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 №5**

**Дисциплина: Обработка больших данных**

Работу выполнил: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ К. А. Корнилов

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_В. И. Шиян

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

**Цель**: Научиться работать извлекать информацию сWEB-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-oflife/rankings\_by\_country.jsp?title=2021 с 2014 по 2021гг согласно варианту:



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, прокомментировать в отчете результат.
2. 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 (с вашей точки зрения) способ визуалазации.
3. С одной из страниц (по варианту) : 8) <https://ru.wikipedia.org/>wiki/ Список\_музеев\_Ростовской\_области (-//-), собрать информацию в data.frame, которая содержала бы: Название музея, его адрес и ссылку для перехода при клике на фото / ссылке на музей.

**Ход работы**:

1. Была собрана информация об уровне жизни стран с 2014 по 2021гг из таблиц с предложенного сайта согласно варианту. На основе данной информации создан датасет в виде датафрейма.

Код создания датасета:

library(rvest)

url\_data = read\_html("https://www.numbeo.com/quality-of-life/rankings\_by\_country.jsp?title=2021")

#Выдернули года

years = html\_nodes(url\_data,"select[name='title'] option") %>% html\_attr("value")

years = years[years>="2014" & years<="2021"]

years

#Функция получения таблицы по ссылке

getTable = function(year){

country\_names = c("Finland", "Denmark", "France", "Germany","Romania")

link = paste("https:// [www.numbeo.com/](http://www.numbeo.com/) quality-of-life/ rankings\_by\_country.jsp?t itle=",year,sep="")

url\_data = read\_html(link)

table = html\_table(html\_nodes(url\_data,"div table"))[[2]] %>% as.data.frame()

table[,1]=c(1:length(table[,1]))

table=table[table$Country %in% country\_names,]

table = cbind(data.frame(Year=year,table))

return(table)

}

#Создаем 1 общую таблицу

data = do.call(rbind,lapply(years,getTable))

rownames(data)=c(1:dim(data)[1])

data[data=="-"]=0

data

Результат создания датасета представлен на рисунке 1:

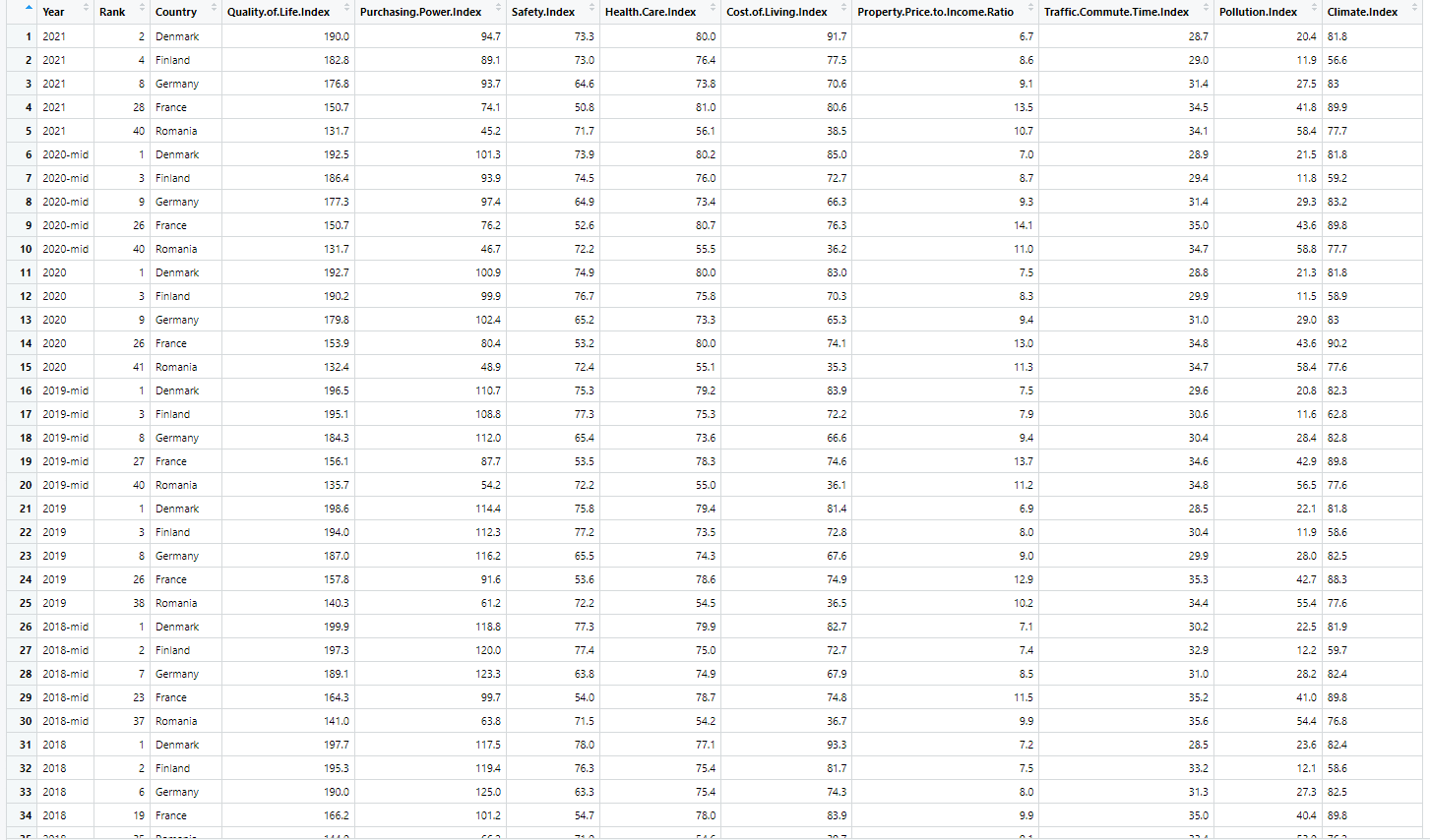


Рисунок 1 – Созданный датасет

1. Был создан график, отображающий изменения рейтинга стран по годам.

Код создания графика:

createPlot=function(number,title){

plot.new()

grid(nx = NULL, ny = NULL,

lty = 1, # Grid line type

col = "gray", # Grid line color

lwd = 1) # Grid line width

par(new=TRUE)

print(c(min(data[,number]),max(data[,number])))

plot(c(1:length(years)),rev(data[,number][data$Country=="Finland"]),ylab=title,xlab="",type='b', lwd=2,col="green",xaxt="n",ylim=c(as.numeric(min(data[,number])),as.numeric(max(data[,number]))),

pch=20,main="Значение выбранного параметра для стран")

lines(c(1:length(years)),rev(data[,number][data$Country=="Denmark"]),col='blue',type='b',pch=20,lwd=2)

lines(c(1:length(years)),rev(data[,number][data$Country=="France"]),col='red',type='b',pch=20,lwd=2)

lines(c(1:length(years)),rev(data[,number][data$Country=="Germany"]),col='purple',type='b',pch=20,lwd=2)

lines(c(1:length(years)),rev(data[,number][data$Country=="Romania"]),col='orange',type='b',pch=20,lwd=2)

#Настроили заголовки

axis.labels = rev(years)

axis(1, at = c(1:length(years)), labels = axis.labels,las=2,cex.axis=0.8)

title(xlab="Year", line=4, cex.lab=1.2)

return("")

}

warnings()

createPlot(2,"Rank")

legend('top',cex=1,title='Страны',c("Finland", "Denmark", "France", "Germany","Romania"),lty=c(1,1,1,1,1),pch=c(20,20,20,20,20),col=c('green','blue','red','purple','orange'))

Результат построения графика представлен на рисунке 2. Согласно данному графику видно,что Румыния стабильно занимала более низкие места в рейтинге по сравнению с остальными странами. Франция относительно стабильно держится в районе 20-30 мест. А Финляндия, Германия и Дания стабильно входят в десятку лучший стран для жизни.

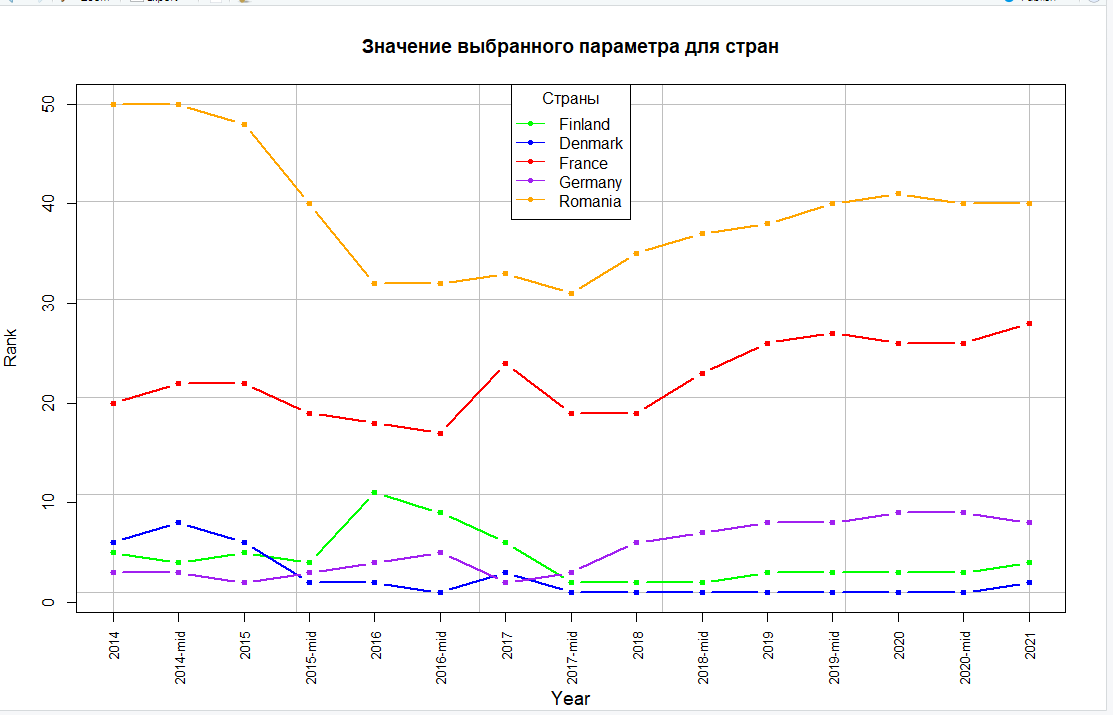


Рисунок 2 – Результат построения графика изменения рейтинга стран по годам

1. Были построены графики, отображающие изменения остальных 10 показателей для данных стран.

Код построения графиков:

#Рисуем график

layout(matrix(1:6,nrow=2,ncol=3))

createPlot(4,"Country Quality.of.Life.Index")

createPlot(5,"Purchasing.Power.Index")

createPlot(6,"Safety.Index")

createPlot(7,"Health.Care.Index")

#Легенда

plot(1:2,1:2,xaxt="n",yaxt="n",main="Легенда",xlab="",ylab="")

legend('topright',cex=2.7,title='Страны',c("Finland", "Denmark", "France", "Germany","Romania"),lty=c(1,1,1,1,1),pch=c(20,20,20,20,20),col=c('green','blue','red','purple','orange'))

layout(matrix(1:6,nrow=2,ncol=3))

createPlot(8,"Cost.of.Living.Index")

createPlot(9,"Property.Price.to.Income.Ratio")

createPlot(10,"Traffic.Commute.Time.Index")

createPlot(11,"Pollution.Index")

createPlot(12,"Climate.Index")

plot(1:2,1:2,xaxt="n",yaxt="n",main="Легенда",xlab="",ylab="")

legend('topright',cex=2,title='Страны',c("Finland", "Denmark", "France", "Germany","Romania"),lty=c(1,1,1,1,1),pch=c(20,20,20,20,20),col=c('green','blue','red','purple','orange'))

Результаты построения графиков представлены на рисунках 3 и 4.

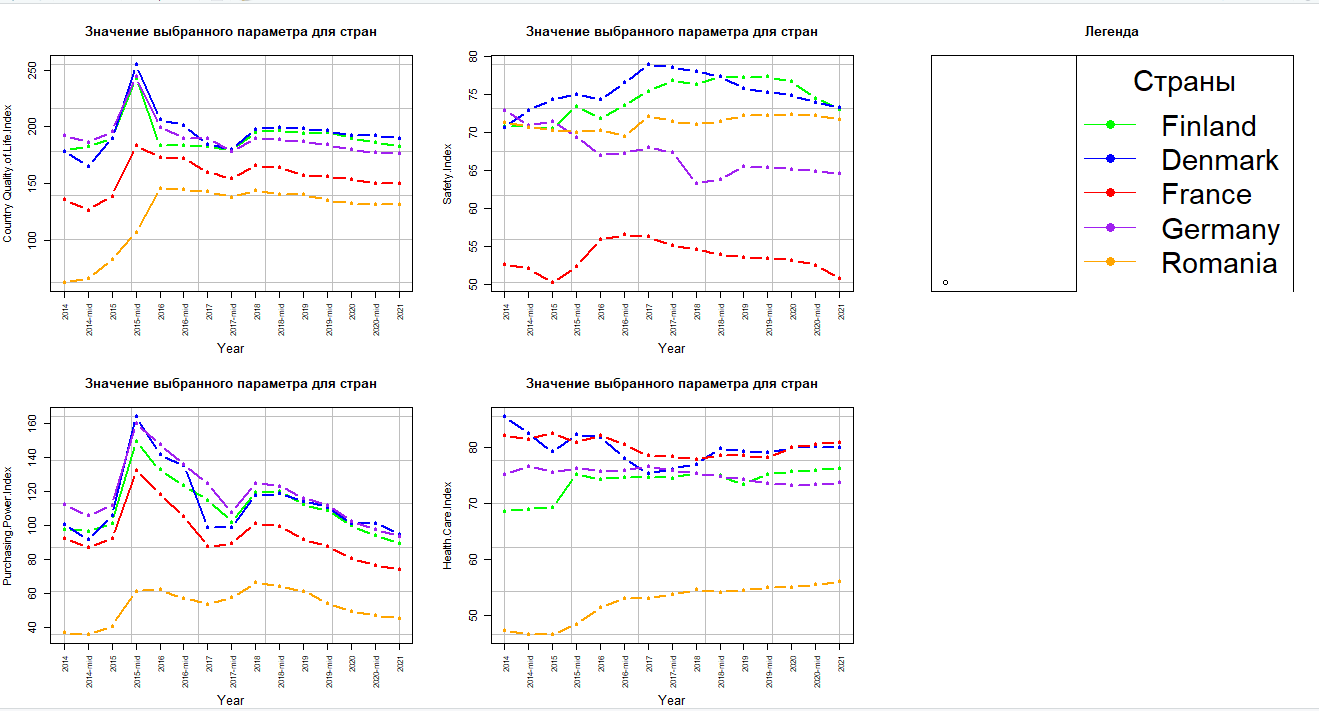


Рисунок 3 – Результат построения для первых 4 показателей

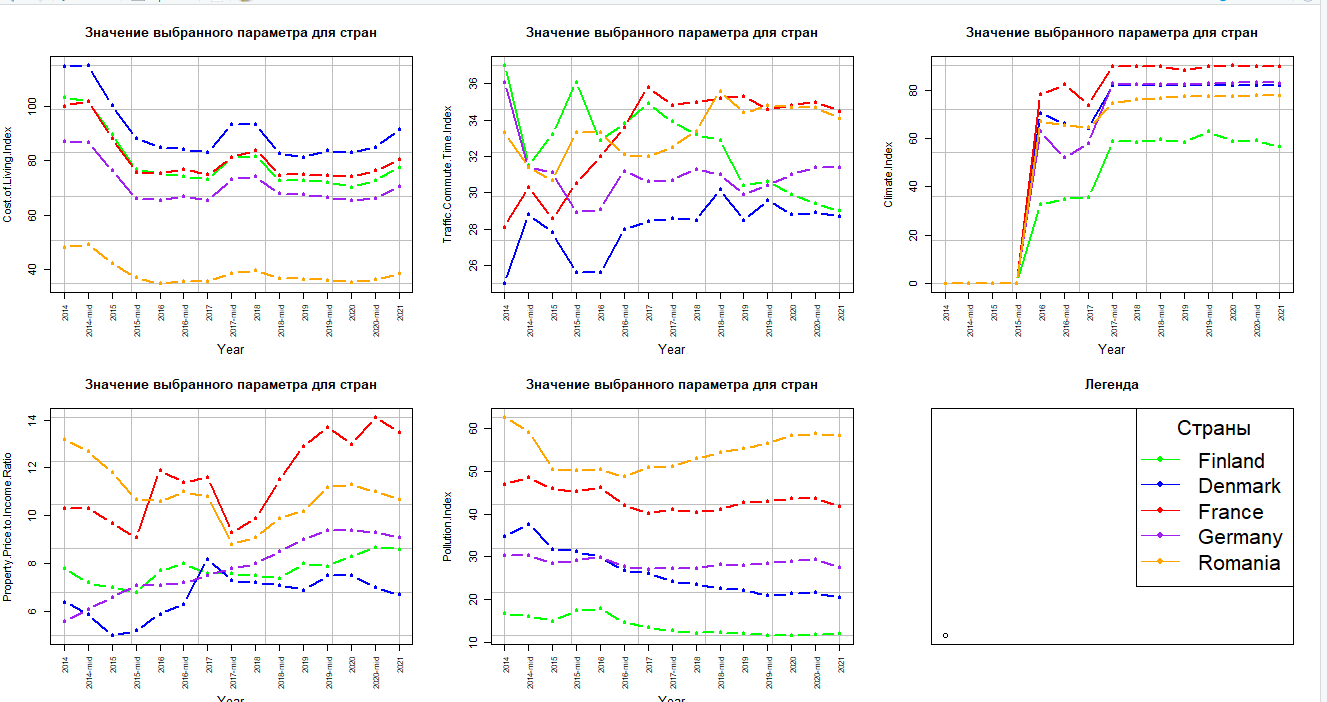


Рисунок 4 – Результат построения для остальных показателей

Согласно 1-ому графику Финляндия, Дания и Германия имеют достаточно высокий индекс уровня жизни, однако его показатели начинают постепенно падать. Аналогичная ситуация с уменьшением показателей наблюдается у Румынии и Франции.

Согласно 2-ому графику Финляндия и Дания имеют самые высокие индексы зарплат. У Германии данный индекс стабильно снижается. У Румынии индекс повышается. У Франции стабильно находится внизу.

Согласно 3-ьему графику покупательская способность в каждой из стран с каждым годом снижается у всех стран. Его пик приходится на 2015-2016 года.

Согласно 4-ому графику индекс здравоохранения у всех, кроме Румынии, колеблется примерно в одном положении. У Румынии он стабильно растет.

Согласно 5-ому графику индекс стоимости жизни у всех стран сначала стабильно снижался, затем в 2018 наблюдался не большой подъем, затем снова снижение и повышение с 2020 года.

Согласно 6-ому графику индекс времени в пути у всех стран с переменным успехом растет. Исключение – Финляндия, у нее он падает с 2018 года.

Согласно 7-ому графику индекс климата растет постепенно у всех стран, кроме Финляндии, в которой он постепенно падает.

Согласно 8-ому графику соотношение цены на недвижимость к доходу во всех стран периодически резко падает и растет.

Согласно 9-ому графику индекс загрязнений постепенно падает во Франции, Дании и Финляндии. В Румынии и Франции он растет.

1. С одной из страниц была собрана информацию в data.frame, которая содержала бы: Название музея, его адрес и ссылку для перехода при клике на фото / ссылке на музей.

Код создания фрейма:

url = "https://ru.m.wikipedia.org/ wiki/ %D0%A1% D0%BF%D0%B 8%D1%81 %D0%BE% D0%BA\_%D0%BC%D1%83%D0%B7%D0%B5%D0%B5%D0%B2\_%D0%A0%D0% BE%D1%81%D1%82%D0%BE%D0%B2%D1%81%D0%BA%D0%B E%D0 %B9\_%D 0%BE%D 0%B1%D0%BB%D0%B0%D1%81%D1%82%D0%B8"

data = read\_html(url) %>% html\_nodes("div section table") %>% html\_table() %>% as.data.frame()

data = data[c(-1,-5,-4)]

dim(data)

data

#Ссылки на музеи

museum = read\_html(url) %>% html\_nodes("div>section>table>tbody>tr>td")

museum=museum[seq(2,length(museum),by=5)]

length(museum)

ex=html\_nodes(museum,"sup>a")

allNameLinks = html\_nodes(museum,"a")

Links=allNameLinks[!(allNameLinks %in% ex)] %>% html\_attr("href")

Links= Links[!duplicated(Links)]

Links = data.frame(Links=paste(c("https://ru.m.wikipedia.org/"),Links,sep=""))

print(Links)

#Ссылки на картинки

getPictLink = function(x){

if(length(html\_nodes(x,"span>a"))>0){

link = (html\_nodes(x,"span>a") %>% html\_attr("href"))

return(paste(c("https://ru.m.wikipedia.org/"),link,sep=""))

}

else{

return(NA)

}

}

pictures = read\_html(url) %>% html\_nodes("div>section>table>tbody>tr>td")

pictures= pictures[seq(4,length(pictures),by=5)]

pictures= sapply(pictures,getPictLink)

pictures = data.frame(PictureLink=pictures)

print(pictures)

result = cbind(data,pictures) %>% cbind(Links)

result

Результат создания датафрейма представлен на рисунке 5.

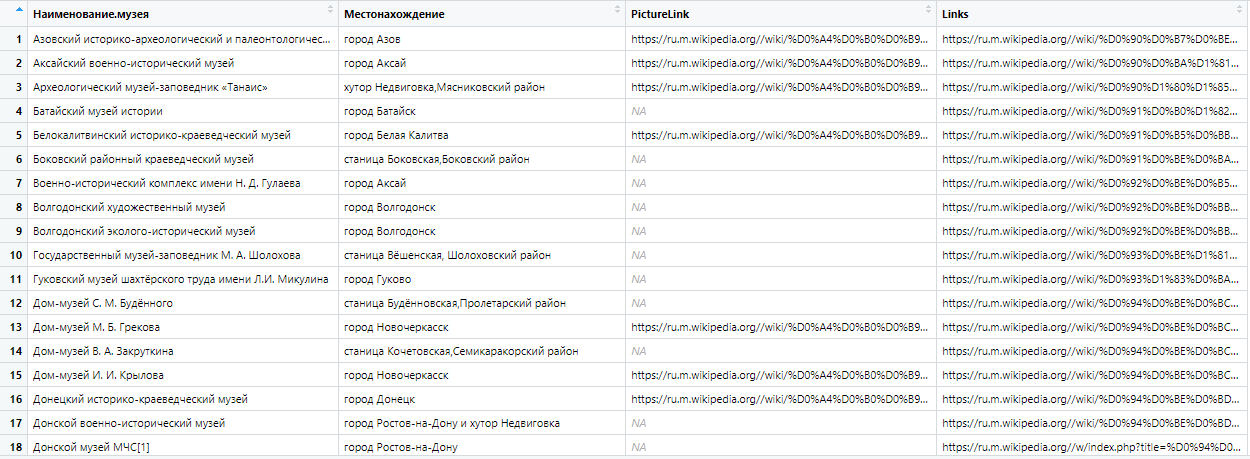


Рисунок 5 – Результат создания датафрейма.

**Вывод:** Была освоена работа c технологией webscrapping с помощью инструментов языка R.