Лабораторная работа №7

Компьютерный практикум по статистическому анализу данных

Николаев Дмитрий Иванович

Содержание

# 1 Цель работы

Основной целью работы является освоение специализированных пакетов Julia для обработки данных.

# 2 Выполнение лабораторной работы

## 2.1 Повторение примеров

Повторим примеры, представленные в лабораторной работе ([1]).

### 2.1.1 Julia для науки о данных

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

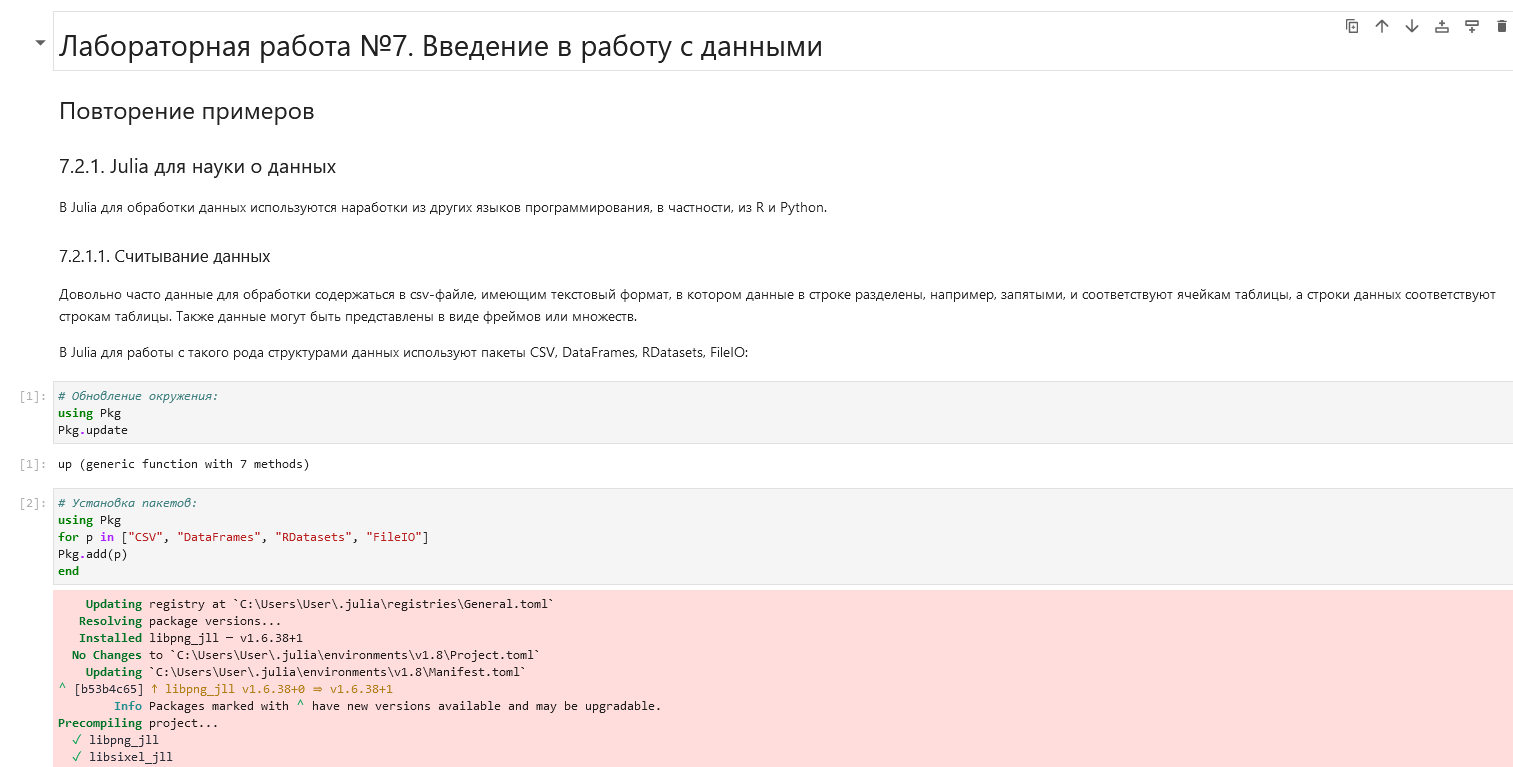
#### 2.1.1.1 Считывание данных

Перед тем, как начать проводить какие-либо операции над данными, необходимо их откуда-то считать и возможно сохранить в определённой структуре.

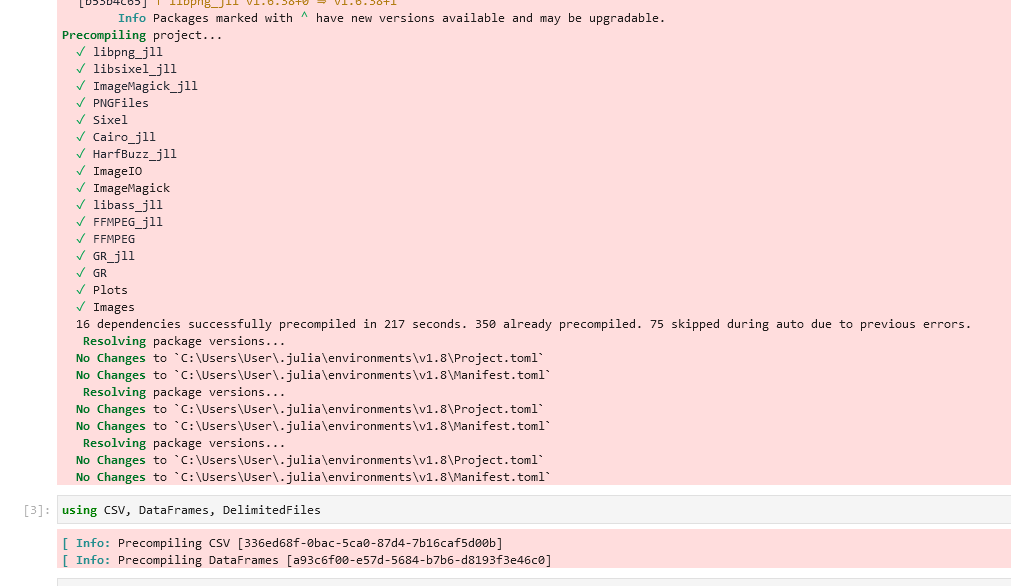
Довольно часто данные для обработки содержаться в csv-файле, имеющим текстовый формат, в котором данные в строке разделены, например, запятыми, и соответствуют ячейкам таблицы, а строки данных соответствуют строкам таблицы. Также данные могут быть представлены в виде фреймов или множеств.

В Julia для работы с такого рода структурами данных используют пакеты CSV, DataFrames, RDatasets, FileIO:

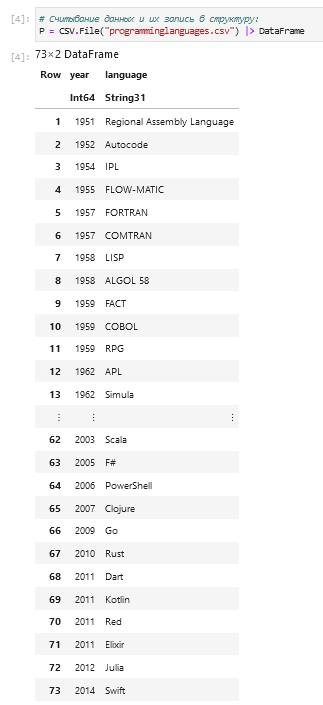
Рассмотрим возможности пакетов Julia по считыванию данных ([??-??]).



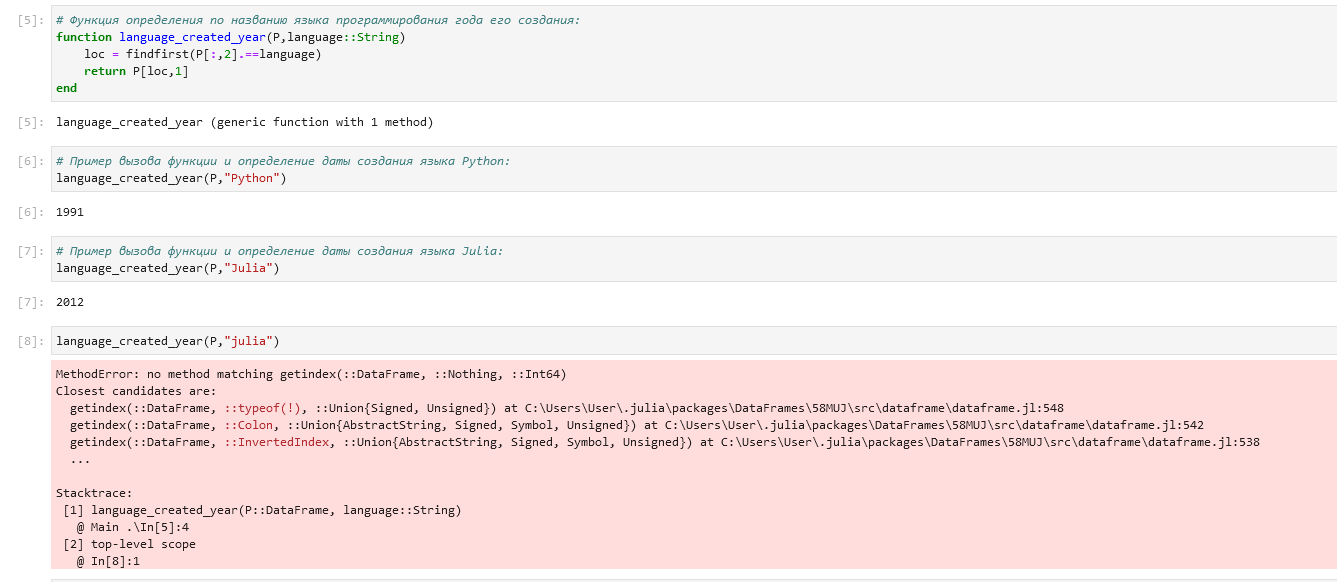
Считывание данных (1)



Считывание данных (2)



Считывание данных (3)



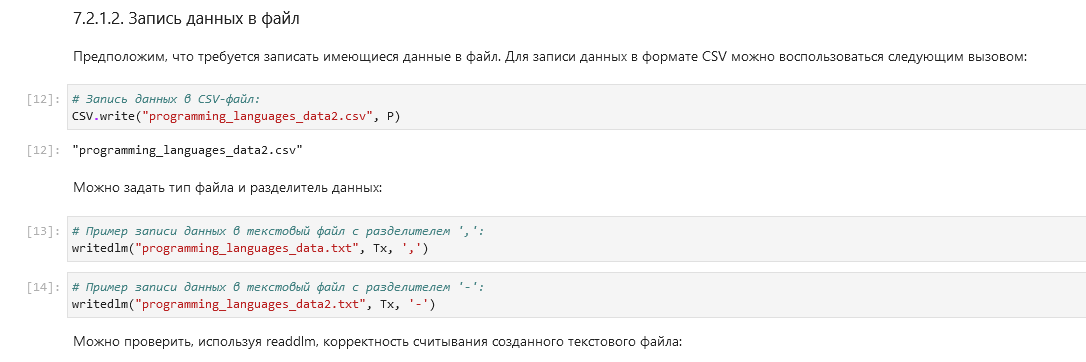
Считывание данных (4)



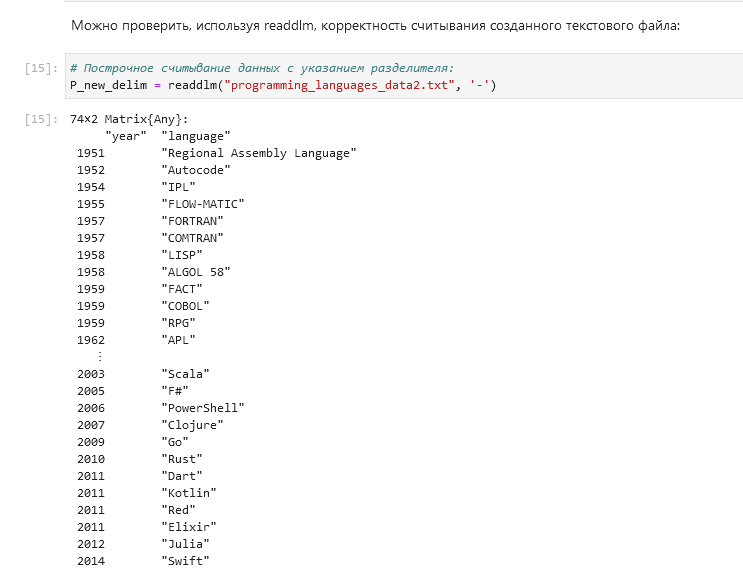
Считывание данных (5)

#### 2.1.1.2 Запись данных в файл

Рассмотрим возможности пакетов Julia по записи данных в файл ([??,??]).



Запись данных в файл (1)



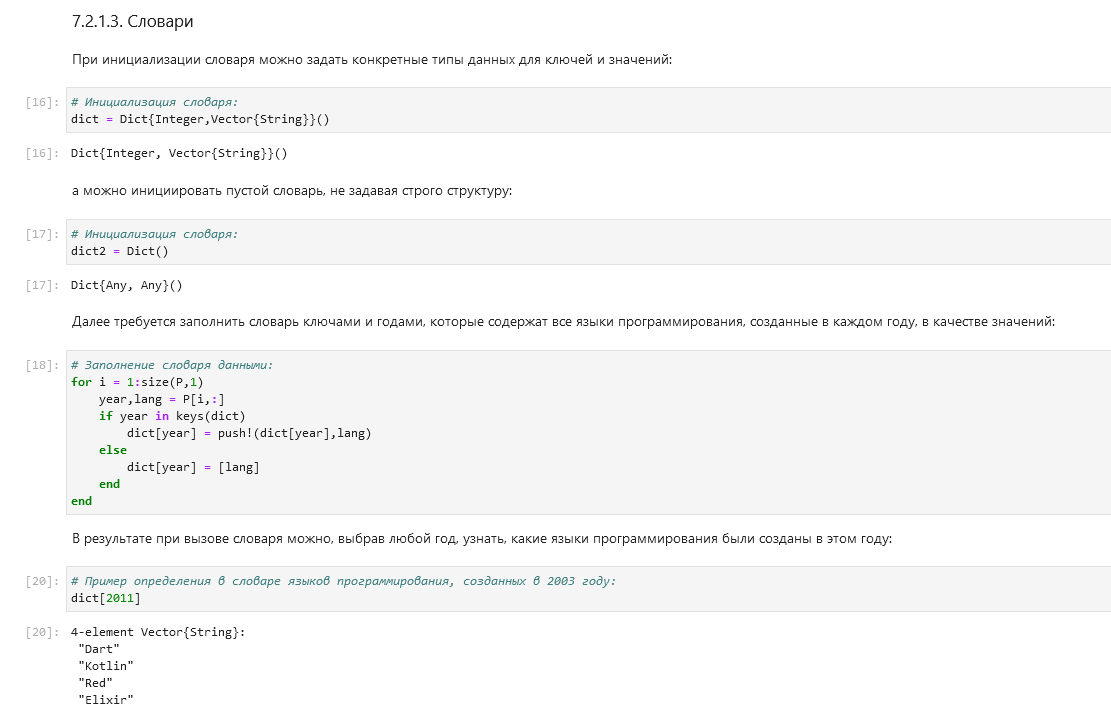
Запись данных в файл (2)

#### 2.1.1.3 Словари

При работе с данными бывает удобно записать их в формате словаря.

Предположим, что словарь должен содержать перечень всех языков программирования и года их создания, при этом при указании года выводить все языки программирования, созданные в этом году.

Создадим словарь с теми же данными ([??]).

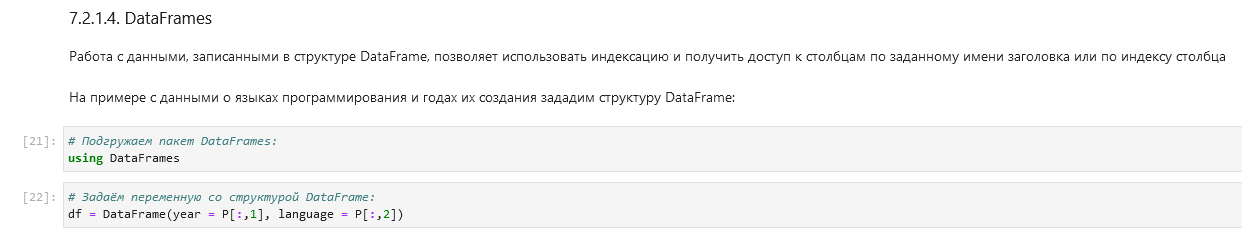


Словари

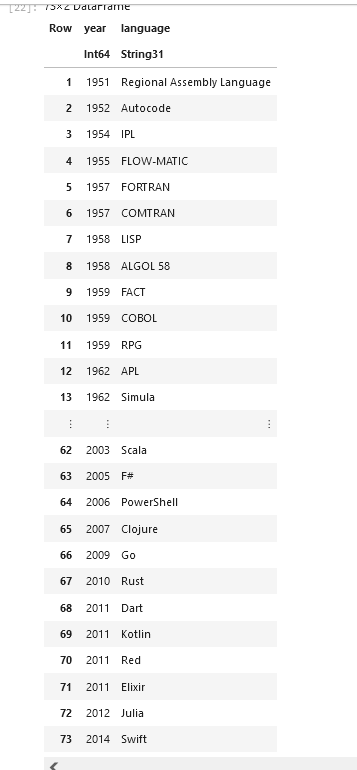
#### 2.1.1.4 DataFrames

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

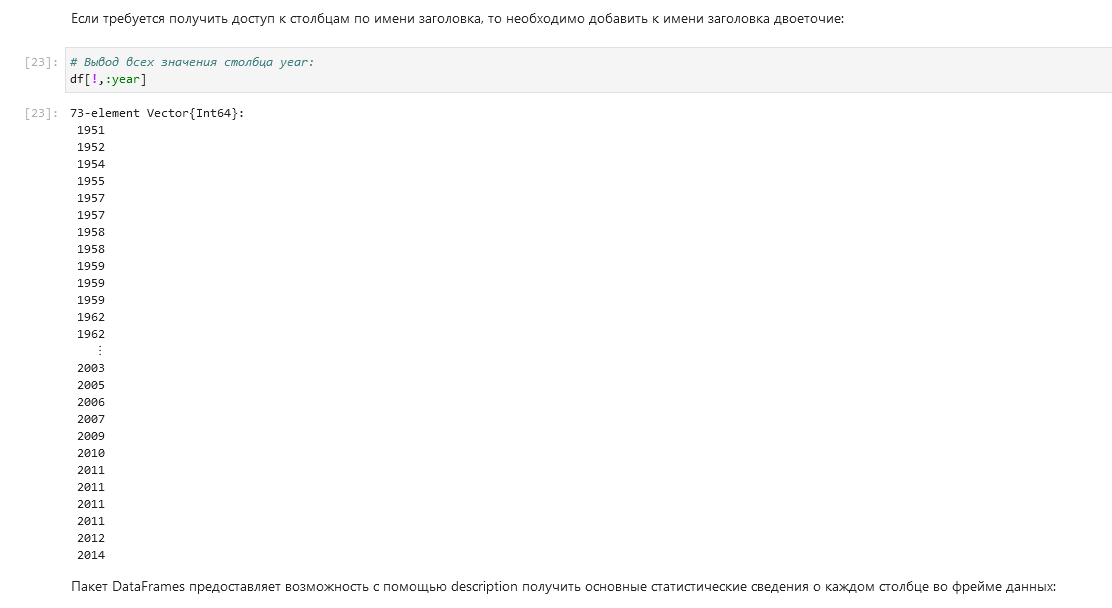
Рассмотрим работу с DataFrames ([??-??]).



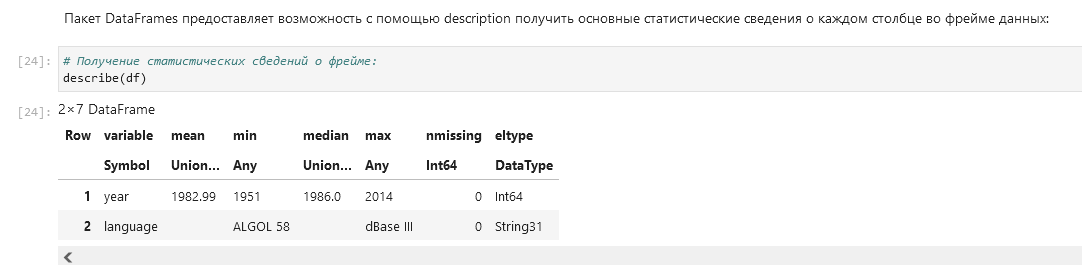
DataFrames (1)



DataFrames (2)



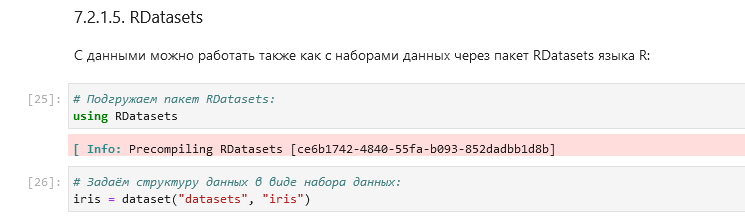
DataFrames (3)



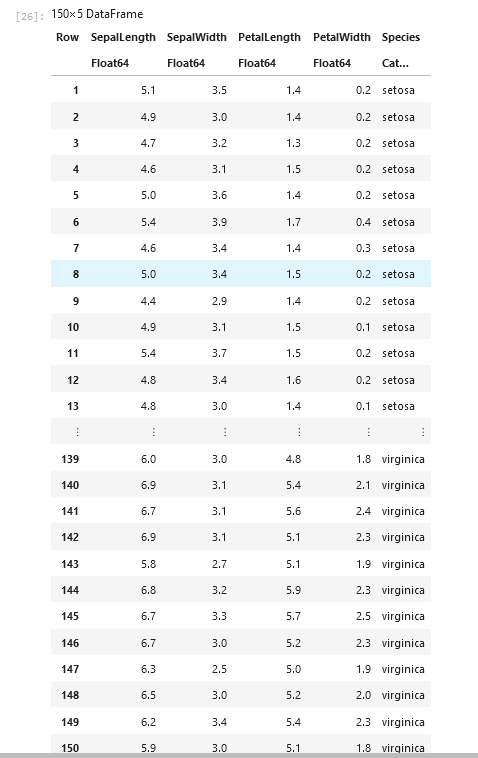
DataFrames (4)

#### 2.1.1.5 RDatasets

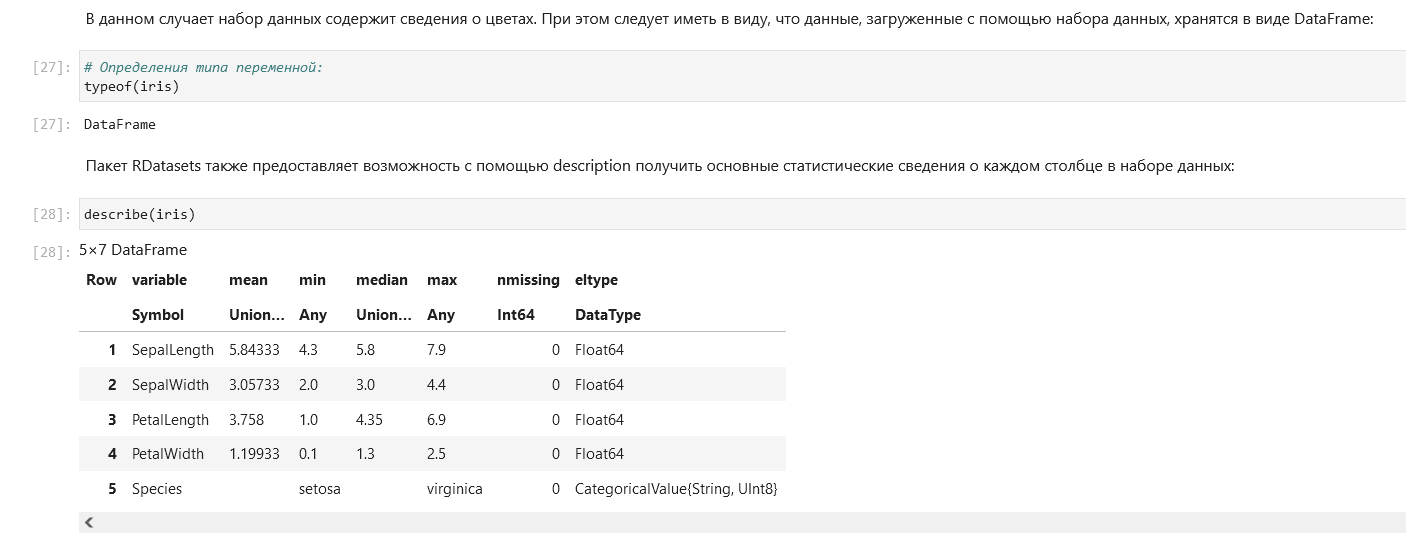
Рассмотрим работу с RDatasets ([??-??]).



RDatasets (1)



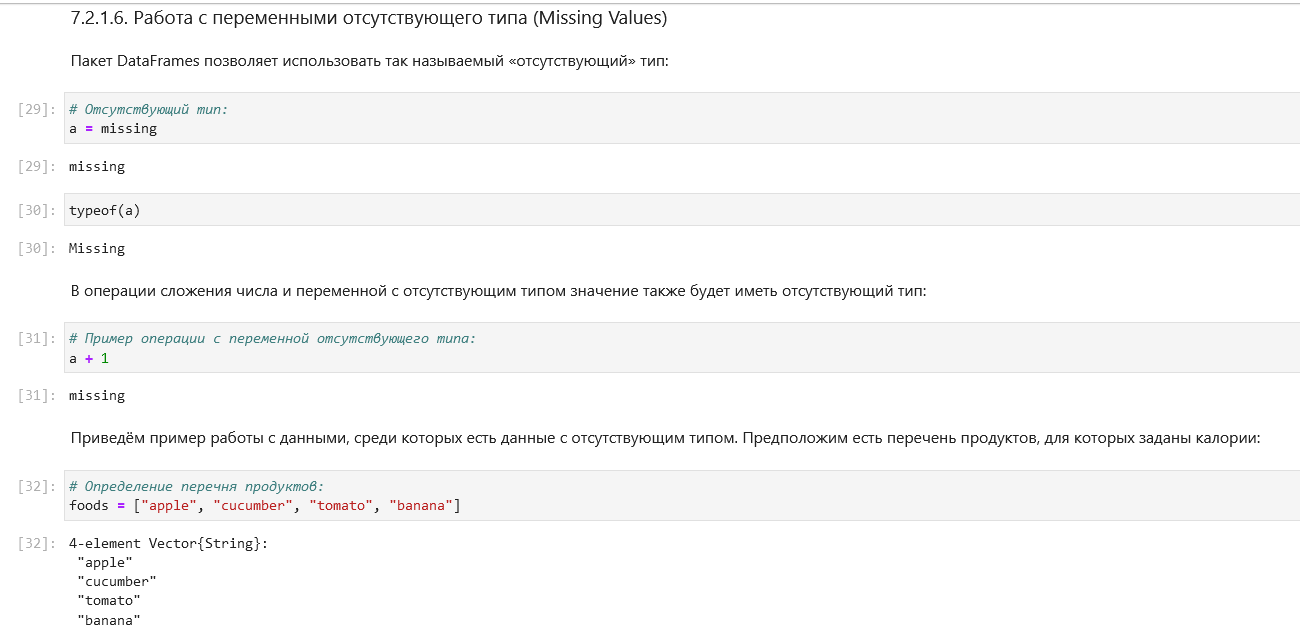
RDatasets (2)



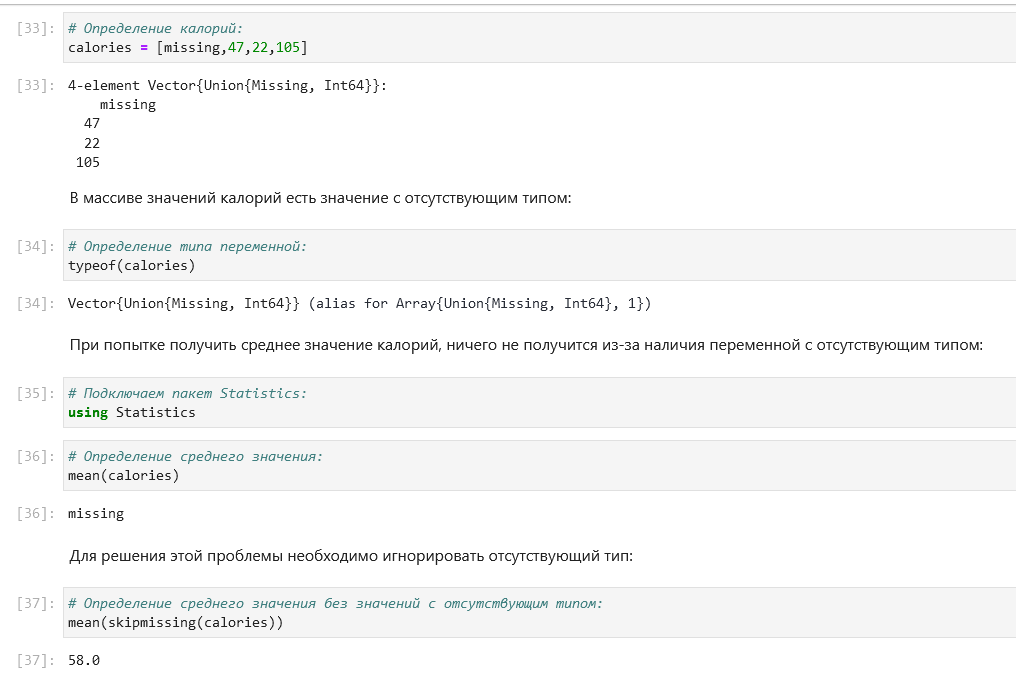
RDatasets (3)

#### 2.1.1.6 Работа с переменными отсутствующего типа (Missing Values)

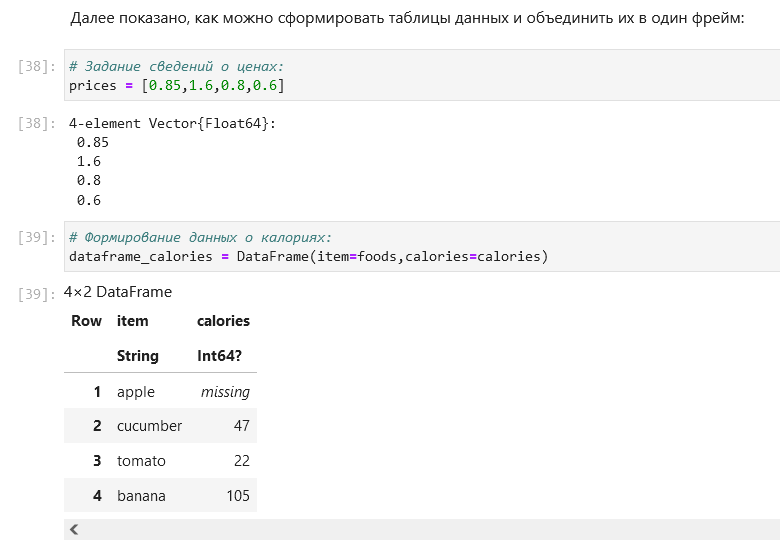
Рассмотрим особенности работы с переменными отсутствующего типа ([??-??]).



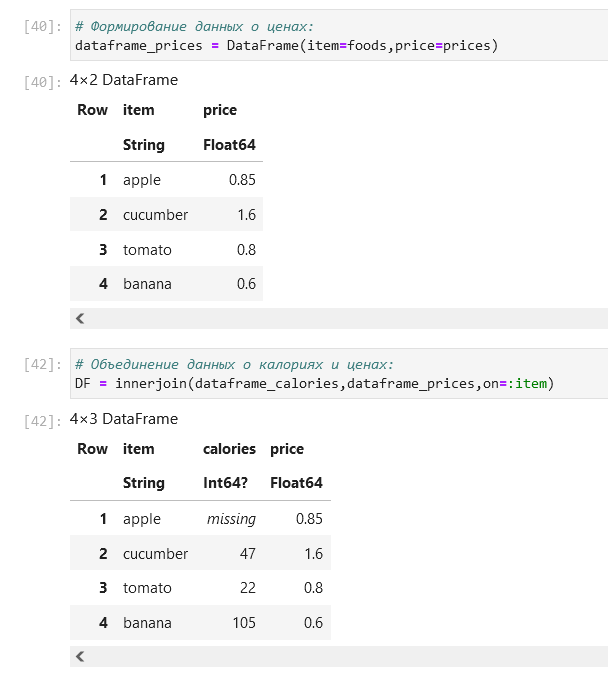
Работа с Missing Values (1)



Работа с Missing Values (2)



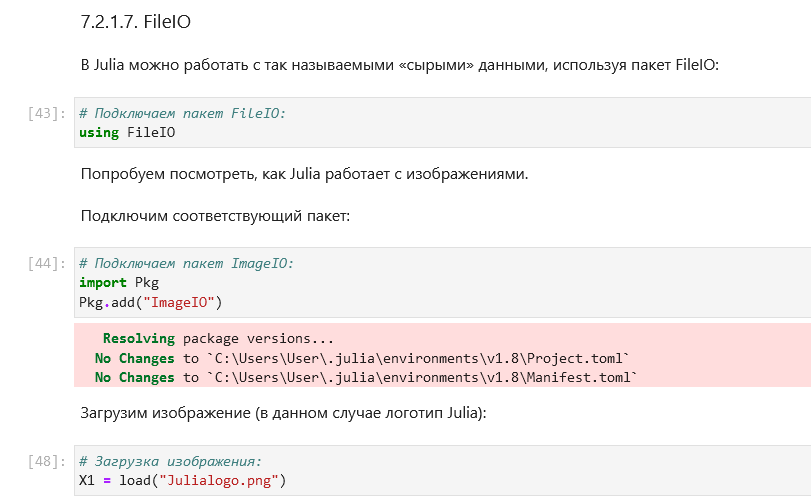
Работа с Missing Values (3)



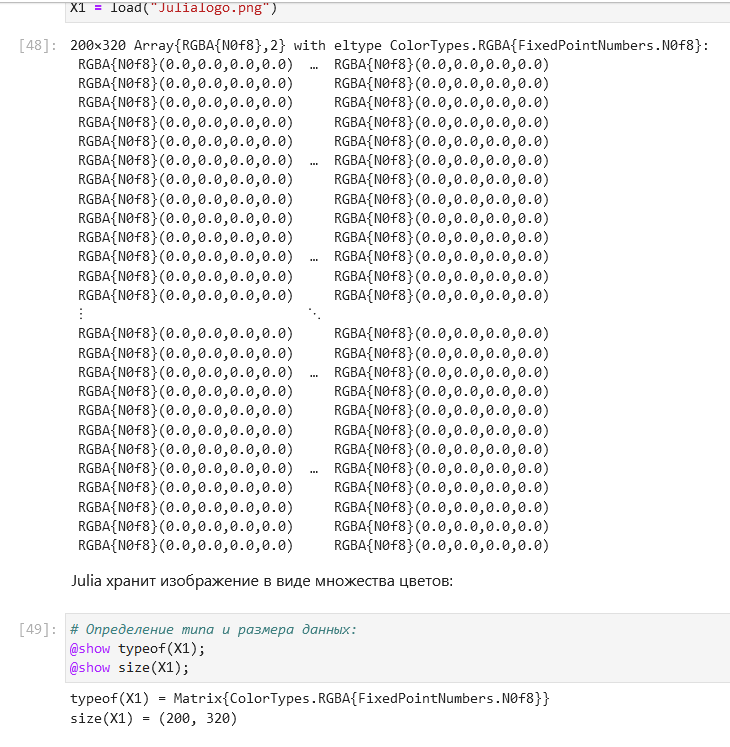
Работа с Missing Values (4)

#### 2.1.1.7 FileIO

Рассмотрим работу с так называемыми “сырыми данными” с помощью пакета FileIO ([??,??]).



FileIO (1)



FileIO (2)

### 2.1.2 Обработка данных: стандартные алгоритмы машинного обучения в Julia

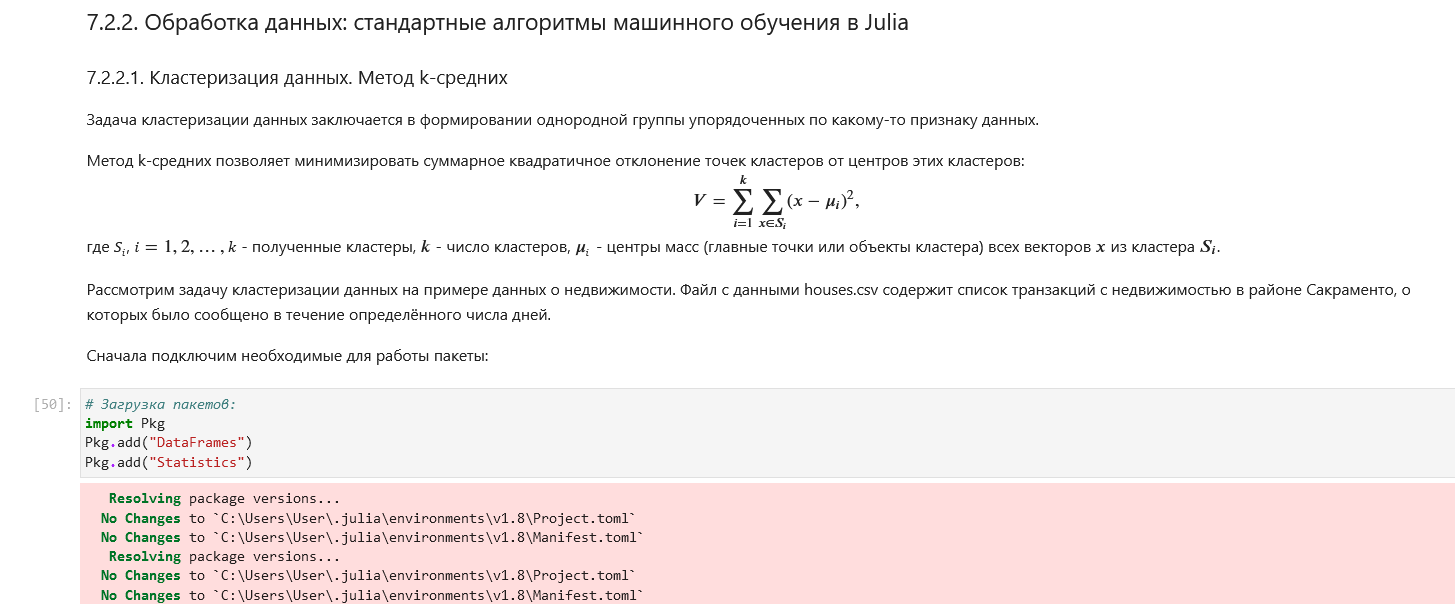
#### 2.1.2.1 Кластеризация данных. Метод k-средних

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:

где , — полученные кластеры, — число кластеров, — центры масс (главные точки или объекты кластера) всех векторов из кластера .

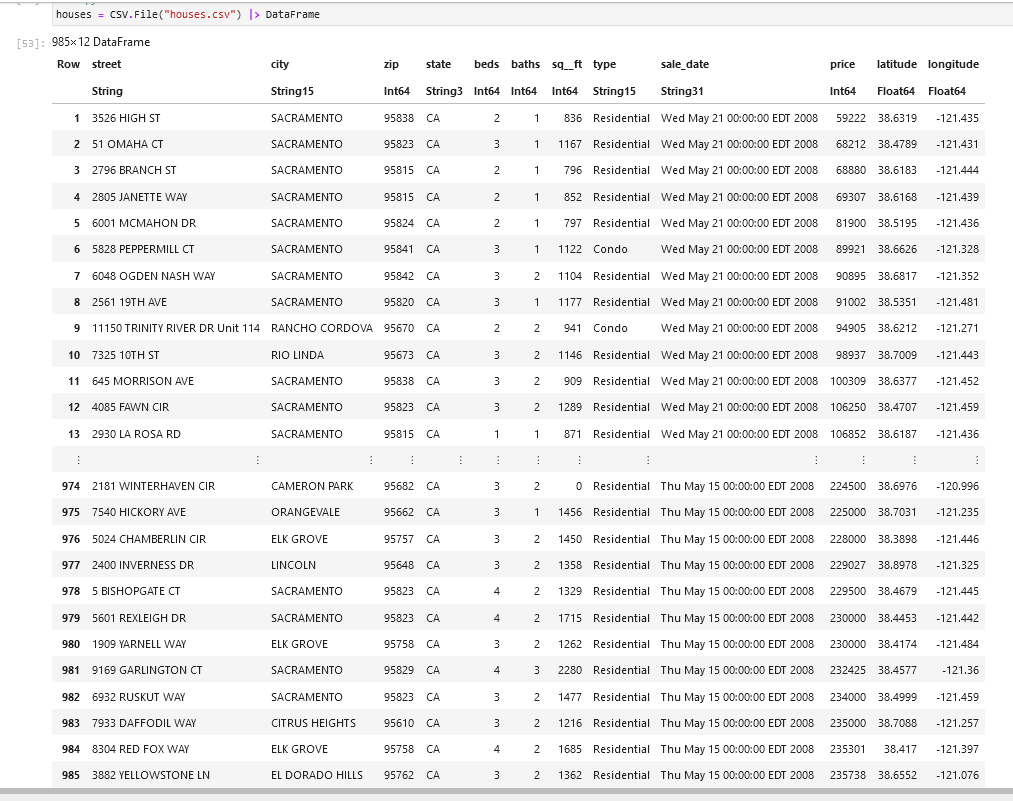
Реализуем кластеризация данных методом k-средних ([??-??]).



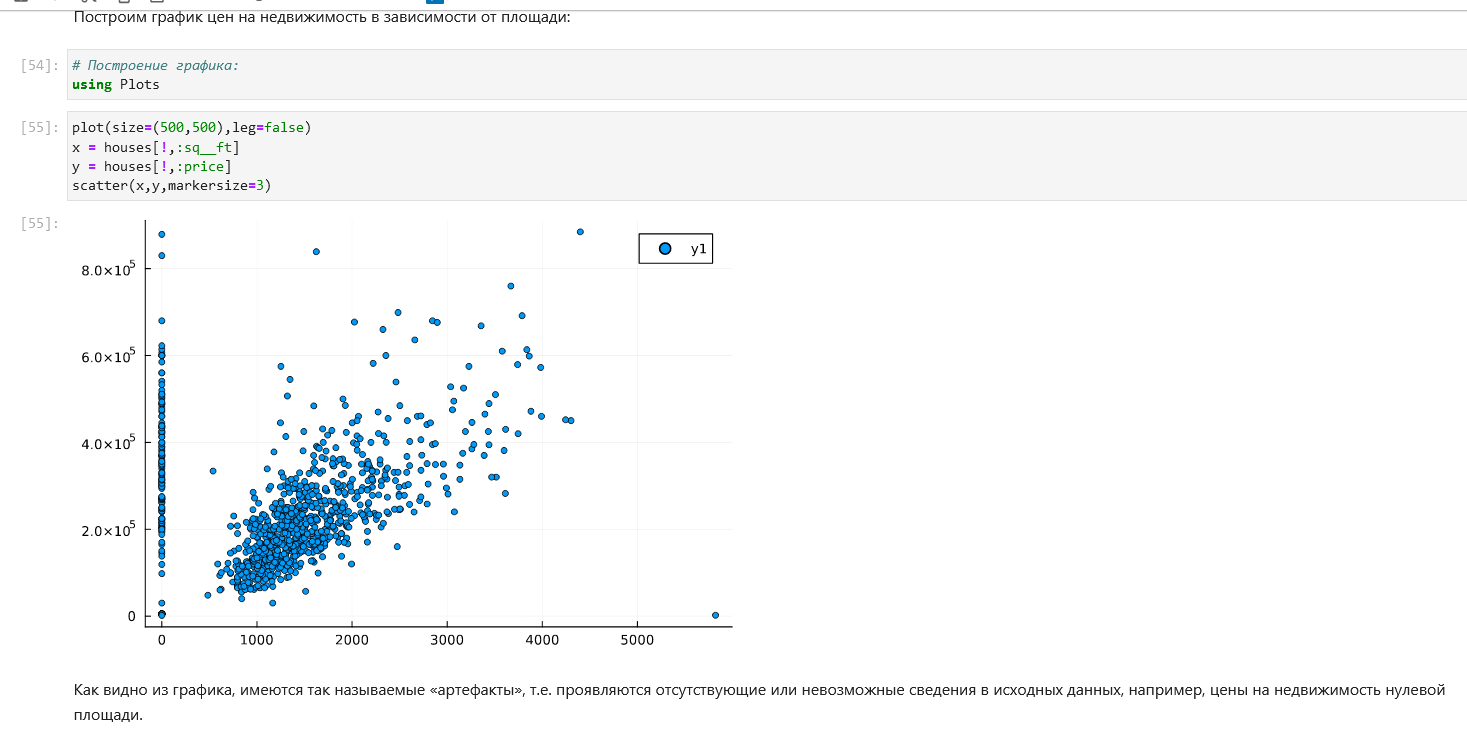
Кластеризация данных. Метод k-средних (1)



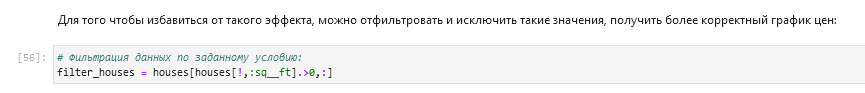
Кластеризация данных. Метод k-средних (2)



Кластеризация данных. Метод k-средних (3)



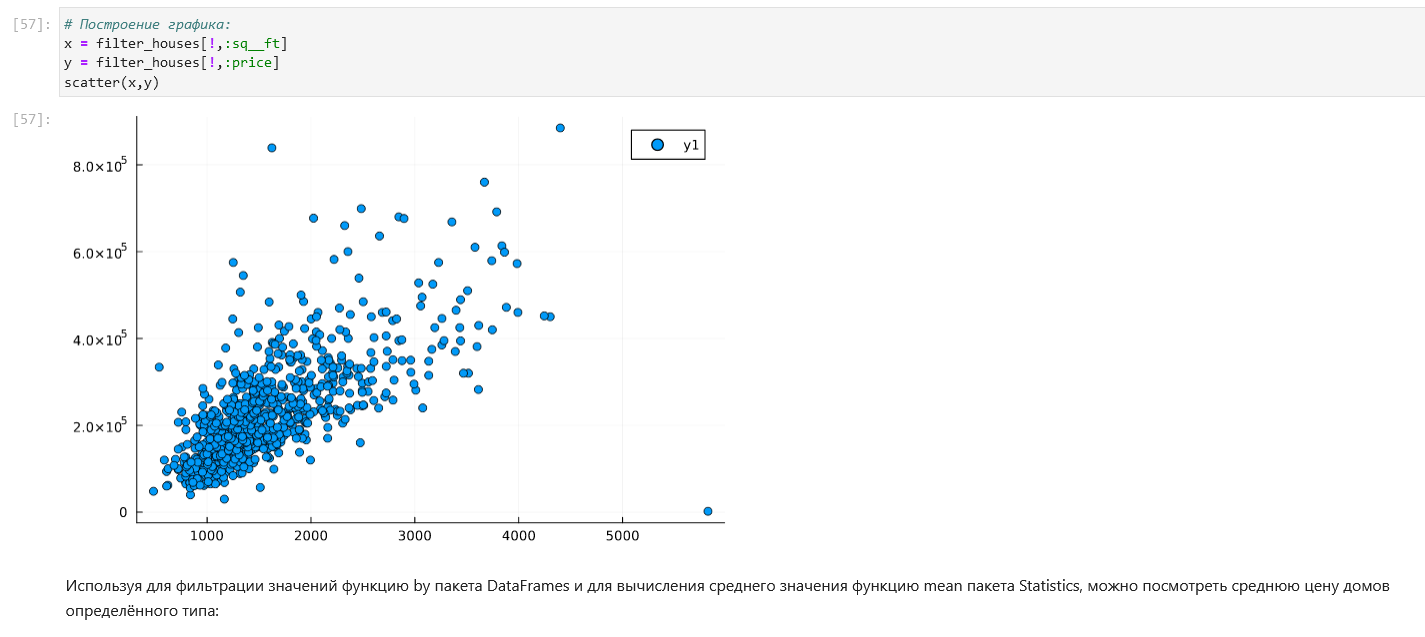
Кластеризация данных. Метод k-средних (4)



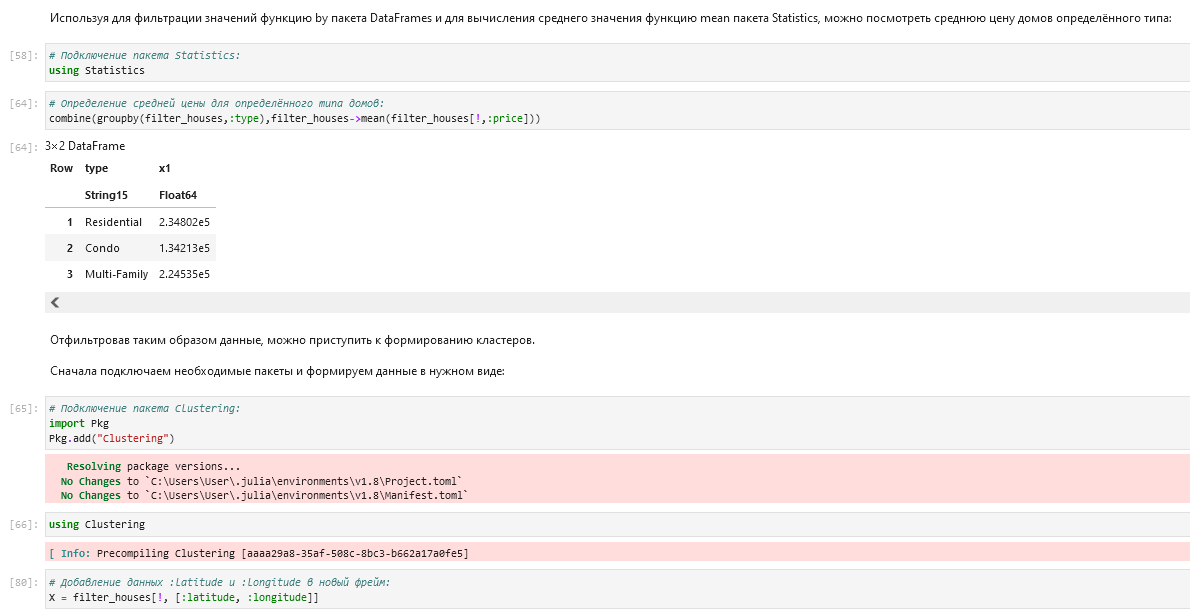
Кластеризация данных. Метод k-средних (5)



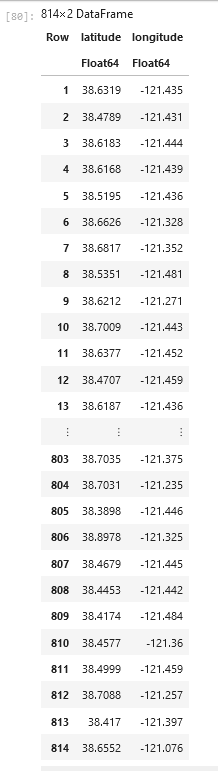
Кластеризация данных. Метод k-средних (6)



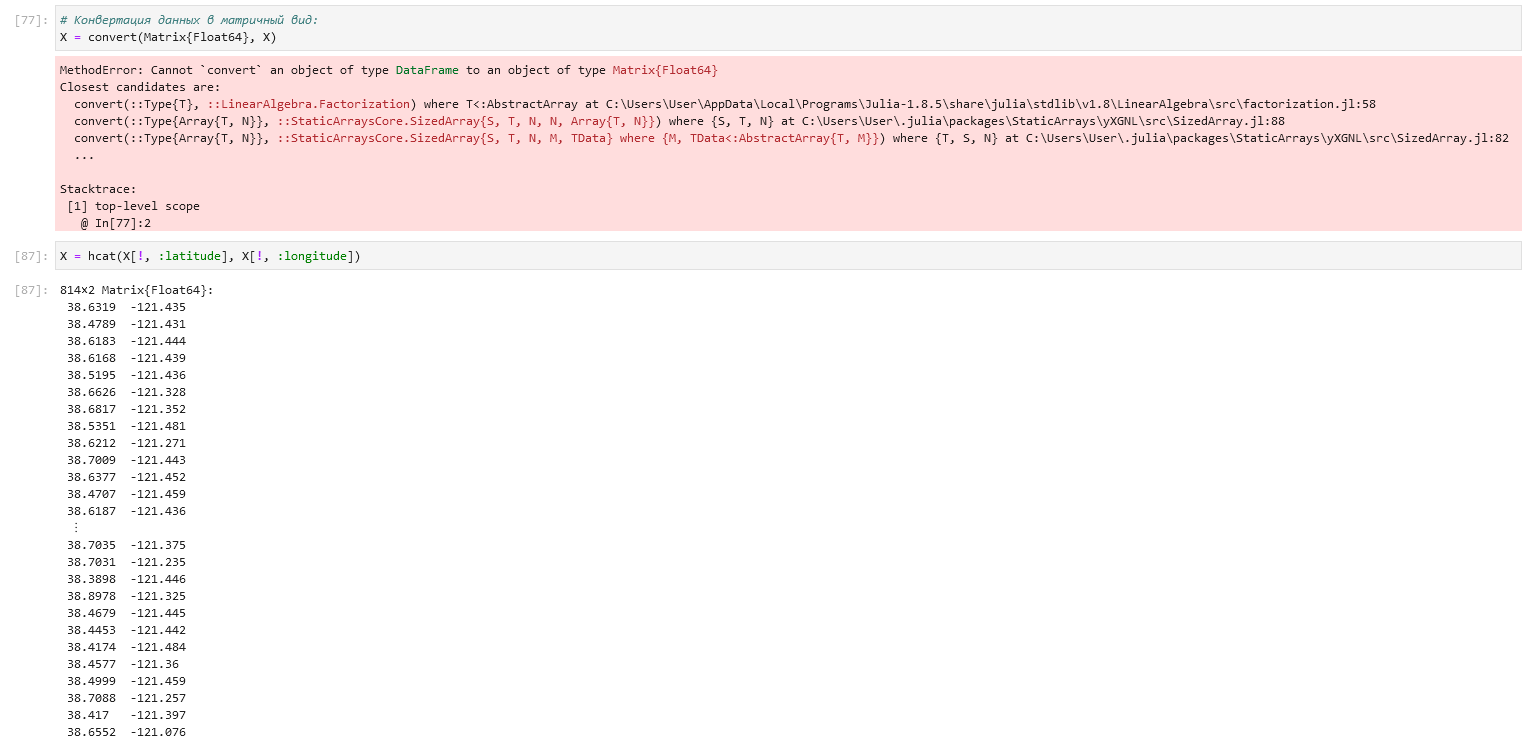
Кластеризация данных. Метод k-средних (7)



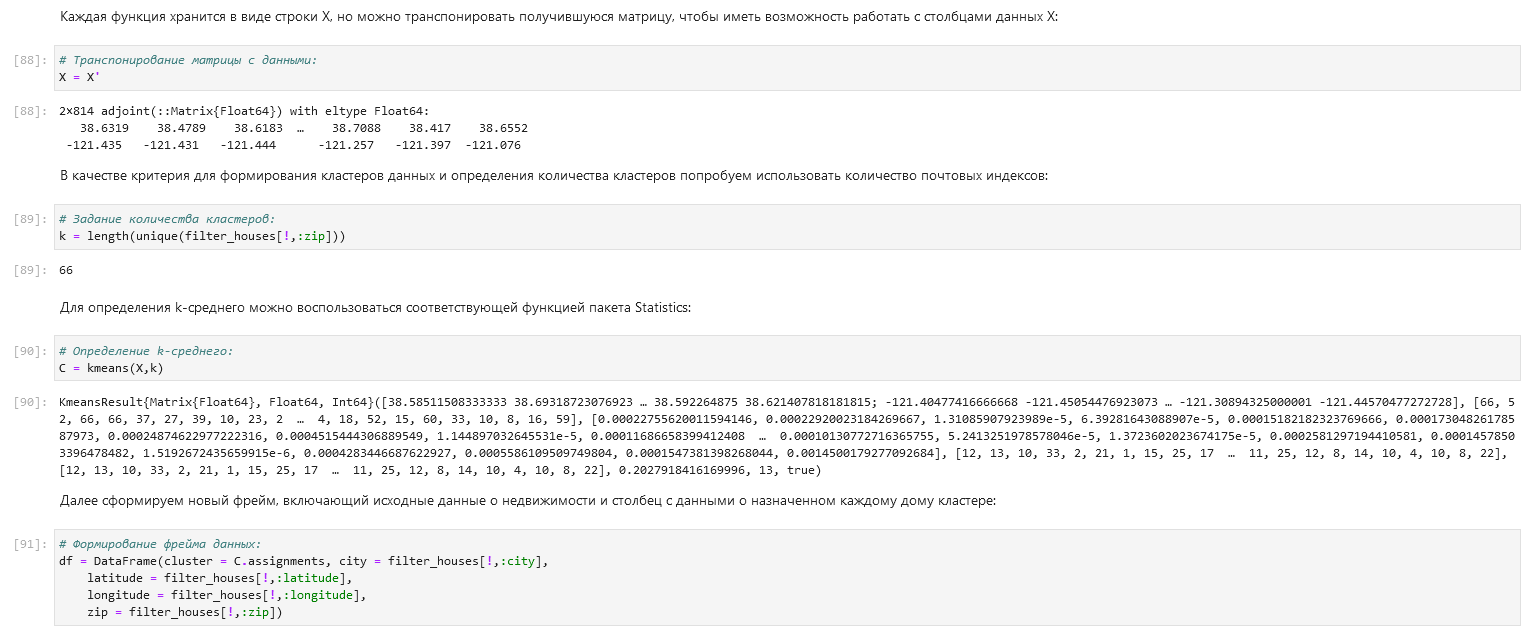
Кластеризация данных. Метод k-средних (8)



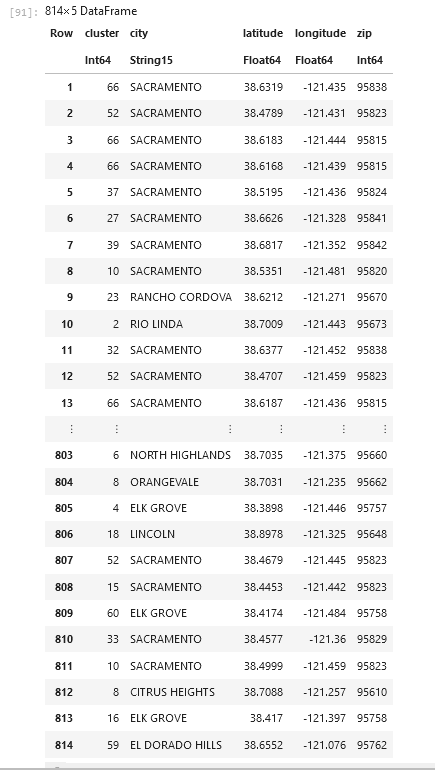
Кластеризация данных. Метод k-средних (9)



Кластеризация данных. Метод k-средних (10)



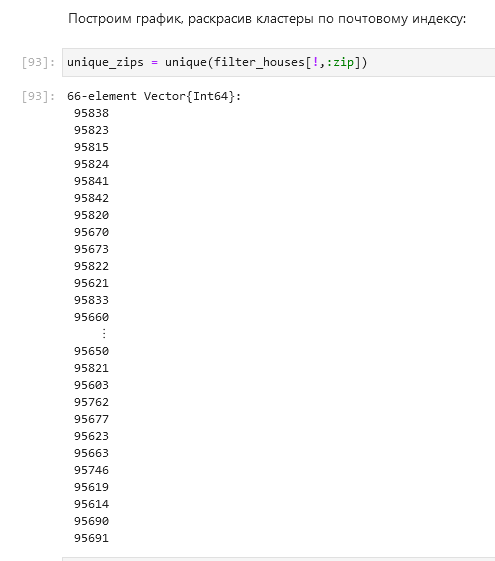
Кластеризация данных. Метод k-средних (11)



Кластеризация данных. Метод k-средних (12)



Кластеризация данных. Метод k-средних (13)



Кластеризация данных. Метод k-средних (14)

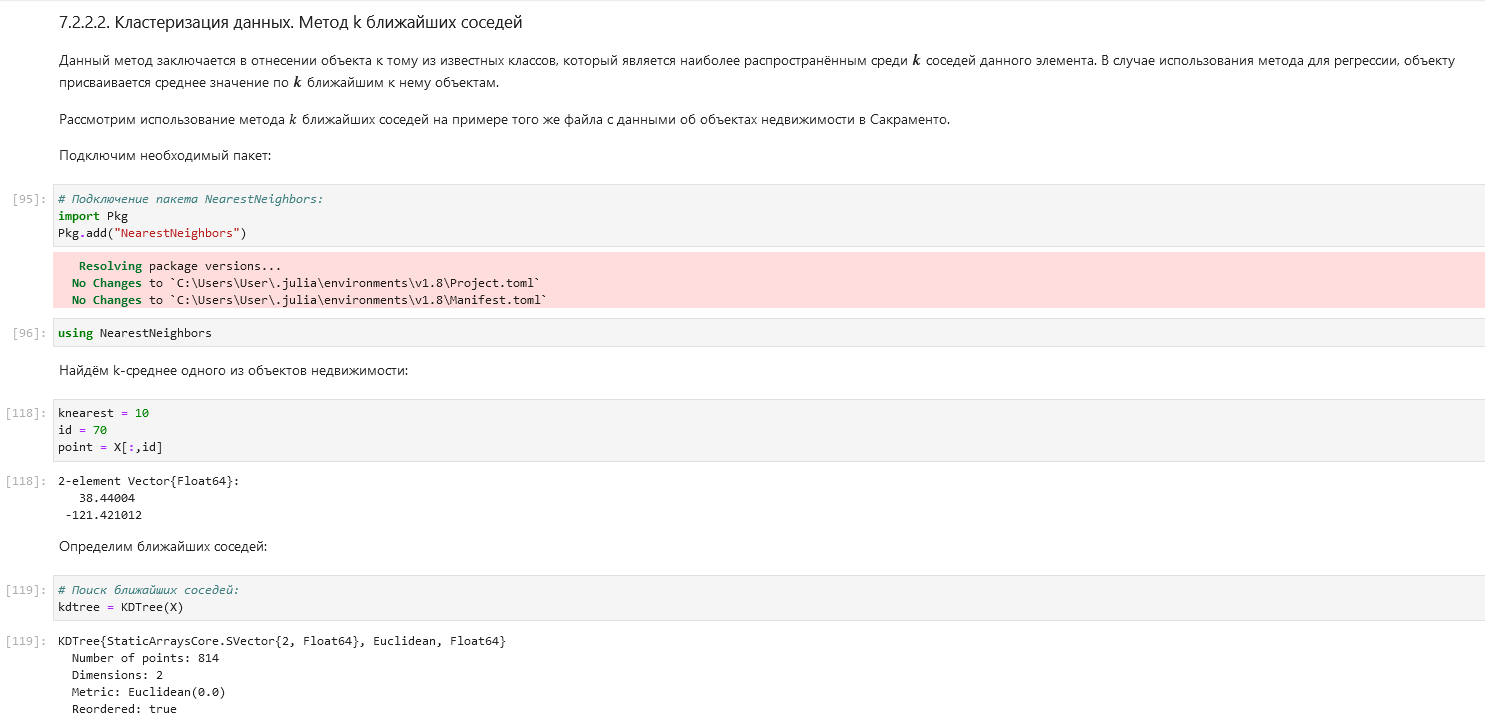


Кластеризация данных. Метод k-средних (15)

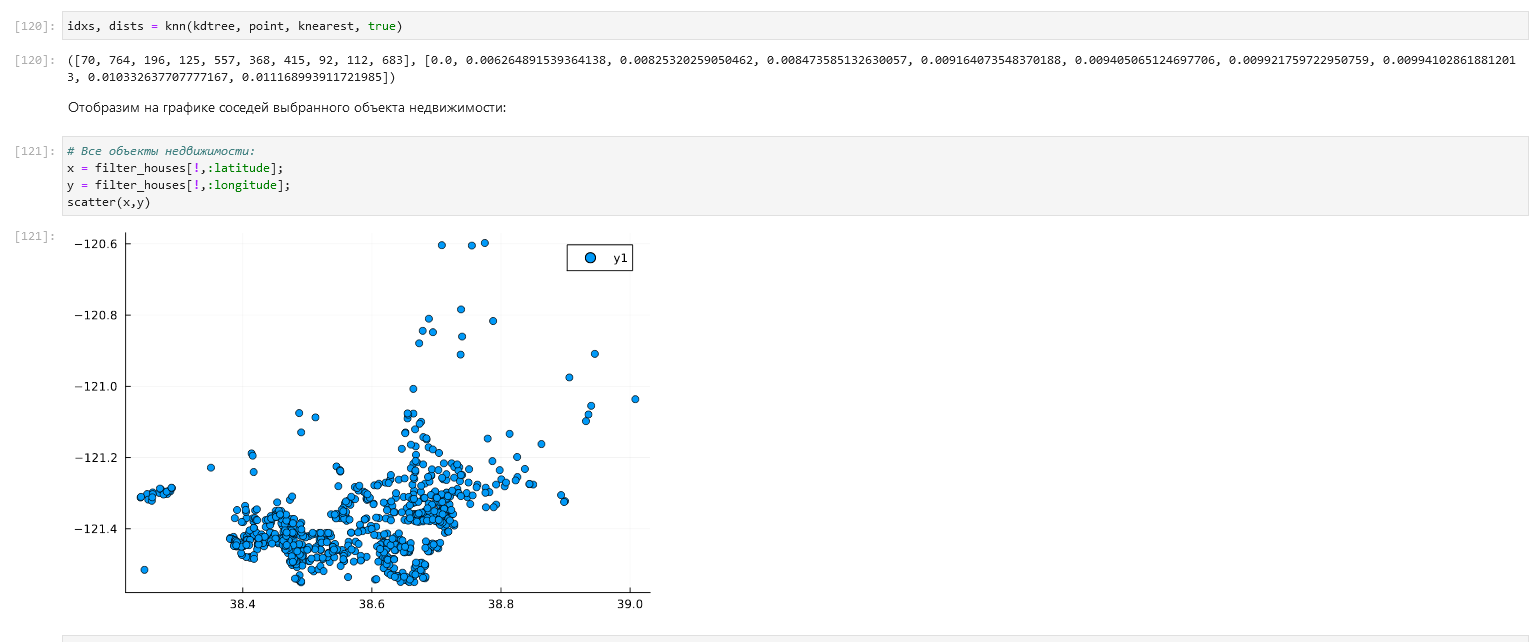
#### 2.1.2.2 Кластеризация данных. Метод k ближайших соседей

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.

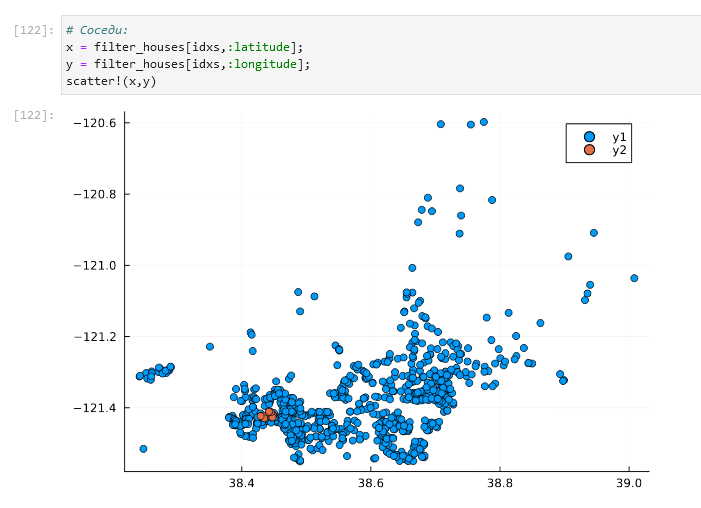
Реализуем кластеризация данных методом k ближайших соседей ([??-??]).



Кластеризация данных. Метод k ближайших соседей (1)



Кластеризация данных. Метод k ближайших соседей (2)



Кластеризация данных. Метод k ближайших соседей (3)

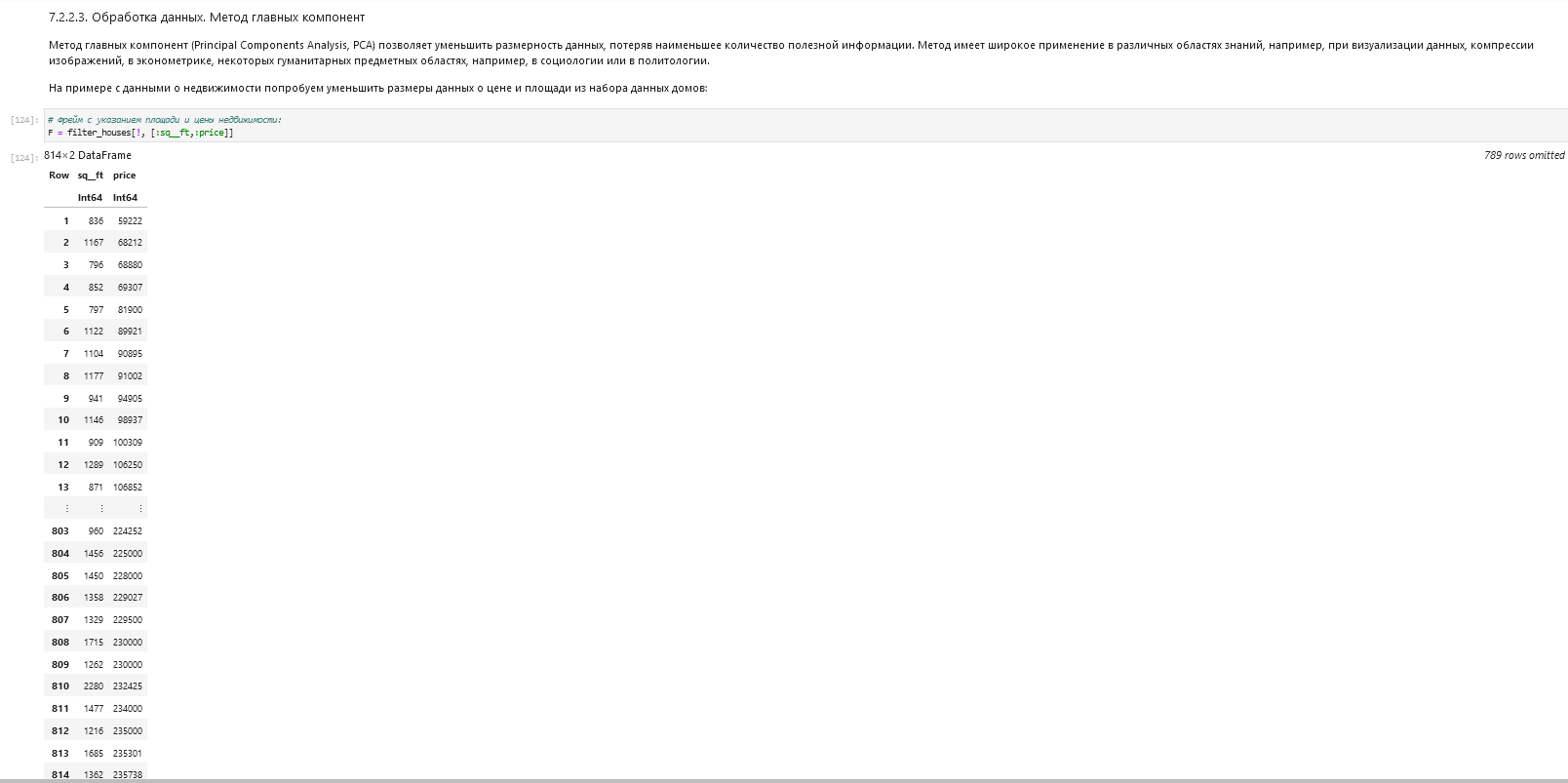


Кластеризация данных. Метод k ближайших соседей (4)

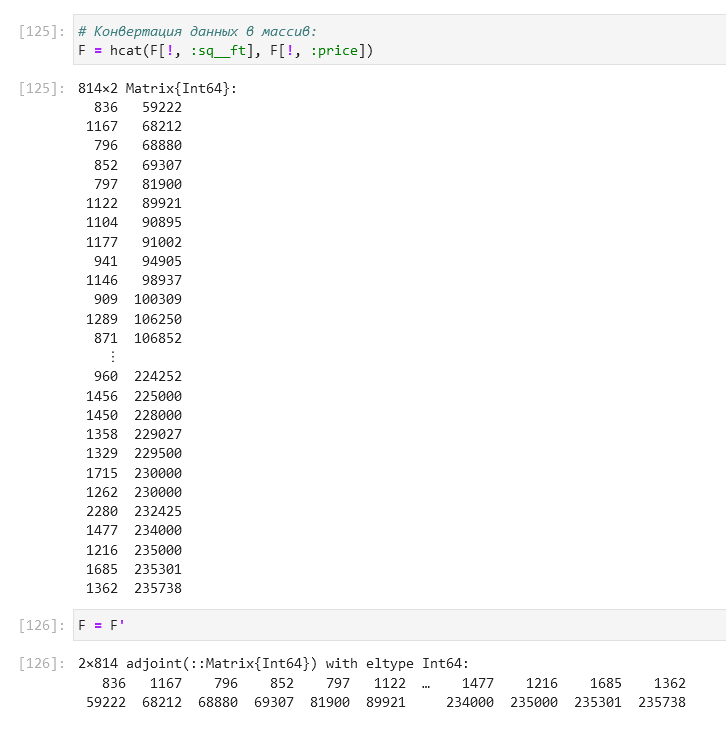
#### 2.1.2.3 Обработка данных. Метод главных компонент

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.

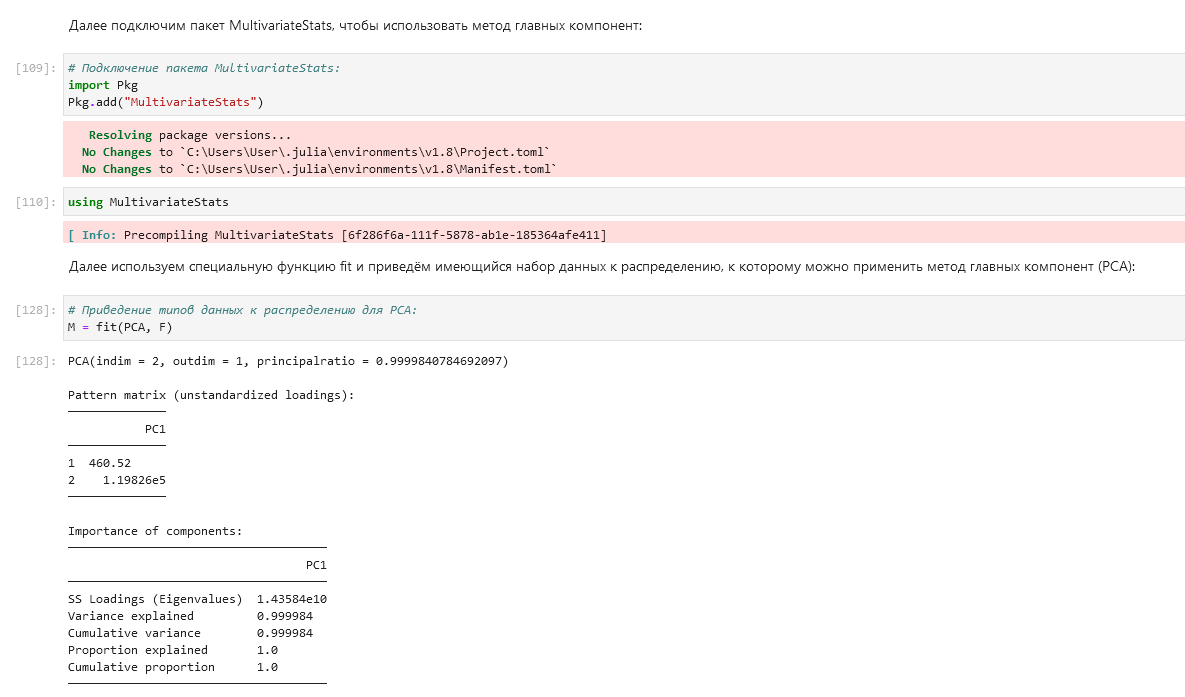
Реализуем метод главных компонент ([??-??]).



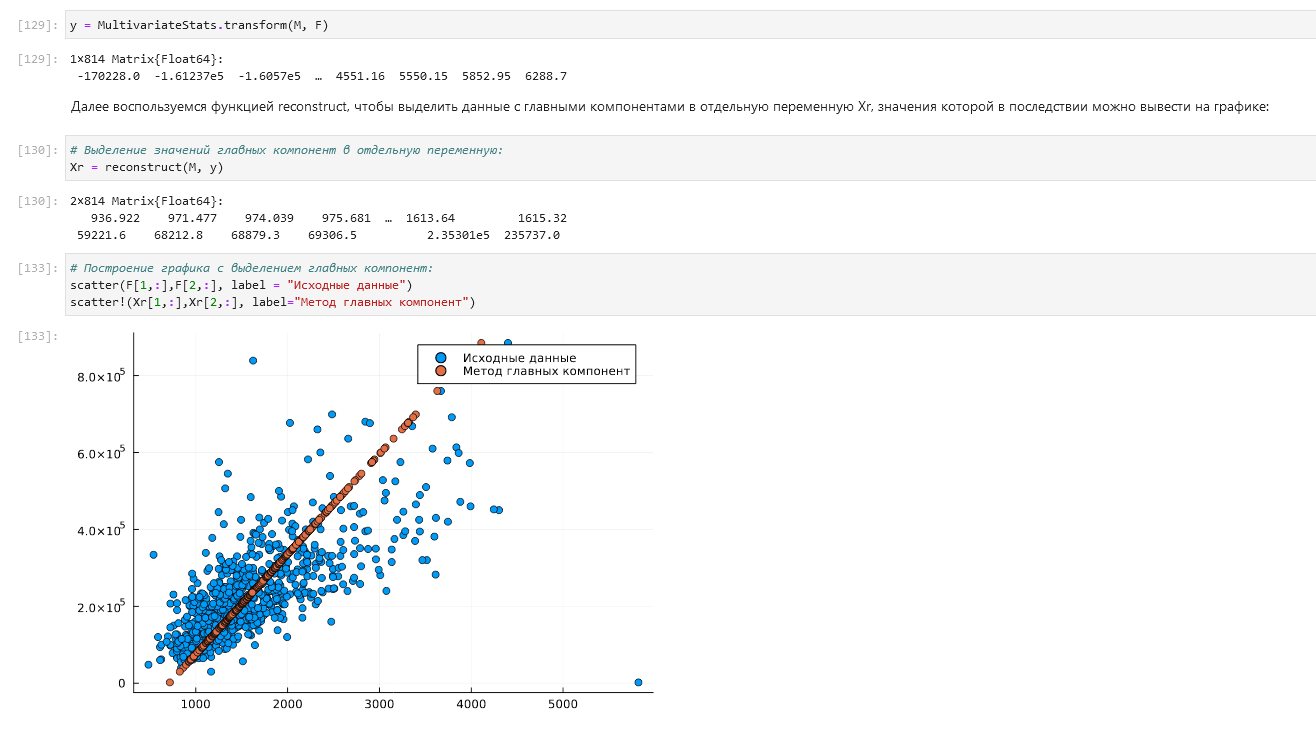
Обработка данных. Метод главных компонент (1)



Обработка данных. Метод главных компонент (2)



Обработка данных. Метод главных компонент (3)



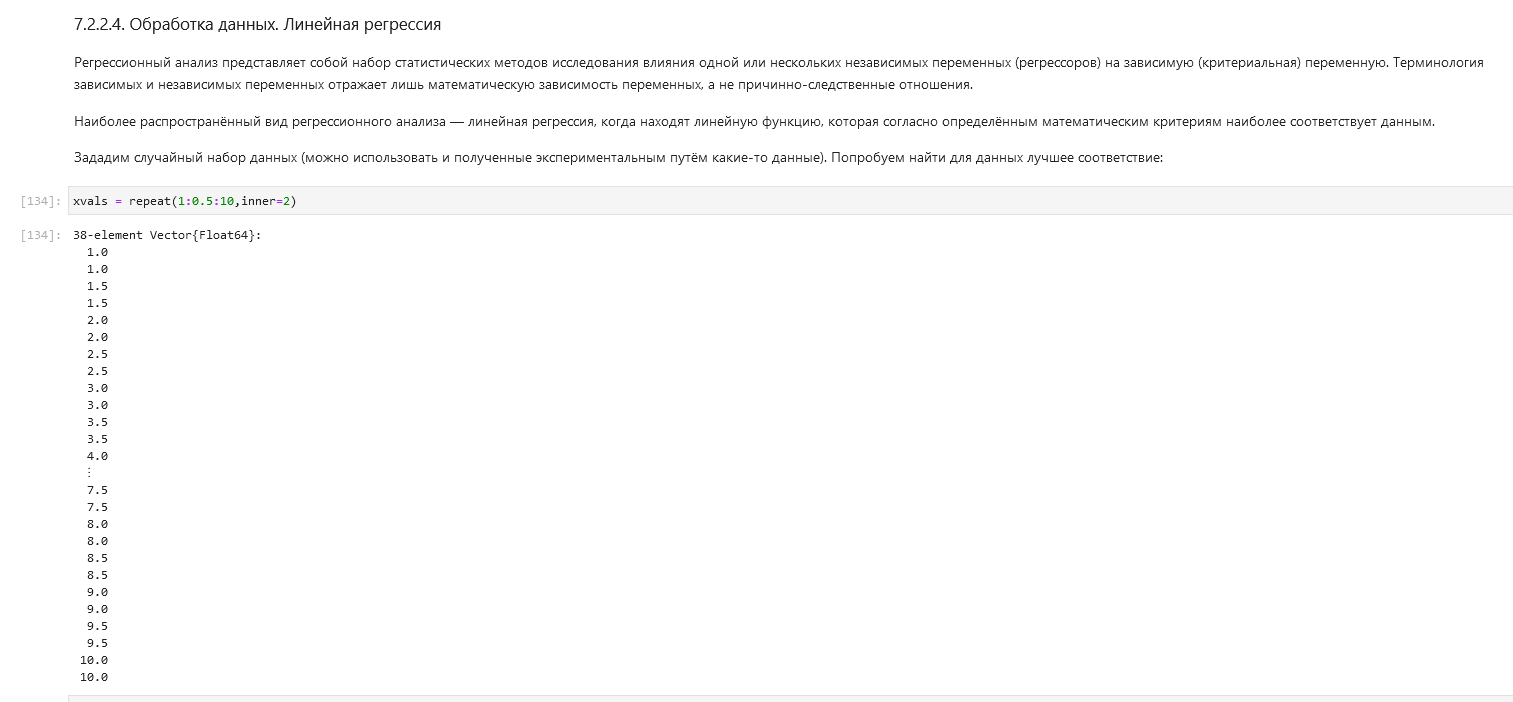
Обработка данных. Метод главных компонент (4)

#### 2.1.2.4 Обработка данных. Линейная регрессия

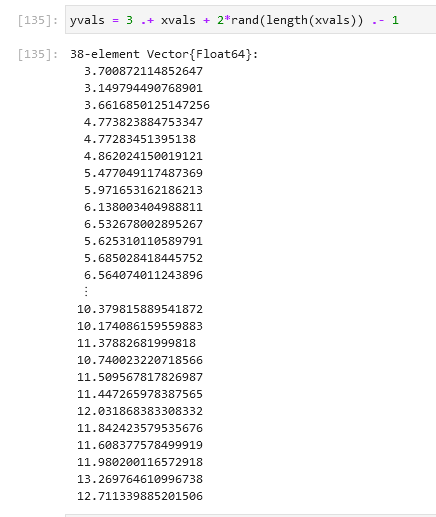
Регрессионный анализ представляет собой набор статистических методов исследования влияния одной или нескольких независимых переменных (регрессоров) на зависимую (критериальная) переменную. Терминология зависимых и независимых переменных отражает лишь математическую зависимость переменных, а не причинно-следственные отношения.

Наиболее распространённый вид регрессионного анализа — линейная регрессия, когда находят линейную функцию, которая согласно определённым математическим критериям наиболее соответствует данным.

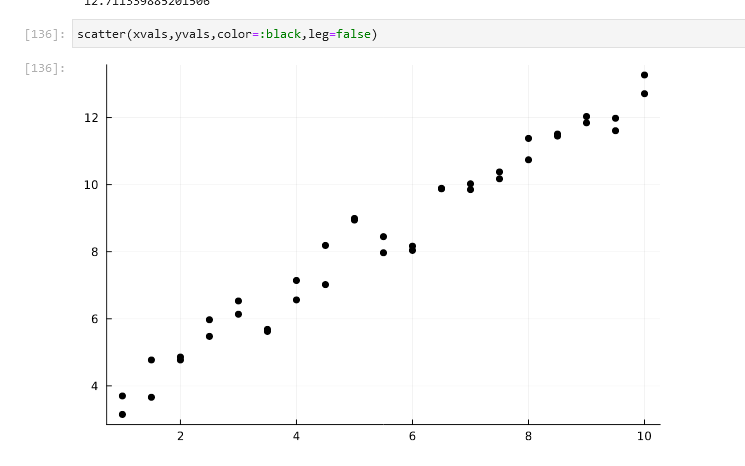
Реализуем линейную регрессию ([??-??]).



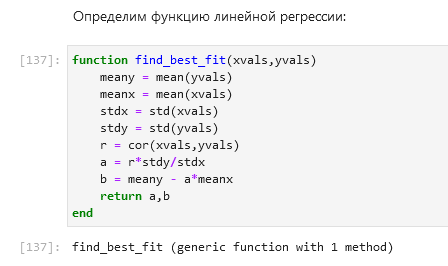
Обработка данных. Линейная регрессия (1)



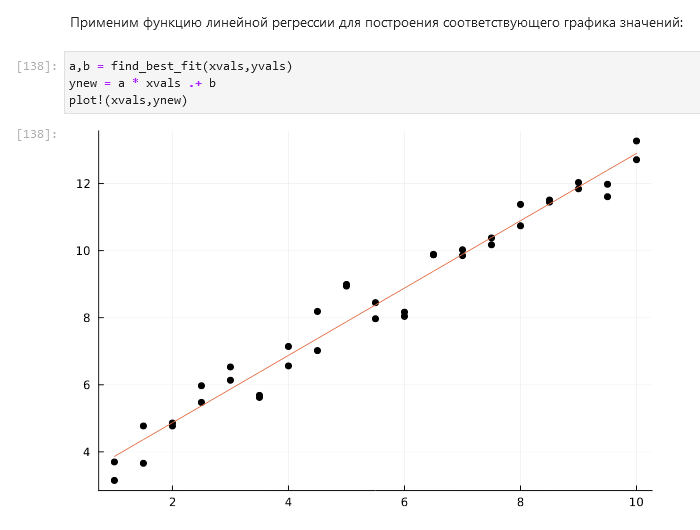
Обработка данных. Линейная регрессия (2)



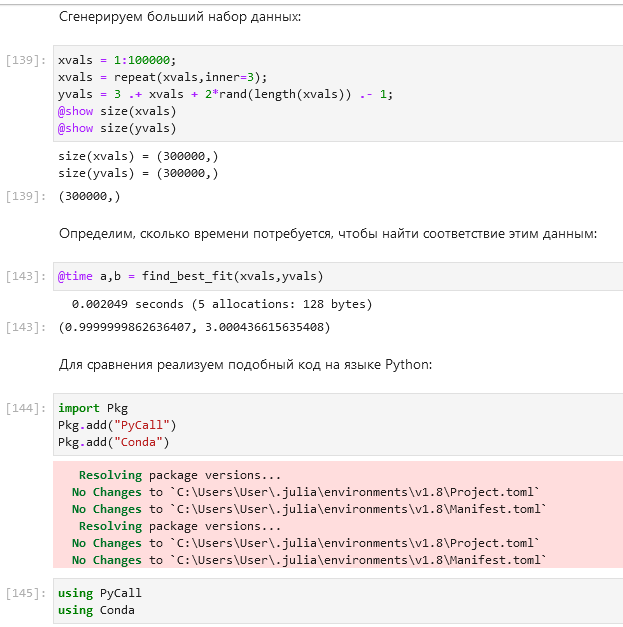
Обработка данных. Линейная регрессия (3)



Обработка данных. Линейная регрессия (4)



Обработка данных. Линейная регрессия (5)



Обработка данных. Линейная регрессия (6)



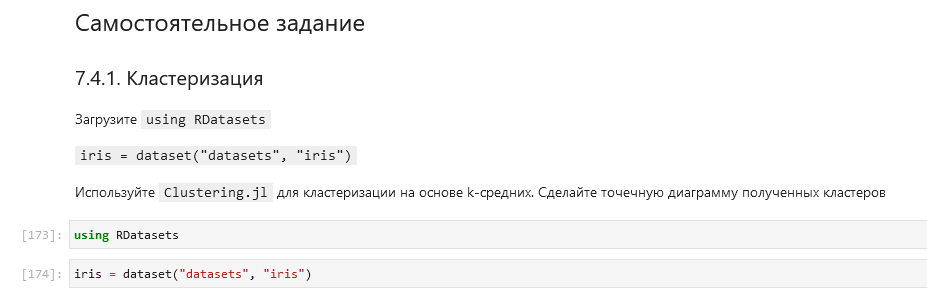
Обработка данных. Линейная регрессия (7)

## 2.2 Самостоятельная работа

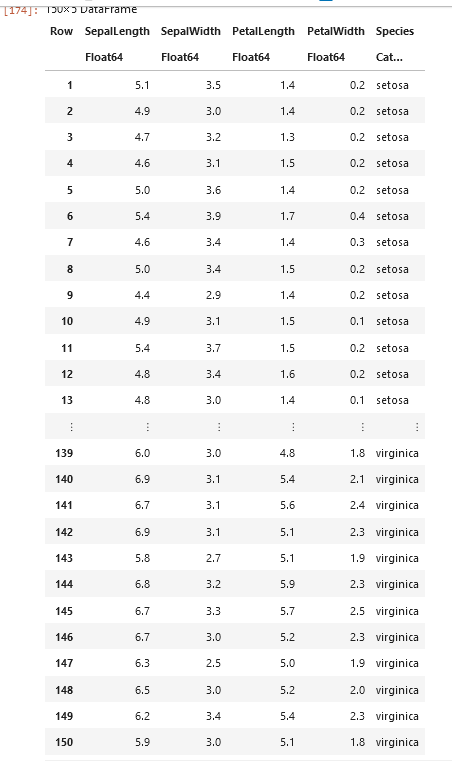
### 2.2.1 Кластеризация

Используя Clustering.jl для кластеризации на основе k-средних, сделаем точечную диаграмму полученных кластеров.

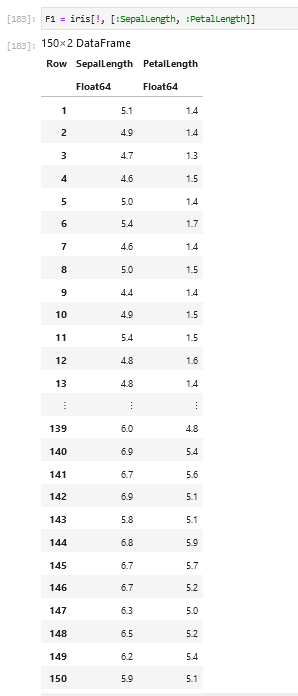
Для таблицы данных об ирисах реализуем кластеризацию по количеству видов ирисов в зависимости от длин стебля и лепестков, а также ширины стебля и лепестков ([??-??]). Итоговая кластеризация, где справа изображены кластеры по исходным видам ирисов (из данных таблицы), а слева результат кластеризации на основе k-средних ([??])



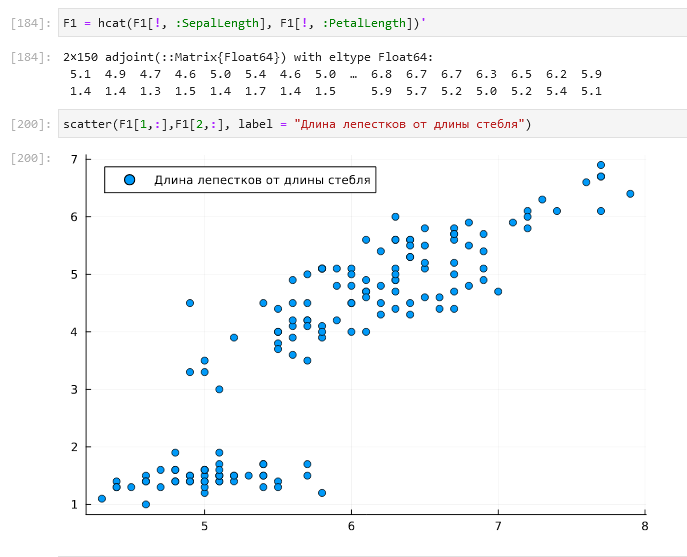
Задание 7.4.1. Кластеризация (1)



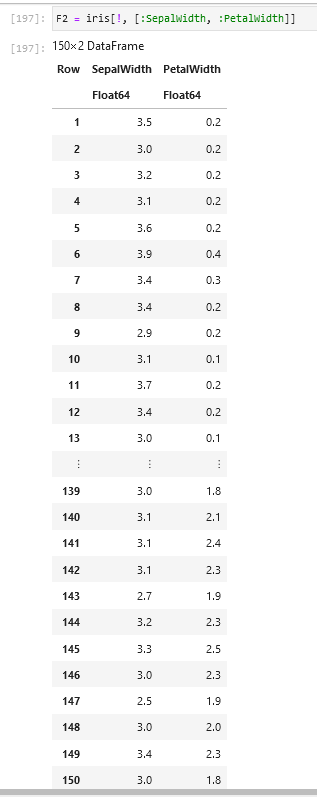
Задание 7.4.1. Кластеризация (2)



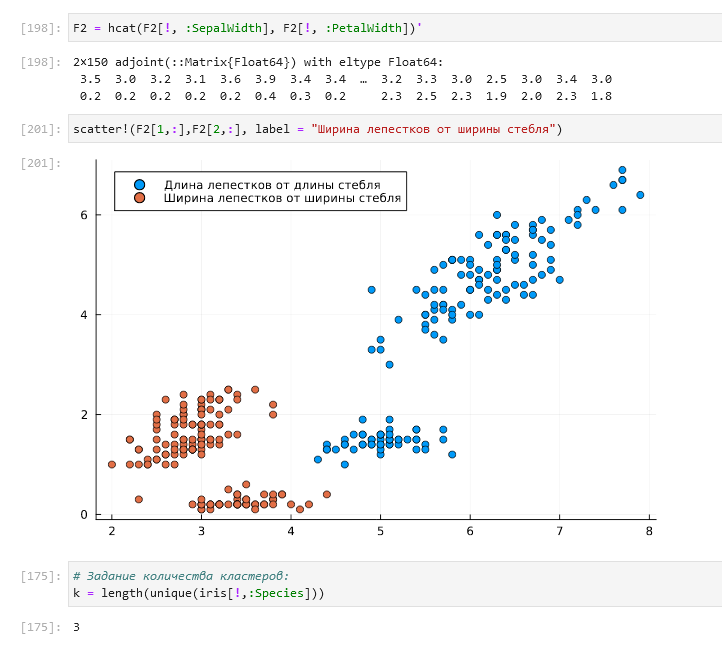
Задание 7.4.1. Кластеризация (3)



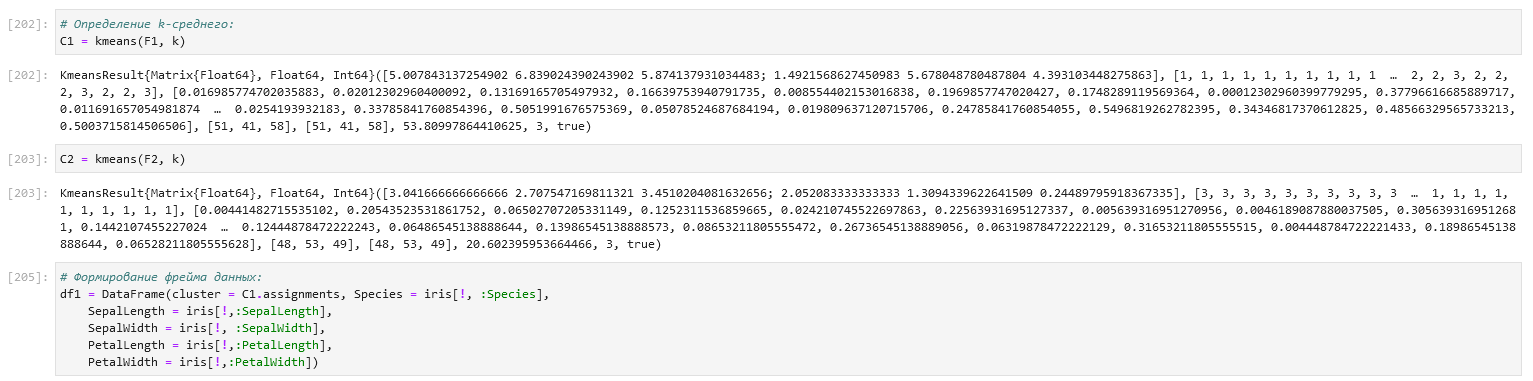
Задание 7.4.1. Кластеризация (4)



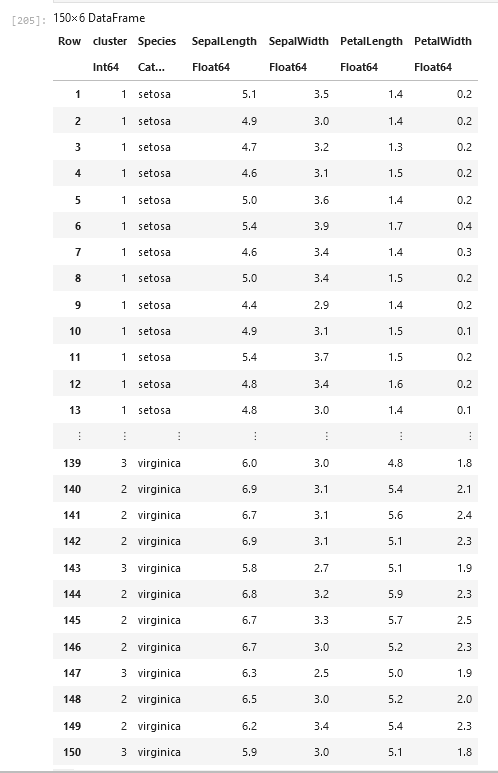
Задание 7.4.1. Кластеризация (5)



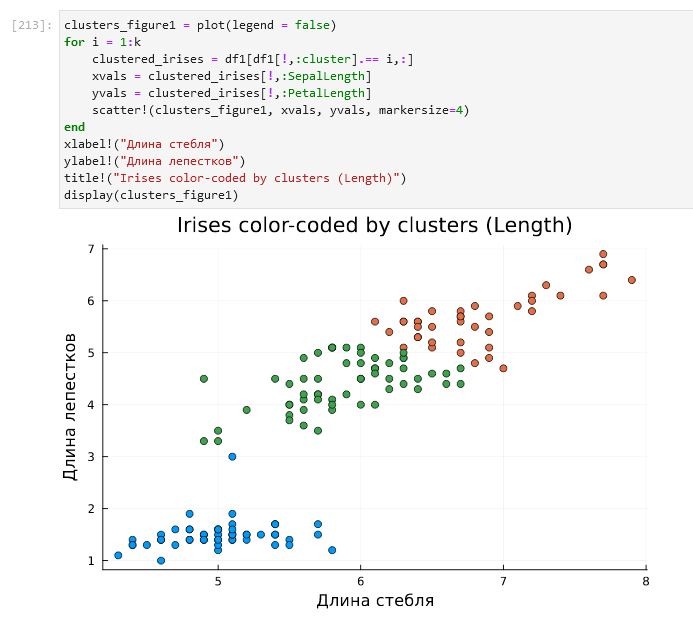
Задание 7.4.1. Кластеризация (6)



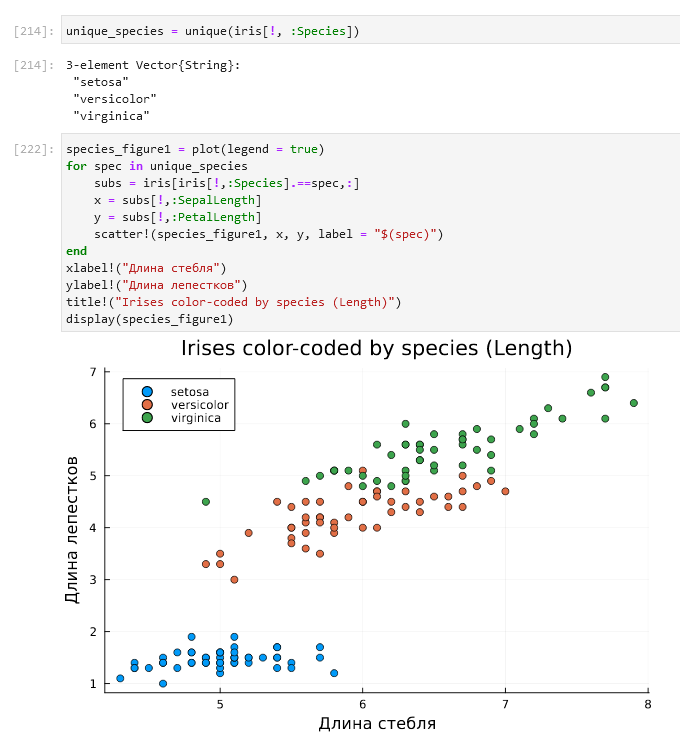
Задание 7.4.1. Кластеризация (7)



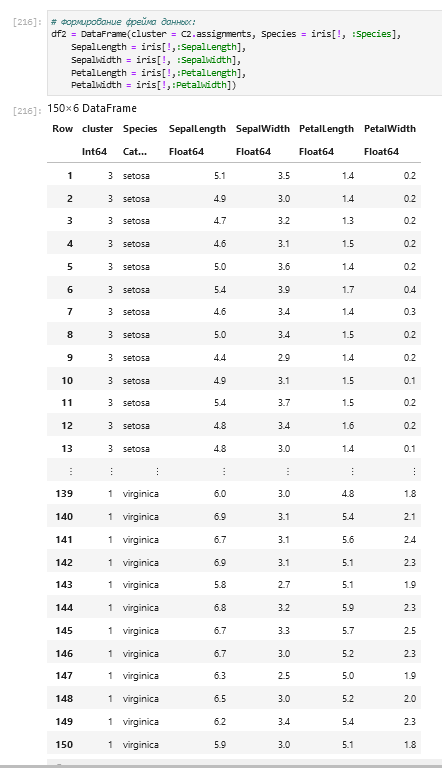
Задание 7.4.1. Кластеризация (8)



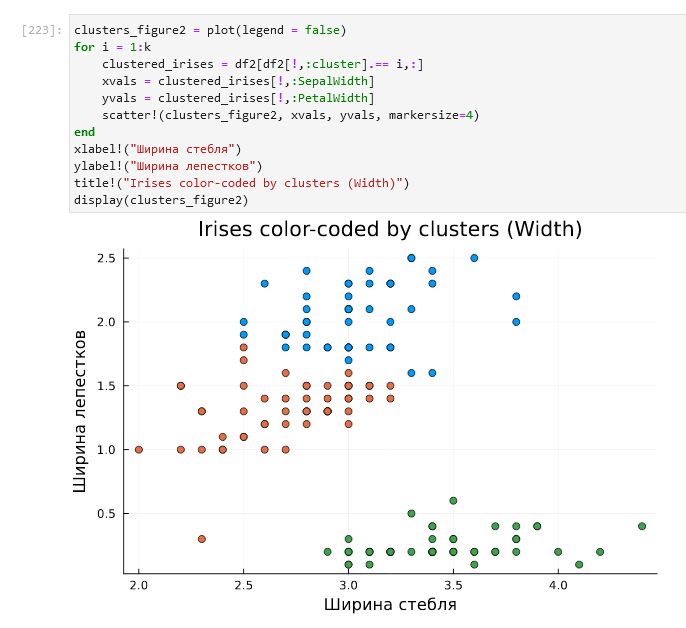
Задание 7.4.1. Кластеризация (9)



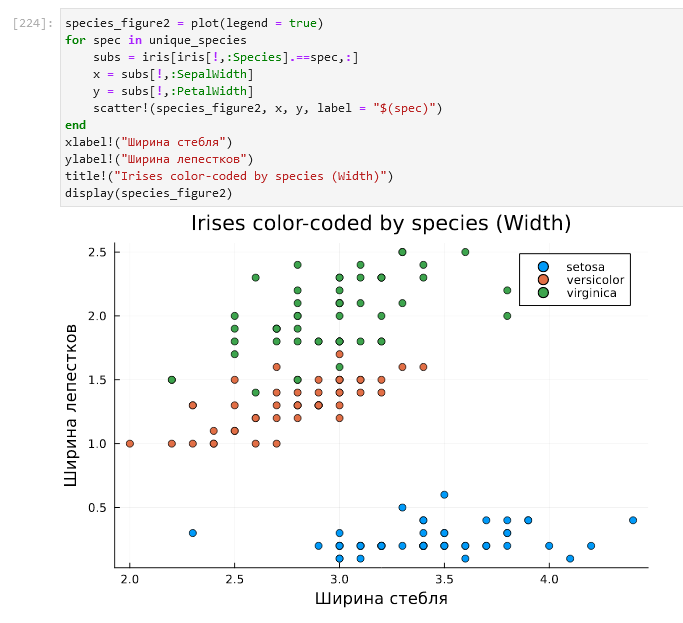
Задание 7.4.1. Кластеризация (10)



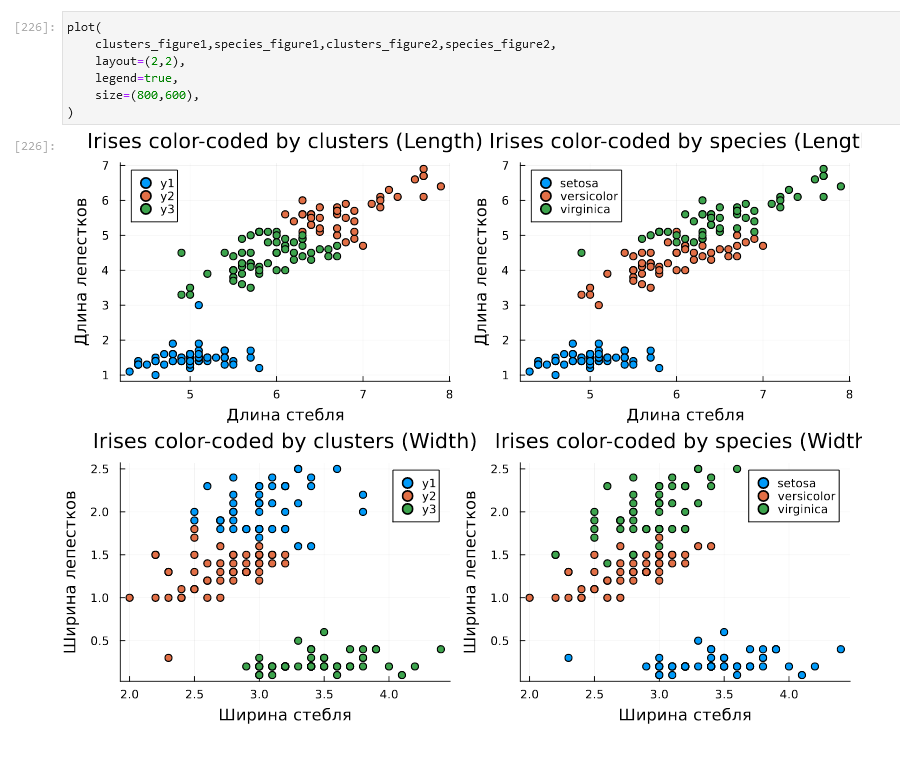
Задание 7.4.1. Кластеризация (11)



Задание 7.4.1. Кластеризация (12)



Задание 7.4.1. Кластеризация (13)

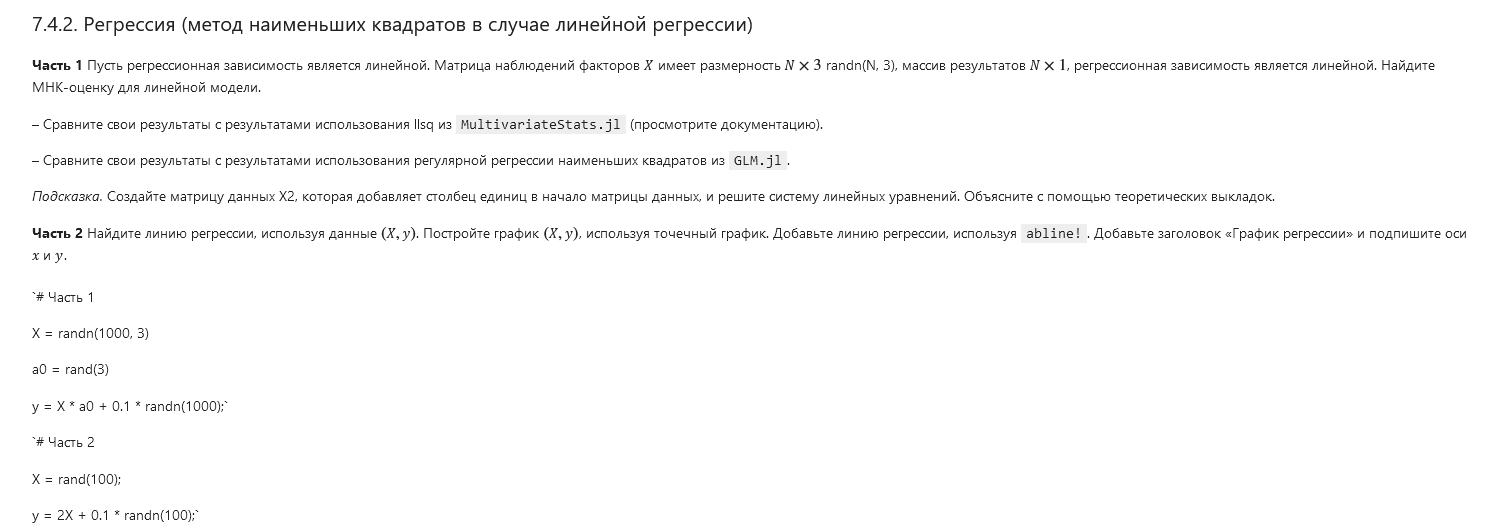


Задание 7.4.1. Кластеризация (14)

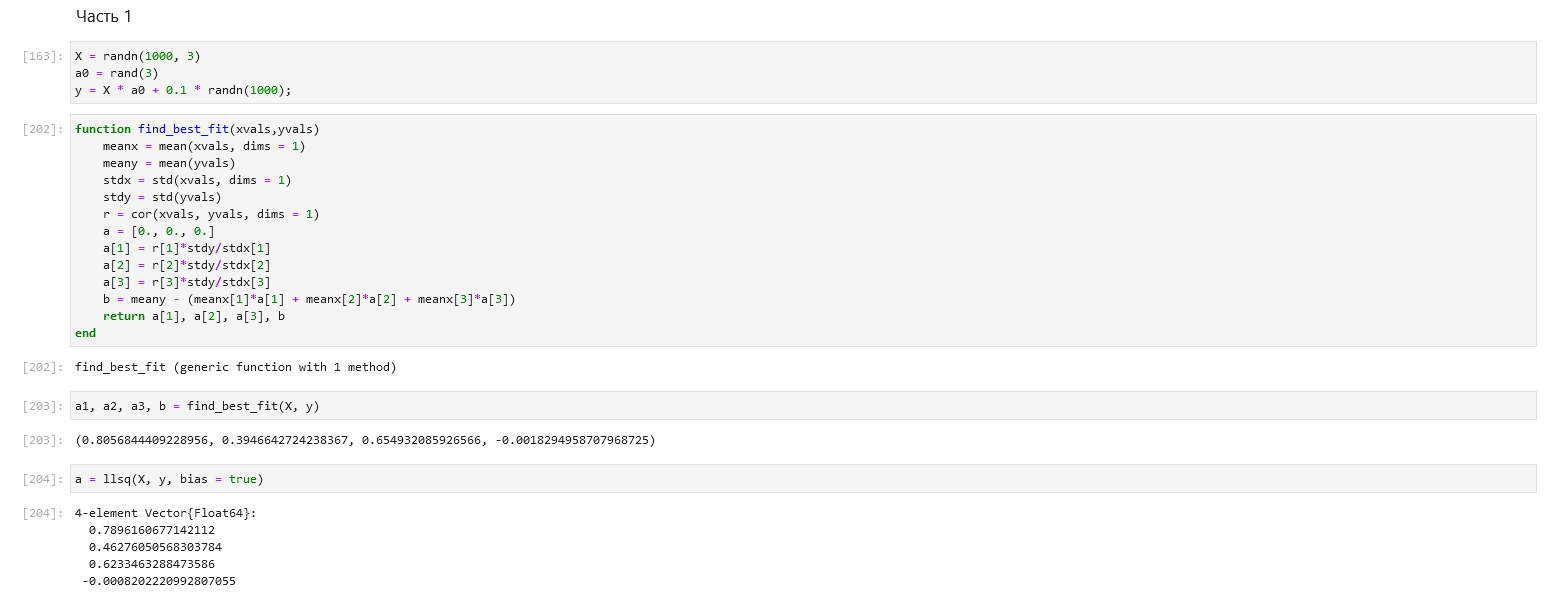
### 2.2.2 Регрессия (метод наименьших квадратов в случае линейной регрессии)

**Часть 1:**

В условиях задания ([??]) выполним линейную регрессию своим методом и методом llsq из пакета MultivariateStats.jl ([??]), загрузим пакет GLM.jl ([??]) и решим ту же задачу методом регулярной регрессии из данного пакета ([??,??]).



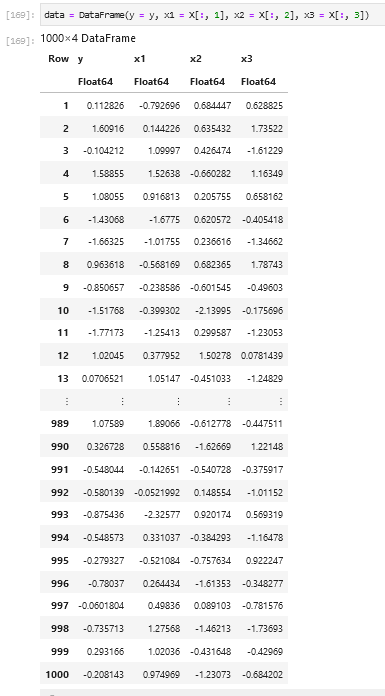
Задание 7.4.2. Часть 1. Регрессия (метод наименьших квадратов в случае линейной регрессии) (1)



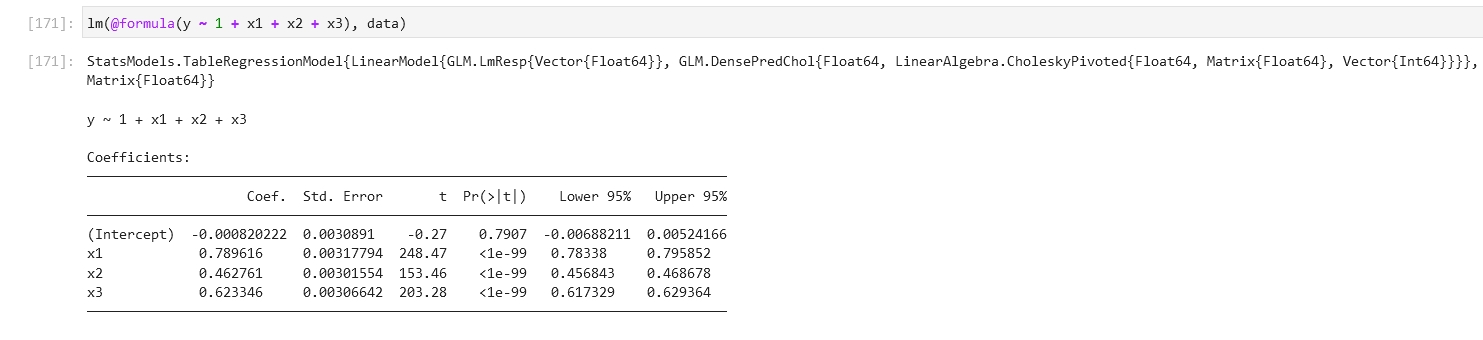
Задание 7.4.2. Часть 1. Регрессия (метод наименьших квадратов в случае линейной регрессии) (2)



Задание 7.4.2. Часть 1. Регрессия (метод наименьших квадратов в случае линейной регрессии) (3)



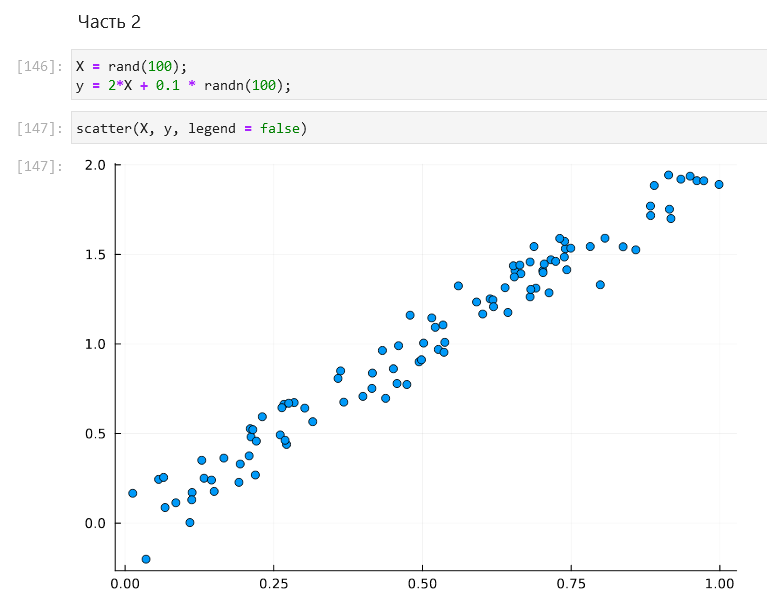
Задание 7.4.2. Часть 1. Регрессия (метод наименьших квадратов в случае линейной регрессии) (4)



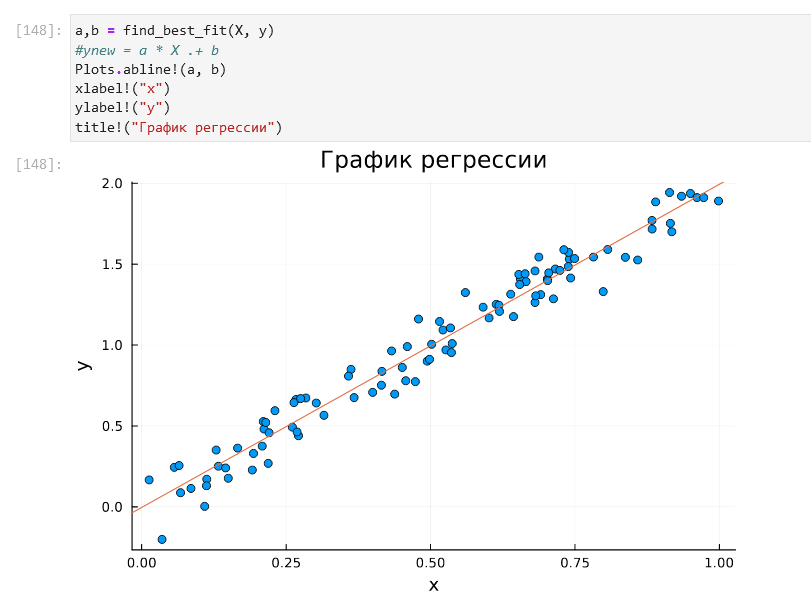
Задание 7.4.2. Часть 1. Регрессия (метод наименьших квадратов в случае линейной регрессии) (5)

**Часть 2:**

Построим точечный график и осуществим линейную регрессию, построив регресионную прямую с помощью команды abline ([??,??]).



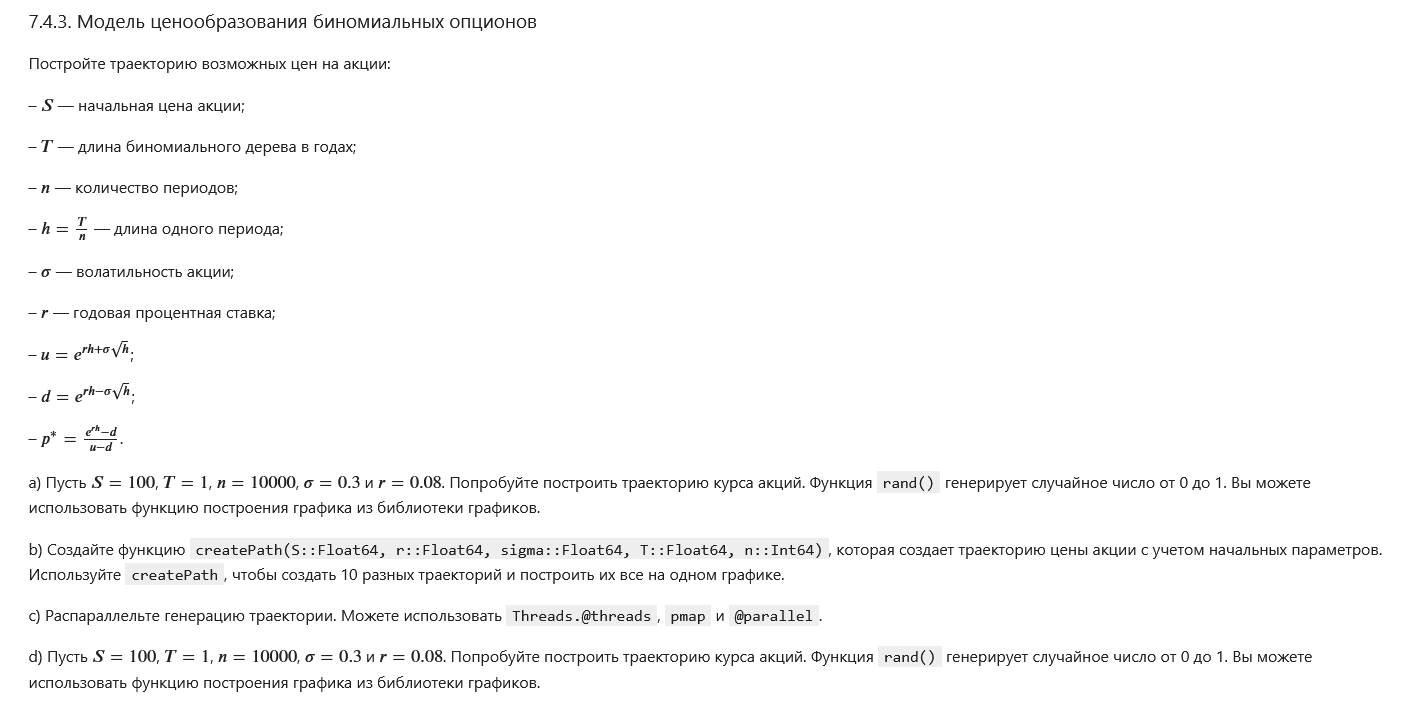
Задание 7.4.2. Часть 2. Регрессия (метод наименьших квадратов в случае линейной регрессии) (1)



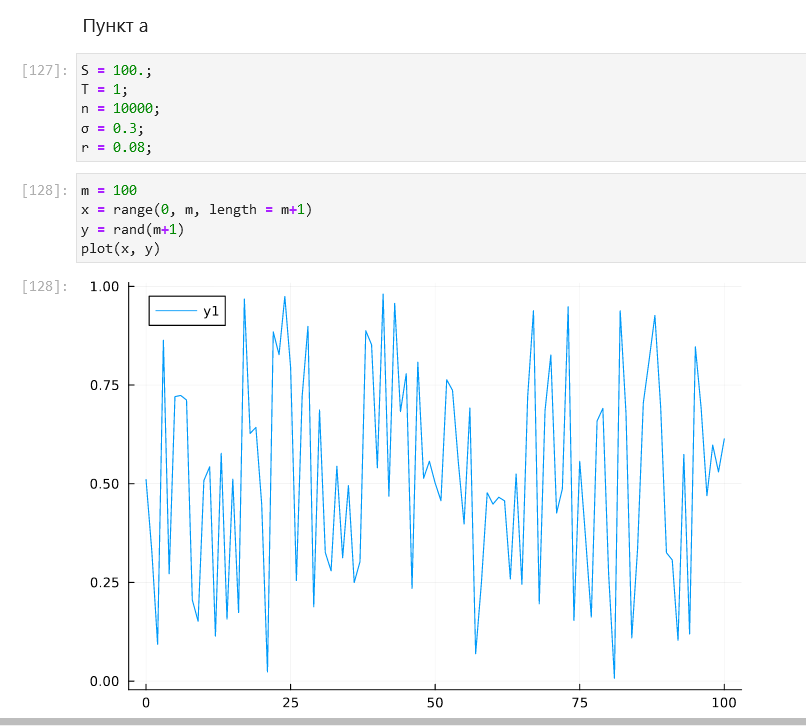
Задание 7.4.2. Часть 2. Регрессия (метод наименьших квадратов в случае линейной регрессии) (2)

### 2.2.3 Модель ценообразования биномиальных опционов

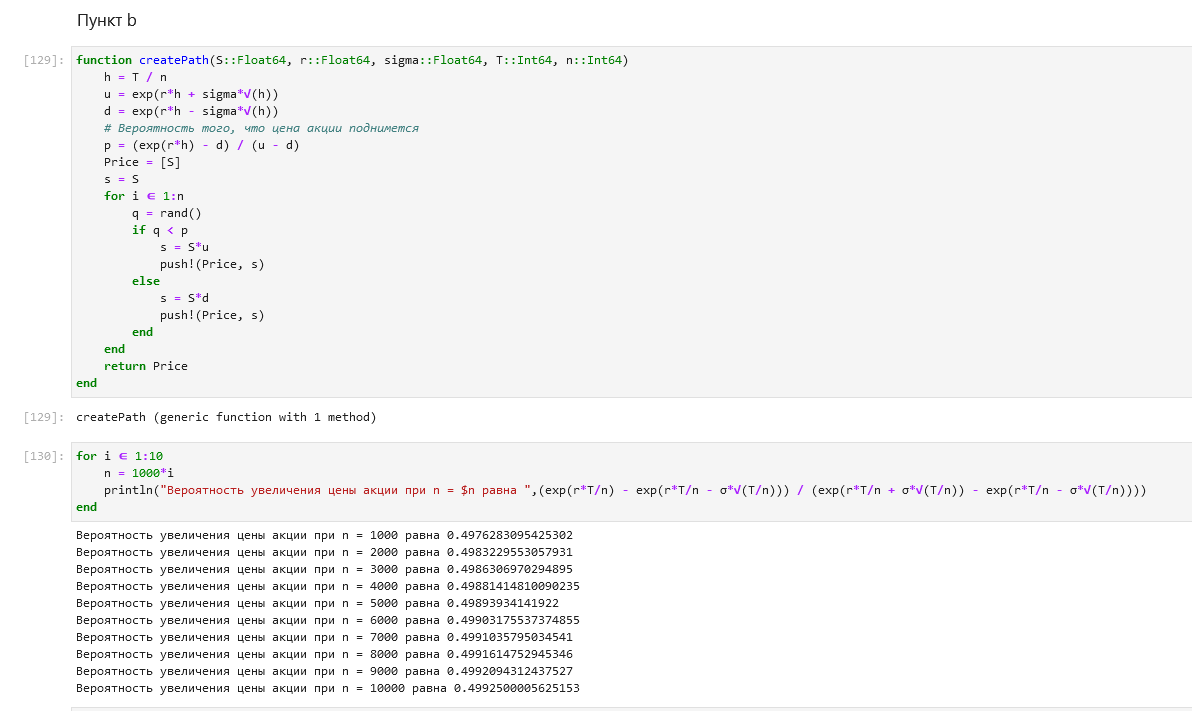
В условиях задачи ([??]), после построения случайной кривой ([??]), создадим функцию, которая высчитывает вероятность подъёма цены акции и множители при увеличении и сокращении её цены, после чего, разыгрывая вероятность подъема цены акции, составим список значений цен акции за заданный промежуток ([??]). Далее построим траектории цены акции в зависимости от числа периодов ([??,??]).



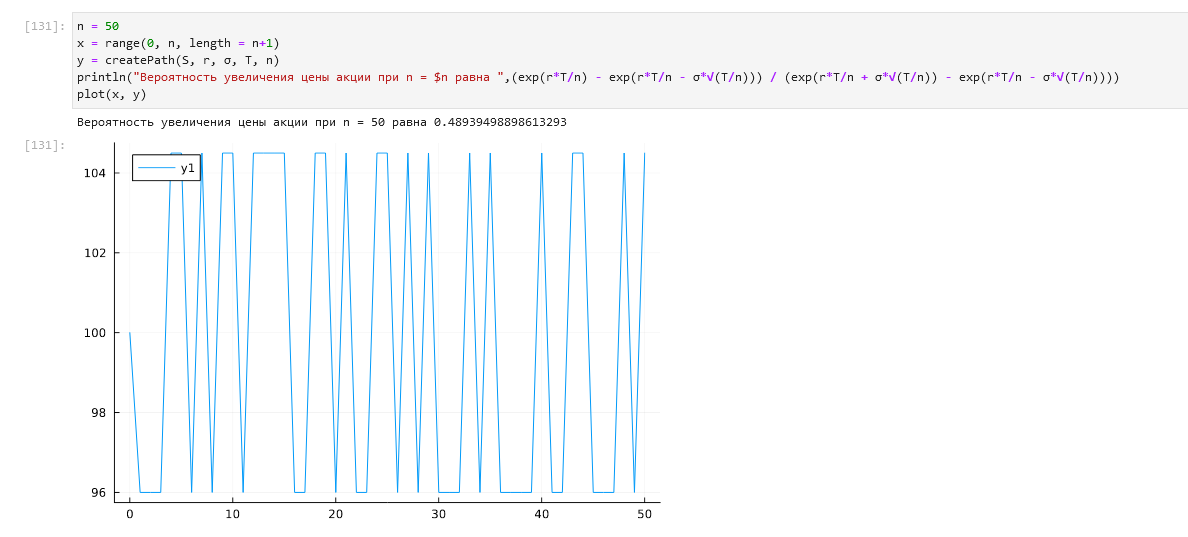
Задание 7.4.3. Модель ценообразования биномиальных опционов (1)



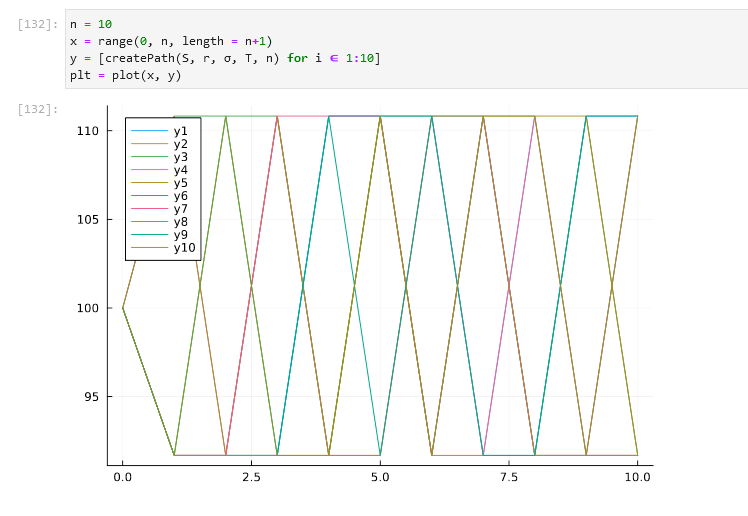
Задание 7.4.3. Модель ценообразования биномиальных опционов (2)



Задание 7.4.3. Модель ценообразования биномиальных опционов (3)



Задание 7.4.3. Модель ценообразования биномиальных опционов (4)

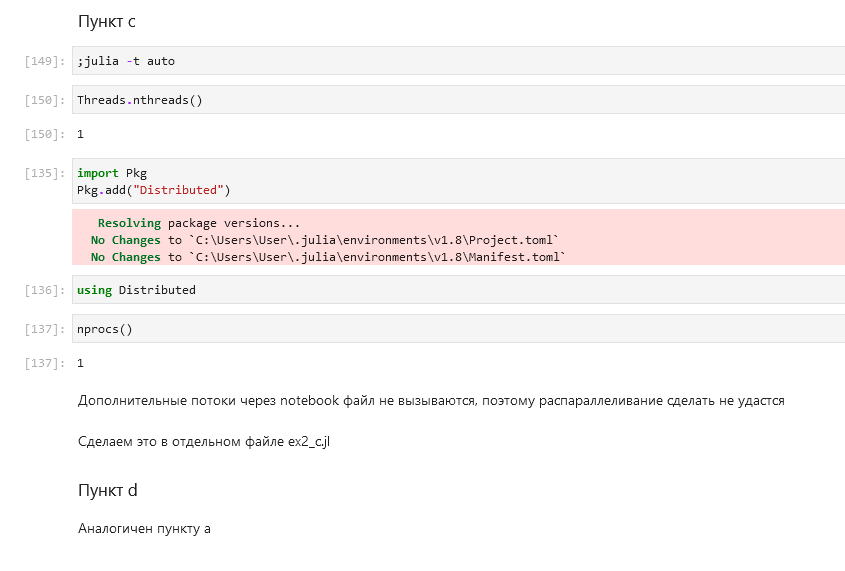


Задание 7.4.3. Модель ценообразования биномиальных опционов (5)

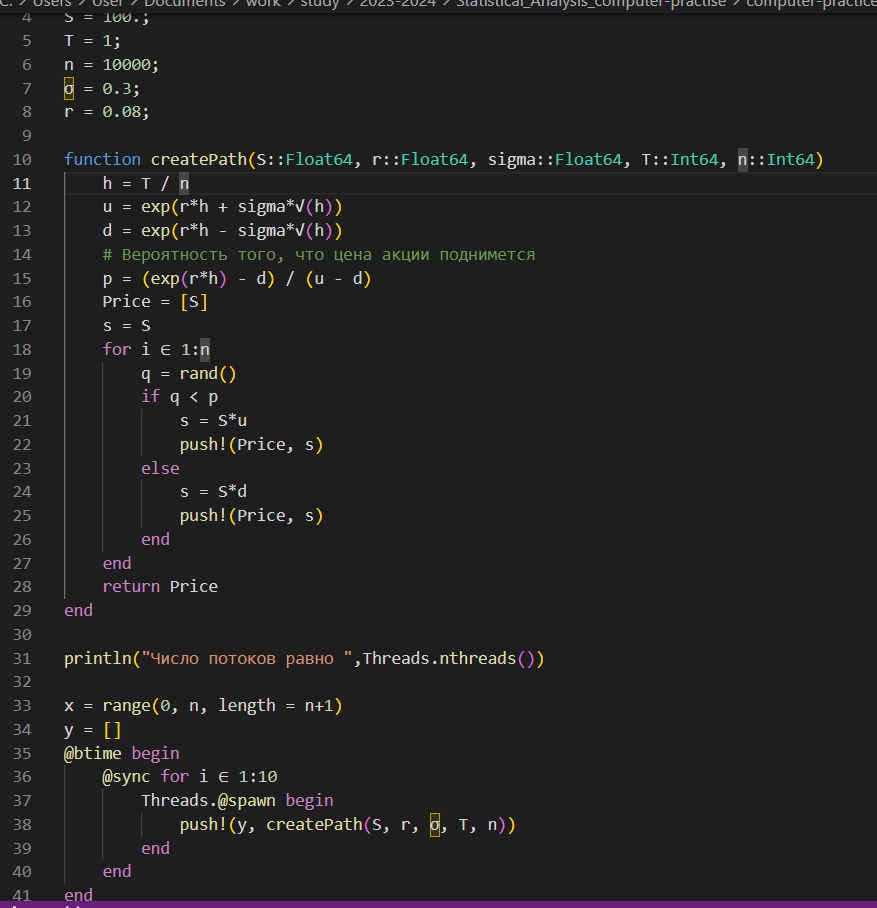
Так как параллелизация невозможна внутри Jupyter ([??]), то сделаем данный пункт в отдельном файле ([??,??]).

Код для пункта c:

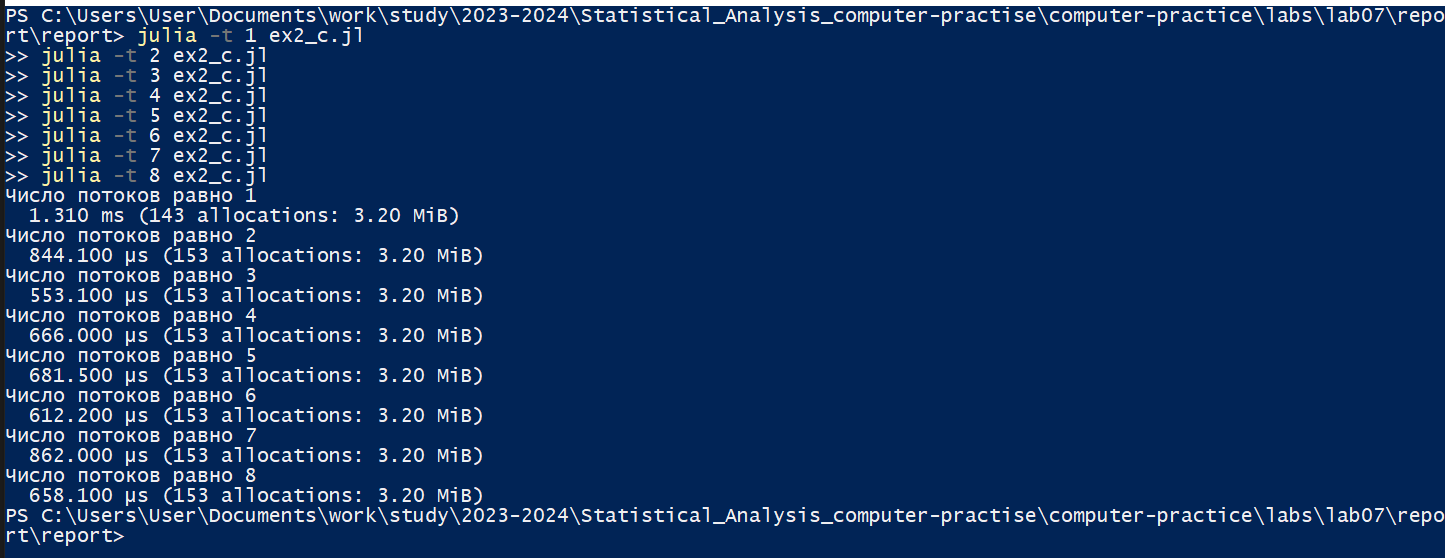
using Base.Threads  
using BenchmarkTools  
  
S = 100.;  
T = 1;  
n = 10000;  
sigma = 0.3;  
r = 0.08;  
  
function createPath(S::Float64, r::Float64, sigma::Float64, T::Int64, n::Int64)  
 h = T / n  
 u = exp(r\*h + sigma\*sqrt(h))  
 d = exp(r\*h - sigma\*sqrt(h))  
 # Вероятность того, что цена акции поднимется  
 p = (exp(r\*h) - d) / (u - d)  
 Price = [S]  
 s = S  
 for i in 1:n  
 q = rand()  
 if q < p  
 s = S\*u  
 push!(Price, s)  
 else  
 s = S\*d  
 push!(Price, s)  
 end  
 end  
 return Price  
end  
  
println("Число потоков равно ",Threads.nthreads())  
  
x = range(0, n, length = n+1)  
y = []  
@btime begin  
 @sync for i in 1:10  
 Threads.@spawn begin  
 push!(y, createPath(S, r, sigma, T, n))  
 end  
 end  
end



Задание 7.4.3. Модель ценообразования биномиальных опционов (6)



Задание 7.4.3. Модель ценообразования биномиальных опционов (7)



Задание 7.4.3. Модель ценообразования биномиальных опционов (8)

# 3 Выводы

В ходе выполнения лабораторной работы я освоил специализированные пакеты в Julia для обработки данных.

# Список литературы

1. Королькова А. В., Кулябов Д. С. Лабораторная работа № 7. Введение в работу с данными [Электронный ресурс]. RUDN, 2023. URL: <https://esystem.rudn.ru/pluginfile.php/2231361/mod_resource/content/2/007-lab_data-science.pdf>.