

Лабораторная работа № 7

Ли Тимофей Александрович, НФИбд-01-18

Изучить специализированные пакеты Julia для обработки данных.

Ход работы. Примеры

```
# 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
```

```
P = CSV.File("programminglanguages.csv") |> DataFrame
```

8 rows × 2 columns

| | year | language |
|---|-------|----------|
| | Int64 | String7 |
| 1 | 1950 | basic |
| 2 | 1951 | fortran |
| 3 | 1952 | pascal |
| 4 | 1953 | c |
| 5 | 1954 | java |
| 6 | 1955 | prolog |
| 7 | 1956 | r |
| 8 | 1957 | python |

```
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)

```
language_created_year(P, "python")
```

1957

```
language_created_year(P, "prolog")
```

1955

```
language_created_year(P, "Prolog")
```

MethodError: no method matching getindex(::DataFrame, ::Nothing, ::Int64)

```
function language_created_year_v2(P, language::String)
    loc = findfirst(lowercase.(P[:,2]).==lowercase.(language))
    return P[loc,1]
end
```

language_created_year_v2 (generic function with 1 method)

```
language_created_year_v2(P, "Prolog")
```

1955

```
Tx = readdlm("programminglanguages.csv", ',',')
```

```
9×1 Matrix{Any}:
"year;language"
"1950;basic"
"1951;fortran"
"1952;pascal"
"1953;c"
"1954;java"
"1955;prolog"
"1956;r"
"1957;python"
```

Рис. 1: примеры1

```
In [11]: CSV.write("programming_languages_data2.csv", P)
```

```
Out[11]: "programming_languages_data2.csv"
```

```
In [12]: writedlm("programming_languages_data.txt", Tx, ',')
```

```
In [13]: # Пример записи данных в текстовый файл с разделителем '-'  
writedlm("programming_languages_data2.txt", Tx, '-')
```

```
In [14]: P_new_delim = readdlm("programming_languages_data2.txt", '-')
```

```
Out[14]: 9×1 Matrix{Any}:  
  "year;language"  
  "1950;basic"  
  "1951;fortran"  
  "1952;pascal"  
  "1953;c"  
  "1954;java"  
  "1955;prolog"  
  "1956;r"  
  "1957;python"
```

```
In [15]: dict = Dict{Integer,Vector{String}}()  
dict2 = Dict()
```

```
Out[15]: Dict{Any, Any}()
```

```
In [16]: 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
```

```
In [18]: dict[1954]
```

```
Out[18]: 1-element Vector{String}:  
 "java"
```

```
using DataFrames
```

```
df = DataFrame(year = P[:,1], language = P[:,2])
```

8 rows × 2 columns

| | year | language |
|---|-------|----------|
| | Int64 | String7 |
| 1 | 1950 | basic |
| 2 | 1951 | fortran |
| 3 | 1952 | pascal |
| 4 | 1953 | c |
| 5 | 1954 | java |
| 6 | 1955 | prolog |
| 7 | 1956 | r |
| 8 | 1957 | python |

```
df[:, :year]
```

8-element Vector{Int64}:

1950
1951
1952
1953
1954
1955
1956
1957

```
describe(df)
```

2 rows × 7 columns

| | variable | mean | min | median | max | nmissing | eltype |
|---|----------|----------|-------|----------|------|----------|-----------|
| | Symbol | Union... | Any | Union... | Any | Int64 | Data Type |
| 1 | year | 1953.5 | 1950 | 1953.5 | 1957 | 0 | Int64 |
| 2 | language | | basic | | r | 0 | String7 |

Рис. 4: примеры4

```
using RDatasets
```

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

150 rows × 5 columns

| | 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 |

```
typeof(iris)
```

DataFrame

```
describe(iris)
```

5 rows × 7 columns

| | variable | mean | min | median | max | nmissing | eltype |
|---|-------------|----------|-----|-----------|-----|----------|---------------------------------|
| | Symbol | Union... | Any | Union... | Any | Int64 | Data Type |
| 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} |

Рис. 5: примеры

```
a = missing
typeof(a)

Missing

a+1

missing

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

4-element Vector{Union{Missing, Int64}}:
 missing
  47
  22
 105

typeof(calories)

Vector{Union{Missing, Int64}} (alias for Array{Union{Missing, Int64}, 1})

using Statistics

mean(calories)

missing

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 = innerjoin(dataframe_calories,dataframe_prices,on=:item)

4 rows × 3 columns

   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
```

Рис. 6: примеры


```
import Pkg
Pkg.add("ImageIO")
```

```
Installed Imath_jll ----- v3.1.2+0
Installed AbstractFFTs ----- v1.0.1
```

```
776x1200 Array[RGBA{0.0f,0.0f},2] with eltype ColorTypes.RGBA{FixedPointNumbers.NoF8}:
RGBA{NoF8}{0.0,0.0,0.0,0.0} ... RGBA{NoF8}{0.0,0.0,0.0,0.0}
RGBA{NoF8}{0.0,0.0,0.0,0.0} RGBA{NoF8}{0.0,0.0,0.0,0.0}
RGBA{NoF8}{0.0,0.0,0.0,0.0} RGBA{NoF8}{0.0,0.0,0.0,0.0}
RGBA{NoF8}{0.0,0.0,0.0,0.0} RGBA{NoF8}{0.0,0.0,0.0,0.0}
RGBA{NoF8}{0.0,0.0,0.0,0.0} RGBA{NoF8}{0.0,0.0,0.0,0.0}
RGBA{NoF8}{0.0,0.0,0.0,0.0} ... RGBA{NoF8}{0.0,0.0,0.0,0.0}
RGBA{NoF8}{0.0,0.0,0.0,0.0} RGBA{NoF8}{0.0,0.0,0.0,0.0}
RGBA{NoF8}{0.0,0.0,0.0,0.0} RGBA{NoF8}{0.0,0.0,0.0,0.0}
RGBA{NoF8}{0.0,0.0,0.0,0.0} RGBA{NoF8}{0.0,0.0,0.0,0.0}
RGBA{NoF8}{0.0,0.0,0.0,0.0} RGBA{NoF8}{0.0,0.0,0.0,0.0}
RGBA{NoF8}{0.0,0.0,0.0,0.0} ... RGBA{NoF8}{0.0,0.0,0.0,0.0}
RGBA{NoF8}{0.0,0.0,0.0,0.0} RGBA{NoF8}{0.0,0.0,0.0,0.0}
RGBA{NoF8}{0.0,0.0,0.0,0.0} RGBA{NoF8}{0.0,0.0,0.0,0.0}
```



```
typeof(X1) = Matrix{ColorTypes.RGBA{FixedPointNumbers.N0f8}}
size(X1) = (776, 1200)
```

Рис. 7: примеры7

Ход работы. Примеры

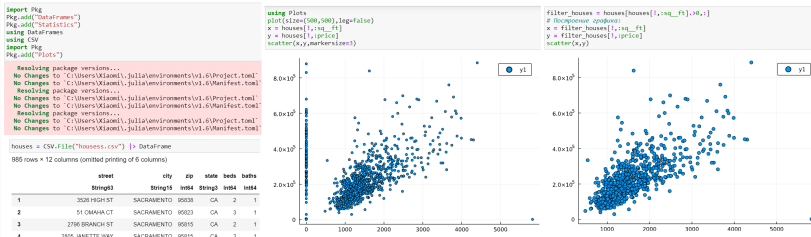


Рис. 8: примеры8

Ход работы. Примеры

```
using Statistics
# Операнды: cpechd unary dan opechdchdno mna dnoad:
combine(filter_houses[:type], filter_houses[:mean(filter_houses[1,:price])])

814 rows × 2 columns
```

| | type | x1 |
|---|-------------|-----------|
| | String16 | Float64 |
| 1 | Residential | 2.29449e5 |
| 2 | Residential | 2.29449e5 |
| 3 | Residential | 2.29449e5 |
| 4 | Residential | 2.29449e5 |
| 5 | Residential | 2.29449e5 |

```
import Plots
Plots.add!("Clustering")
using Clustering

# Resolving package versions...

# Jupyter Notebook:
X = filter_houses[1, [:latitude, :longitude]]
# K-means clustering:
X = Matrix{X}
```

```
814x2 Matrix{Float64}:
38.6319 -121.435
38.4789 -121.431
38.6183 -121.444
38.6168 -121.439
38.5195 -121.436
38.6626 -121.328
38.6817 -121.352
38.5351 -121.481
38.6312 -121.271
38.7089 -121.443
38.6377 -121.452
38.4787 -121.459
38.6187 -121.436
```

```
X'X'
```

```
29814 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
```

```
k = length(unique(filter_houses[1,:zip]))
C = kmeans(X,k)
```

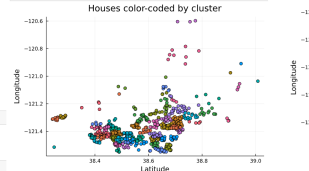
```
MeansResult{Matrix{Float64}, Float64, Int64}([38.56742925 38.490447 - 38.53068435714287 38.69875490476191, -121.49139358333333
-121.129337 -121.46543535714284 -121.37139285714287, [60, 42, 53, 60, 11, 4, 21, 29, 33 - 6, 39, 42, 14, 43, 54, 59, 2
8, 30, 13], [6.734555427101441e-5, 0.00037363141382229514, 0.0001296234316758052, 0.00017246662901015952, 0.000155398280185181
65, 0.000116170828954192, 0.0002536112095971003, 0.000147048147482256467, 0.344488792819902e-5, 0.00011686658399412408 - 5.7
84670761318538e-5, 2.413251978578046e-5, 9.49839880743575e-6, 0.0001942090842757628, 0.00013418837747743535, 0.00014433953582
75962, 7.579613156849518e-5, 8.884414273779839e-5, 0.000186035837845411, 0.000109576998918783], [12, 1, 11, 18, 12, 28, 0, 1
9, 14, 28 - 6, 3, 17, 18, 6, 4, 14, 1, 14, 21], [12, 1, 11, 18, 12, 28, 8, 19, 14, 28 - 6, 3, 17, 18, 6, 4, 14, 1, 14, 21],
0.203318177624358214, 16, true)
```

```
df = DataFrame(cluster = C.assignments, city = filter_houses[1,:city],
latitude = filter_houses[1,:latitude], longitude = filter_houses[1,:longitude],
zip = filter_houses[1,:zip])
```

```
814 rows × 6 columns
```

| | cluster | city | latitude | longitude | zip |
|---|---------|------------|----------|-----------|-------|
| | Int64 | String16 | Float64 | Float64 | Int64 |
| 1 | 60 | SACRAMENTO | 38.6319 | -121.435 | 95838 |
| 2 | 42 | SACRAMENTO | 38.4789 | -121.431 | 95823 |
| 3 | 53 | SACRAMENTO | 38.6183 | -121.444 | 95815 |
| 4 | 60 | SACRAMENTO | 38.6168 | -121.439 | 95815 |
| 5 | 11 | SACRAMENTO | 38.5195 | -121.436 | 95812 |

```
clusters_figure = plot(legend = false)
for i = 1:k
    clustered_houses = df[df[1,:cluster] == i, :]
    xvals = clustered_houses[1,:latitude]
    yvals = clustered_houses[1,:longitude]
    scatter!(clusters_figure, xvals, yvals, markersize=4)
end
xlabel!("Latitude")
ylabel!("Longitude")
title!("Houses color-coded by cluster")
display(clusters_figure)
```



```
unique_zips = unique(filter_houses[1,:zip])
zips_figure = plot(legend = false)
for izip in unique_zips
    subs = filter_houses[filter_houses[1,:zip] == izip, :]
    x = subs[1,:latitude]
    y = subs[1,:longitude]
    scatter!(zips_figure, x, y)
end
xlabel!("Latitude")
ylabel!("Longitude")
title!("Houses color-coded by zip code")
display(zips_figure)
```

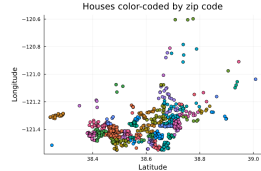


Рис. 9: примеры

Ход работы. Примеры

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

Resolving package versions...

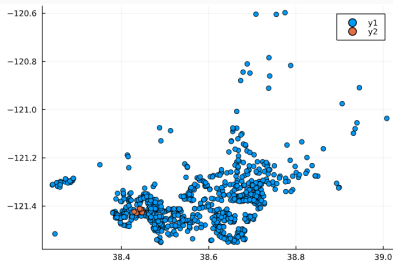
Updating `C:\Users\Xiaomi\.julia\environments\v1.6\Project.toml`
[b8a86587] + NearestNeighbors v0.4.9

No Changes to `C:\Users\Xiaomi\.julia\environments\v1.6\Manifest.toml`

```
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)
# Все объекты neighborhoods:
x = filter_houses[1,:latitude];
y = filter_houses[1,:longitude];
scatter(x,y)
# Соседи:
x = filter_houses[idxs,:latitude];
y = filter_houses[idxs,:longitude];
scatter!(x,y)
```



```
cities = filter_houses[idxs,:city]
```

```
10-element PooledArrays.PooledVector{String15, UInt32, Vector{UInt32}}:
"SACRAMENTO"
"ELK GROVE"
"SACRAMENTO"
"SACRAMENTO"
"SACRAMENTO"
"SACRAMENTO"
"ELK GROVE"
"ELK GROVE"
"ELK GROVE"
"ELK GROVE"
```

Рис. 10: примеры10

Ход работы. Примеры

```
F = filter_houses[1,[:sq_ft,:price]]  
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
```

```
convert(Array{Float64,2}, F)
```

```
2×814 Matrix{Float64}:  
 836.0  1167.0   796.0   852.0   797.0  ...  1216.0   1685.0   1362.0  
59222.0 68212.0 68880.0 69307.0 81900.0 ... 235000.0 235301.0 235738.0
```

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

```
Resolving package versions...
```

```
No Changes to `C:\Users\Xiaomi\.julia\environments\v1.6\Project.toml`  
No Changes to `C:\Users\Xiaomi\.julia\environments\v1.6\Manifest.toml`
```

```
M = fit(PCA, F)
```

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

```
# Выделение значений главных компонент в отдельную переменную:
```

```
Xr = reconstruct(M, y')
```

```
# Построение графика с выделением главных компонент:
```

```
scatter(F[1,:],F[2,:])  
scatter!(Xr[1,:],Xr[2,:])
```

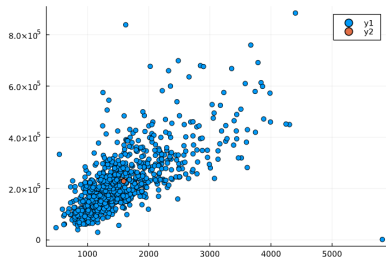
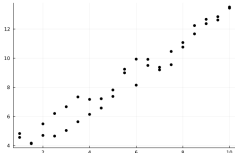


Рис. 11: примеры11

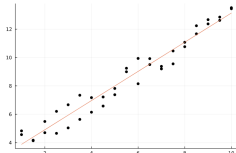
```
xvals = repeat(1:0.5:10, inner=2)
yvals = 3 + xvals + 2*rand(length(xvals)) - 1
scatter(xvals, yvals, color='black', legend=false)
```



```
function find_fit(xvals, yvals)
    means = mean(xvals)
    means_y = mean(yvals)
    stdx = std(xvals)
    stdy = std(yvals)
    r = cor(xvals, yvals)
    a = r*stdy/stdx
    b = means_y - a*means
    return a, b
end
```

find_fit_fit (generic function with 1 method)

```
a, b = find_fit(xvals, yvals)
ym = a * xvals + b
plot!(xvals, yvals)
```



```
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_fit(xvals, yvals)

0.116022 seconds (103.56 k allocations: 6.250 MiB, 08.48% compilation time)
(0.9999999583660921, 3.086238500166666)
```

```
import Plots
Plots.add!("Corda")
using Plots
using Conda
```

Resolving package versions...

```
py"""
import numpy
def find_fit_python(xvals, yvals):
    means = numpy.mean(xvals)
    means_y = numpy.mean(yvals)
    stdx = numpy.std(xvals)
    stdy = numpy.std(yvals)
    r = numpy.corrcoef(xvals, yvals)[0][1]
    a = r*stdy/stdx
    b = means_y - a*means
    return a, b
...
find_fit_fit_python=find_fit_fit_python
xpy = PyObj(jc(xvals))
ypr = PyObj(jc(yvals))
@time a, b = find_fit_fit_python(xpy, ypy)
```

```
0.007533 seconds (237.03 k allocations: 15.021 MiB, 23.34% gc time, 84.55% compilation time)
(0.9999999583660921, 3.086238500166666)
```

```
import Plots
Plots.add!("BenchmarkTools")
using BenchmarkTools
```

Resolving package versions...

```
@time a, b = find_fit_fit_python(xvals, yvals)
@time a, b = find_fit_fit(xvals, yvals)

10.267 ms (17 allocations: 1.49 KiB)
1.508 ms (1 allocation: 32 bytes)
(0.9999999583660921, 3.086238500166666)
```

Рис. 12: примеры12

Ход работы. K-Means

150 rows x 5 columns

150 rows \times 4 columns

| | | | | |
|---|---|-----|-----|--------|
| 1 | 1 | 5.1 | 3.5 | setosa |
| 2 | 1 | 4.9 | 3.0 | setosa |
| 3 | 1 | 4.7 | 3.2 | setosa |
| 4 | 1 | 4.6 | 3.1 | setosa |
| 5 | 1 | 5.0 | 3.6 | setosa |

A scatter plot showing the relationship between Sepal.Length (x-axis) and Sepal.Width (y-axis) for three species: Setosa (blue), Versicolour (green), and Virginica (red). The x-axis ranges from approximately 4.5 to 8.5, and the y-axis ranges from 2.0 to 4.5. Setosa points are clustered in the lower-left region (Sepal.Length < 6.5, Sepal.Width < 3.5). Versicolour points are clustered in the middle region (Sepal.Length between 5.5 and 7.0, Sepal.Width between 2.5 and 3.5). Virginica points are clustered in the upper-right region (Sepal.Length > 6.5, Sepal.Width > 3.0).

Рис. 13: k-means

Ход работы. Регрессия (метод наименьших квадратов в случае линейной регрессии)

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

```
function mnklin(X,y)
    x=fli(1:0,size(X)[1])
    X2=[x X]
    x=X2\y
    return x
end
```

mnklin (generic function with 1 method)

```
a=mnklin(X,y)
```

```
4-element Vector{Float64}:
-0.0031974281796004903
 0.70204612149304233
 0.37818205181339587
 0.2488419694557077
```

```
using MultivariateStats
Pkg.add("GLM")
using GLM
```

Resolving package versions...

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

```
4-element Vector{Float64}:
 0.70204612149304235
 0.37818205181339587
 0.2488419694557076
-0.0031974281796004783
```

```
X1=X[:,1]
X2=X[:,2]
X3=X[:,3]
reg=DataFrame(y=y,x1=X1,x2=X2,x3=X3)
lm=@formula(y~x1+x2+x3),reg)
```

```
StatsModels.TableRegressionModel{LinearModel{GLM.LmResp{Vector{Float64}}}, GLM.Fitted{Float64, Matrix{Float64}}}, Matrix{Float64}}
```

$y \sim 1 + x_1 + x_2 + x_3$

Coefficients:

| | Coeff. | Std. Error | t | Pr(> t) | Lower 95% | Upper 95% |
|-------------|-------------|------------|--------|----------|------------|------------|
| (Intercept) | -0.00319743 | 0.00313093 | -1.02 | 0.3074 | -0.0093414 | 0.00294655 |
| x1 | 0.7020461 | 0.00312824 | 224.42 | <1e-99 | 0.695903 | 0.708019 |
| x2 | 0.378182 | 0.0031445 | 120.27 | <1e-99 | 0.372011 | 0.384353 |
| x3 | 0.248842 | 0.00302775 | 82.19 | <1e-99 | 0.2429 | 0.254783 |

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

```
a=mnklin(X,y)
b=a[2]
c=a[1]
b,c
```

(1.960615878718642, 0.02412528002865031)

```
scatter(X,y)
Plots.abline(b,c, title="График регрессии", xlabel="x", ylabel="y")
```

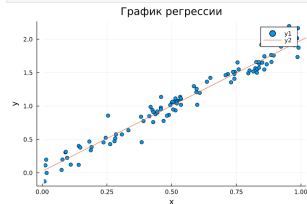


Рис. 14: МНК в случае лин.регрессии

Ход работы. Модель ценообразования биномиальных опционов

```
S=100
T=1
n=10000
h=T/n
sigma=0.3
r=0.06
u=exp(r*h + sigma*sqrt(h))
d=exp(r*h - sigma*sqrt(h))
p=(exp(r*h)-d)/(u-d)
0.4992500005425153
```

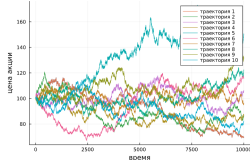
```
line=[]
c=0
append(line,S)
for i in 1:n
    k=rand()
    if k<p
        append(line,S*u*(1-c)^d*c)
    else
        append(line,S*u*(1-c-1)^d*(c+1))
    end
end
plot(line, title="траектория курса акций", xlabel="время", ylabel="цена акции")
```



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

CreatePath (generic function with 1 method)

```
for i in 1:10
    IJulia.clear_output(true)
    path=CreatePath(100,0.06,0.3,1,10000)
    if i==1
        plot(path, label=false)
    end
    plot!(path, label="траектория $i", xlabel="время", ylabel="цена акции")
    display(p)
end
```



using Base.Threads

```
Threads.@threads for i in 1:10
    IJulia.clear_output(true)
    path=CreatePath(100,0.06,0.3,1,10000)
    if i==1
        plot(path, label=false)
    end
    plot!(path, label="траектория $i", xlabel="время", ylabel="цена акции")
    display(p)
end
```

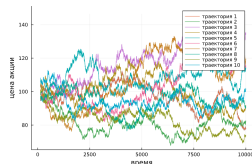


Рис. 15: модель ценообразования бином.опц.

Изучил специализированные пакеты Julia для обработки данных.