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»

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

Выполнил: ст. гр. 36/2

Сгонник Н.С.

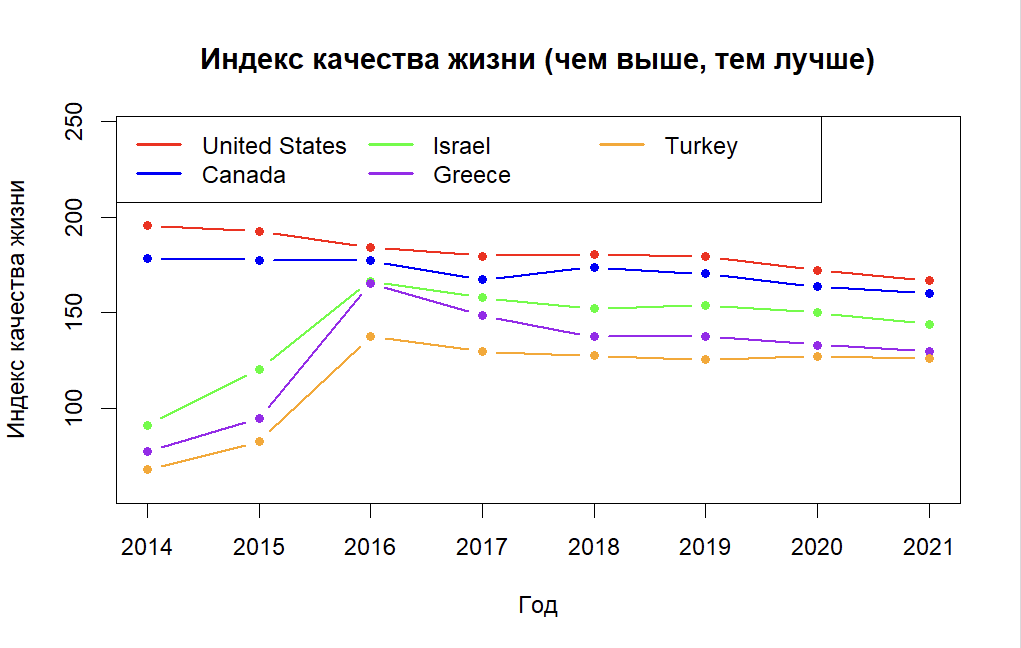
Проверил: преп. кафедры ВТ

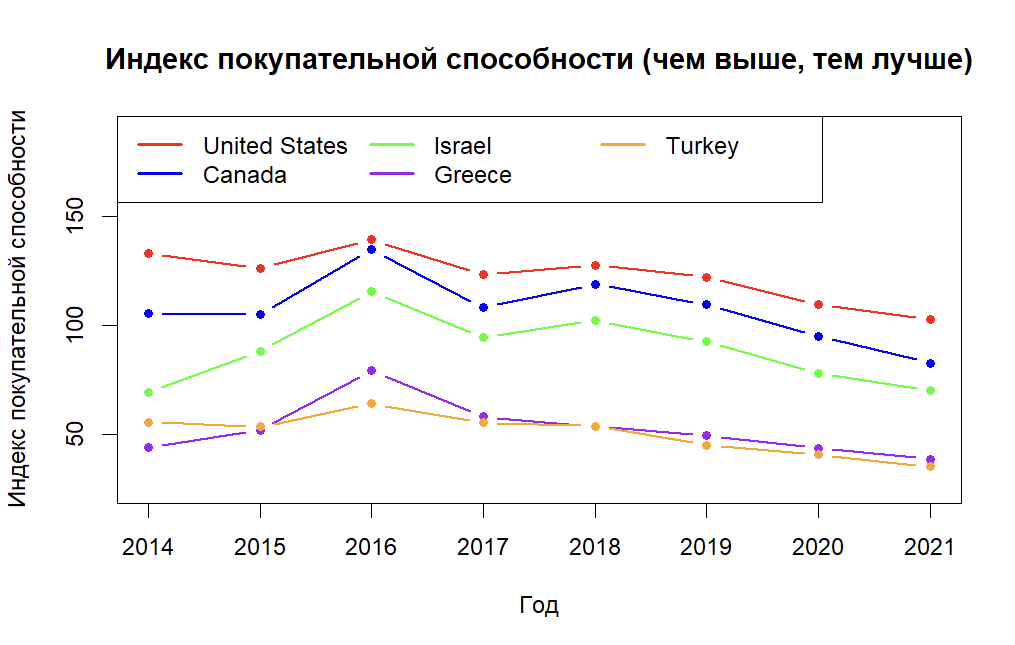
Яхонтов А. А.

Краснодар

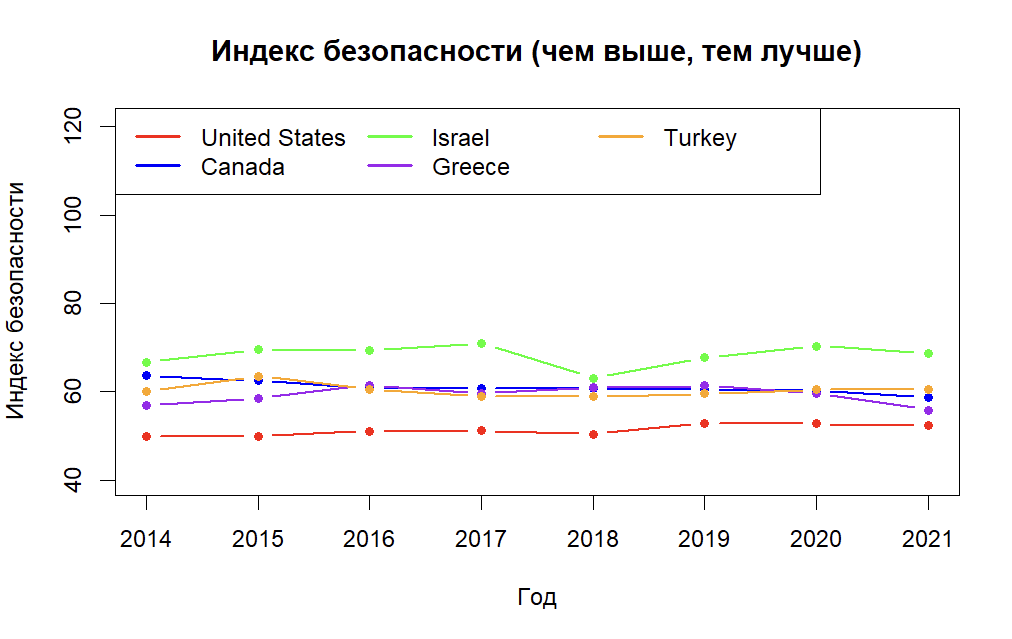
2025

Вариант 31: Канада, США, Турция, Греция, Израиль.

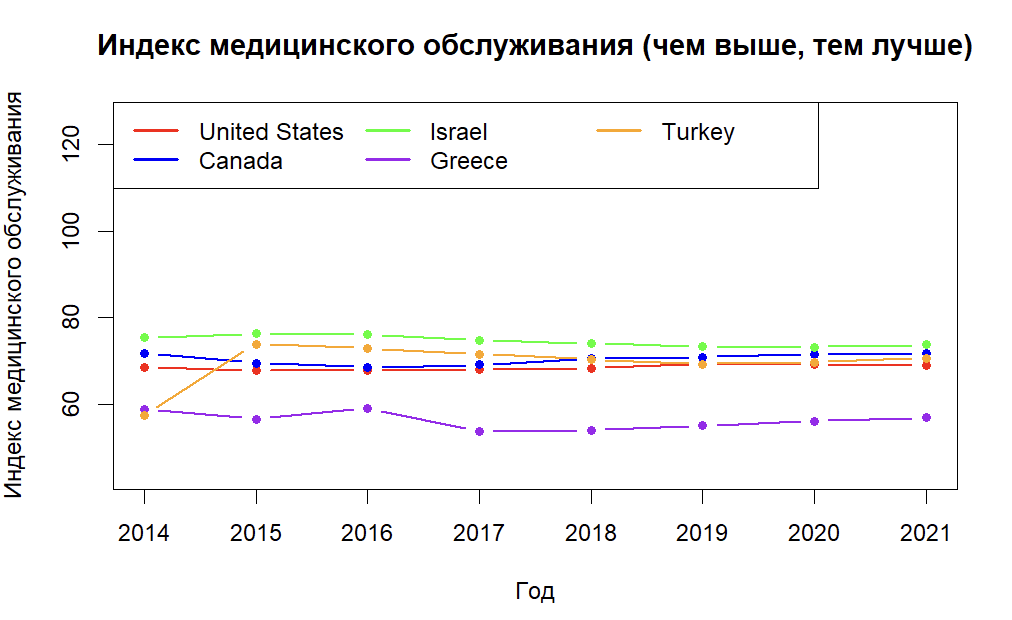
Индекс качества жизни вычисляется на основании нескольких факторов, графики которых приведены ниже. Чем выше данный показатель, тем лучше. Исходя из полученных данных можно сделать вывод, что у США и Канады этот показатель довольно высокий, в то время как как у остальных стран он будет на порядок ниже, хотя также можно подметить, что в 2016 год у отстающих стран он неплохо вырос.



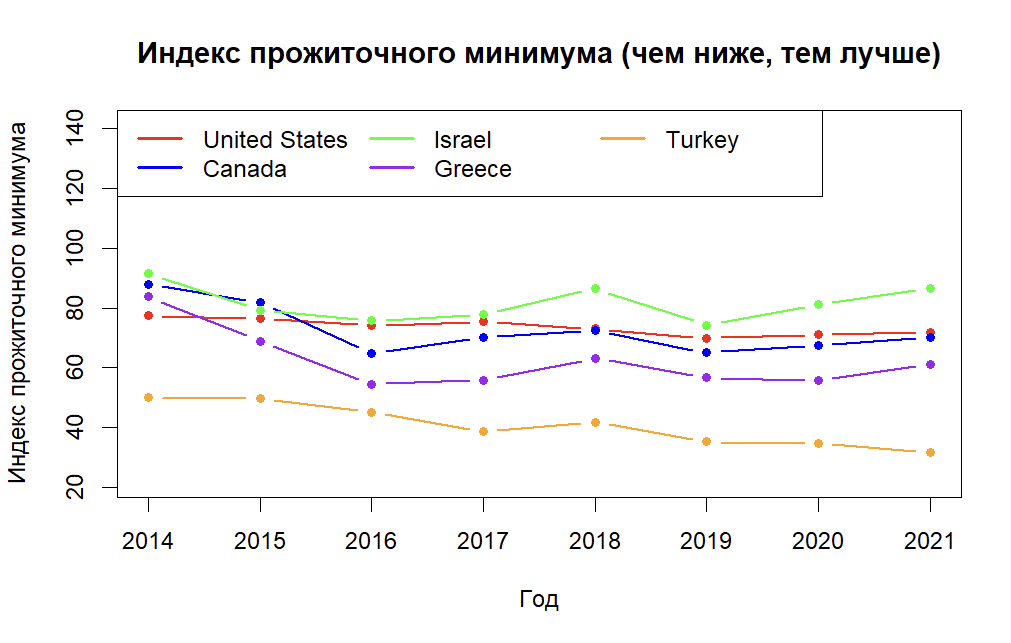
По индексу покупательной способности себя также хорошо показывают США и Канада, но и Израиль не сильно отстаёт. Но если говорить про Турцию и Грецию, то здесь индекс покупательной способности в три раза меньше чем у США.



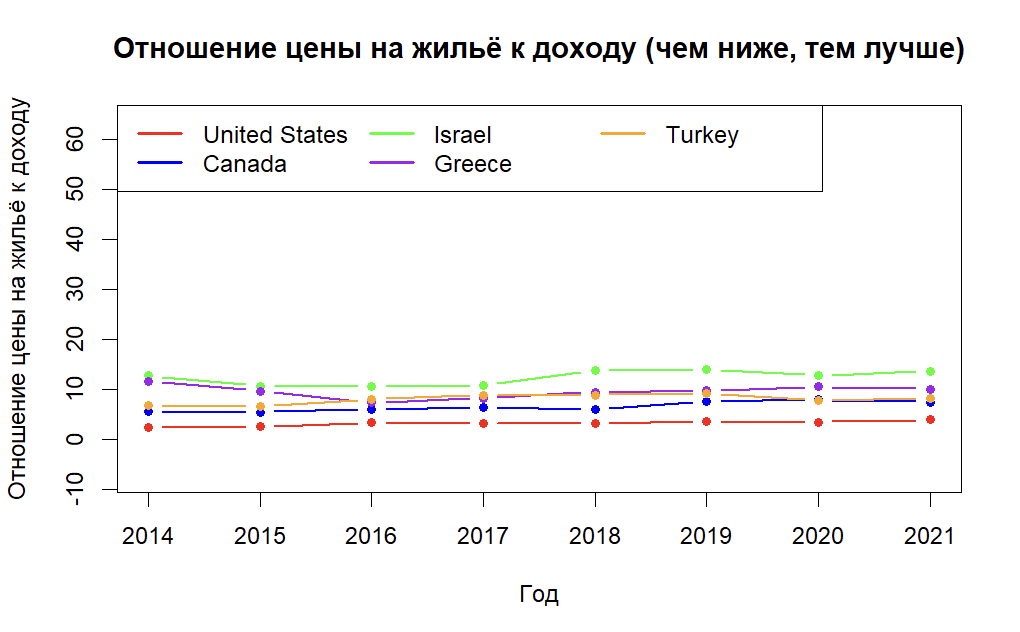
По индексу безопасности все страны держатся близко друг к другу, однако показатель Израиль немного выше, чем остальные. Также видно, что в странах нет никаких резких изменений по данному индексу.



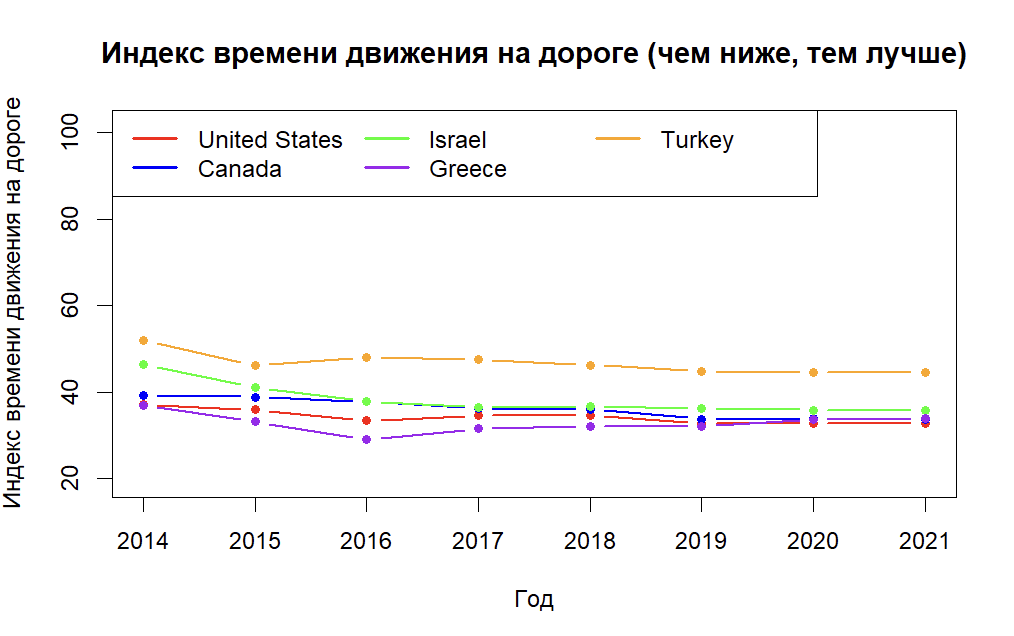
По индексу медицинского обслуживания лучше всего себя показывает Израиль, а также именно здесь себя неплохо показывает Турция. Сильно от остальных здесь отстаёт Греция.



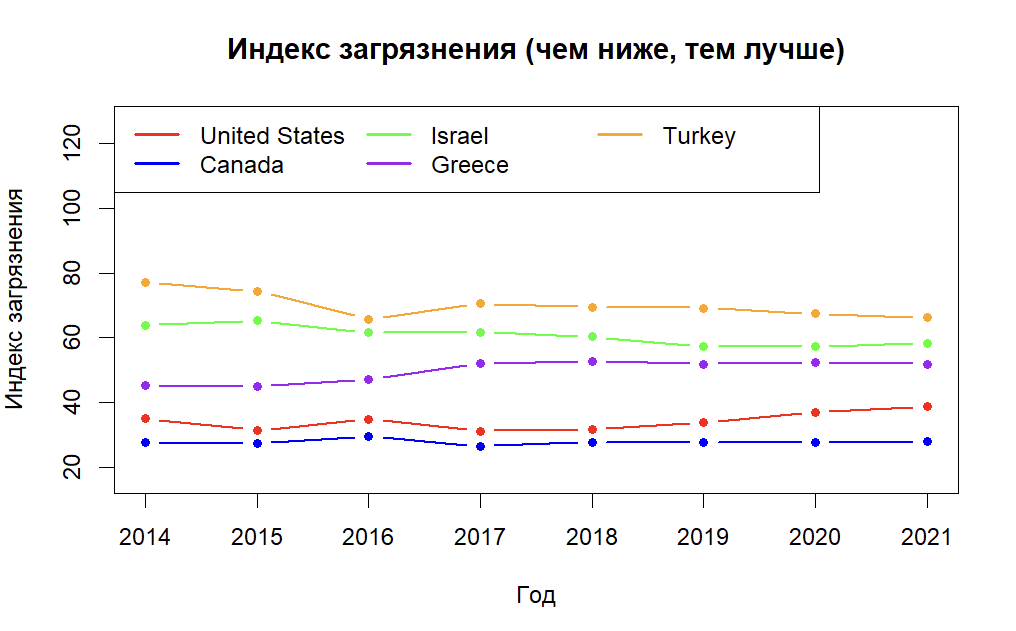
Чем ниже индекс прожиточного минимума, тем лучше. Исходя из графика, можно сделать вывод, что самый высокий индекс наблюдается в Израиле на протяжении всех лет, но в целом показатели США, Канады и Израиля держатся близко друг к другу. Наилучшим показателем обладает Турция, после неё также неплохо показывает себя Греция.

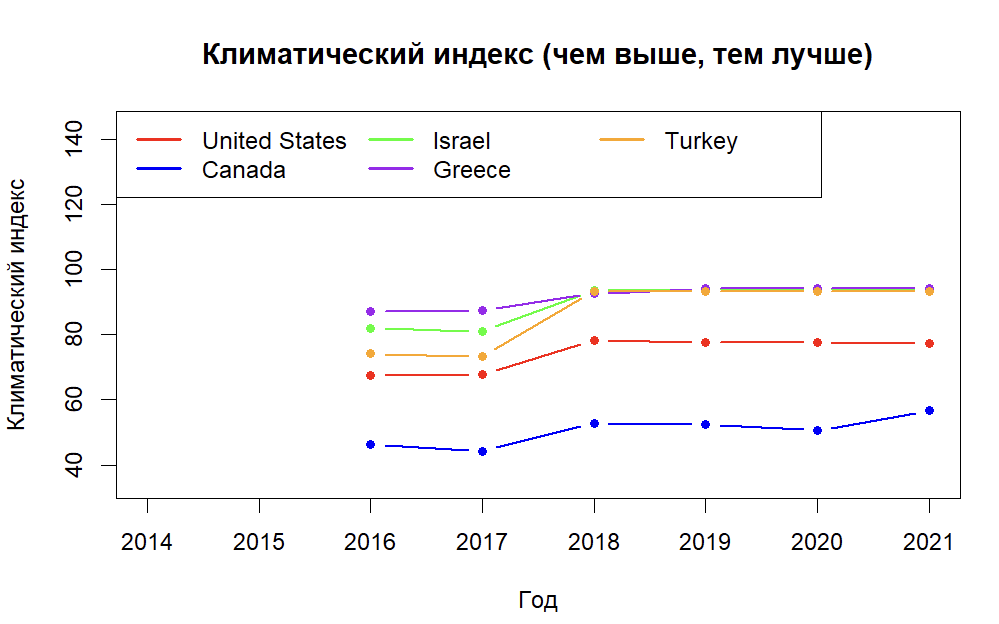


На данном графике видно, что все страны показывают примерно одинаковые показатели. Лучший он у США и худший у Израиля.



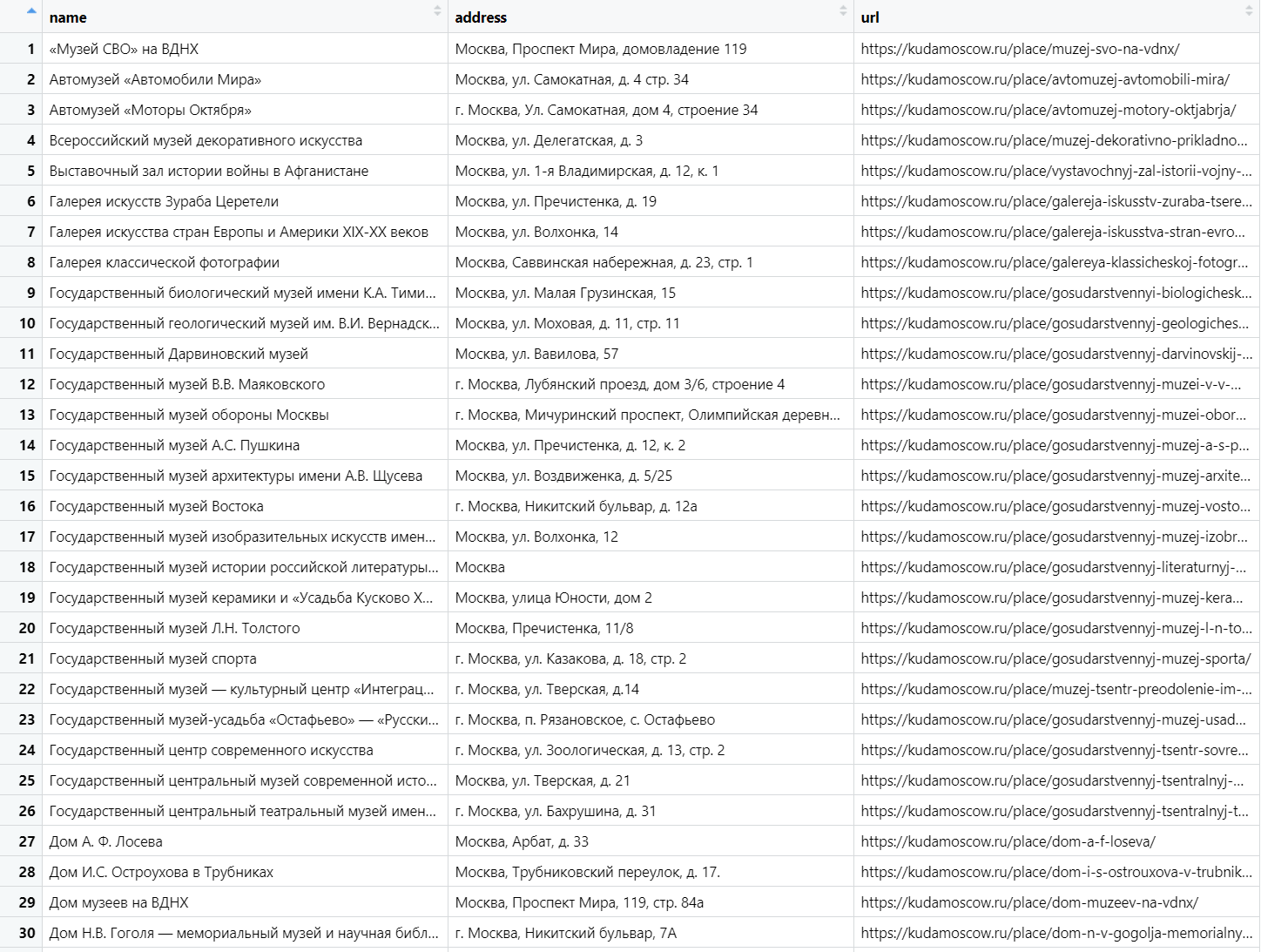
Здесь также чем ниже, тем лучше. Худший показатель нахождения в дороге в Турции, а остальные показывают себя примерно одинаково.

Можно заметить, что самой грязной страной из этой пятёрки является Турция, а самой чистой Канада, хотя и США от Канады не сильно отстаёт. Греция по индексу загрязнения находится где-то между Израилем и Канадой.



Данные по климату до 2016 года отсутствуют. Турция, Израиль и Греция показывают себя одинаково хорошо. У США этот показатель будет похуже чем у остальных, а в Канаде он совсем низкий. Канада уступает здесь почти в три раза.

С помощью пакета rvest были собраны дата фрейм о 233 музеях в Москве, в котором содержатся названия музеев, их местонахождения и ссылки этих музеев на этом сайте.



Вывод: В ходе лабораторной работы были освоены основные возможности библиотеки rvest для веб-скрапинга. Эта библиотека предоставляет удобные инструменты для извлечения данных с веб-страниц, включая работу с CSS-селекторами, парсинг HTML-таблиц и другие функции.

На основе полученных данных были построены графики, что позволило провести сравнительный анализ различных социально-экономических показателей стран. Применение методов веб-скрапинга значительно упростило сбор и обработку информации, продемонстрировав эффективность rvest для анализа открытых интернет-источников. Таким образом, работа подтвердила, что rvest - это мощный инструмент для автоматизации сбора и обработки веб-данных в R.

Листинг (lab4.R):

library(stringi)

library(rvest)

years <- 2014:2021

countries <- c("Canada", "United States", "Turkey", "Greece", "Israel")

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

getAllData <- function(year){

cat(sprintf("Загрузка данных о странах за %d год ... ", year))

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

page <- read\_html(url)

table\_node <- html\_nodes(page, "table#t2")[[1]]

table <- as.data.frame(html\_table(table\_node, na.strings="-"))

table <- table[-1]

cat("Готово\n")

return(table)

}

allData <- lapply(years, getAllData)

names(allData) <- years

CountriesData <- lapply(allData,subset, Country %in% countries)

#Подготовка данных для

quality\_assessment\_list <- lapply(2:ncol(CountriesData[[1]]),function(i){

df = data.frame(

)

result\_list <- lapply(years, function(year){

row <- t(CountriesData[[as.character(year)]][i])

rownames(row) = year

colnames(row) = t(CountriesData[[as.character(year)]][1])

row

})

df <- do.call(rbind, result\_list)

df

})

mains = c('Индекс качества жизни (чем выше, тем лучше)',

'Индекс покупательной способности (чем выше, тем лучше)',

'Индекс безопасности (чем выше, тем лучше)',

'Индекс медицинского обслуживания (чем выше, тем лучше)',

'Индекс прожиточного минимума (чем ниже, тем лучше)',

'Отношение цены на жильё к доходу (чем ниже, тем лучше)',

'Индекс времени движения на дороге (чем ниже, тем лучше)',

'Индекс загрязнения (чем ниже, тем лучше)',

'Климатический индекс (чем выше, тем лучше)'

)

quality\_assessment\_plot <- function(quality\_df, main){

min <- min(as.matrix(quality\_df), na.rm = TRUE)

max <- max(as.matrix(quality\_df), na.rm = TRUE)

print(main)

print(quality\_df)

matplot(

years,

quality\_df,

type="b",

pch=16,

lty=1,

lwd=1.8,

cex=0.8,

col=colors,

ylim=c(min - 10, max + 50),

main= main,

xlab='Год',

ylab=strsplit(main, " (", fixed = TRUE)[[1]][1]

)

legend('topleft', colnames(quality\_df), ncol=3, lty=1, lwd=2, col=colors)

}

lapply(seq\_along(quality\_assessment\_list) , function(i){

quality\_assessment\_plot(quality\_assessment\_list[[i]],mains[i])

})

Листинг (lab4\_museum.R):

library(rvest)

library(xml2)

url <- "https://kudamoscow.ru/place/museum/"

page <- read\_html(url)

place\_items <- html\_elements(page, ".place\_item")

urls <- html\_attr(html\_elements(place\_items, "a"),"href")

places\_pages <- lapply(seq\_along(urls), function(i) {

cat(sprintf("Загрузка %d страницы c url '%s' ...", i, urls[[i]]))

page <- read\_html(urls[i])

cat(" Готово!\n")

page

})

places\_names <- lapply(places\_pages, function(page){

html\_text(html\_elements(page, "h1"))

})

places\_addresses <- lapply(places\_pages, function(page){

html\_text(html\_elements(page,".info p"))[1]

})

museums\_df <- data.frame(

name = unlist(places\_names),

address = unlist(places\_addresses),

url = unlist(urls),

stringsAsFactors = FALSE

)