(output\_dir, paste0(country, ".csv"))

write.csv(country\_data, file\_path, row.names = FALSE)

} else {

message("Нет данных для страны: ", country)

}

} **Рисунок 3 – скриншот списка сохраненных csv файлов для каждой страны

*2) Преобразуем данные в «длинную версию» в подряд идущие индексы по годам и строим по новому дата-сету графики по каждому показателю отдельно взятому “Index” для каждого года по всем странам и сохраняем в png:*

# Преобразуем данные в длинный формат

long\_data <- gather(filtered\_data, key = "Index", value = "Value", -Country, -Year)

if ("Rank" %in% unique(long\_data$Index)) {

long\_data <- long\_data[long\_data$Index != "Rank", ]

}

# Создадим папку для графиков, если её нет

plot\_dir <- file.path(output\_dir, "plots")

if (!dir.exists(plot\_dir)) {

dir.create(plot\_dir, recursive = TRUE)

}

# Построение и сохранение отдельных графиков

unique\_indexes <- unique(long\_data$Index)

for (index in unique\_indexes) {

plot\_data <- long\_data[long\_data$Index == index, ]

if (nrow(plot\_data) > 0) {

p <- ggplot(plot\_data, aes(x = Year, y = Value, color = Country, group = Country)) +

geom\_line() +

geom\_point() +

theme\_minimal() +

labs(title = paste("Изменение", index, "по годам"), x = "Год", y = "Значение индекса")

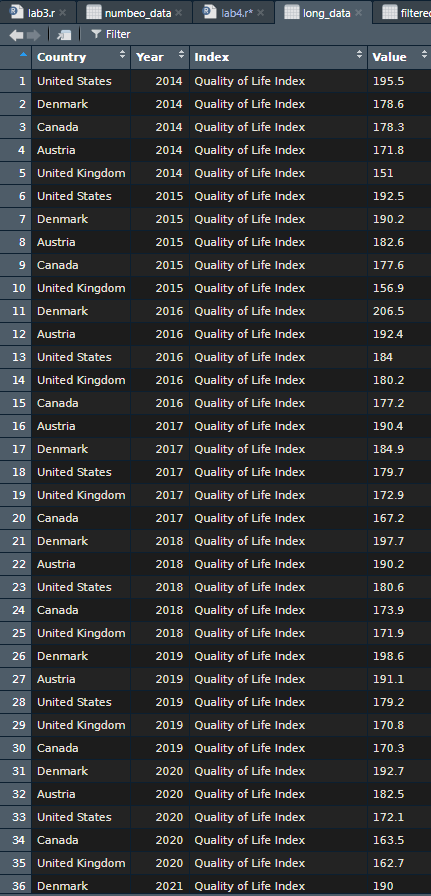
print(p) # Вывод графика в окно

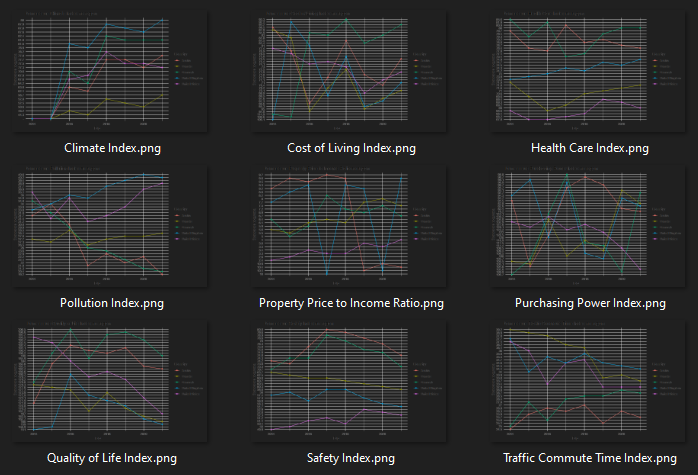
# Сохранение графика в файл

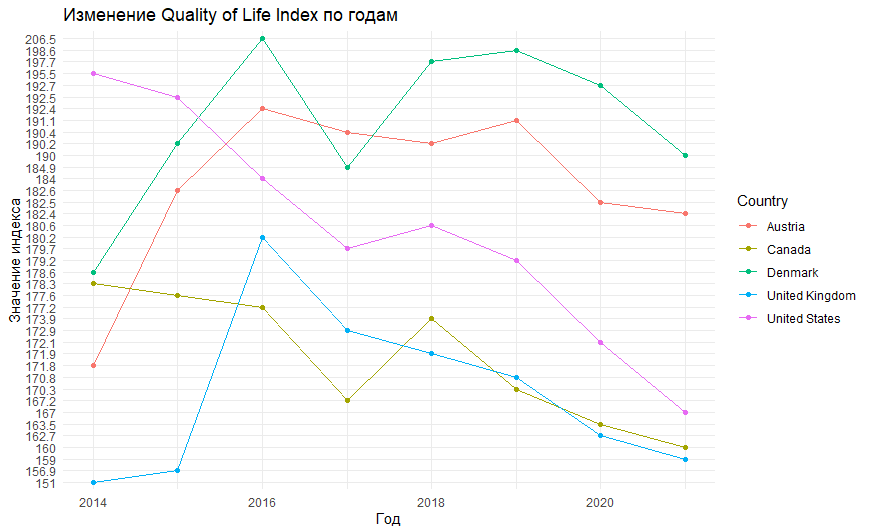
ggsave(filename = file.path(plot\_dir, paste0(index, ".png")), plot = p, width = 8, height = 5)

}

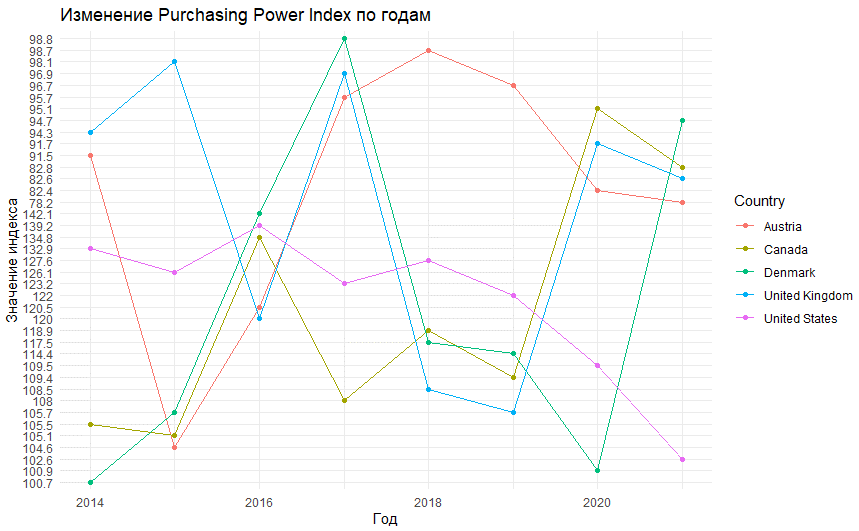
}

**Рисунок 4 – новый дата-сет с подряд выстроенными показателями для каждой страны от 2014 до 2021 года

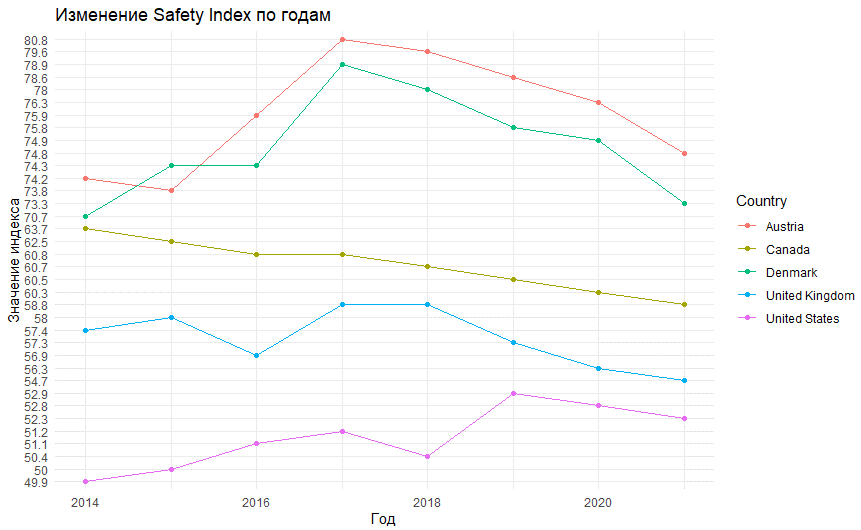
**}Рисунок 5 – скриншот сохраненных в локальной директории всех графиков в формате png.

*3) Анализ графиков:*Рисунок 6 – скриншот сохраненных в локальной директории всех графиков в формате png.

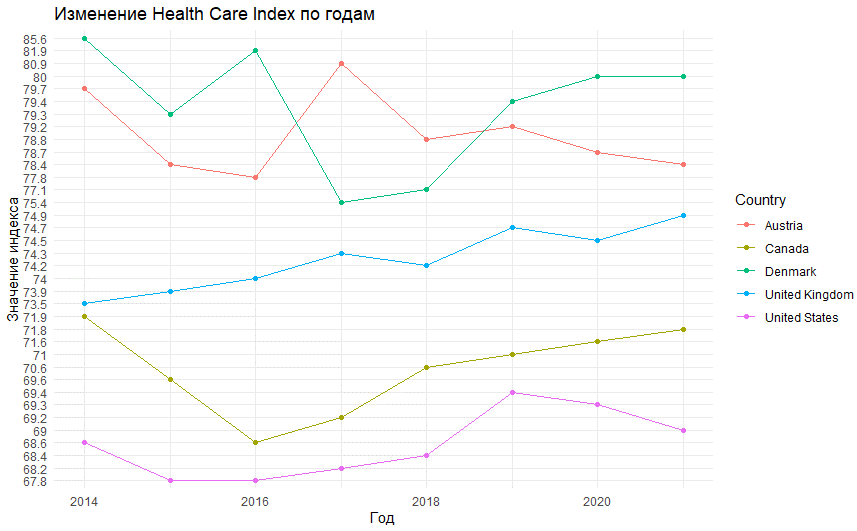
Линейный график иллюстрирует динамику индекса качества жизни (Quality of Life Index) для пяти стран с 2014 по 2021 годы, демонстрируя, что Дания и Канада стабильно удерживают лидирующие позиции (190–200 пунктов) с незначительными колебаниями, тогда как Австрия, Великобритания и США показывают более выраженные спады, особенно в 2017–2018 годах, с последующим снижением к 2021 году до 160–170 пунктов. Анализ выявляет общую тенденцию к снижению индекса качества жизни для большинства стран, что может быть связано с ухудшением таких факторов, как индекс загрязнения или времени движения на дороге.

Рисунок 7 – скриншот сохраненных в локальной директории всех графиков в формате png.

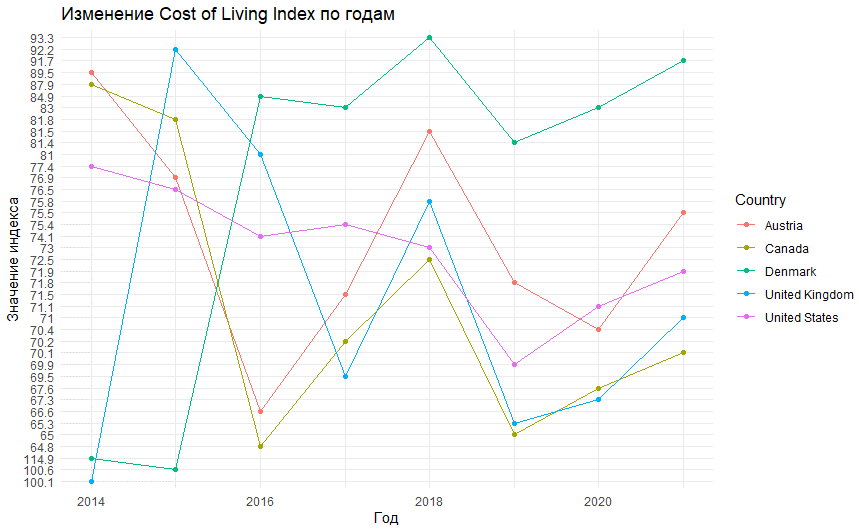
Линейный график отражает динамику индекса покупательной способности (Purchasing Power Index) для пяти стран с 2014 по 2021 годы, показывая значительные колебания у всех стран, с пиковыми значениями у Австрии и Великобритании в 2016 году (около 980 пунктов), но общей тенденцией к снижению к 2021 году (до 900–950 пунктов), что может указывать на экономические изменения, влияющие на доходы населения. Дания и Канада демонстрируют более стабильные показатели с умеренным снижением, тогда как США показывают наибольший спад, что может быть связано с ростом стоимости жизни или снижением реальных доходов.

Рисунок 8 – скриншот сохраненных в локальной директории всех графиков в формате png.

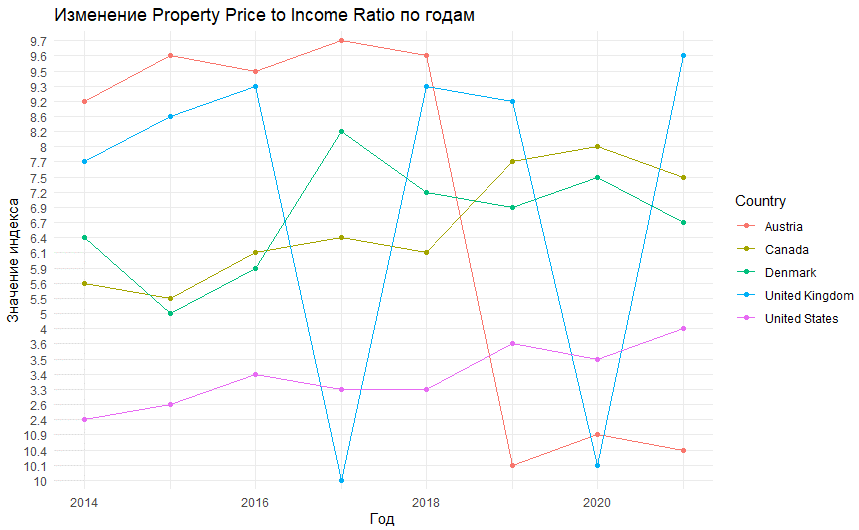
Линейный график демонстрирует динамику индекса безопасности (Safety Index) для пяти стран с 2014 по 2021 годы, показывая, что Австрия и Дания удерживают высокие позиции (70–80 пунктов) с небольшим снижением к 2021 году, тогда как Канада, Великобритания и США демонстрируют более низкие и устойчивые значения (50–65 пунктов) с общей тенденцией к снижению, что может указывать на рост уровня преступности или снижение общественной безопасности. Наиболее заметное снижение наблюдается у США, где индекс упал с 50 до 45 пунктов, что подчёркивает ухудшение восприятия безопасности в этой стране.

Рисунок 9 – скриншот сохраненных в локальной директории всех графиков в формате png.

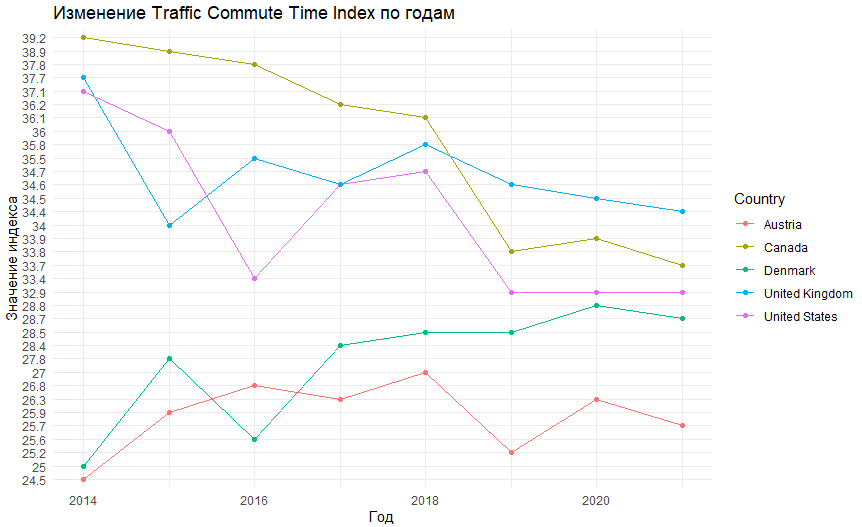
Линейный график отражает динамику индекса медицинского обслуживания (Health Care Index) для пяти стран с 2014 по 2021 годы, демонстрируя, что Дания и Австрия стабильно удерживают высокие показатели (78–85 пунктов) с небольшими колебаниями, тогда как Канада, Великобритания и США показывают более низкие значения (66–74 пункта) с тенденцией к росту у Канады и Великобритании, что может указывать на улучшение доступности и качества медицинских услуг в этих странах. США демонстрируют наименьший прогресс, оставаясь на уровне 66 пунктов, что может быть связано с высокими затратами на здравоохранение и неравномерным доступом к услугам.

Рисунок 10 – скриншот сохраненных в локальной директории всех графиков в формате png.

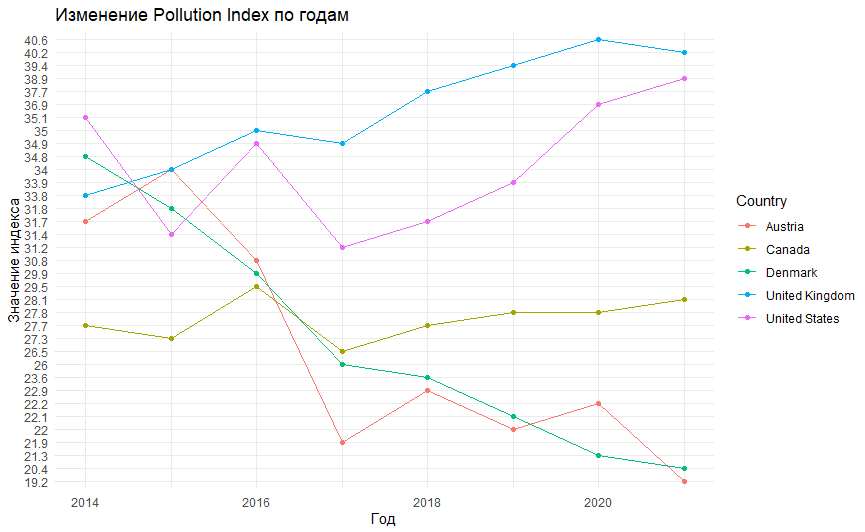
Линейный график иллюстрирует динамику индекса стоимости жизни (Cost of Living Index) для пяти стран с 2014 по 2021 годы, показывая, что Дания и Австрия имеют самые высокие значения (80–90 пунктов) с тенденцией к росту, тогда как Канада, Великобритания и США демонстрируют более низкие показатели (65–75 пунктов) с колебаниями, что может указывать на рост цен на основные товары и услуги в странах с более высоким индексом. Наиболее заметный рост наблюдается у Дании, где индекс увеличивается с 80 до 90 пунктов, что может быть связано с инфляцией или увеличением стоимости жилья.

Рисунок 11 – скриншот сохраненных в локальной директории всех графиков в формате png.

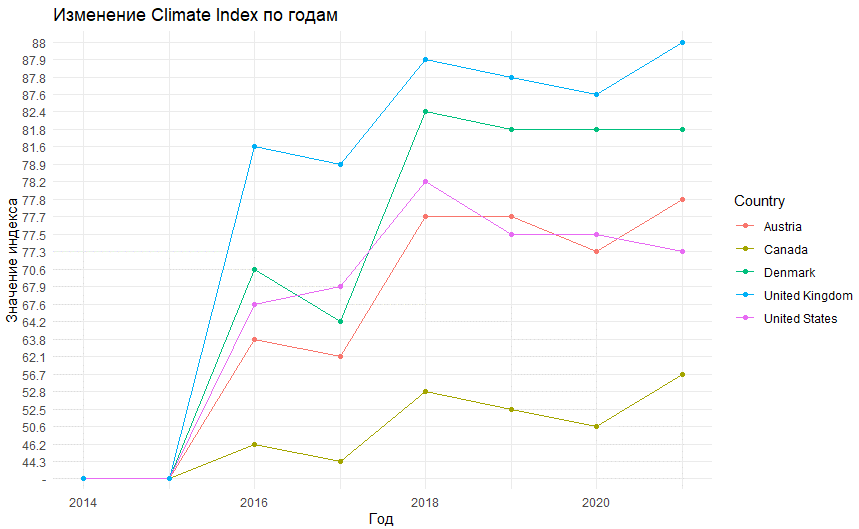
Линейный график отражает динамику отношения цены на жилье к доходу (Property Price to Income Ratio) для пяти стран с 2014 по 2021 годы, показывая, что Австрия и Великобритания имеют самые высокие значения (8–9) с пиками в 2016 и 2021 годах, тогда как Дания, Канада и США демонстрируют более низкие показатели (3–7) с тенденцией к снижению у Австрии и ростом у Великобритании, что может указывать на рост недоступности жилья в этих странах. Наиболее заметное снижение наблюдается у Австрии, где показатель упал с 9.5 до 7, что может быть связано с улучшением доходов или снижением цен на жильё.

Рисунок 12 – скриншот сохраненных в локальной директории всех графиков в формате png.

Линейный график иллюстрирует динамику индекса времени в пути (Traffic Commute Time Index) для пяти стран с 2014 по 2021 годы, показывая, что Канада и США имеют самые высокие значения (34–39 минут) с тенденцией к снижению, тогда как Австрия и Дания демонстрируют более низкие показатели (24–28 минут) с колебаниями, что может указывать на улучшение транспортной инфраструктуры или снижение загруженности дорог в этих странах. Великобритания показывает умеренное снижение с 37 до 32 минут, что может быть связано с оптимизацией транспортных потоков или изменением рабочих привычек населения.

Рисунок 13 – скриншот сохраненных в локальной директории всех графиков в формате png.

Линейный график демонстрирует динамику индекса загрязнения (Pollution Index) для пяти стран с 2014 по 2021 годы, показывая, что Великобритания и США имеют самые высокие значения (34–40 пунктов) с тенденцией к росту, тогда как Австрия, Дания и Канада демонстрируют более низкие показатели (20–30 пунктов) с общим снижением, что может указывать на успешные меры по улучшению экологической ситуации в этих странах. Наиболее заметное снижение наблюдается у Австрии, где индекс упал с 31 до 20 пунктов, что может быть связано с внедрением экологических программ и снижением выбросов.

Рисунок 14 – скриншот сохраненных в локальной директории всех графиков в формате png.

Линейный график отражает динамику климатического индекса (Climate Index) для пяти стран с 2014 по 2021 годы, показывая, что Великобритания и Дания имеют самые высокие значения (80–88 пунктов) с незначительными колебаниями, тогда как Австрия, США и Канада демонстрируют более низкие показатели (44–70 пунктов) с тенденцией к росту у Канады, что может указывать на более благоприятные климатические условия в странах с высокими значениями. Наиболее заметный рост наблюдается у Канады, где индекс увеличился с 44 до 56 пунктов, что может быть связано с изменением климатических условий или методологии расчёта индекса.

4. Сбор данных с сайта по индивидуальному варианту: название музея, его адрес и ссылку для перехода на музей

# Функция для получения адреса со страницы музея

main\_page <- read\_html("https://www.afisha.ru/krasnodar/museum/")

#извлекаем ссылки

selector\_link <- "div.\_yjkz > .CjnHd.y8A5E.L0ZCf.LRhja"

links <- html\_nodes(main\_page, selector\_link) %>% html\_attr("href")

links <- paste0("https://www.afisha.ru", links)

print(links)

# Извлекаем названия

selector\_name <- "div.\_yjkz > .CjnHd.y8A5E.vVS2J"

titles <- html\_nodes(main\_page, selector\_name) %>% html\_text()

print(titles)

#изыв

selector\_adress <- "div.\_yjkz > span.hmVRD.DiLyV"

adresses <- html\_nodes(main\_page, selector\_adress) %>% html\_text()

print(adresses)

# Собираем данные в датафрейм

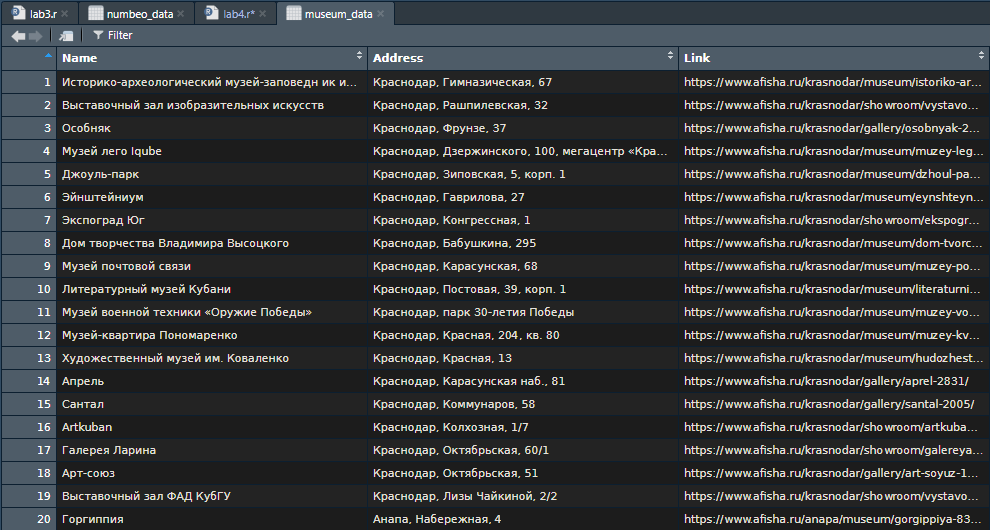
museum\_data <- data.frame(

Name = titles,

Address = adresses,

Link = links

)

**Рисунок 15 – скриншот сохраненных в локальной директории всех графиков в формате png.

**Код программы:**

library(rvest)

library(ggplot2)

library(dplyr)

library(tidyr)

library(stringr)

library(purrr)

get\_numbeo\_data <- function(year) {

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

webpage <- read\_html(url)

tables <- html\_nodes(webpage, "table")

if (length(tables) < 2) {

stop("Не удалось найти нужную таблицу на странице: ", url)

}

data <- html\_table(tables[2], fill = TRUE)[[1]]

data$Year <- year

if ("Rank" %in% colnames(data)) {

data <- data[, !(names(data) %in% c("Rank"))]

}

return(data)

}

years <- 2014:2021

numbeo\_data <- do.call(rbind, lapply(years, get\_numbeo\_data))

if ("Rank" %in% colnames(numbeo\_data)) {

numbeo\_data <- numbeo\_data[, !(names(numbeo\_data) %in% c("Rank"))]

}

# Фильтруем данные для выбранных стран

countries <- c("United States", "Canada", "Austria", "United Kingdom", "Denmark")

filtered\_data <- numbeo\_data[numbeo\_data$Country %in% countries, ]

# Указываем путь к папке для сохранения файлов

output\_dir <- "C:/Users/twink/Desktop/vse/ycheba)))/6 СЕМ/Обработка БД Rstudio (Яхонтов - Приходько)/lab4/"

if (!dir.exists(output\_dir)) {

dir.create(output\_dir, recursive = TRUE)

}

if ("Rank" %in% colnames(filtered\_data)) {

filtered\_data <- filtered\_data[, !(names(filtered\_data) %in% c("Rank"))]

}

# Сохраняем данные для каждой страны в отдельный CSV-файл

for (country in countries) {

country\_data <- filtered\_data[filtered\_data$Country == country, ]

if (nrow(country\_data) > 0) {

file\_path <- file.path(output\_dir, paste0(country, ".csv"))

write.csv(country\_data, file\_path, row.names = FALSE)

} else {

message("Нет данных для страны: ", country)

}

}

print("Файлы успешно сохранены в папку")

# Преобразуем данные в длинный формат

long\_data <- gather(filtered\_data, key = "Index", value = "Value", -Country, -Year)

if ("Rank" %in% unique(long\_data$Index)) {

long\_data <- long\_data[long\_data$Index != "Rank", ]

}

# Создадим папку для графиков, если её нет

plot\_dir <- file.path(output\_dir, "plots")

if (!dir.exists(plot\_dir)) {

dir.create(plot\_dir, recursive = TRUE)

}

# Построение и сохранение отдельных графиков

unique\_indexes <- unique(long\_data$Index)

for (index in unique\_indexes) {

plot\_data <- long\_data[long\_data$Index == index, ]

if (nrow(plot\_data) > 0) {

p <- ggplot(plot\_data, aes(x = Year, y = Value, color = Country, group = Country)) +

geom\_line() +

geom\_point() +

theme\_minimal() +

labs(title = paste("Изменение", index, "по годам"), x = "Год", y = "Значение индекса")

print(p) # Вывод графика в окно

# Сохранение графика в файл

ggsave(filename = file.path(plot\_dir, paste0(index, ".png")), plot = p, width = 8, height = 5)

}

}

print("Графики выведены и сохранены")

# URL страницы с музеями

# Функция для получения адреса со страницы музея

main\_page <- read\_html("https://www.afisha.ru/krasnodar/museum/")

#извлекаем ссылки

selector\_link <- "div.\_yjkz > .CjnHd.y8A5E.L0ZCf.LRhja"

links <- html\_nodes(main\_page, selector\_link) %>% html\_attr("href")

links <- paste0("https://www.afisha.ru", links)

print(links)

# Извлекаем названия

selector\_name <- "div.\_yjkz > .CjnHd.y8A5E.vVS2J"

titles <- html\_nodes(main\_page, selector\_name) %>% html\_text()

print(titles)

#изыв

selector\_adress <- "div.\_yjkz > span.hmVRD.DiLyV"

adresses <- html\_nodes(main\_page, selector\_adress) %>% html\_text()

print(adresses)

# Собираем данные в датафрейм

museum\_data <- data.frame(

Name = titles,

Address = adresses,

Link = links

)

# Удаляем строки с отсутствующими данными

museum\_data <- museum\_data[complete.cases(museum\_data), ]

# Просмотр результата

View(museum\_data)

**Вывод:** лабораторная работа, посвящённая извлечению данных с веб-страниц с использованием пакета rvest в R, позволила успешно собрать информацию об уровне жизни в пяти странах за 2014–2021 годы с сайта Numbeo, а также данные о музеях Краснодара с сайта Afisha, сформировав data.frame для анализа и визуализации динамики показателей. В ходе выполнения заданий были освоены методы парсинга веб-страниц, включая извлечение таблиц и обработку динамического контента, что дало возможность проанализировать рейтинги стран и представить результаты в виде графиков, выявив ключевые тенденции в изменении качества жизни и характеристик музеев