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:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Вавакин В.О.

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 3:

|  |  |
| --- | --- |
| Вавакин Владислав Олегович | 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.

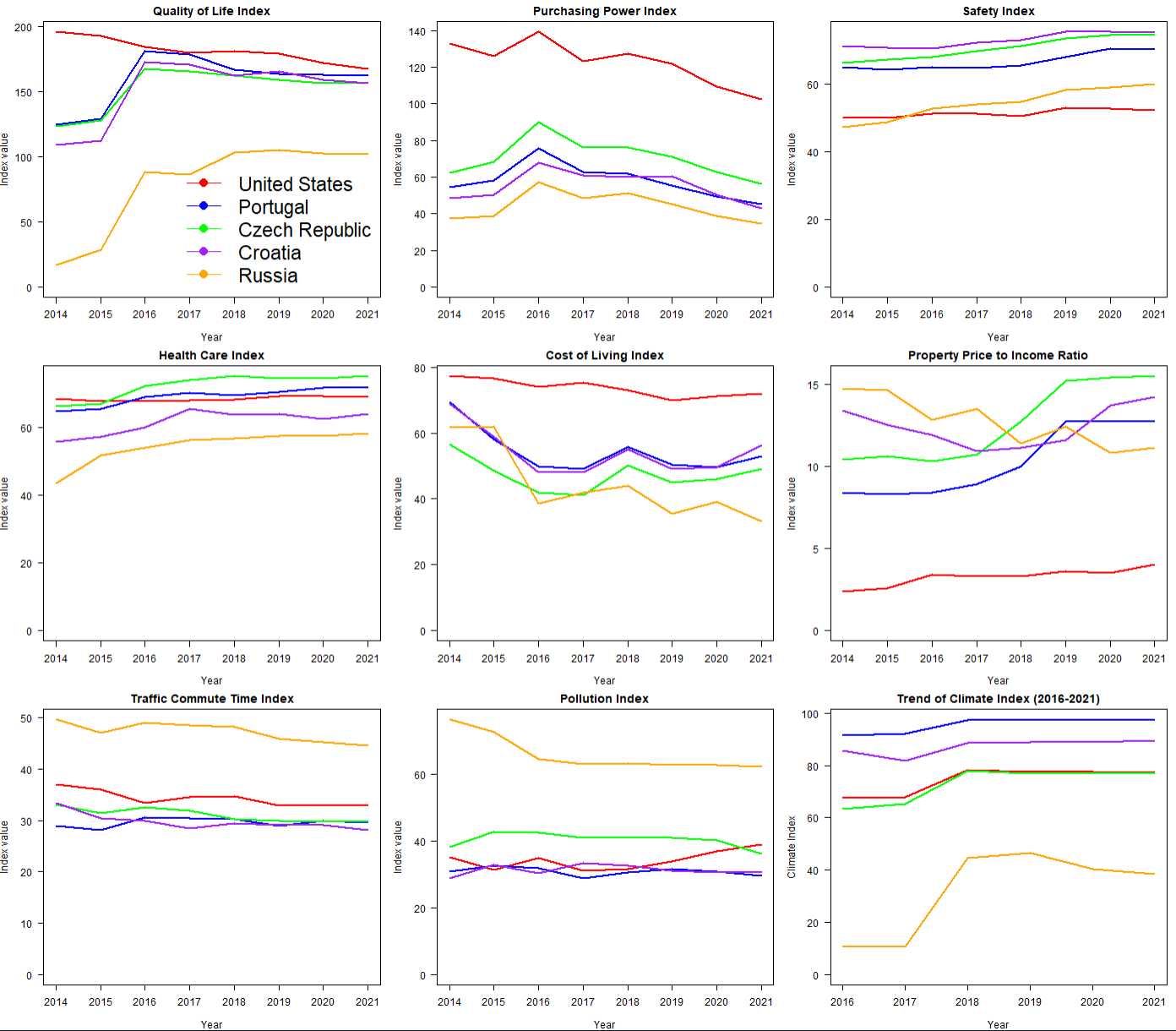


Рисунок 1. Графики динамики значений каждого из индексов относительно временной шкалы для каждой из пяти стран в наборе данных

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

1. Уровень жизни: несмотря на резкий рост в 2016 году, худшей страной по уровню жизни среди рассматриваемых остаётся Россия, где по окончании выборки он остаётся ниже, чем у второй худшей страны – Хорватии – в начале. Лучшей по данному показателю на всём протяжении остаётся США, хотя там он стабильно снижается.
2. Покупательная способность: динамика данного показателя практически совпадает для всех рассматриваемых стран – разница между значениями почти не меняется, за исключением небольших изменений у Хорватии и Португалии. Лидером по индексу является США, где он превосходит остальные страны из выборки в 2-3 раза. На последнем месте находится Россия.
3. Безопасность: все рассматриваемые страны находятся на достаточно высоком уровне безопасности. К 2016 году Россия обошла по этому показателю США, в результате чего Соединённые Штаты стали худшими по индексу среди рассматриваемых стран. Наилучший уровень безопасности наблюдается в Хорватии, близкие к ней показатели имеет Чехия.
4. Медицинское обслуживание: наилучший уровень мед. обслуживания наблюдается в Чехии, наихудший – в России, но там он наиболее стабильно растёт. В США, в отличие от остальных участников выборки, показатель практически не возрастает.
5. Стоимость жизни: на всём протяжении практически совпадает в Португалии и Хорватии, с небольшим отставанием в Чехии. Наибольшее значение имеет в США, наименьшее – в России, где он резко упал в 2016 году и продолжает снижаться.
6. Отношение цены на жильё к доходу: наиболее стабильно и мало в США, где оно в 4-5 раз меньше по сравнению с другими странами. Показатели у остальных участников выборки варьируется – в России – с колебаниями падает, в Чехии и Португалии динамика похожа, с резким ростом в 2019 году, в Хорватии также наблюдается рост к концу выборки после падения в начале.
7. Время движения на дороге: наибольшее значение в России, где оно превышает значения остальных участников выборки в 1,5-2 раза. В остальных странах показатель примерно одинаков, с небольшим превышением в США.
8. Загрязнение: худший результат показателя – в России, где он больше больше остальных в среднем в 2 раза. К концу выборки отмечается снижение уровня загрязнения в Чехии и повышение в США.
9. Климат: в России наблюдается значительное улучшение данного показателя в 2018 году, но там он всё равно остаётся худшим среди всех рассматриваемых государств. В остальных странах значение индекса практически не меняется, лучшее – в Португалии, далее – в Хорватии, затем – в США и Чехии, где оно практически совпадает.

2. Со страницы <https://kudamoscow.ru/place/avtomuzej-motory-oktjabrja/> 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 элементе типа *a* со значением параметра *title*, равным *“Сайт организаторов”*:



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.