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:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Лотарев С.Ю.

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

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

Краснодар

2025

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

**Цель работы**: Научиться извлекать информацию с WEB-страниц с помощью инструментов языка R.

Вариант 9:

|  |  |
| --- | --- |
| Лотарев Сергей Юрьевич | США, Португалия,Чехия, Хорватия, Россия |

Задание:

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.

1. 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, прокомменитровать в отчете результат.

Необходимо нарисовать на одном и том же графике рейтинг всех 5 стран, проанализировать резутьтат, анализ словесно отразить в отчете. Проанализировать изменение во времени всех показателей указанных стран, подобрать наилучший способ визуализации.

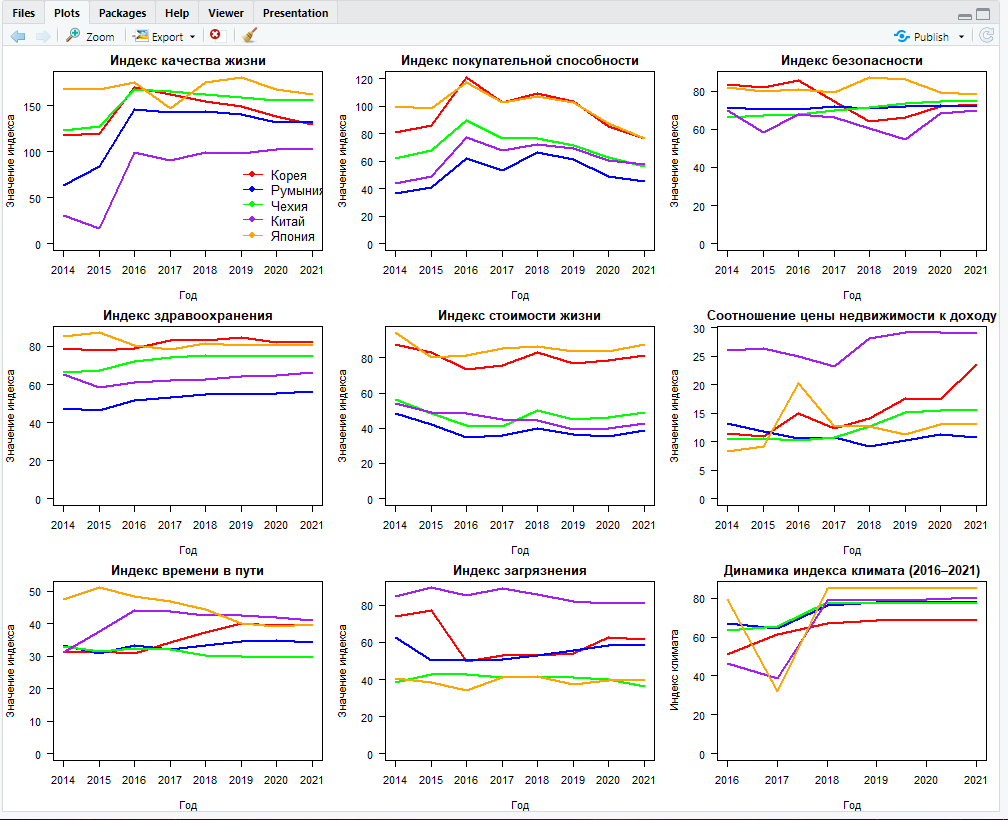


Рисунок 1. Графики динамики значений каждого из индексов относительно временной шкалы для каждой из пяти стран в наборе данных

Набор графиков содержит информацию об изменениях 10 различных индексов из набора данных на протяжении 7 лет с 2014 по 2021 год:

1. Качество жизни**:** на всём протяжении периода лидирует Япония, демонстрируя стабильно высокие значения. Следом идёт Чехия, незначительно уступая, но удерживая устойчиво высокую позицию. Корея занимает среднее положение с постепенным снижением индекса к 2021 году. Румыния демонстрирует скачок в 2016 году, но затем значения стабилизируются на среднем уровне. Китай стабильно занимает последнее место по индексу качества жизни среди всех стран, несмотря на незначительный рост к середине периода.
2. **Покупательная способность:** в 2016 году все страны достигают локальных максимумов, после чего наблюдается общий нисходящий тренд. Япония и Южная Корея существенно выделяются: именно они формируют верхнюю группу по уровню доходов, хотя в Японии после 2016 года прослеживается наиболее резкое снижение. Чехия занимает устойчивое положение в средней зоне, демонстрируя плавную динамику. Китай и особенно Румыния находятся в нижней части графика: несмотря на рост в начале периода, они остаются существенно позади по уровню покупательной способности.На последнем месте находится Россия.
3. **Безопасность:** Япония стабильно удерживает лидирующую позицию, демонстрируя высокий и ровный уровень безопасности на всём протяжении периода. Корея в начале показывала сопоставимые значения, но затем индекс постепенно снижался до минимума в 2018 году, после чего началось восстановление. Чехия и Румыния идут близко друг к другу, с умеренными колебаниями, причём Румыния к концу периода немного выравнивается. Наихудшие показатели на протяжении всех лет у Китая — начиная с самого низкого уровня, страна демонстрирует лишь незначительные колебания без существенного улучшения.
4. **Индекс здравоохранения:** Корея демонстрирует наивысшие значения индекса к концу периода, продемонстрировав плавный рост показателя. Япония идёт следом, несмотря на то, что индекс был выше в начале исследуемого периода, страна смогла сохранить высокий показатель. Чехия показывает ровный рост, постепенно сокращая отставание от лидеров. Китай остаётся в нижней части графика, но также демонстрирует постепенное улучшение. Румыния уверенно занимает последнее место, лишь незначительно увеличив свой индекс здравоохранения.
5. **Стоимость жизни**: на всём протяжении периода Япония демонстрирует наивысший уровень цен, удерживая стабильное лидерство по индексу. Корея идёт сразу за ней, однако показывает больше колебаний — в частности, заметен спад в 2016 году и повторный рост к 2021-му. Остальные страны — Чехия, Китай и Румыния — формируют нижнюю группу с более низкой стоимостью жизни. Среди них Чехия выделяется кратковременным ростом в 2018 году, тогда как Китай и Румыния идут почти параллельно, демонстрируя минимальные значения и умеренную стабильность.
6. **Отношение цены на жильё к доходу:** Китай в течение всего периода уверенно удерживает наивысшие значения, что указывает на наибольшую недоступность жилья среди рассматриваемых стран. Корея демонстрирует резкий рост после 2017 года, и к 2021 году значительно отрывается от остальных, занимая второе место. Япония показывает всплеск в 2016 году, но затем индекс возвращается к прежнему уровню и остаётся относительно стабильным. Румыния удерживает низкие значения на всём интервале, с умеренными колебаниями, что говорит о более сбалансированном соотношении доходов и стоимости жилья.
7. **Время движения на дороге**: Япония на всём протяжении остаётся страной с самым высоким индексом, что свидетельствует о наибольших затратах времени на дорогу. Китай также демонстрирует высокие значения, особенно в 2016–2018 годах, после чего наблюдается небольшое снижение. Корея показывает рост индекса с 2016 года, выходя на уровень Китая к концу периода. В то же время Чехия сохраняет наиболее низкие и стабильные значения, указывая на минимальные временные затраты на передвижение. Румыния держится чуть выше Чехии, с умеренными изменениями на протяжении всего периода.
8. **Загрязнение:** Китай стабильно демонстрирует наихудшие показатели по загрязнённости среды — на протяжении всех лет он остаётся на первом месте с наибольшими значениями. Корея занимает второе место, причём после снижения в 2016 году уровень загрязнения вновь начинает расти и к 2020 году сравнивается с Румынией. Румыния, несмотря на заметное улучшение в 2016 году, затем показывает стабильный рост индекса загрязнения. Чехия и Япония, напротив, на всём интервале сохраняют наиболее благоприятную экологическую ситуацию с наименьшими значениями индекса и минимальными колебаниями.
9. **Климат**: после резких колебаний в 2017 году, особенно выраженных у Японии, где наблюдается кратковременное падение, все страны к 2018 году выходят на более стабильный уровень. С этого момента показатели у всех стран удерживаются на устойчиво высоких значениях.

2. Со страницы https://ru.wikipedia.org/wiki/Музей\_Мирового\_океана

собрать информацию в data.frame, которая содержала бы: название музея, его адрес и ссылку для перехода при клике на фото / ссылке на музей.

1) Название музея находится на странице в первом элементе типа *h1*:



Рисунок 2. Элемент с названием музея.

2) Адрес музея находится в элементе со значением *itemprop*, равным *“address”*:



Рисунок 3. Элемент с адресом музея.

3) Ссылка на сайт музея находится в теге a с классом “external text”



Рисунок 4. Элемент со ссылкой на сайт музея.

Таким образом, получаем следующий датафрейм:

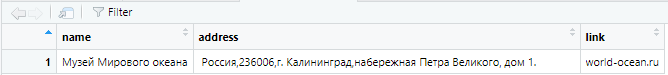


Рисунок 5. Датафрейм с данными музея.

**Вывод**: в процессе выполнения данной лабораторной работы я изучил способы извлечения информации из WEB-страниц при помощи средств языка R и применил полученные знания для выполнения анализа данных.