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**

**ЛАБОРАТОРНАЯ РАБОТА №4**

**Дисциплина: Обработка больших данных**

Работу выполнил(а):\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Иванченко П.А.

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Т. А. Приходько

Краснодар

2025

**Тема**: извлечение данных с веб-страниц.

**Цель работы**: научиться работать с инструментами пакета rvest и извлекать информацию с web-страниц с помощью инструментов языка R.

**Вариант 7:**

|  |  |
| --- | --- |
| Иванченко Павла Андреевна | Турция, Греция, Египет, Австралия, Новая Зеландия |

**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. Используя библиотеку rvest, считаем все нужные данные с таблицу all\_data и проведем дескриптивный анализ данных.

Количество стран: 92;

Годы в данных: 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021;

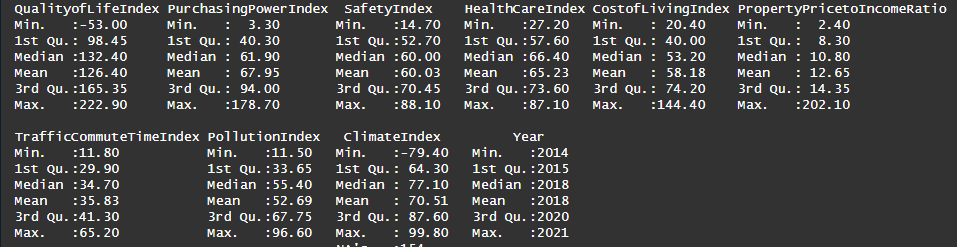


Рис. 1 − Сводная статистика

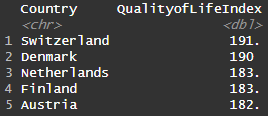


Рис.2 − Топ-5 стран за 2021 год

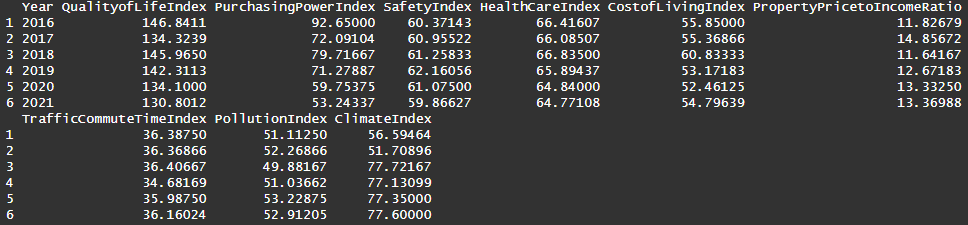


Рис.3 − Средние значения показателей по годам

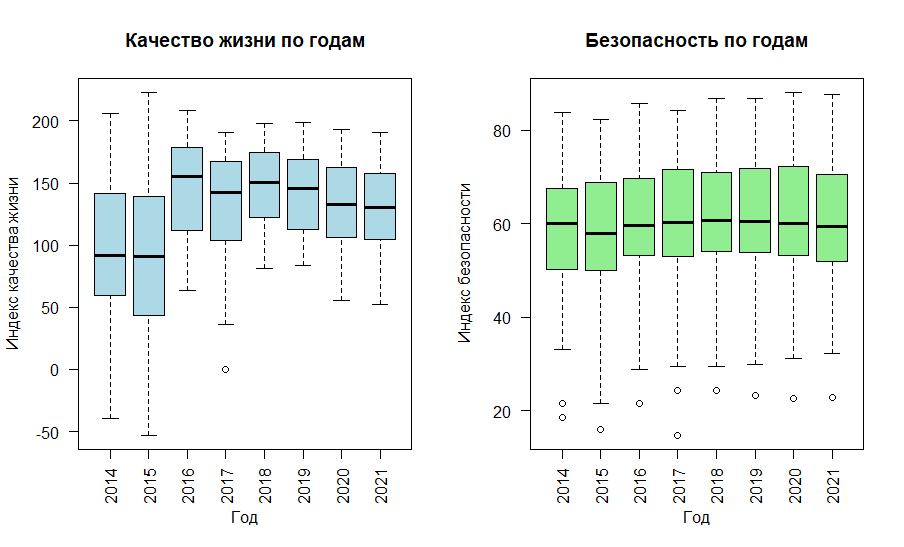


Рис.4 − Боксплоты по основным параметрам выборки

2-3. Из общей выборки проанализируем данные по пяти странам: Турции, Греции, Египту, Австралии и Новой Зеландии. Для этого сначала построим графики по всем показателям:

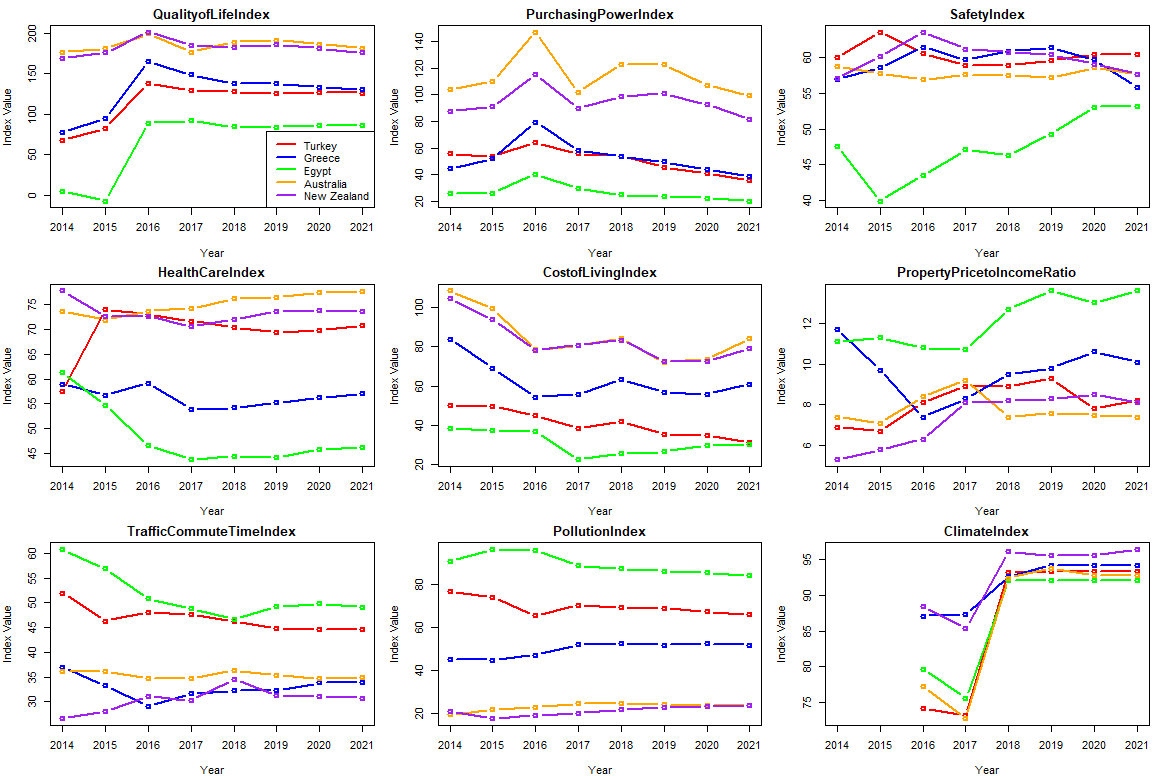


Рис.5 − Матрица графиков для 5 выбранных стран

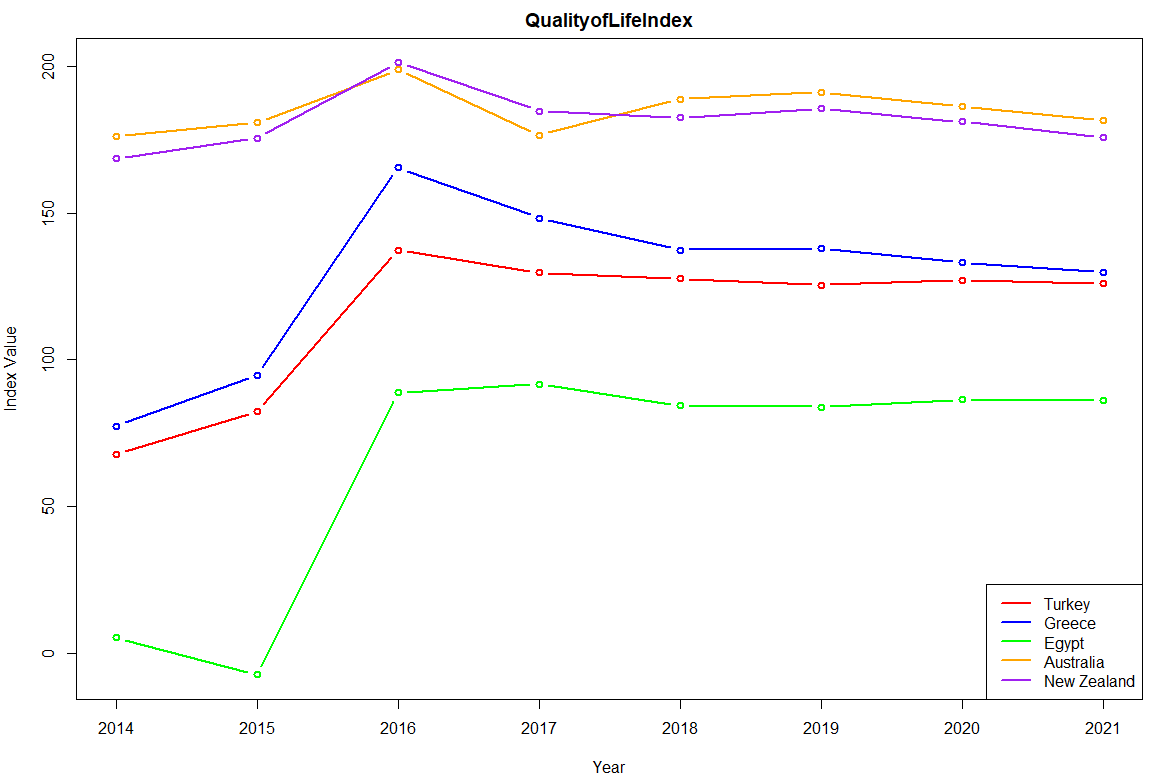


Рис.6 − График качества жизни для 5 выбранных стран

Исходя из этого графика, можем сделать вывод, что самое высокое качество жизни в Австралии и Новой Зеландии; Турция и Греция имеют средние показатели; Египет является худшей страной для жизни среди выбранных. Для всех стран показатель качества жизни возрос в 2016 году, однако это может быть связано не с реальным улучшением, а с изменением метрик сайта Numbeo.

Кроме этого, стоит отметить, что в странах, находящихся близко друг к другу географически, схожие показатели.

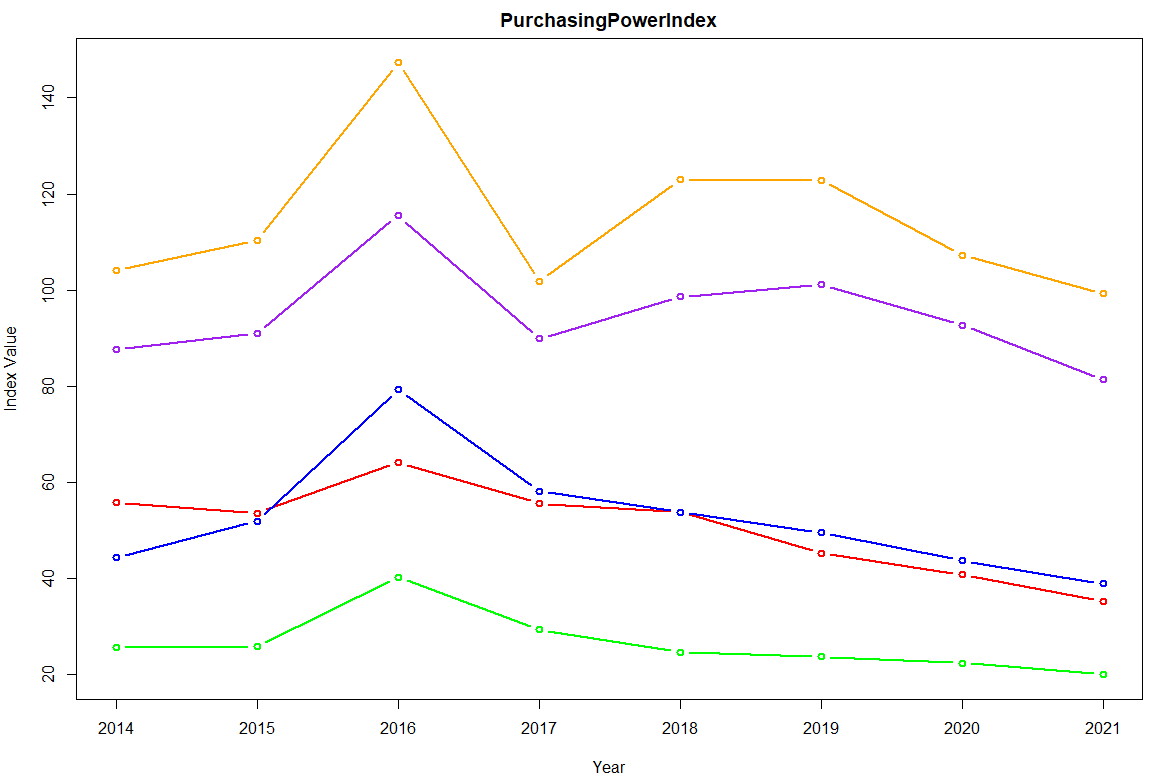


Рис.7 − График покупательной способности по 5 выбранным странам

В данном графике тенденция сохраняется – Австралия и Новая Зеландия также занимают лидирующие места по покупательной способности, однако теперь Австралия четко вырывается вперед. Греция и Турция занимают средние позиции, сохраняя значение индекса между 40 и 60 в среднем, в то время как Египет остается аутсайдером и имеет значение покупательной способности на уровне 20.

Скачок PPI в 2016 — это в первую очередь артефакт методологии, а не реальное удвоение покупательной способности.

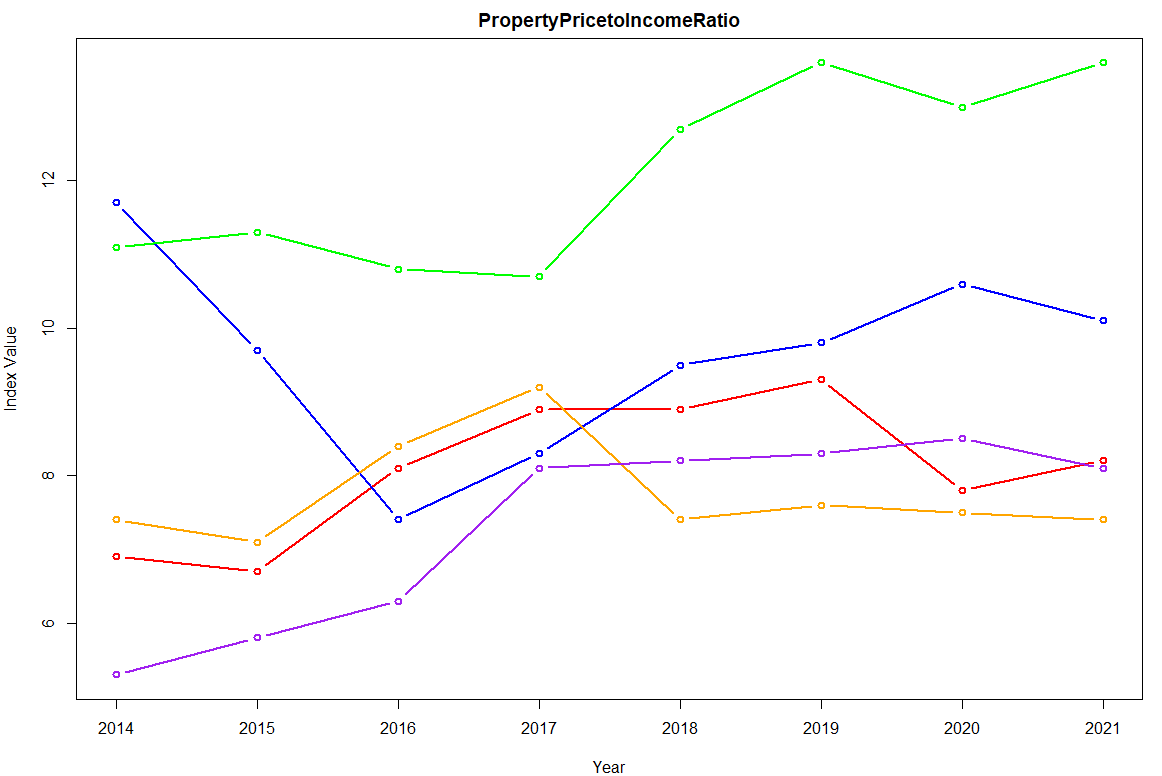


Рис.8 − График соотношения цены жилья к доходу по 5 выбранным странам

По данному графику можно сразу сказать, что Египет является самой доступной страной для покупки жилья в соотношении с доходом. В Греции этот показатель за последние годы стабильно растет, однако после 2020 года слегка опускается, что, скорее всего, связано с пандемией Ковид-19. В Турции и Новой Зеландии показатели примерно одинаковы, но Новая Зеландия является более стабильной. Австралия имеет самый низкий показатель по данной метрике, что означает, что жилье здесь самое дорогое.

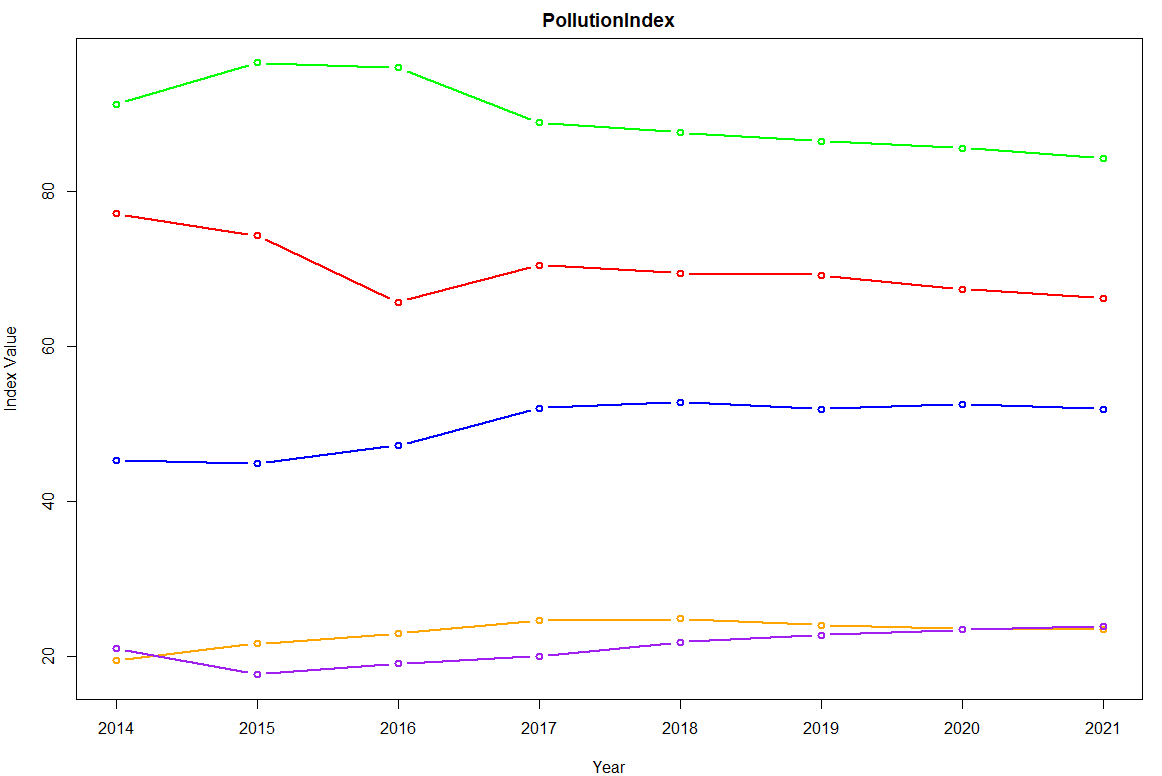


Рис.9 − График загрязнения по 5 выбранным странам

По графику можно заметить, что самой загрязненной страной из выборки является Египет, что может быть связано с природными факторами, например, пыльными бурями. Самыми низкими показателями обладают Австралия и Новая Зеландия, что может быть связано с низкой плотностью населения в этих странах и экологическим законодательством.

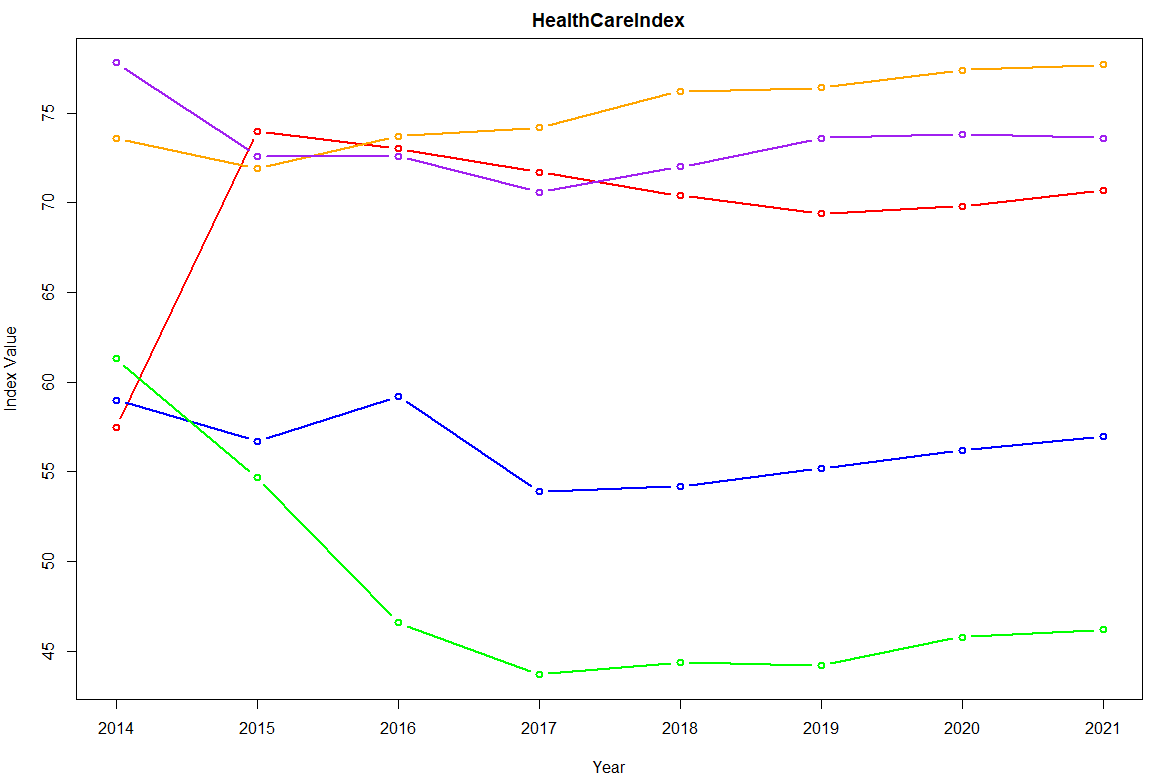


Рис. 10 − График медицины в 5 выбранных странах

На графике видно, что и по медицинским критериям лидерами остаются Австралия и Новая Зеландия, но также можно заметить, что к ним присоединяется Турция. Это может быть связано с развитой областью медицинского туризма в стране и с реформами, которые проводили власти с 2010 по 2020 годы (цифровизация, строительство новых госпиталей). Греция с 2017 показывает стабильный рост в качестве медицины. В Египте все как обычно :(

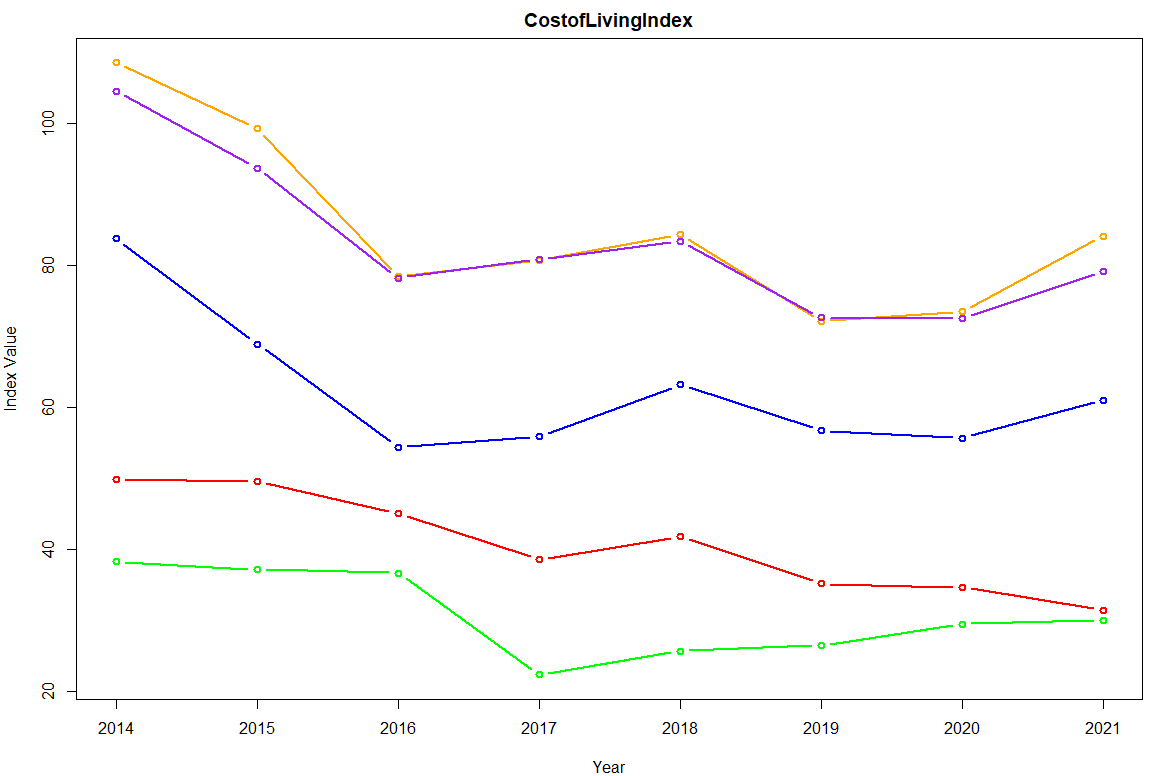


Рис. 11 − График стоимости жизни в 5 выбранных странах

Стоимость жизни в Турции и Египте небольшая, что соответствует низкому уровню жизни в этих странах относительно других из выборки. В 2018 году стоимость жизни во всех странах возросла, т.к. выросла цена нефти. Она стабилизировалась уже к 2019, поэтому показатели уменьшились.

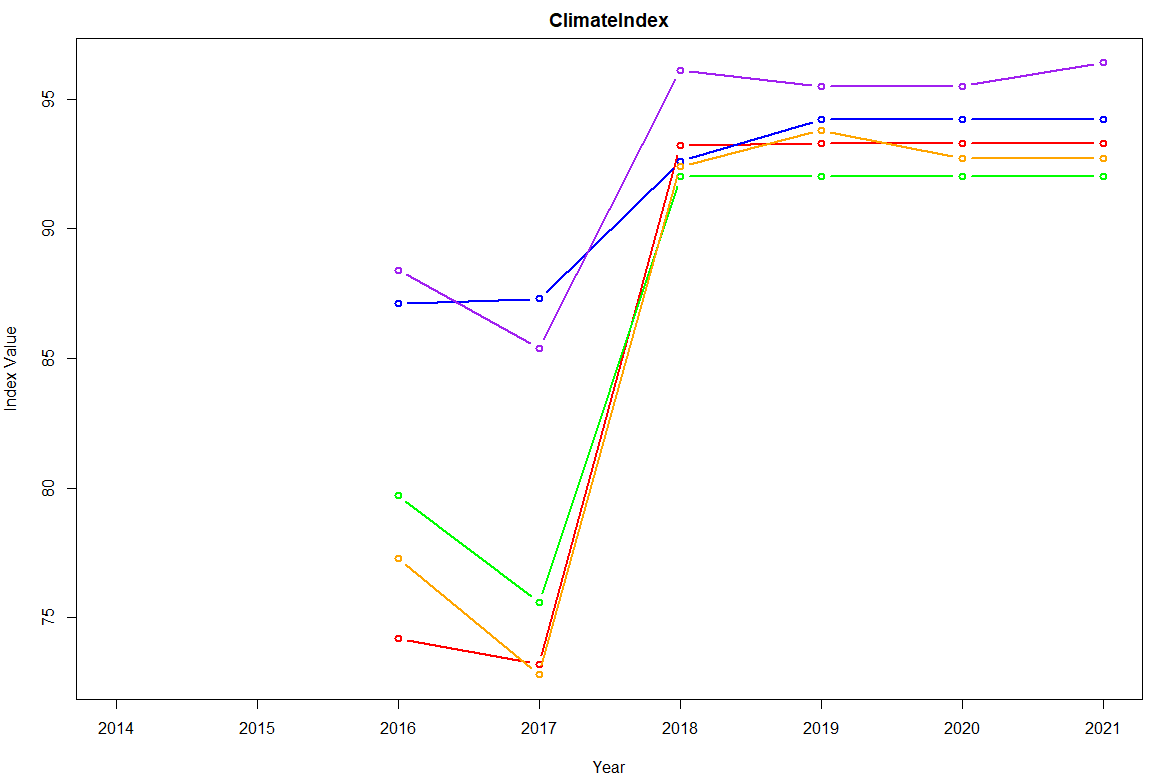


Рис. 12 − График климатических изменений по 5 выбранным странам

В 2018 году все показатели возросли в связи с изменениями в методологии Numbeo.

4. Соберем информацию по ссылке <https://ru.wikipedia.org/wiki/Список_музеев_Ростовской_области>: название музея, его местоположение и ссылку на музей. Получим следующую таблицу:

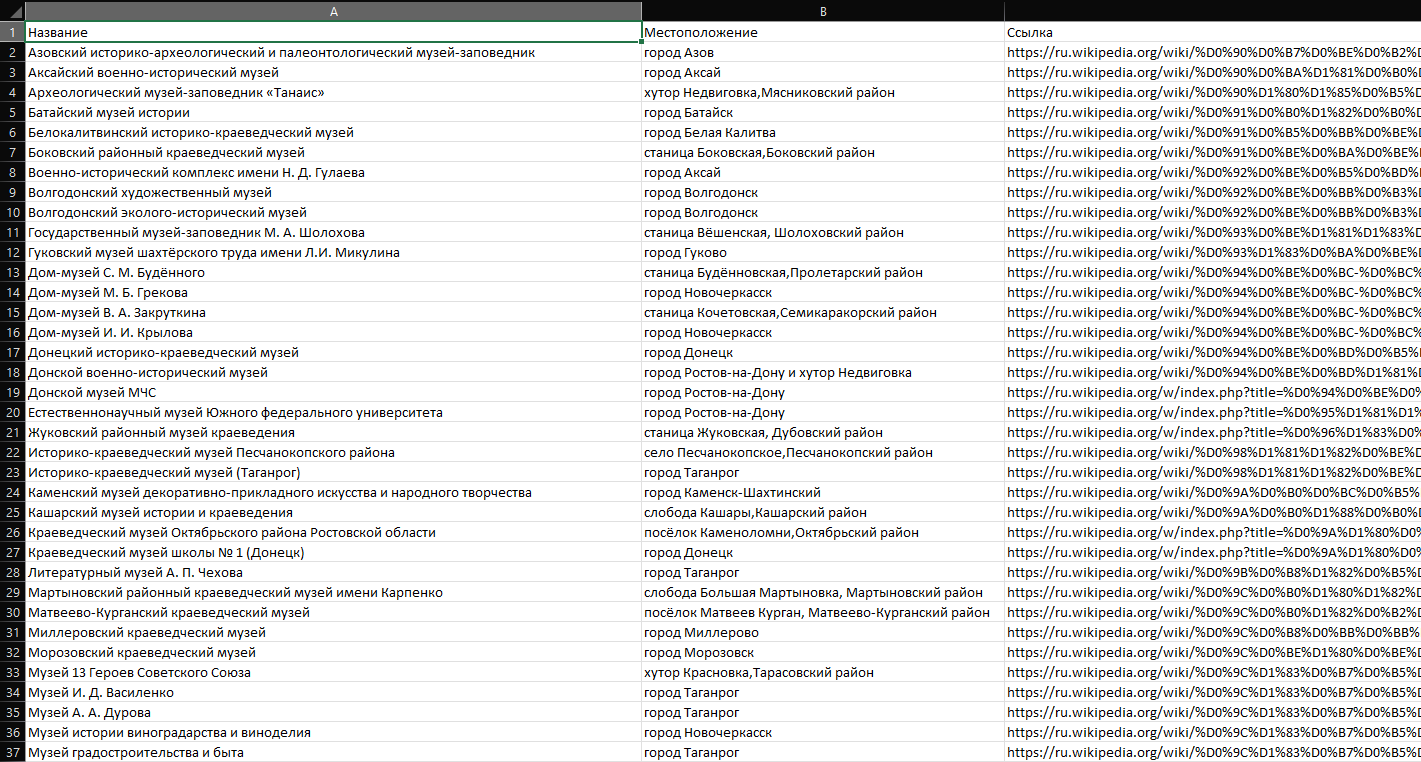


Рис.13 − Таблица собранных данных о музеях

**Вывод**: в процессе выполнения лабораторной работы я научилась работать с инструментами пакета rvest и извлекать информацию с web-страниц с помощью инструментов языка R.

**ПРИЛОЖЕНИЕ А**

Листинг кода на R

library(rvest)

library(dplyr)

library(purrr)

library(stringr)

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

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

tryCatch({

page <- read\_html(url)

table\_node <- page %>% html\_node("table#t2")

if (is.na(table\_node)) {

message(paste("Таблица не найдена на странице за", year))

return(NULL)

}

table\_data <- table\_node %>%

html\_table(header = TRUE, fill = TRUE) %>%

as\_tibble() %>%

mutate(across(-c(Rank, Country), as.numeric)) %>%

rename\_all(~str\_remove\_all(., "\\s+")) %>%

mutate(Year = year)

return(table\_data)

}, error = function(e) {

message(paste("Ошибка при обработке данных за", year, ":", e$message))

return(NULL)

})

}

test\_data <- get\_numbeo\_data(2021)

if (!is.null(test\_data)) {

print("Тестовый запрос успешен:")

print(head(test\_data))

years <- 2014:2021

all\_data <- map\_dfr(years, ~{

Sys.sleep(3)

get\_numbeo\_data(.x)

})

if (!is.null(all\_data) && nrow(all\_data) > 0) {

all\_data <- all\_data %>% select(where(~!all(is.na(.x))))

write.csv(all\_data, "numbeo\_quality\_of\_life\_2014\_2021.csv", row.names = FALSE)

message("\nДанные успешно сохранены в файл 'numbeo\_quality\_of\_life\_2014\_2021.csv'")

message("Всего записей: ", nrow(all\_data))

message("Столбцы: ", paste(names(all\_data), collapse = ", "))

str(all\_data)

} else {

message("Не удалось собрать данные.")

}

} else {

message("Тестовый запрос не удался.")

}

all\_data

# ДЕСКРИПТИВНЫЙ АНАЛИЗ

cat("=== ОБЩАЯ ИНФОРМАЦИЯ ===\n")

str(all\_data)

cat("\nКоличество стран:", length(unique(all\_data$Country)))

cat("\nГоды в данных:", unique(all\_data$Year), "\n")

cat("\n=== СВОДНАЯ СТАТИСТИКА ===\n")

numeric\_cols <- sapply(all\_data, is.numeric)

summary(all\_data[, numeric\_cols])

last\_year <- max(all\_data$Year)

cat("\n=== ТОП-5 СТРАН ЗА", last\_year, "ГОД ===\n")

head(all\_data[all\_data$Year == last\_year, c("Country", "QualityofLifeIndex")] %>%

arrange(desc(QualityofLifeIndex)), 5)

par(mfrow = c(1, 2))

boxplot(QualityofLifeIndex ~ Year, data = all\_data,

main = "Качество жизни по годам",

xlab = "Год", ylab = "Индекс качества жизни",

col = "lightblue", las = 2)

boxplot(SafetyIndex ~ Year, data = all\_data,

main = "Безопасность по годам",

xlab = "Год", ylab = "Индекс безопасности",

col = "lightgreen", las = 2)

par(mfrow = c(1, 1))

cat("\n=== СРЕДНИЕ ЗНАЧЕНИЯ ПОКАЗАТЕЛЕЙ ПО ГОДАМ ===\n")

aggregate(. ~ Year, data = all\_data[, numeric\_cols], mean, na.rm = TRUE)

last\_year\_data <- all\_data[all\_data$Year == last\_year, numeric\_cols]

cat("\n=== КОРРЕЛЯЦИЯ ПОКАЗАТЕЛЕЙ ЗА", last\_year, "ГОД ===\n")

round(cor(last\_year\_data, use = "complete.obs"), 2)

# ВАРИАНТ 7

selected\_countries <- c("Turkey", "Greece", "Egypt", "Australia", "New Zealand")

filtered\_data <- all\_data[all\_data$Country %in% selected\_countries, ]

if (nrow(filtered\_data) == 0) {

stop("Не найдены данные для указанных стран. Проверьте названия.")

} else {

cat("Найдены данные для стран:", toString(unique(filtered\_data$Country)), "\n")

}

metrics <- setdiff(names(filtered\_data), c("Year", "Country", "Rank"))

par(mfrow = c(3, 3), mar = c(4, 4, 2, 1))

for (metric in metrics) {

plot(1, type = "n",

xlim = range(filtered\_data$Year),

ylim = range(filtered\_data[[metric]], na.rm = TRUE),

main = metric,

xlab = "Year", ylab = "Index Value")

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

for (i in seq\_along(selected\_countries)) {

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

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

col = colors[i], lwd = 2, type = "b")

}

if (metric == metrics[1]) {

legend("bottomright", legend = selected\_countries,

col = colors, lwd = 2, cex = 1.2)

}

}

#4 ЗАДАНИЕ

library(rvest)

library(dplyr)

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

page <- read\_html(url)

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

rows <- html\_nodes(table\_node, "tr")[-1]

museum\_data <- lapply(rows, function(row) {

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

if (length(columns) >= 3) {

name\_node <- html\_node(columns[2], "a")

name <- if (!is.na(name\_node)) html\_text(name\_node, trim = TRUE) else html\_text(columns[2], trim = TRUE)

link <- html\_attr(name\_node, "href")

full\_link <- if (!is.na(link)) paste0("https://ru.wikipedia.org", link) else NA

location <- html\_text(columns[3], trim = TRUE)

data.frame(

Название = name,

Местоположение = location,

Ссылка = full\_link,

stringsAsFactors = FALSE

)

} else {

NULL

}

})

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

write.csv(museum\_df, file = "музеи\_ростовской\_области.csv", row.names = FALSE, fileEncoding = "Windows-1251")