Лабораторная работа №7

Введение в работу с данными

Коняева Марина Александровна
НФИбд-01-21
Студ. билет: 1032217044

2024

RUDN

Цель работы

Освоить специализированные пакеты Julia для обработки данных.

Теоретическое введение

Обработка и анализ данных, полученных в результате проведения исследований, — важная и неотъемлемая часть исследовательской деятельности. Большое значение имеет выявление определённых связей и закономерностей в имеющихся неструктурированных данных, особенно в данных больших размерностей. Выявленные в данных связей и закономерностей позволяет строить прогнозные модели с предполагаемым результатом. Для решения таких задач применяют методы из таких областей знаний как математическая статистика, программирование, искусственный интеллект, машинное обучение.

Задачи лабораторной работы

1. Используя Jupyter Lab, повторите примеры из раздела 7.2.
2. Выполните задания для самостоятельной работы (раздел 7.4).

Выполнение лабораторной работы

Julia для науки о данных

В Julia для обработки данных используются наработки из других языков программирования, в частности, из R и Python.

Считывание данных

Установим нужные пакеты, загрузим в систему необходимые файлы с данными, а затем считаем данные из файла и запишем их в структуру:

```
# Обновление окружения:
using Pkg
Pkg.update
# Установка пакетов:
using Pkg
for p in ["CSV", "DataFrames", "RDatasets", "FileIO"]
Pkg.add(p)
end
using CSV, DataFrames, DelimitedFiles
```

```
   Resolving package versions...
  No Changes to `C:\Users\Admin\.julia\environments\v1.10\Project.toml`
  No Changes to `C:\Users\Admin\.julia\environments\v1.10\Manifest.toml`
   Resolving package versions...
  No Changes to `C:\Users\Admin\.julia\environments\v1.10\Project.toml`
  No Changes to `C:\Users\Admin\.julia\environments\v1.10\Manifest.toml`
   Resolving package versions...
  No Changes to `C:\Users\Admin\.julia\environments\v1.10\Project.toml`
  No Changes to `C:\Users\Admin\.julia\environments\v1.10\Manifest.toml`
   Resolving package versions...
  No Changes to `C:\Users\Admin\.julia\environments\v1.10\Project.toml`
  No Changes to `C:\Users\Admin\.julia\environments\v1.10\Manifest.toml`
```

```
# Функция определения по названию языка программирования года его создания:
function language_created_year(P,language::String)
loc = findfirst(P[:,2].==language)
return P[loc,1]
end
```

```
language_created_year (generic function with 1 method)
```

```
# Считывание данных и их запись в структуру:
P = CSV.File("programminglanguages.csv") |> DataFrame
```

73×2 DataFrame

Row	year	language
	Int64	String31
1	1951	Regional Assembly Language
2	1952	Autocode
3	1954	IPL
4	1955	FLOW-MATIC
5	1957	FORTRAN
6	1957	COMTRAN
7	1958	LISP
8	1958	ALGOL 58
9	1959	FACT
10	1959	COBOL
11	1959	RPG
12	1962	APL
13	1962	Simula
⋮	⋮	⋮

Считывание данных

Напишем функцию для определения по названию языка программирования года его создания и протестируем её. Затем построчно считаем данные с указанием разделителя и запишем данные в CSV-файл:

```
]: # Функция определения по названию языка программирования
   # года его создания (без учёта регистра):
   function language_created_year_v2(P,language::String)
   loc = findfirst(lowercase.(P[:,2]).==lowercase.(language))
   return P[loc,1]
   end
   # Пример вызова функции и определение даты создания языка julia:
   language_created_year_v2(P,"julia")
```

```
]: 2012
```

```
]: # Построчное считывание данных с указанием разделителя:
   Tx = readdlm("programminglanguages.csv", ',')
```

```
]: 74×2 Matrix{Any}:
     "year"  "language"
 1951        "Regional Assembly Language"
 1952        "Autocode"
 1954        "IPL"
 1955        "FLOW-MATIC"
 1957        "FORTRAN"
 1957        "COMTRAN"
 1958        "LISP"
 1958        "ALGOL 58"
 1959        "FACT"
 1959        "COBOL"
 1959        "RPG"
 1962        "APL"
    ⋮
 2003        "Scala"
 2005        "F#"
 2006        "PowerShell"
 2007        "Clojure"
 2009        "Go"
 2010        "Rust"
 2011        "Dart"
 2011        "Kotlin"
 2011        "Red"
 2011        "Elixir"
 2012        "Julia"
```

Словари

Я инициализировала словарь и заполнила его данными. При инициализации словаря можно задать конкретные типы данных для ключей и значений:

```
# Инициализация словаря:
dict = Dict{Integer,Vector{String}}()
```

```
Dict{Integer, Vector{String}}()
```

```
# Заполнение словаря данными:
for i = 1:size(P,1)
    year,lang = P[i,:]
    if year in keys(dict)
        dict[year] = push!(dict[year],lang)
    else
        dict[year] = [lang]
    end
end
```

```
# Пример определения в словаре языков программирования, созданных в 2003 году:
dict[2003]
```

```
2-element Vector{String}:
 "Groovy"
 "Scala"
```

DataFrames

Работа с данными, записанными в структуре DataFrame, позволяет использовать индексацию и получить доступ к столбцам по заданному имени заголовка или по индексу столбца.

Подгрузила необходимый пакет DataFrames, задала переменную со структурой DataFrame и вывела ее:

```
# Подгружаем пакет DataFrames:
using DataFrames
# Задаём переменную со структурой DataFrame:
df = DataFrame(year = P[:,1], language = P[:,2])
```

73×2 DataFrame

Row	year	language
	Int64	String31
1	1951	Regional Assembly Language
2	1952	Autocode
3	1954	IPL
4	1955	FLOW-MATIC
5	1957	FORTRAN
6	1957	COMTRAN
7	1958	LISP
8	1958	ALGOL 58
9	1959	FACT
10	1959	COBOL
11	1959	RPG
12	1962	APL
13	1962	Simula
⋮	⋮	⋮
62	2003	Scala
63	2005	F#
64	2006	PowerShell
65	2007	Clojure
66	2009	Go
67	2010	Rust
68	2011	Dart
69	2011	Kotlin
70	2011	Red
71	2011	Elixir

RDatasets

С данными можно работать также как с наборами данных через пакет RDatasets языка R. Подгрузила пакет RDatasets и задала структуру данных в виде набора данных:

```
# Получение статистических сведений о фрейме:
describe(df)
```

2×7 DataFrame

Row	variable	mean	min	median	max	nmissing	eltype
	Symbol	Union...	Any	Union...	Any	Int64	DataType
1	year	1982.99	1951	1986.0	2014	0	Int64
2	language		ALGOL 58		dBase III	0	String31

```
# Подгружаем пакет RDatasets:
using RDatasets
# Задаём структуру данных в виде набора данных:
iris = dataset("datasets", "iris")
```

150×5 DataFrame

Row	SepalLength	SepalWidth	PetalLength	PetalWidth	Species
	Float64	Float64	Float64	Float64	Cat...
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa

Работа с переменными отсутствующего типа (Missing Values)

Пакет DataFrames позволяет использовать так называемый «отсутствующий» тип.

Выполняем действия с переменными отсутствующего типа:

```
# Определения типа переменной:
typeof(iris)
```

```
DataFrame
```

```
describe(iris)
```

5×7 DataFrame

Row	variable	mean	min	median	max	nmissing	eltype
	Symbol	Union...	Any	Union...	Any	Int64	DataType
1	SepalLength	5.84333	4.3	5.8	7.9	0	Float64
2	SepalWidth	3.05733	2.0	3.0	4.4	0	Float64
3	PetalLength	3.758	1.0	4.35	6.9	0	Float64
4	PetalWidth	1.19933	0.1	1.3	2.5	0	Float64
5	Species		setosa		virginica	0	CategoricalValue{String, UInt8}

```
# Отсутствующий тип:
a = missing
typeof(a)
```

```
Missing
```

```
# Пример операции с переменной отсутствующего типа:
a + 1
```

```
missing
```

```
# Определение перечня продуктов:
foods = ["apple", "cucumber", "tomato", "banana"]
# Определение калорий:
calories = [missing,47,22,105]
```

Работа с переменными отсутствующего типа (Missing Values)

```
# Подключаем пакет Statistics:
using Statistics
```

```
# Определение среднего значения:
mean(calories)
```

```
missing
```

```
# Определение среднего значения без значений с отсутствующим типом:
mean(skipmissing(calories))
```

```
58.0
```

```
# Задание сведений о ценах:
prices = [0.85,1.6,0.8,0.6]
# Формирование данных о калориях:
dataframe_calories = DataFrame(item=foods,calories=calories)
# Формирование данных о ценах:
dataframe_prices = DataFrame(item=foods,price=prices)
# Объединение данных о калориях и ценах:
DF = outerjoin(dataframe_calories,dataframe_prices,on=:item)
```

4×3 DataFrame

Row	item	calories	price
	String	Int64?	Float64?
1	apple	*missing*	0.85
2	cucumber	47	1.6
3	tomato	22	0.8
4	banana	105	0.6

Кластеризация данных. Метод k-средних

Задача кластеризации данных заключается в формировании однородной группы упорядоченных по какому-то признаку данных. Метод k-средних позволяет минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров.

Рассмотрим задачу кластеризации данных на примере данных о недвижимости. Файл с данными houses.csv содержит список транзакций с недвижимостью в районе Сакраменто, о которых было сообщено в течение определённого числа дней.

Кластеризация данных. Метод k-средних

Загружаю пакеты, а также данные из файла и строю графики, а затем произвожу кластеризацию данных:

```
# Загрузка пакетов:
import Pkg
Pkg.add("DataFrames")
Pkg.add("Statistics")
using DataFrames
using CSV
import Pkg
Pkg.add("Plots")
```

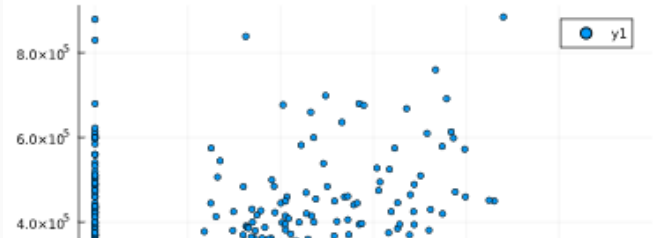
```
  Resolving package versions...
 No Changes to `C:\Users\Admin\.julia\environments\v1.10\Project.toml`
 No Changes to `C:\Users\Admin\.julia\environments\v1.10\Manifest.toml`
  Resolving package versions...
 No Changes to `C:\Users\Admin\.julia\environments\v1.10\Project.toml`
 No Changes to `C:\Users\Admin\.julia\environments\v1.10\Manifest.toml`
  Resolving package versions...
 No Changes to `C:\Users\Admin\.julia\environments\v1.10\Project.toml`
 No Changes to `C:\Users\Admin\.julia\environments\v1.10\Manifest.toml`
```

```
# Загрузка данных:
houses = CSV.File("houses.csv") |> DataFrame
```

985×12 DataFrame — *960 rows omitted*

Row	street	city	zip	state	beds	baths	sq__ft	type	sale_date	price	latitude	longitude
	String	String15	Int64	String3	Int64	Int64	Int64	String15	String31	Int64	Float64	Float64
1	3526 HIGH ST	SACRAMENTO	95838	CA	2	1	836	Residential	Wed May 21 00:00:00 EDT 2008	59222	38.6319	-121.435
2	51 OMAHA CT	SACRAMENTO	95823	CA	3	1	1167	Residential	Wed May 21 00:00:00 EDT 2008	68212	38.4789	-121.431
3	2796 BRANCH ST	SACRAMENTO	95815	CA	2	1	796	Residential	Wed May 21 00:00:00 EDT 2008	68880	38.6183	-121.444
4	2805 JANETTE WAY	SACRAMENTO	95815	CA	2	1	852	Residential	Wed May 21 00:00:00 EDT 2008	69307	38.6168	-121.439
5	6001 MCMAHON DR	SACRAMENTO	95824	CA	2	1	797	Residential	Wed May 21 00:00:00 EDT	81900	38.5195	-121.436

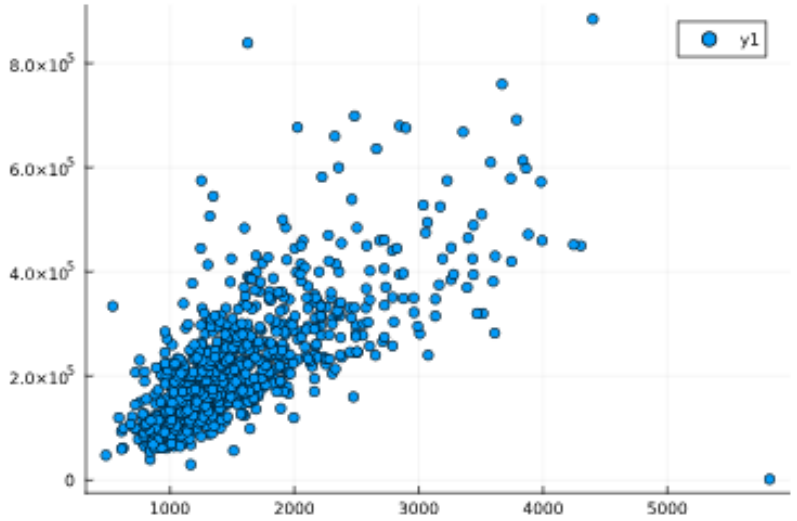
```
# Построение графика:
using Plots
plot(size=(500,500),leg=false)
x = houses[!,:sq__ft]
y = houses[!,:price]
scatter(x,y,markersize=3)
```



Кластеризация данных. Метод k-средних

Используя для фильтрации значений функцию by пакета DataFrames и для вычисления среднего значения функцию mean пакета Statistics, можно посмотреть среднюю цену домов определённого типа:

```
# Фильтрация данных по заданному условию:
filter_houses = houses[houses[!,:sq__ft].>0,:]
# Построение графика:
x = filter_houses[!,:sq__ft]
y = filter_houses[!,:price]
scatter(x,y)
```




```
using DataFrames
using CSV
import Pkg
Pkg.add("Plots")
```

```
  Resolving package versions...
 No Changes to `C:\Users\Admin\.julia\environments\v1.10\Project.toml`
 No Changes to `C:\Users\Admin\.julia\environments\v1.10\Manifest.toml`
```

```
using Statistics
# Определение средней цены для определённого типа домов:
combine(groupby(filter_houses,:type),filter_houses->mean(filter_houses[!,:price]))
```

3×2 DataFrame

Кластеризация данных. Метод k-средних

Отфильтровав таким образом данные, можно приступить к формированию кластеров. Сначала подключаем необходимые пакеты и формируем данные в нужном виде:

```
import Pkg
Pkg.add("Clustering")
using Clustering
```

```
  Resolving package versions...
  No Changes to `C:\Users\Admin\.julia\environments\v1.10\Project.toml`
  No Changes to `C:\Users\Admin\.julia\environments\v1.10\Manifest.toml`
```

```
# Добавление данных :latitude и :longitude в новый фрейм:
X = filter_houses[!,[:latitude,:longitude]]
```

814×2 DataFrame

789 ro

Row	latitude	longitude
	Float64	Float64
1	38.6319	-121.435
2	38.4789	-121.431
3	38.6183	-121.444
4	38.6168	-121.439
5	38.5195	-121.436
6	38.6626	-121.328
7	38.6817	-121.352
8	38.5351	-121.481
9	38.6212	-121.271
10	38.7009	-121.443
11	38.6377	-121.452
12	38.4707	-121.459
13	38.6187	-121.436
⋮	⋮	⋮
803	38.7035	-121.375
804	38.7031	-121.235
805	38.3898	-121.446
806	38.8978	-121.325
807	38.4679	-121.445
808	38.4453	-121.442
809	38.4174	-121.484
810	38.4577	-121.36
811	38.4999	-121.459
812	38.7088	-121.257
813	38.417	-121.397
814	38.6552	-121.076

Кластеризация данных. Метод k-средних

Каждая функция хранится в виде строки X, но можно транспонировать получившуюся матрицу, чтобы иметь возможность работать с столбцами данных X:

```
# Транспонирование матрицы с данными:
X = X'
```

```
2×814 adjoint(::Matrix{Float64}) with eltype Float64:
   38.6319    38.4789    38.6183  …    38.7088    38.417    38.6552
 -121.435   -121.431   -121.444      -121.257   -121.397  -121.076
```

```
# Задание количества кластеров:
k = length(unique(filter_houses[!,:zip]))
```

```
66
```

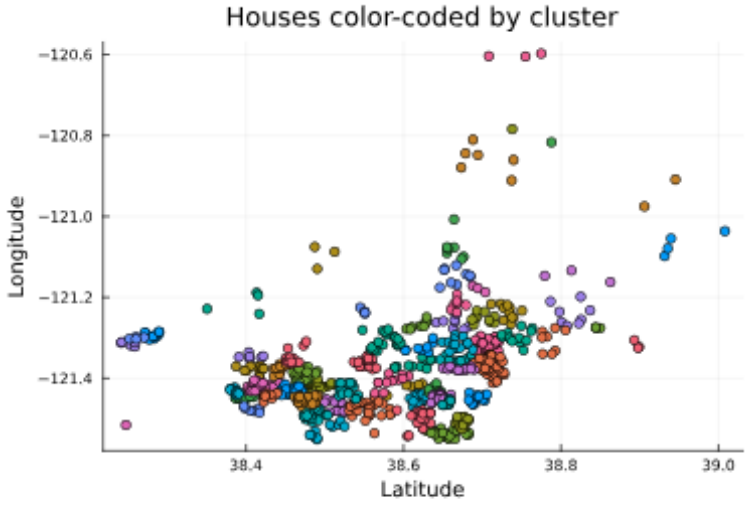
```
# Определение k-среднего:
C = kmeans(X,k)
```

```
KmeansResult{Matrix{Float64}, Float64, Int64}([38.28195836363636 38.7155373076923 … 38.69318723076923 38.786281875; -121.29388763636362 -121.39096746153845 … -121.45054476923073 -121.306017], [51, 26, 11, 11, 38, 59, 6, 18, 64, 65  …  19, 14, 56, 33, 45, 46, 4, 42, 53, 3], [8.741454075789079e-5, 0.0004009756121376995, 2.1663516236003488e-5, 9.154340659733862e-5, 0.00011844087202916853, 0.00017304826178587973, 0.00011862715109600686, 0.00030225235241232244, 9.256233170162886e-6, 0.00011686658399412408  …  4.192032793071121e-5, 5.2413251978578046e-5, 0.00011806063048425 13, 0.00023417692500515 84, 0.0001341583774774335 3, 0.00014433951582759 62, 0.00016121321459650062, 9.205109017784707e-5, 0.00037928301389911 21, 7.061329961288 72e-5], [11, 13, 7, 14, 11, 19, 3, 2, 11, 1  …  21, 20, 17, 5, 4, 22, 15, 8, 13, 8], [11, 13, 7, 14, 11, 19, 3, 2, 11, 1  …  21, 20, 17, 5, 4, 22, 15, 8, 13, 8], 0.2177436743368161618, 11, true)
```

```
# Формирование фрейма данных:
df = DataFrame(cluster = C.assignments,city = filter_houses[!,:city],
latitude = filter_houses[!,:latitude],longitude = filter_houses[!,:longitude],zip = filter_houses[!,:zip])
clusters_figure = plot(legend = false)
for i = 1:k
clustered_houses = df[df[!,:cluster].== i,:]
xvals = clustered_houses[!,:latitude]
yvals = clustered_houses[!,:longitude]
scatter!(clusters_figure,xvals,yvals,markersize=4)
end
xlabel!("Latitude")
ylabel!("Longitude")
title!("Houses color-coded by cluster")
```

Кластеризация данных. Метод k-средних

```
display(clusters_figure)
unique_zips = unique(filter_houses[!,:zip])
zips_figure = plot(legend = false)
for uzip in unique_zips
subs = filter_houses[filter_houses[!,:zip].==uzip,:]
x = subs[!,:latitude]
y = subs[!,:longitude]
scatter!(zips_figure,x,y)
end
xlabel!("Latitude")
ylabel!("Longitude")
title!("Houses color-coded by zip code")
display(zips_figure)
```



Кластеризация данных. Метод k ближайших соседей

Данный метод заключается в отнесении объекта к тому из известных классов, который является наиболее распространённым среди □ соседей данного элемента. В случае использования метода для регрессии, объекту присваивается среднее значение по □ ближайшим к нему объектам.

Решаю задачу методом k ближайших соседей и строю график:

```
# Подключение пакета NearestNeighbors:
import Pkg
Pkg.add("NearestNeighbors")
```

```
   Resolving package versions...
    Updating `C:\Users\Admin\.julia\environments\v1.10\Project.toml`
  [b8a86587] + NearestNeighbors v0.4.21
  No Changes to `C:\Users\Admin\.julia\environments\v1.10\Manifest.toml`
```

```
using NearestNeighbors
knearest = 10
id = 70
point = X[:,id]
```

```
2-element Vector{Float64}:
   38.44004
 -121.421012
```

```
# Поиск ближайших соседей:
kdtree = KDTree(X)
idxs, dists = knn(kdtree, point, knearest, true)
```

```
([70, 764, 196, 125, 557, 368, 415, 92, 112, 683], [0.0, 0.006264891539364138, 0.008253202590508462, 0.008473585132630057, 0.009164073548
5065124697706, 0.009921759722950759, 0.009941028618812013, 0.010332637707777167, 0.011168993911721985])
```

```
# Все объекты недвижимости:
x = filter_houses[!,:latitude];
y = filter_houses[!,:longitude];
scatter(x,y)
# Соседи:
x = filter_houses[idxs,:latitude];
y = filter_houses[idxs,:longitude];
scatter!(x,y)
```

Обработка данных. Метод главных компонент

Метод главных компонент (Principal Components Analysis, PCA) позволяет уменьшить размерность данных, потеряв наименьшее количество полезной информации. Метод имеет широкое применение в различных областях знаний, например, при визуализации данных, компрессии изображений, в эконометрике, некоторых гуманитарных предметных областях, например, в социологии или в политологии.

На примере с данными о недвижимости попробуем уменьшить размеры данных о цене и площади из набора данных домов.

Подключаю необходимые пакеты и строю график с выделением главных компонент:

```
# Фрейм с указанием площади и цены недвижимости:
F = filter_houses[!,[:sq__ft,:price]]
# Конвертация данных в массив:
F = Array(F)'
```

```
2×814 adjoint(::Matrix{Int64}) with eltype Int64:
   836   1167    796    852    797   1122  …    1477    1216    1685    1362
 59222  68212  68880  69307  81900  89921     234000  235000  235301  235738
```

```
# Подключение пакета MultivariateStats:
import Pkg
Pkg.add("MultivariateStats")
```

```
   Resolving package versions...
  No Changes to `C:\Users\Admin\.julia\environments\v1.10\Project.toml`
  No Changes to `C:\Users\Admin\.julia\environments\v1.10\Manifest.toml`
```

```
using MultivariateStats
# Приведение типов данных к распределению для PCA:
M = fit(PCA, F)
```

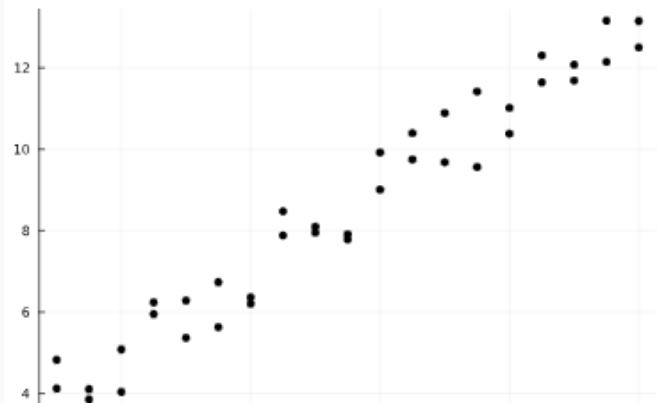
```
PCA(indim = 2, outdim = 1, principalratio = 0.9999840784692097)

Pattern matrix (unstandardized loadings):
──────────────
          PC1
──────────────
1  460.52
2    1.19826e5
──────────────
```

Обработка данных. Линейная регрессия

Регрессионный анализ представляет собой набор статистических методов исследования влияния одной или нескольких независимых переменных (регрессоров) на зависимую (критериальная) переменную. Терминология зависимых и независимых переменных отражает лишь математическую зависимость переменных, а не причинноследственные отношения. Наиболее распространённый вид регрессионного анализа — линейная регрессия, когда находят линейную функцию, которая согласно определённым математическим критериям наиболее соответствует данным.

Применяю функцию линейной регрессии для построения соответствующего графика::

```
xvals = repeat(1:0.5:10,inner=2)
yvals = 3 .+ xvals + 2*rand(length(xvals)) .- 1
scatter(xvals,yvals,color=:black,leg=false)
```



Обработка данных. Линейная регрессия

```
xvals = 1:100000;
xvals = repeat(xvals,inner=3);
yvals = 3 .+ xvals + 2*rand(length(xvals)) .- 1;
@show size(xvals)
@show size(yvals)
```

```
size(xvals) = (300000,)
size(yvals) = (300000,)
(300000,)
```

```
@time a,b = find_best_fit(xvals,yvals)
```

```
  0.066963 seconds (20.15 k allocations: 1.315 MiB, 95.55% compilation time)
(0.999999986637046, 3.0007307429841603)
```

Рис. 16: Функция линейной регрессии

Задания для самостоятельного выполнения

1. Кластеризация:

Загрузите using RDatasets iris = dataset(“datasets”, “iris”) Используйте Clustering.jl для кластеризации на основе k-средних. Сделайте точечную диаграмму полученных кластеров. Подсказка: вам нужно будет проиндексировать фрейм данных, преобразовать его в массив и транспонировать

```
using RDatasets
iris = dataset("datasets", "iris")
```

150×5 DataFrame

Row	SepalLength	SepalWidth	PetalLength	PetalWidth	Species
	Float64	Float64	Float64	Float64	Cat...
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
⋮	⋮	⋮	⋮	⋮	⋮
139	6.0	3.0	4.8	1.8	virginica

Задания для самостоятельного выполнения

```
using Clustering
using DataFrames
X = iris[!,[:SepalLength, :PetalWidth]]
```

150×2 DataFrame

Row	SepalLength	PetalWidth
	Float64	Float64
1	5.1	0.2
2	4.9	0.2
3	4.7	0.2
4	4.6	0.2
5	5.0	0.2
6	5.4	0.4
7	4.6	0.3
8	5.0	0.2
9	4.4	0.2
10	4.9	0.1
11	5.4	0.2
12	4.8	0.2
13	4.8	0.1
⋮	⋮	⋮
139	6.0	1.8
140	6.9	2.1
141	6.7	2.4
142	6.9	2.3
143	5.8	1.9
144	6.8	2.3
145	6.7	2.5
146	6.7	2.3
147	6.3	1.9
148	6.5	2.0
149	6.2	2.3
150	5.9	1.8

Задания для самостоятельного выполнения

```
X = Matrix(X)
```

```
150×2 Matrix{Float64}:
 5.1  0.2
 4.9  0.2
 4.7  0.2
 4.6  0.2
 5.0  0.2
 5.4  0.4
 4.6  0.3
 5.0  0.2
 4.4  0.2
 4.9  0.1
 5.4  0.2
 4.8  0.2
 4.8  0.1
  ⋮
 6.0  1.8
 6.9  2.1
 6.7  2.4
 6.9  2.3
 5.8  1.9
 6.8  2.3
 6.7  2.5
 6.7  2.3
 6.3  1.9
 6.5  2.0
 6.2  2.3
 5.9  1.8
```

```
X = X'
```

```
2×150 adjoint(::Matrix{Float64}) with eltype Float64:
 5.1  4.9  4.7  4.6  5.0  5.4  4.6  5.0  …  6.8  6.7  6.7  6.3  6.5  6.2  5.9
 0.2  0.2  0.2  0.2  0.2  0.4  0.3  0.2     2.3  2.5  2.3  1.9  2.0  2.3  1.8
```

```
k = length(unique(iris[!,:Species]))
```

```
3
```

```
C = kmeans(X, k)
```

```
KmeansResult{Matrix{Float64}, Float64, Int64}([5.0055555555555555 6.857142857142857 5.892592592592593; 0.3037037037037036 2.0119047619047614 1.4629629629629628], [1, 1, 1, 1, 1, 1, 1, 1, 1, 1  …  2, 2, 3, 2, 2, 2, 2, 2, 2, 3], [0.01967421124828661, 0.02189643347050918, 0.18411865569273004, 0.17522976680384517, 0.010785322359396332, 0.16485939643347088, 0.16448902606310156, 0.010785322359396332, 0.3774519890260635, 0.052637174211248805  …  0.17531179138320852, 0.08483560090702724, 0.19957475994512208, 0.08626417233558925, 0.26293083900225156, 0.10769274376416149, 0.32293083900225383, 0.12769274376417172, 0.5148356089070199, 0.11364883401920167], [54, 42, 54], [54, 42, 54], 32.73746031746019, 4, true)
```

Задания для самостоятельного выполнения

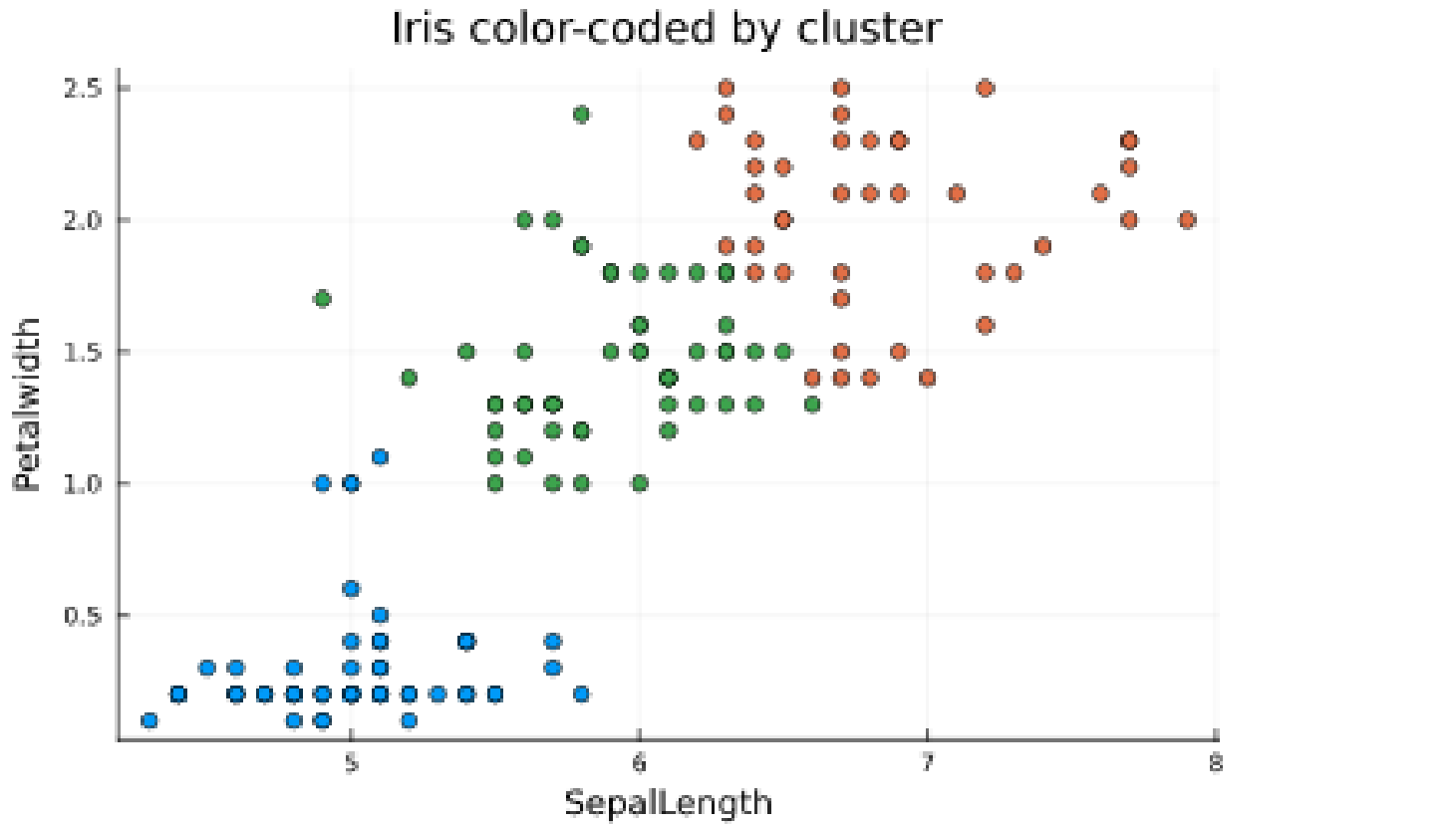
```
df = DataFrame(cluster = C.assignments, SepalLength = iris[!,:SepalLength],
SepalWidth = iris[!,:SepalWidth], Petallength = iris[!,:PetalLength],
Petalwidth = iris[!, :PetalWidth])
```

150×5 DataFrame

Row	cluster	SepalLength	SepalWidth	Petallength	Petalwidth
	Int64	Float64	Float64	Float64	Float64
1	1	5.1	3.5	1.4	0.2
2	1	4.9	3.0	1.4	0.2
3	1	4.7	3.2	1.3	0.2
4	1	4.6	3.1	1.5	0.2
5	1	5.0	3.6	1.4	0.2
6	1	5.4	3.9	1.7	0.4
7	1	4.6	3.4	1.4	0.3
8	1	5.0	3.4	1.5	0.2
9	1	4.4	2.9	1.4	0.2
10	1	4.9	3.1	1.5	0.1
11	1	5.4	3.7	1.5	0.2
12	1	4.8	3.4	1.6	0.2
13	1	4.8	3.0	1.4	0.1
⋮	⋮	⋮	⋮	⋮	⋮
139	3	6.0	3.0	4.8	1.8
140	2	6.9	3.1	5.4	2.1
141	2	6.7	3.1	5.6	2.4
142	2	6.9	3.1	5.1	2.3
143	3	5.8	2.7	5.1	1.9
144	2	6.8	3.2	5.9	2.3
145	2	6.7	3.3	5.7	2.5
146	2	6.7	3.0	5.2	2.3
147	2	6.3	2.5	5.0	1.9
148	2	6.5	3.0	5.2	2.0
149	2	6.2	3.4	5.4	2.3
150	3	5.9	3.0	5.1	1.8

Задания для самостоятельного выполнения

```
using Plots
clusters_figure = plot(legend = false)
for i = 1:k
clustered_iris = df[df[!,:cluster].== i,:]
xvals = clustered_iris[!,:SepalLength]
yvals = clustered_iris[!,:Petalwidth]
scatter!(clusters_figure, xvals, yvals, markersize=4)
end
xlabel!("SepalLength")
ylabel!("Petalwidth")
title!("Iris color-coded by cluster")
display(clusters_figure)
```



Задания для самостоятельного выполнения

2. Регрессия (метод наименьших квадратов в случае линейной регрессии):

Часть 1 X = randn(1000, 3) a0 = rand(3) y = X * a0 + 0.1 * randn(1000);

Часть 2 X = rand(100); y = 2X + 0.1 * randn(100);

Часть 1: Пусть регрессионная зависимость является линейной. Матрица наблюдений факторов □ имеет размерность □ × 3 randn (N, 3), массив результатов □ × 1, регрессионная зависимость является линейной. Найдите МНК-оценку для линейной модели. – Сравните свои результаты с результатами использования llsq из MultivariateStats.jl (просмотрите документацию). – Сравните свои результаты с результатами использования регулярной регрессии наименьших квадратов из GLM.jl. Подсказка. Создайте матрицу данных X2, которая добавляет столбец единиц в начало матрицы данных, и решите систему линейных уравнений. Объясните с помощью теоретических выкладок.

Задания для самостоятельного выполнения

```
#Часть1
X = randn(1000, 3)
a0 = rand(3)
y = X * a0 + 0.1 * randn(1000);

x = fill(1, 1000)
X2 = [x X]
```

```
1000×4 Matrix{Float64}:
 1.0   1.28112    -0.0316204  -0.947284
 1.0  -0.110067   -0.119291    0.673453
 1.0   0.688834   -1.19778    -0.374792
 1.0  -1.73119    -1.08714    -1.32602
 1.0   0.745369   -0.530478   -1.34856
 1.0  -0.916241   -2.58251    -0.626703
 1.0  -0.0394084  -1.69806     0.430476
 1.0  -0.930756   -0.0455237  -0.58211
 1.0  -0.768932    1.64782     2.20056
 1.0   0.488797   -0.631105   -0.315235
 1.0  -0.582776    1.81536     0.795926
 1.0  -1.04076     0.911972    0.399517
 1.0   0.0957037   0.641517    1.01673
 ⋮
 1.0  -1.06476    -1.36288     0.58173
 1.0  -0.776634   -0.743936   -1.0969
 1.0   2.15173     0.820385   -0.752216
 1.0   0.39244    -1.49542    -0.289433
 1.0   0.64184     1.50497    -0.0325943
 1.0   0.465325    1.65261     0.249237
 1.0  -0.545082   -1.5491      1.13889
 1.0   1.93885     0.0827646   1.70034
 1.0  -0.143606   -0.802992    0.653906
 1.0   0.23558     0.483529    1.31503
 1.0  -1.32577     0.245828   -1.09478
 1.0   1.2651     -0.556048   -2.96912
```

```
X2 = X2\y
```

```
4-element Vector{Float64}:
 -0.0023792586158943584
  0.7058850961270203
  0.6975917014202073
  0.01719257407532062
```

```
using MultivariateStats
llsq(X,y)
```

```
4-element Vector{Float64}:
  0.7058850961270204
  0.6975917014202078
  0.01719257407532062
```

Задания для самостоятельного выполнения

```
import Pkg
Pkg.add("GLM")
using GLM
DF = DataFrame(y=y, x1=X[:,1], x2=X[:,2], x3=X[:,3])
lm(@formula(y ~ x1 + x2 + x3), DF)
```

```
   Resolving package versions...
   Installed GLM ──────────── v1.9.0
   Installed ShiftedArrays ─ v2.0.0
   Installed StatsModels ──── v0.7.4
    Updating `C:\Users\Admin\.julia\environments\v1.10\Project.toml`
  [38e38edf] + GLM v1.9.0
    Updating `C:\Users\Admin\.julia\environments\v1.10\Manifest.toml`
  [38e38edf] + GLM v1.9.0
  [1277b4bf] + ShiftedArrays v2.0.0
  [3eaba693] + StatsModels v0.7.4
Precompiling project...
  ✓ ShiftedArrays
  ✓ StatsModels
  ✓ GLM
  3 dependencies successfully precompiled in 17 seconds. 569 already precompiled.
```

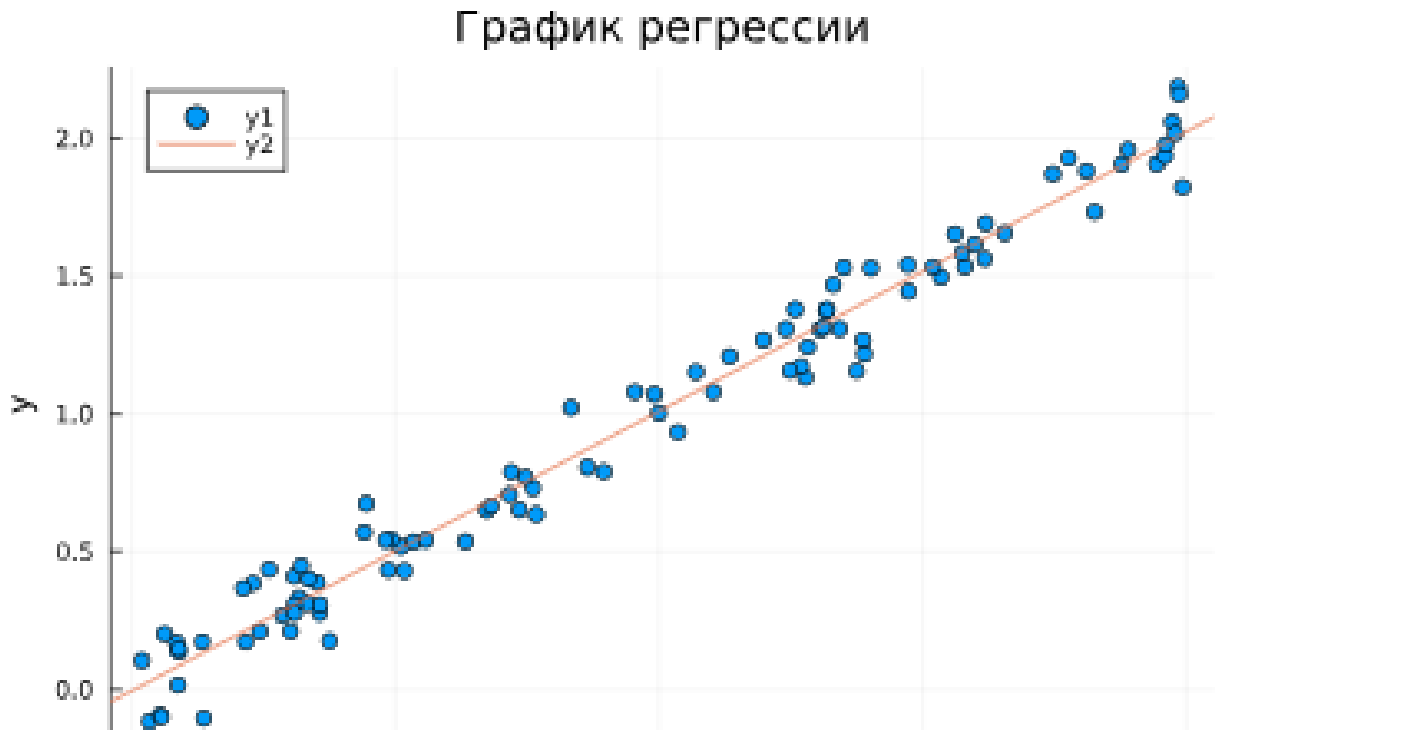
```
StatsModels.TableRegressionModel{LinearModel{GLM.LmResp{Vector{Float64}}, GLM.DensePredChol{Float64, LinearAlgebra.CholeskyPivoted{Float64, Matrix{Float64}, Vector{Int64}}}}, Matrix{Float64}}

y ~ 1 + x1 + x2 + x3

Coefficients:
──────────────────────────────────────────────────────────────────────────────────
                  Coef.  Std. Error       t  Pr(>|t|)    Lower 95%   Upper 95%
──────────────────────────────────────────────────────────────────────────────────
(Intercept)  -0.00237925  0.00315398   -0.75    0.4508  -0.00856845  0.00380995
x1            0.705885    0.00319307  221.07    <1e-99   0.699619    0.712151
x2            0.697592    0.00316063  220.71    <1e-99   0.691389    0.703794
x3            0.0171926   0.00310787    5.53    <1e-07   0.0110938   0.0232913
──────────────────────────────────────────────────────────────────────────────────
```

```
using Statistics
function find_best_fit(xvals,yvals)
    meanx = mean(xvals)
    meany = mean(yvals)
    stdx = std(xvals)
    stdy = std(yvals)
    r = cor(xvals, yvals)
    a = r*stdy/stdx
    b = meany - a*meanx
    return a, b
end
```

Задания для самостоятельного выполнения

Часть 2: Найдите линию регрессии, используя данные (□, □). Постройте график (□, □), используя точечный график. Добавьте линию регрессии, используя abline!. Добавьте заголовок «График регрессии» и подпишите оси □ и □.

```
#Часть2
using Plots
X = rand(100);
y = 2X + 0.1 * randn(100);
a,b = find_best_fit(X,y)
scatter(X,y, title = "График регрессии", xlabel="x", ylabel="y")
Plots.abline!(a,b)
```



Задания для самостоятельного выполнения

3. Модель ценообразования биномиальных опционов:

Описание модели ценообразования биномиальных опционов можно найти на стр. https://en.wikipedia.org/wiki/Binomial_options_pricing_model.

a-d. Построение графика траектории курса акций (Пусть □ = 100, □ = 1, □ = 10000, □ = 0.3 и □ = 0.08. Попробуйте построить траекторию курса акций. Функция rand () генерирует случайное число от 0 до 1. Вы можете использовать функцию построения графика из библиотеки графиков.):

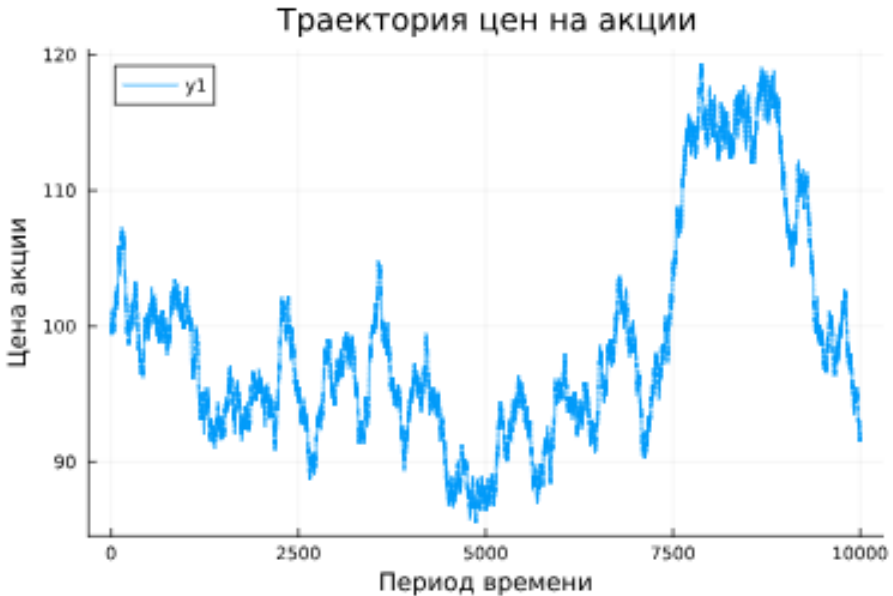
Задания для самостоятельного выполнения

```
S=100
T=1
n=10000
h=T/n
sigma=0.3
r=0.08
u=exp(r*h+sigma*sqrt(h))
d=exp(r*h-sigma*sqrt(h))
p=(exp(r*h)-d)/(u-d)
```

```
0.4992500005625153
```

```
Pr = []
c = 0
append!(Pr, S)
for i in 1:n
    prob=rand()
    if (prob<p)
        append!(Pr,S*u^(i-c)*d^c)
    else
        append!(Pr,S*u^(i-c-1)*d^(c+1))
        c+=1
    end
end
```

```
plot(Pr, xlabel = "Период времени", ylabel = "Цена акции", title = "Траектория цен на акции")
```



Задания для самостоятельного выполнения

b. Построение графика 10 разных траекторий цен на акции (Создайте функцию createPath (S :: Float64, r :: Float64, sigma :: Float64, T :: Float64, n :: Int64), которая создает траекторию цены акции с учетом начальных параметров. Используйте createPath, чтобы создать 10 разных траекторий и построить их все на одном графике.):

```
function createPath(S, r, sigma, T, n)
    h=T/n
    u=exp(r*h+sigma*sqrt(h))
    d=exp(r*h-sigma*sqrt(h))
    p=(exp(r*h)-d)/(u-d)
    Pr=[]
    c=0
    append!(Pr,S)
    for i in 1:n
        prob=rand()
        if(prob<p)
            append!(Pr,S*u^(i-c)*d^c)
        else
            append!(Pr,S*u^(i-c-1)*d^(c+1))
            c+=1
        end
    end
    return Pr
end
```

```
createPath (generic function with 1 method)
```

```
p=plot()
@time for i in 1:10
    plot!(createPath(S,r,sigma,T,n), xlabel = "Период времени",
```

Задания для самостоятельного выполнения

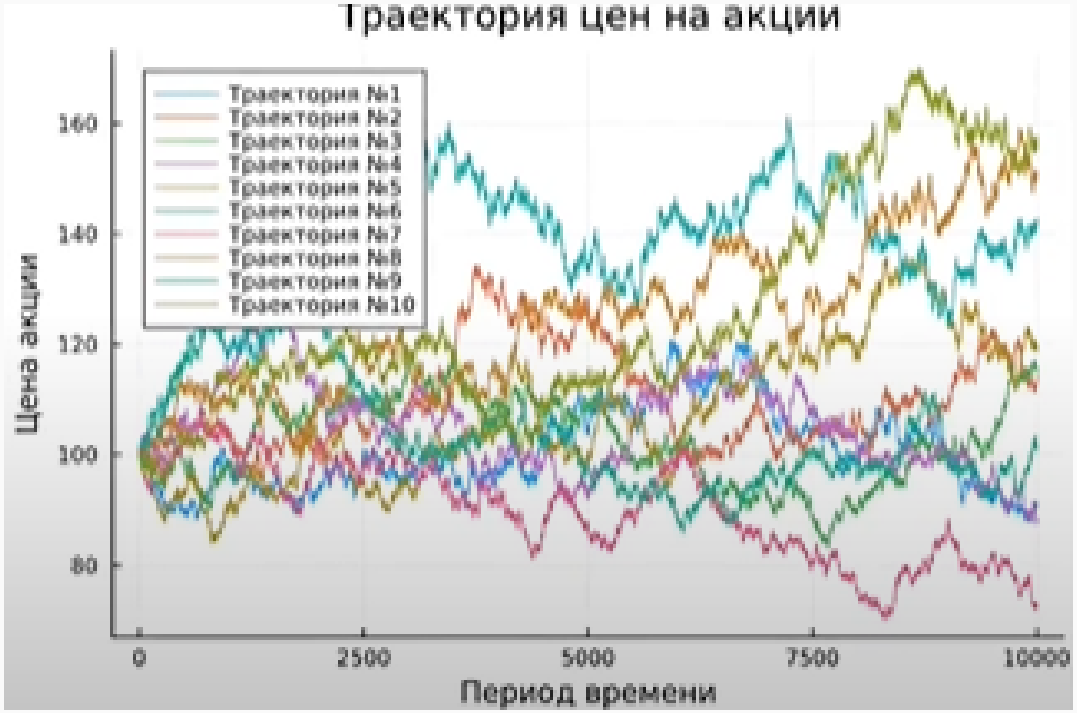


Рис. 27: График 10 разных траекторий цен на акции

Задания для самостоятельного выполнения

c. То же задание, что и в предыдущем пункте, только с распараллеливанием генерации траекторий (Распараллельте генерацию траектории. Можете использовать Threads.@threads,pmap и @parallel.):

```
using Distributed
p=plot()
@time @distributed for i in 1:10
    plot!(createPath(S,r,sigma,T,n), xlabel = "Период времени",
    ylabel = "Цена акции", title = "Траектория цен на акции",
    label = "Траектория W$i")
end
p
```

```
  0.010352 seconds (3.80 k allocations: 273.828 KiB, 99.15% compilation time)
```

Рис. 28: Код

Задания для самостоятельного выполнения

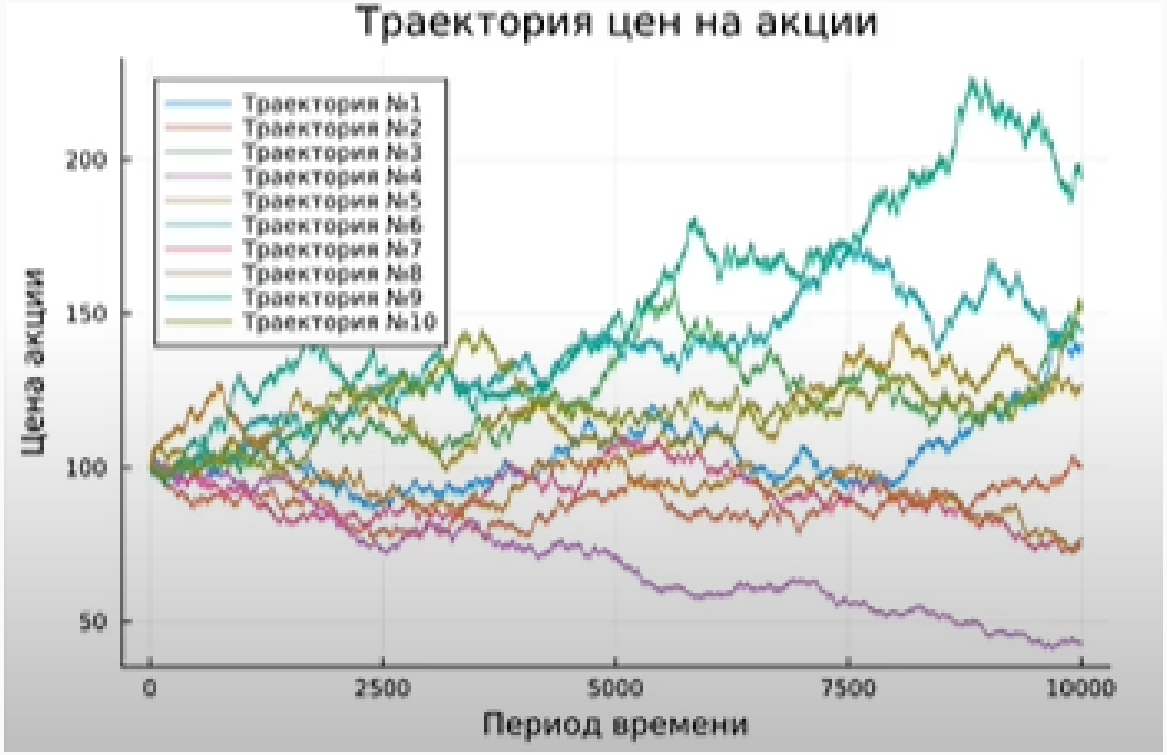


Рис. 29: График 10 разных траекторий цен на акции

Вывод

Я выполнила лабораторную работу №7 и успешно освоила специализированные пакеты Julia для обработки данных.

Список литературы

- Julia: https://ru.wikipedia.org/wiki/Julia
- https://julialang.org/packages/
- https://juliahub.com/ui/Home
- https://juliaobserver.com/
- https://github.com/svaksha/Julia.jl