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 больших данных**

Работу выполнил \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Гаранина Л.В.

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 с WEB-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 2021 год.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 (чем ниже, тем 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).

1. Каждый студент должен взять 5 стран (по варианту).
2. 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.
3. С одной из страниц (по варианту) собрать информацию в data.frame, которая содержала бы: название музея, его адрес и ссылку для перехода при клике на фото / ссылке на музей.

Вариант: 22.

Страны: Россия, Украина, Беларусь, Грузия, Армения.

Страница для 4 задания: <https://tonkosti.ru/Музеи_Санкт-Петербурга>.

**Ход работы**:

Соберем информацию об уровне жизни стран с сайта <https://www.numbeo.com/quality-of-life/rankings_by_country.jsp?title=2021> с 2014 по 2021 год. Данные берутся из таблицы сайта, изображенной на рисунке 1.

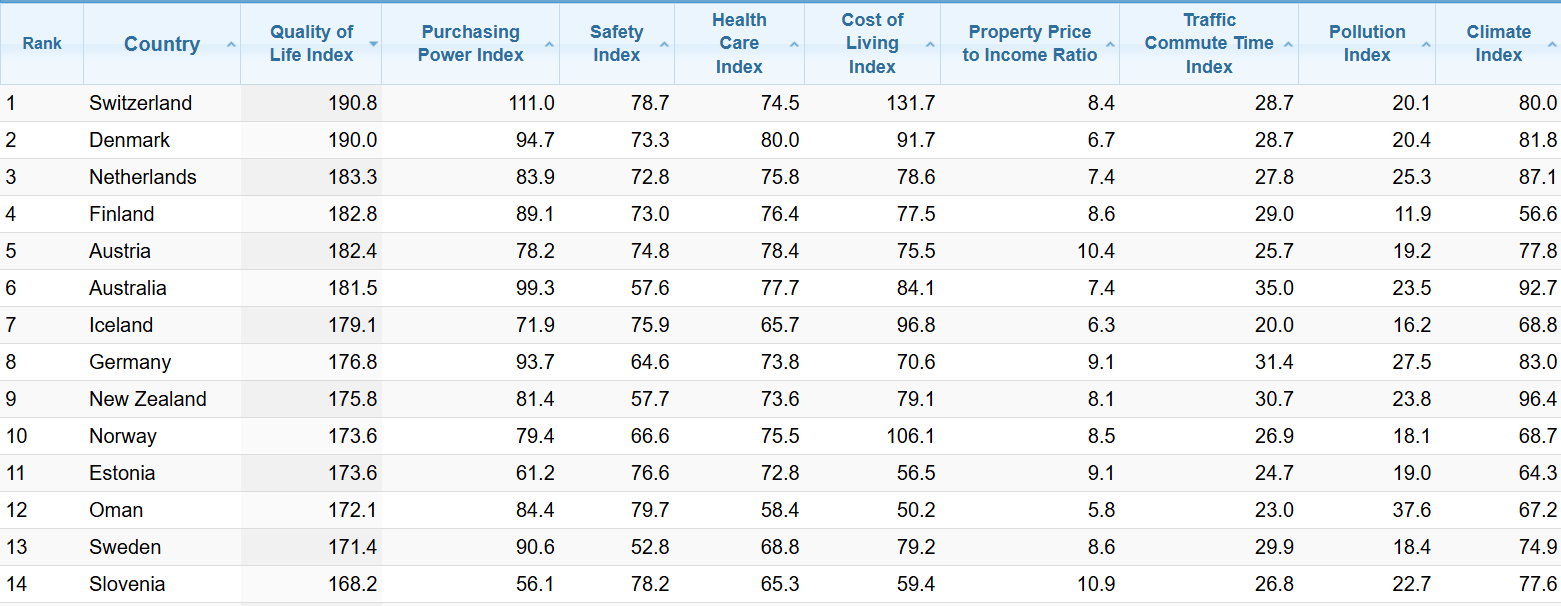


Рисунок 1 – Таблица данных

Создав data.frame со всей информацией, создадим еще несколько, но уже с указанными в варианте странами и по оцениваемому критерию, чтобы можно было наглядно сравнить показатели между выбранными странами и провести более детальный анализ по каждому критерию.

На рисунке 2 изображен график сравнения по индексу качества жизни.

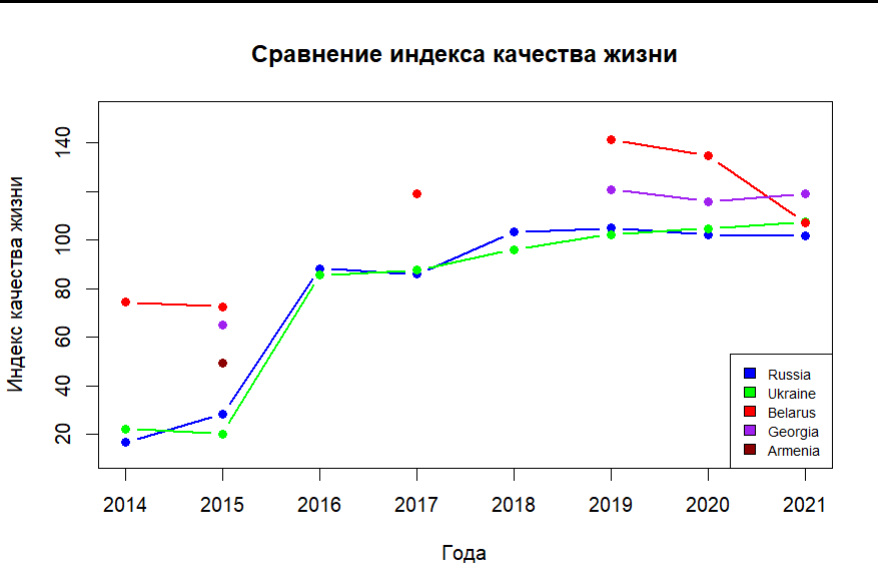


Рисунок 2 – График сравнения стран по индексу качества жизни

Здесь и на других графиках можно заметить, что встречаются пробелы — периоды, когда некоторые страны отсутствуют в рейтинге. Это связано с особенностями формирования данных на платформе.

По первому графику видно, что с каждым годом качество жизни улучшается, а в период с 2015 по 2016 был резки скачок вверх. Россия и Украина имеют схожие показатели и темпы роста, в то время как в Беларуси (кроме 2021 года) и Грузии индексы были всегда немного больше остальных.

Далее сравним по индексу покупательной способности, что показано на рисунке 3.

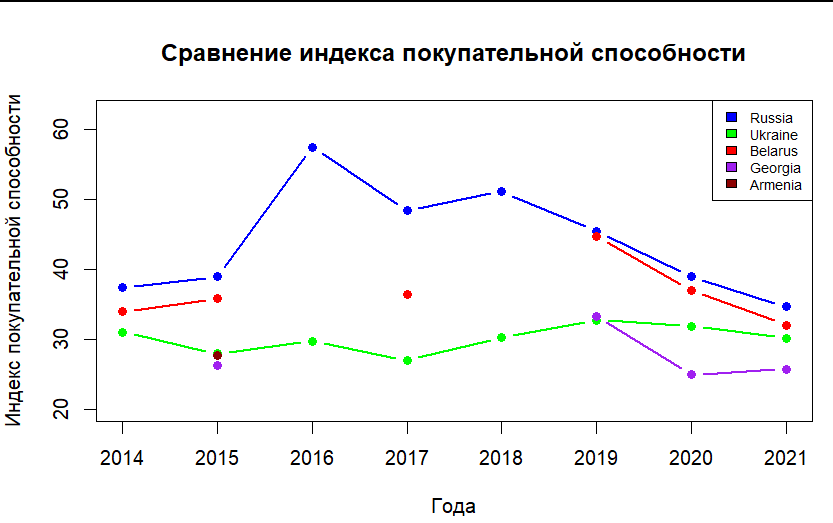


Рисунок 3 – График сравнения стран по индексу покупательной способности

Можно сказать, что в России этот индекс больше, чем в остальных странах, что связано с более низкими ценами на товары и услуги при определенном уровне доходов. Однако к 2021 году он пошел на спад в каждой стране. Это возможно связано с инфляцией, изменением курса валют или другими экономическими факторами.

Далее рассмотри индекс безопасности, изображенный на рисунке 4.

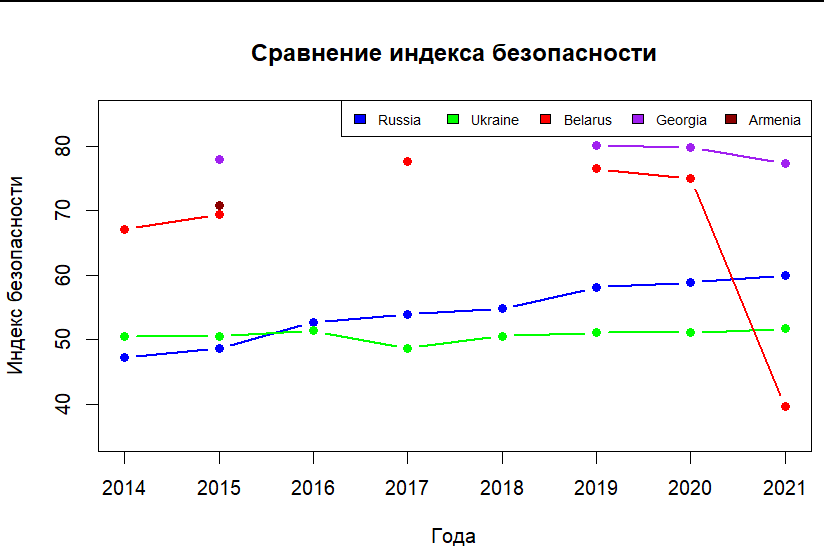


Рисунок 4 – График сравнения стран по индексу безопасности

Индекс безопасности для России и Украины стабильно держится около 50-60, в то время как остальные стран имели индексы выше среднего. Однако к 2021 году в Беларуси индекс безопасности резко опустился ниже уровня других.

Рассмотрим индекс здравоохранения для стран.

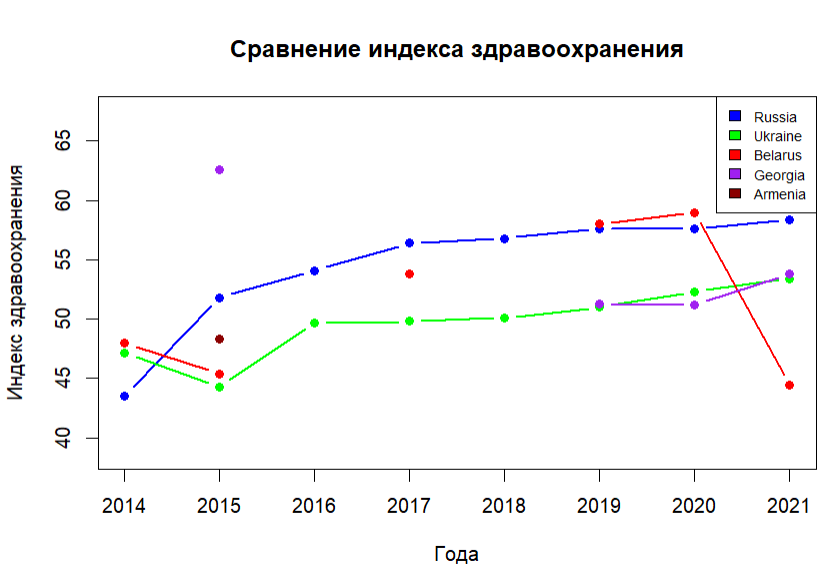


Рисунок 5 – График сравнения стран по индексу здравоохранения

Каждый год индекс здравоохранения стабильно растет в странах, что свидетельствует о постоянном улучшении качества медицинского обслуживания, доступности медицинских услуг и общем здоровья населения. Однако к 2021 году в Беларуси наблюдается резкое падение индекса ниже среднего, как и на предыдущих графиках. Это может быть связано с политической нестабильностью, последствиями пандемии. Грузия в 2015 году имела самый высокий индекс.

Рисунок 6 демонстрирует график сравнения индекса стоимости жизни.

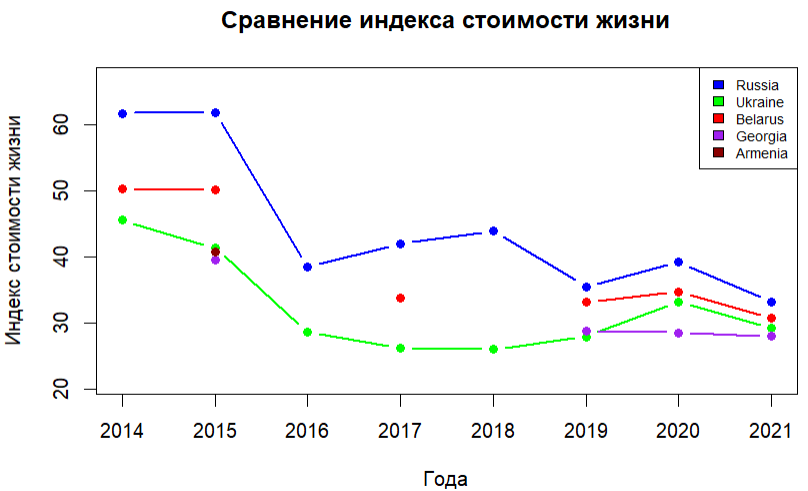


Рисунок 6 – График сравнения стран по индексу стоимости жизни

Как мы видим, с каждым годом индекс падает, показывая, что расходы на еду, транспорт, услуги и другие базовые нужды становятся меньше. Резкий спад бал зафиксирован в 2016 году.

Рассмотрим соотношение цены недвижимости к доходу.

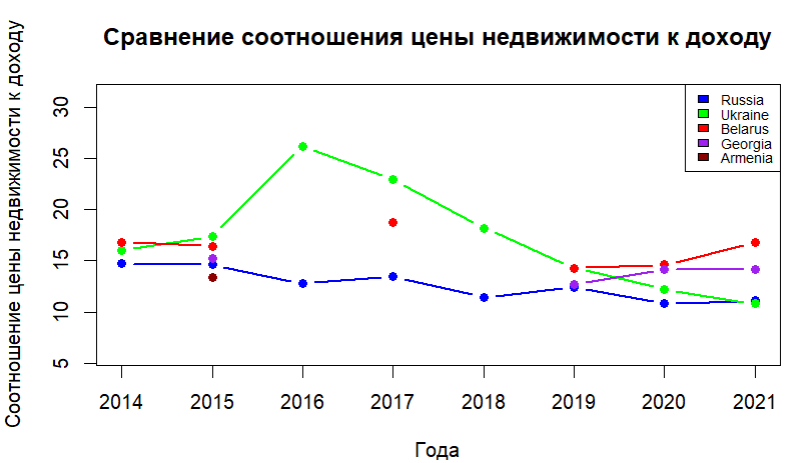


Рисунок 7 – График сравнения стран по соотношению цена недвижимости к доходу

Смотря на график, можно сказать, что самый высокий индекс долгое время имела Украина, пик был в 2016 году, равный 26, после чего пошел стабильный спад. У других стран индекс не поднимался выше 20.

Сравним индекс времени в пути по трафику, график которого изображен на рисунке 8.

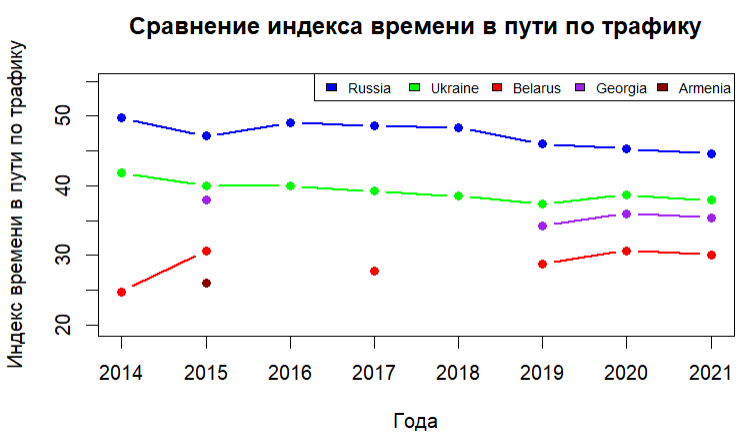
****

Рисунок 8 – График сравнения стран по индексу времени в пути по трафику

Видим, что Россия имеет самые высокие показатели, что указывает на значительные проблемы с пробками, длительными поездками, плохим состоянием дорог. Самые низкие у Армении и Беларуси, что свидетельствует о более быстрых маршрутах и более эффективной транспортной инфраструктуре.

Посмотрим на индекс загрязнения для предложенных стрна.

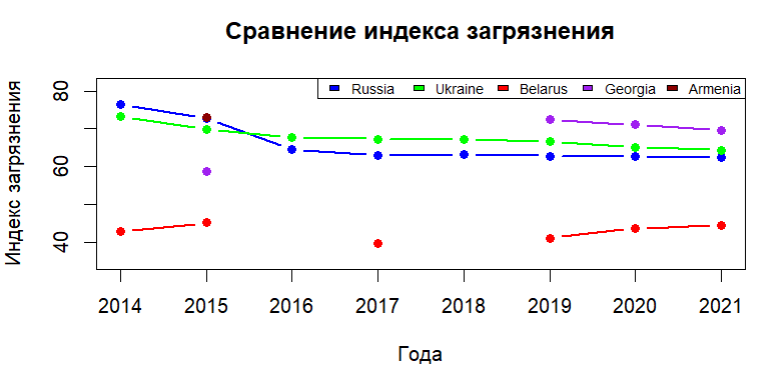
****

Рисунок 9 – График сравнения стран по индексу загрязнения

По графику видно, что самый маленький показатель у Беларуси – не больше 45, что говорит о достаточно чистой окружающей среде и низком уровне загрязнения воздуха. У остальных стран не опускается ниже 62, что может быть следствием более интенсивной промышленной активности, автомобильного трафика и других факторов.

Рассмотрим климатический индекс в странах.

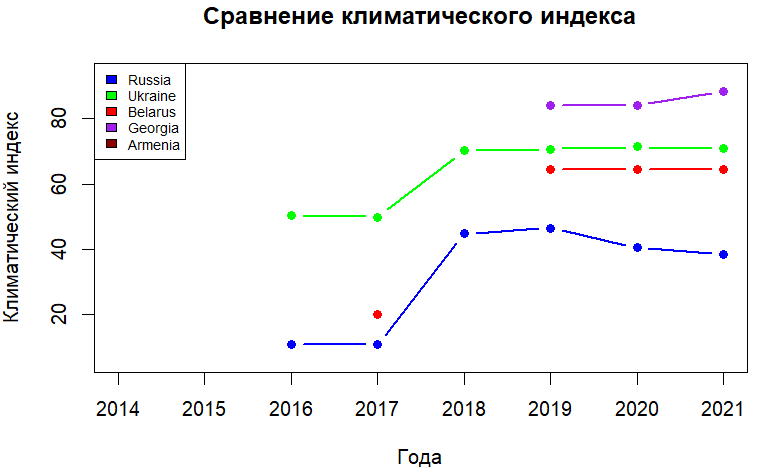
****

Рисунок 10 – График сравнения стран по климатическому индексу

Как видим, показатели во всех странах немного растут, что означает, что климатические условия становятся более комфортными для проживания. Это может быть связано с уменьшением экстремальных погодных явлений, таких как сильные морозы, затяжные дожди или высокие летние температуры. В России самые низкие показатели, потому что страна имеет значительные климатические колебания, что затрудняет благоприятное проживание в некоторых регионах. Данных по Армении нет.

Далее со страницы <https://kudago.com/spb/list/33-luchshih-muzeya-peterburga/> соберем информацию в дата фрейм, которая содержит названия музеев, адрес, описание и ссылку при ссылке на музей. Таблица, содержащая все эти данные, изображена на рисунках 11-12.

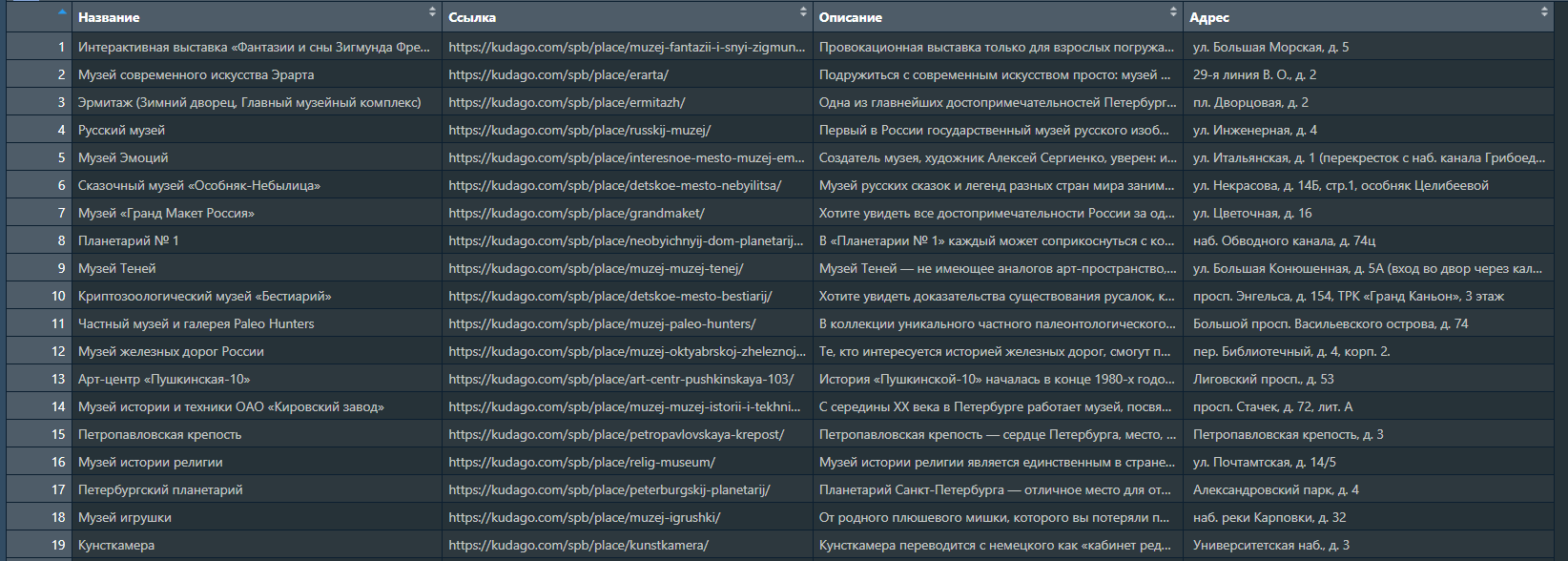
****

Рисунок 11 – 1 часть таблицы с музеями

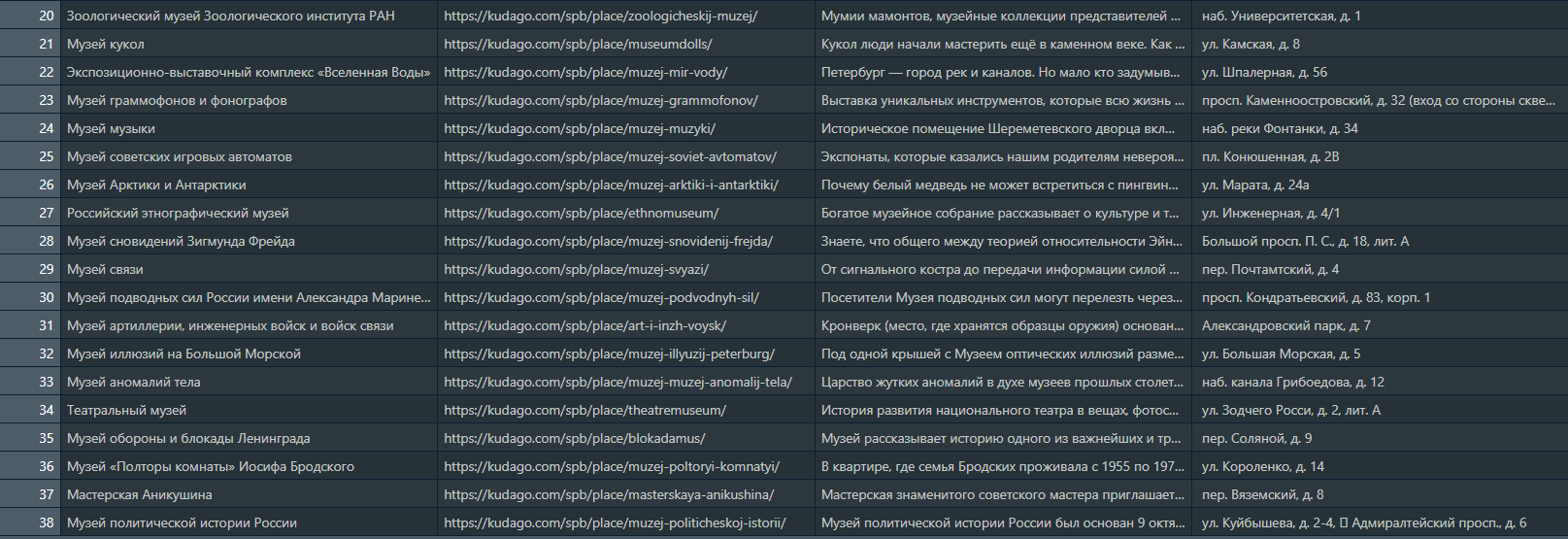
****

Рисунок 12 – 2 часть таблицы с музеями

**Вывод:** проделав лабораторную работу, были изучены способы извлечения информации с WEB-страниц с помощью инструментов языка R.

**Листинг программы:**

library(rvest)

years <- 2014:2021

get\_data <- function(year) {

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

page <- read\_html(url)

nodes <- html\_nodes(page, 'table#t2')

df <- html\_table(nodes) %>% as.data.frame()

rownames(df) <- df[, 2]

df <- df[, 3:11]

return(df)

}

data\_list <- lapply(years, get\_data)

names(data\_list) <- years

countries <- c('Russia', 'Ukraine', 'Belarus', 'Georgia', 'Armenia')

extract\_index <- function(df, country, column\_name ) {

sapply(country, function(ctry) {

if (ctry %in% rownames(df)) {

value <- df[ctry, column\_name]

if (value == "-") {

return(NA)

} else {

return(as.numeric(value))

}

} else {

return(NA)

}

})

}

colors = c('blue', 'green', 'red', 'purple', 'darkred')

index\_names <- colnames(data\_list[[1]])

index\_names\_rus <- c(

"Quality.of.Life.Index" = "Индекс качества жизни",

"Purchasing.Power.Index" = "Индекс покупательной способности",

"Safety.Index" = "Индекс безопасности",

"Health.Care.Index" = "Индекс здравоохранения",

"Cost.of.Living.Index" = "Индекс стоимости жизни",

"Property.Price.to.Income.Ratio" = "Соотношение цен на жилье и доходов",

"Traffic.Commute.Time.Index" = "Индекс времени в пути",

"Pollution.Index" = "Индекс загрязнения",

"Climate.Index" = "Индекс климата"

)

index\_names\_rus <- index\_names\_rus[index\_names]

par(mar = c(5, 5, 4, 8), xpd = TRUE)

for (i in 1:length(index\_names)) {

index <- index\_names[i]

matrix\_data <- t(sapply(data\_list, extract\_index, countries, index))

y\_min <- min(matrix\_data, na.rm = TRUE) - 5

y\_max <- max(matrix\_data, na.rm = TRUE) + 5

plot(NA, xlim = range(years), ylim = c(y\_min, y\_max),xlab = 'Года', ylab = index\_names\_rus[i],

main = paste('Динамика', index\_names\_rus[i], 'по странам'))

for (j in 1:length(countries)) {

lines(years, matrix\_data[, j], type = 'b', col = colors[j], lty = 1, pch = 16, lwd = 2)

}

legend("topright", inset = c(-0.25, 0), legend = countries, fill = colors, bty = 'n')

}

# 4

url<-read\_html('https://kudago.com/spb/list/33-luchshih-muzeya-peterburga/')

articles <- url %>% html\_nodes("article.post-list-item")

titles <- articles %>% html\_nodes("a.post-list-item-title-link") %>% html\_text(trim = TRUE)

links <- articles %>% html\_nodes("a.post-list-item-title-link") %>% html\_attr("href")

descriptions <- articles %>% html\_nodes("div.post-list-item-description div p") %>% html\_text(trim = TRUE)

addresses <- articles %>%

html\_nodes("address.post-list-item-info") %>%

html\_text(trim = TRUE) %>%

gsub("[\n\r]", " ", .) %>%

gsub("\\s+", " ", .) %>%

substr(., 2, nchar(.))

descriptions <- articles %>%html\_nodes("div.post-list-item-description") %>%

.[sapply(., function(x) {

children <- html\_children(x)

length(children) >= 2 && length(html\_children(children[2])) == 1 && html\_name(html\_children(children[2])) == "p"

})] %>%

html\_nodes("div p") %>%

html\_text(trim = TRUE)

museum\_data <- data.frame(

Название = titles,

Ссылка = links,

Описание = descriptions,

Адрес = addresses

)