Отчет по лабораторной работе №7

Введение в работу с данными

Легиньких Галина Андреевна

Содержание

# 1 Цель работы

Основной целью работы является специализированных пакетов Julia для обработки данных.

# 2 Задание

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

# 3 Выполнение лабораторной работы

**1.** Для начала я повторила примеры по данной теме. Начала с раздела считывание данных. В Julia для работы с такого рода структурами данных используют пакеты CSV, DataFrames, RDatasets, FileIO. Скачала эти пакеты и попробовала прочитать данные из скаченного датафрейма. (рис. 1)

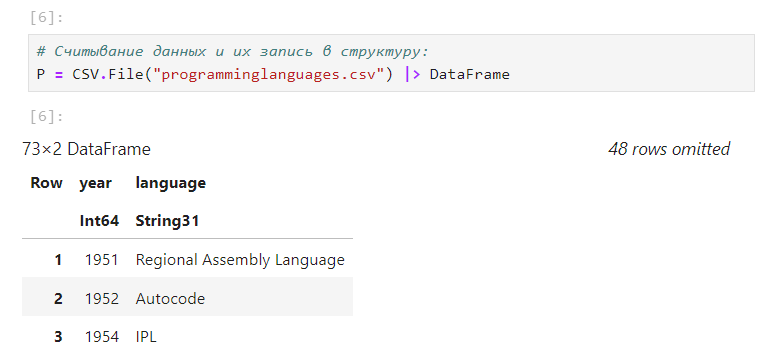


Рис. 1: Считывание данных

Далее приведём пример функции, в которой на входе указывается название языка программирования, а на выходе — год его создания. (рис. 2)

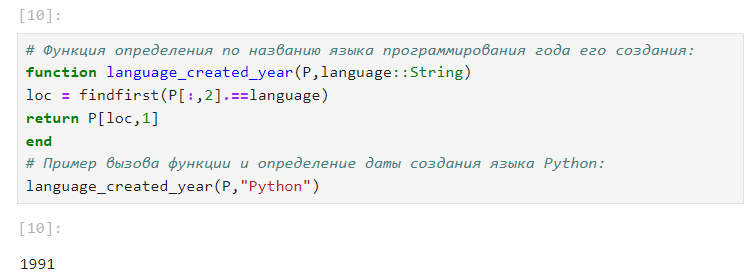


Рис. 2: Функция язык-год

При этом, чтобы убрать в функции зависимость данных от регистра, необходимо изменить исходную функцию следующим образом. (рис. 3)

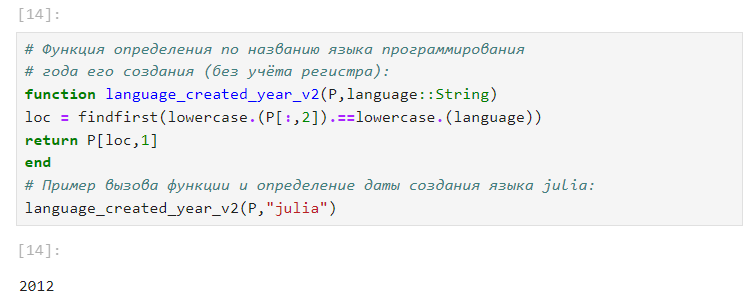


Рис. 3: Зависимость от регистра

**2.** Далее попробовала записать данные в файл, а так же задала при сохранении тип файла и разделитель.

**3.** При работе с данными бывает удобно записать их в формате словаря.При инициализации словаря можно задать конкретные типы данных для ключей и значений,а можно инициировать пустой словарь, не задавая строго структуру. (рис. 4)

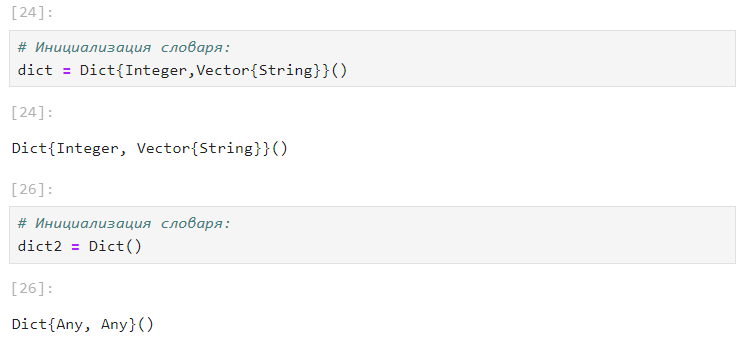


Рис. 4: Инициализация словаря

В результате при вызове словаря можно, выбрав любой год, узнать, какие языки программирования были созданы в этом году. (рис. 5)



Рис. 5: Работа со словарем

**4.** На примере с данными о языках программирования и годах их создания зададим структуру DataFrame. Попробовала вывести один столбец. (рис. 6)

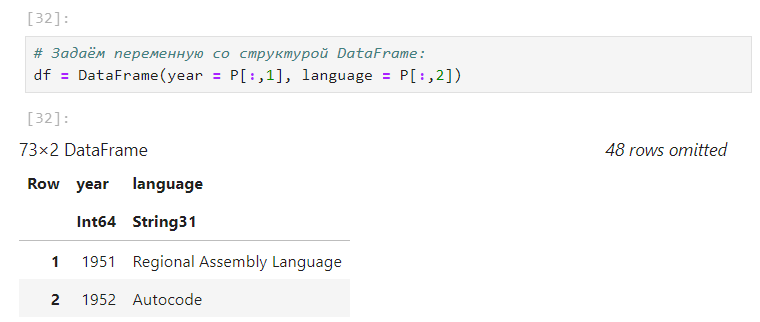


Рис. 6: Переменная со структурой датафрейм

А так же получила статистические сведения о фрейме. (рис. 7)

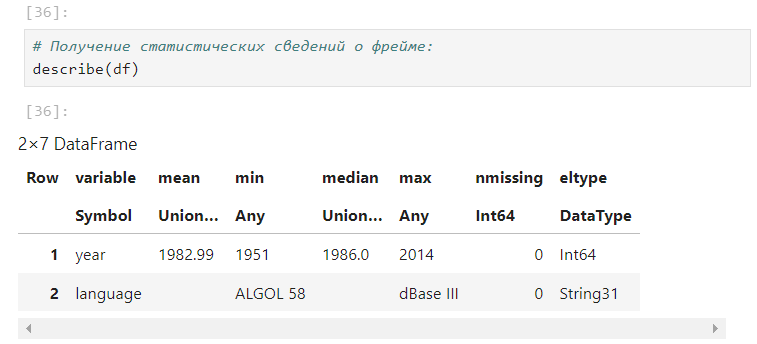


Рис. 7: Статистические сведения о фрейме

**5.** С данными можно работать также как с наборами данных через пакет RDatasets языка R.

**6.** Попробовала порабоать с переменными отсутствующего типа. (рис. 8)

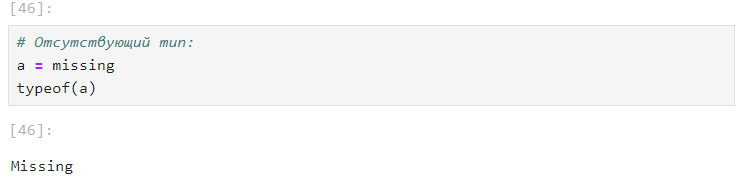


Рис. 8: Переменная отсутсвующего типа

**7.** Перешла к обработке данных. И первый метод это 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. Рассмотрела задачу кластеризации данных на примере данных о недвижимости. (рис. 9) (рис. 10) (рис. 11)

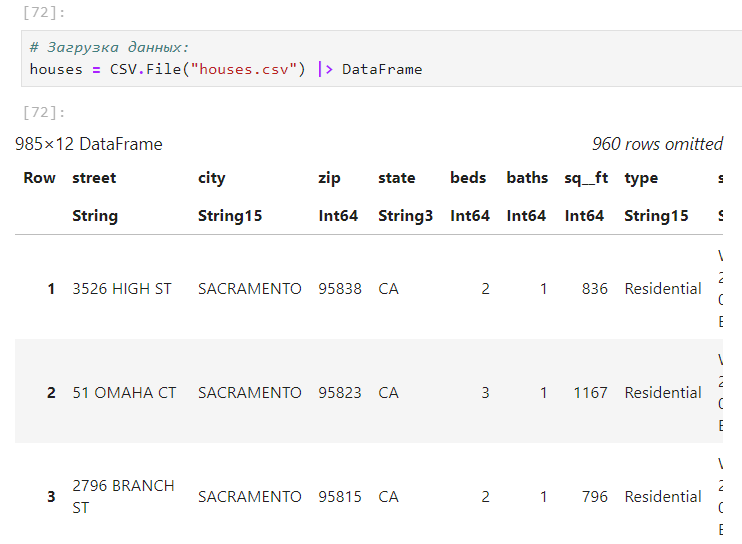


Рис. 9: Кластеризация данных на примере данных о недвижимости

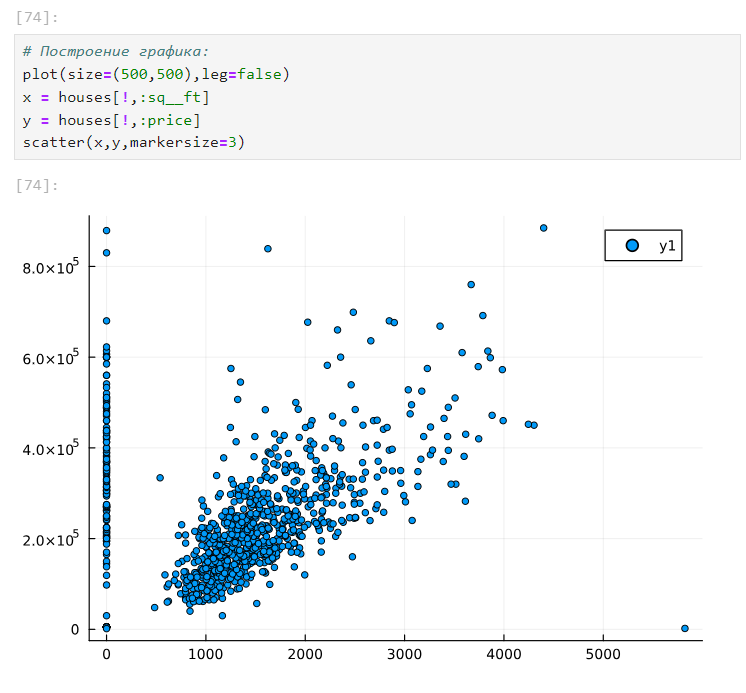


Рис. 10: Цены на недвижимость в зависимости от площади

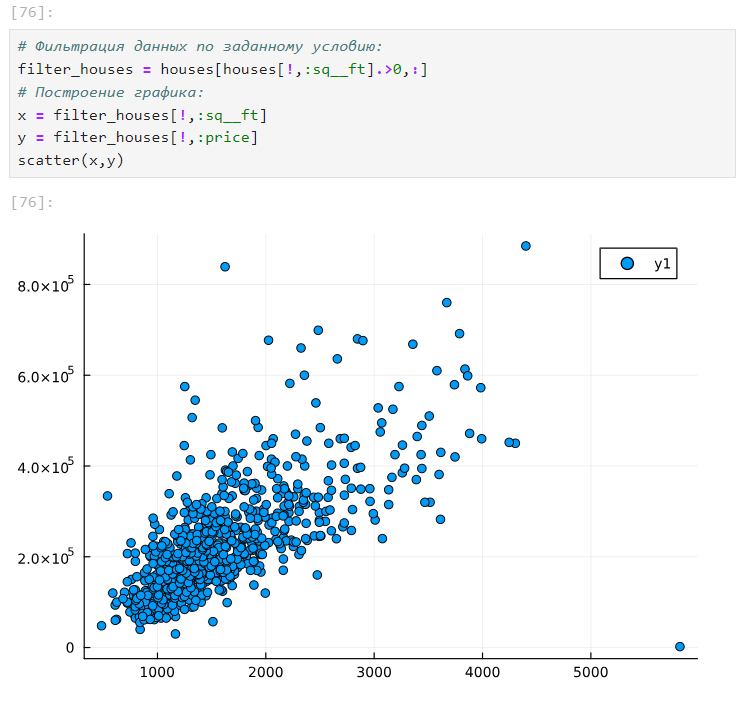


Рис. 11: Цены на недвижимость в зависимости от площади (исключены артефакты данных)

Сначала подключаем необходимые пакеты и формируем данные в нужном виде: (рис. 12) (рис. 13)

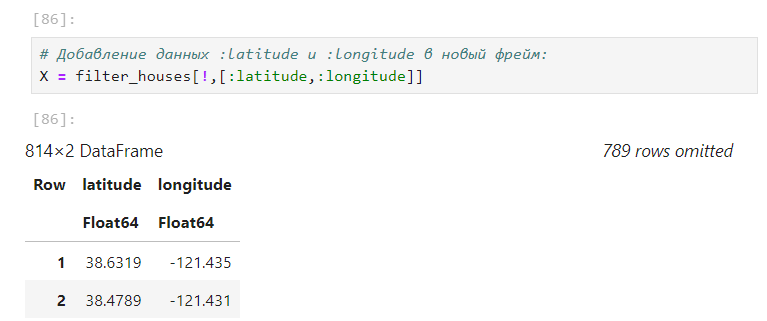


Рис. 12: Добавление данных

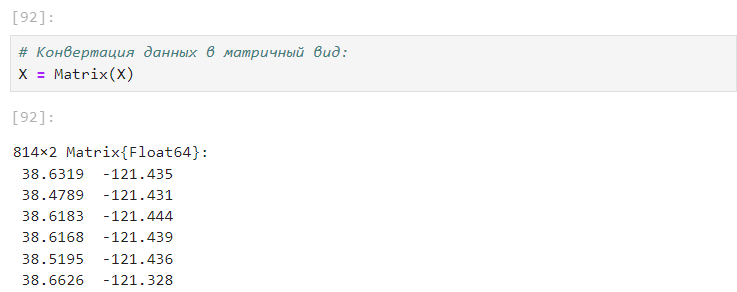


Рис. 13: Конвертация данных в матричный вид

В качестве критерия для формирования кластеров данных и определения количества кластеров попробуем использовать количество почтовых индексов: (рис. 14)

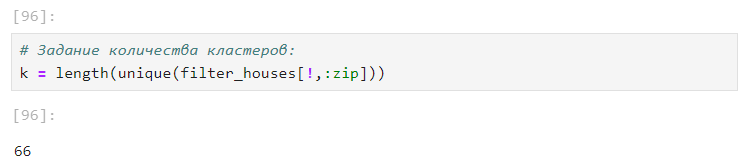


Рис. 14: Задание количества кластеров

Далее сформируем новый фрейм, включающий исходные данные о недвижимости и столбец с данными о назначенном каждому дому кластере (рис. 15)

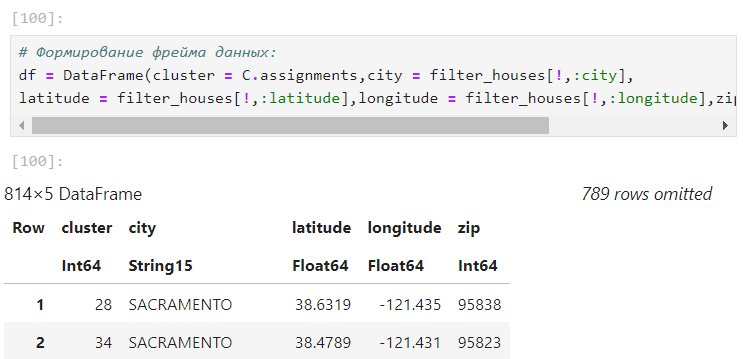


Рис. 15: Формирование фрейма данных

Плстроила график, обозначив каждый кластер отдельным цветом (рис. 16)

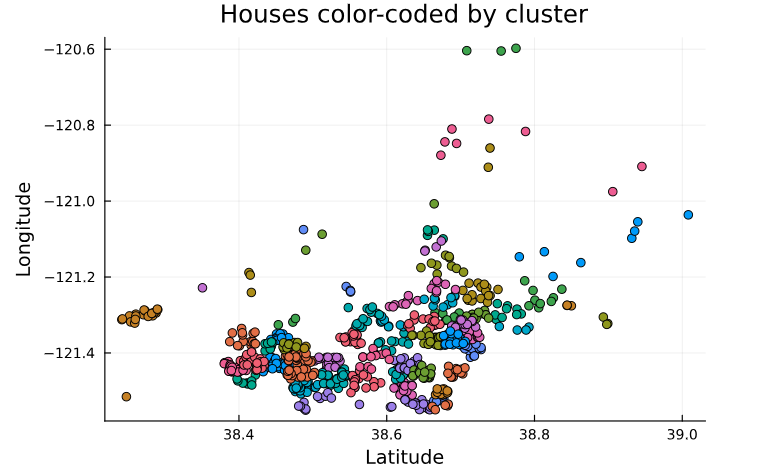


Рис. 16: График кластеров

Построила график, раскрасив кластеры по почтовому индексу (рис. 17)

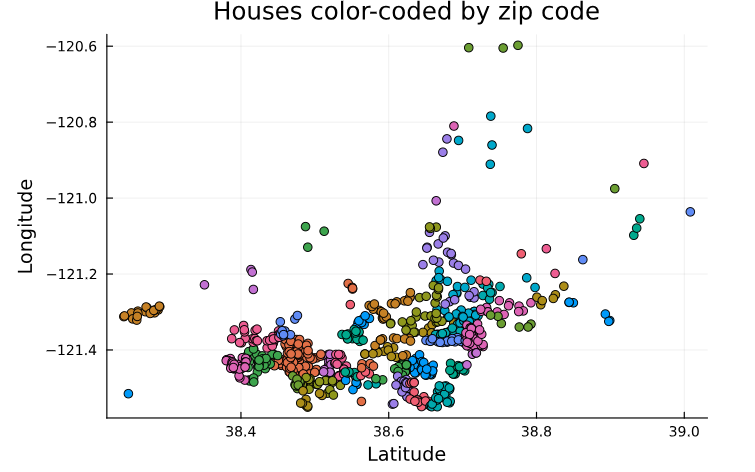


Рис. 17: График кластеров 2

**8.** Кластеризация данных. Метод k ближайших соседей. Данный метод заключается в отнесении объекта к тому из известных классов, который является наиболее распространённым среди 𝑘 соседей данного элемента. В случае использования метода для регрессии, объекту присваивается среднее значение по 𝑘 ближайшим к нему объектам. (рис. 18) (рис. 19)

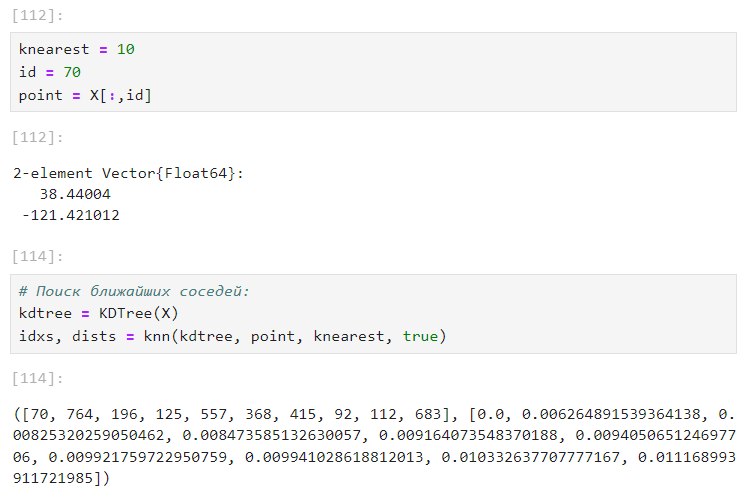


Рис. 18: Метод главных компонент

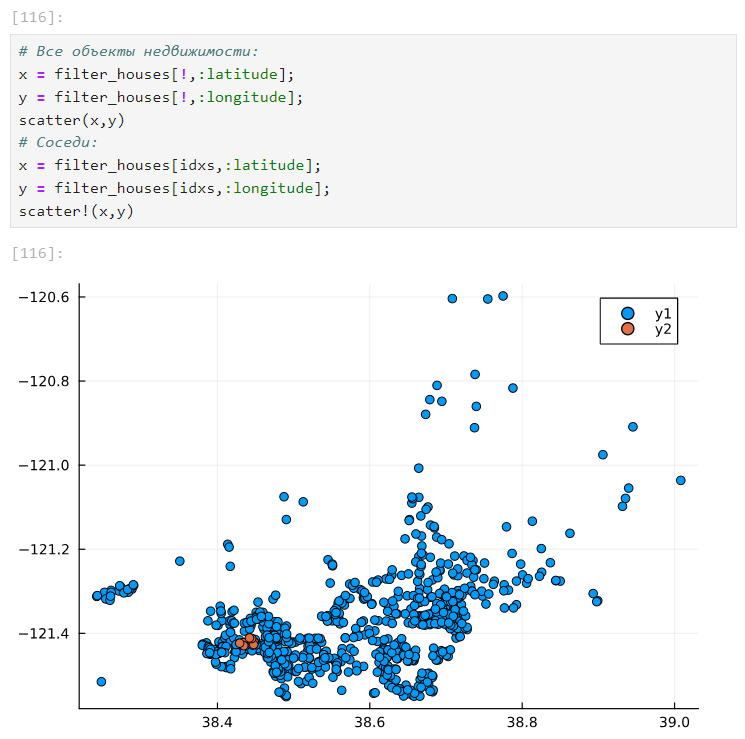


Рис. 19: Определение соседей объекта недвижимости

**9.** 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 (рис. 20) (рис. 21) (рис. 22)

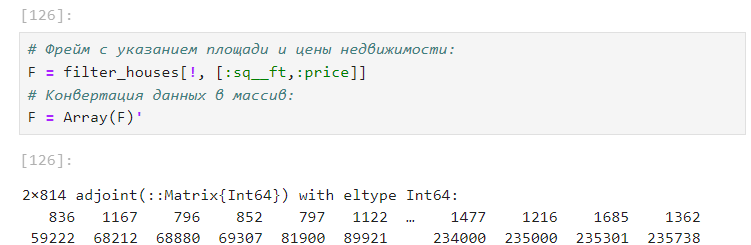


Рис. 20: Создание фрейма

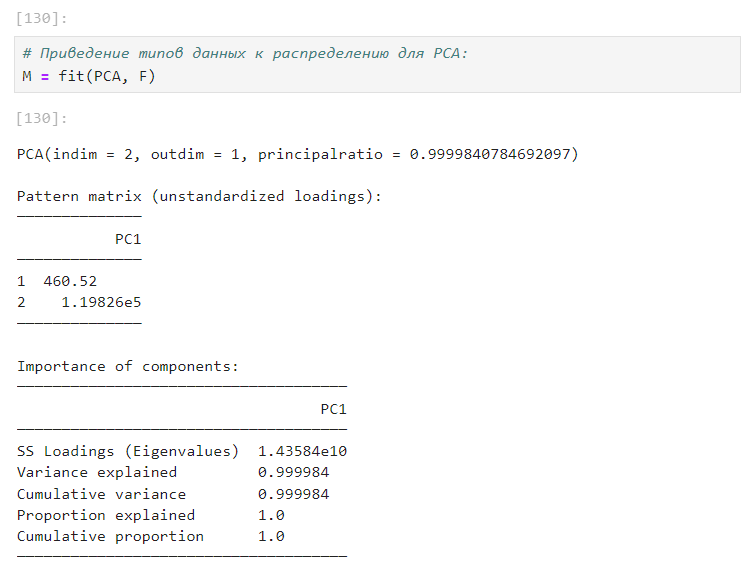


Рис. 21: Приведение типов данных к распределению для PCA

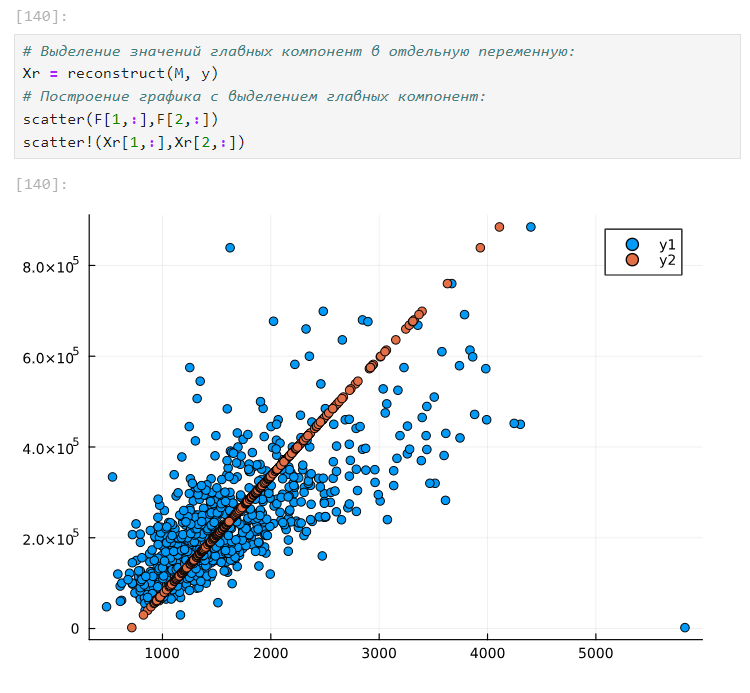


Рис. 22: Определение главных компонент для данных по объектам недвижимости

**10.** 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.

Зададим случайный набор данных (можно использовать и полученные экспериментальным путём какие-то данные). Попробуем найти для данных лучшее соответствие (рис. 23)

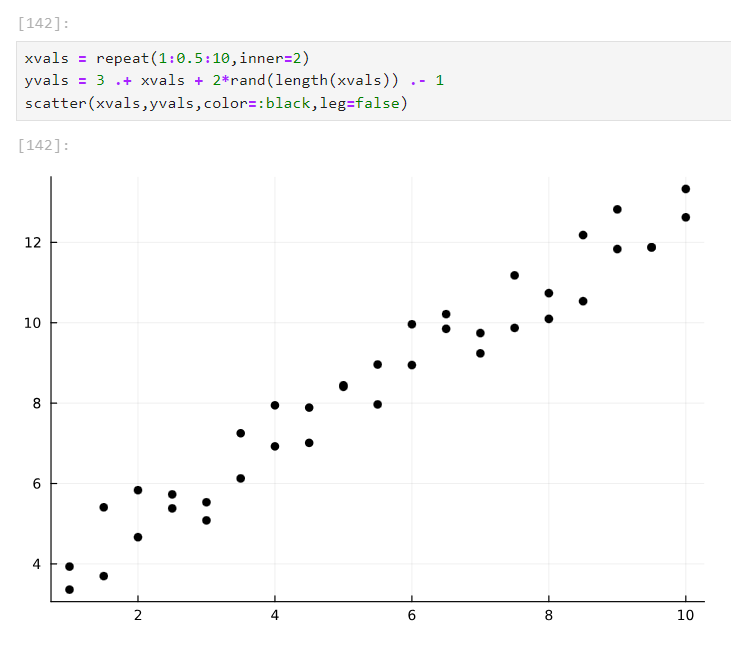


Рис. 23: Исходные данные

Применим функцию линейной регрессии для построения соответствующего графика значений (рис. 24)

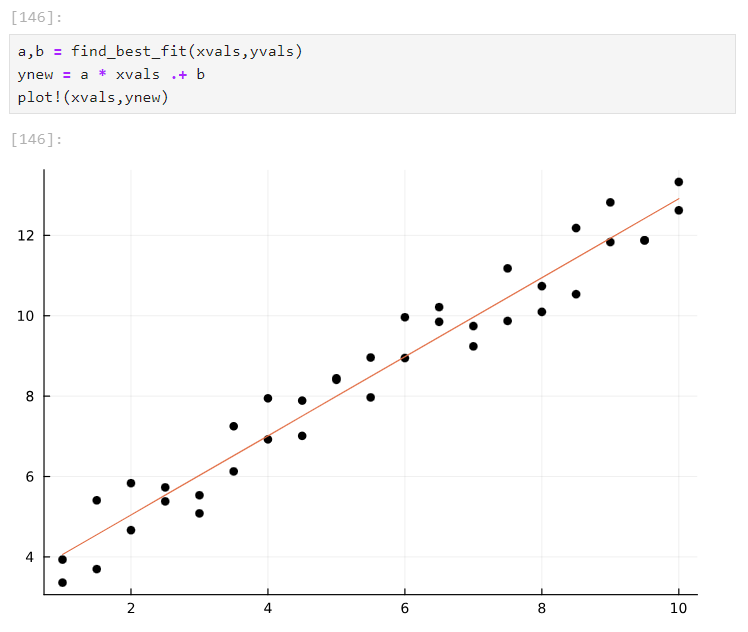


Рис. 24: Линейная регрессия

**11.** Перешла к заданиям для самостоятельного выполнения. Нумерация соотвествует.

* Задание на Кластеризацию. (рис. 25) (рис. 26)



Рис. 25: Задание 1. Код

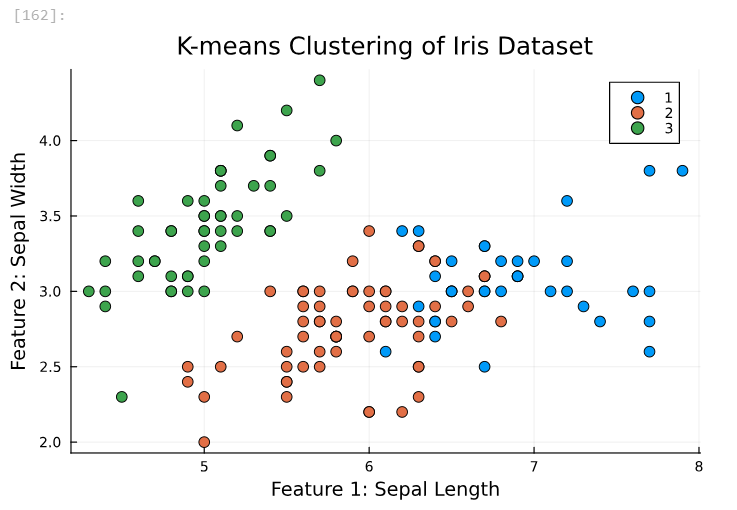


Рис. 26: Задание 1. График

* Задание на регрессию (часть 1) (рис. 27) (рис. 28)



Рис. 27: Задание 2.1. Код

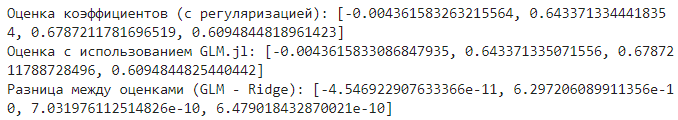


Рис. 28: Задание 2.1. Решение

* Задание на регрессию (часть 2) (рис. 29) (рис. 30)

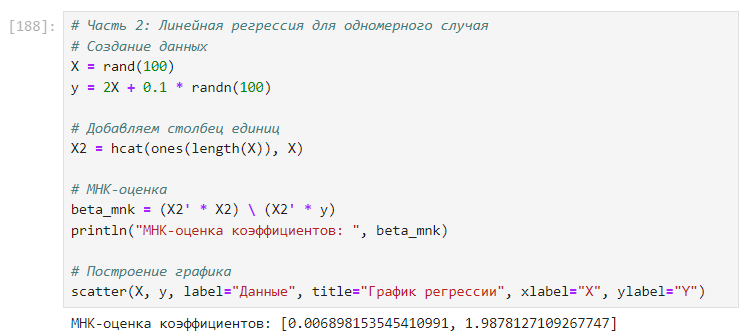


Рис. 29: Задание 2.2. Код

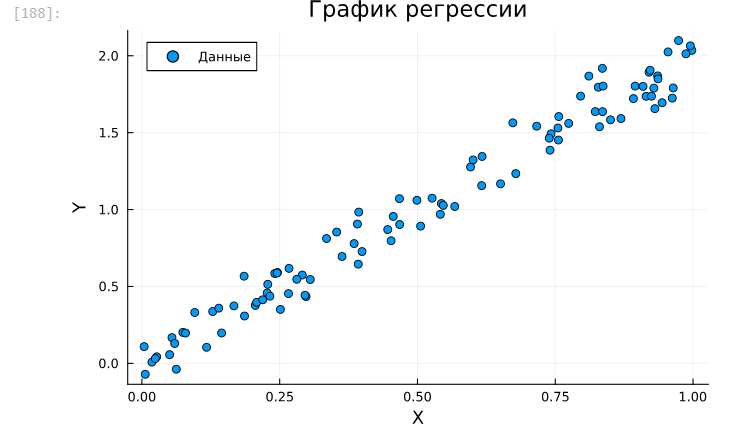


Рис. 30: Задание 2.2. График

* Модель ценообразования биномиальных опционов (a) (рис. 31) (рис. 32)

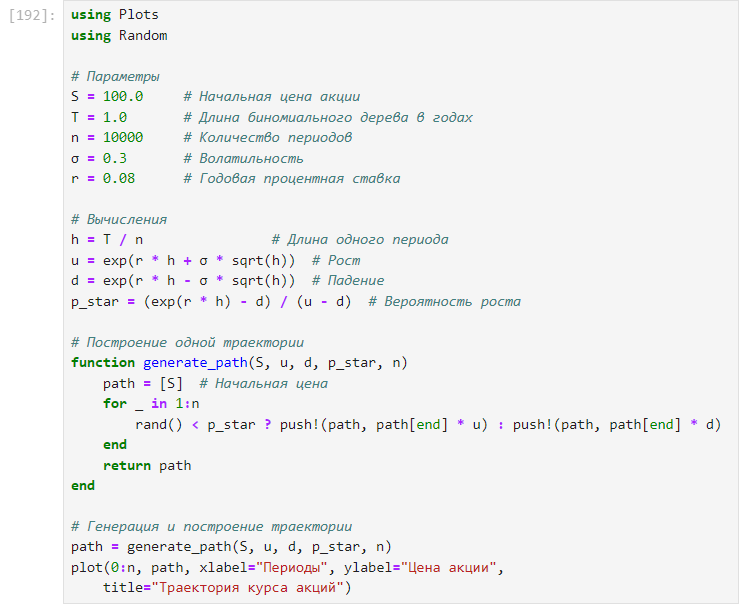


Рис. 31: Задание 3.1. Код

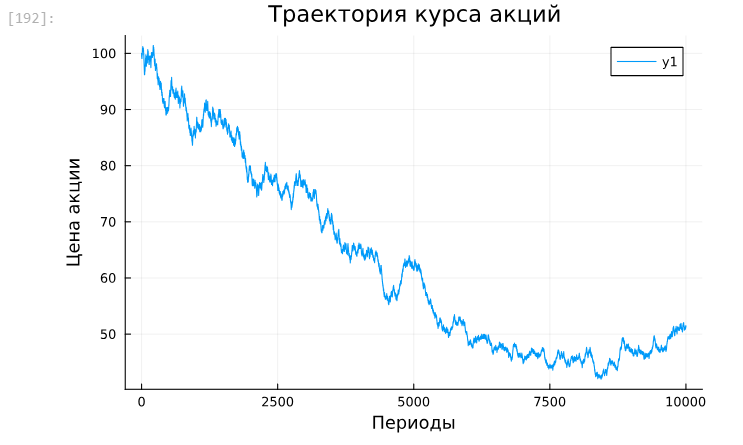


Рис. 32: Задание 3.1. График

* Модель ценообразования биномиальных опционов (b) (рис. 33)

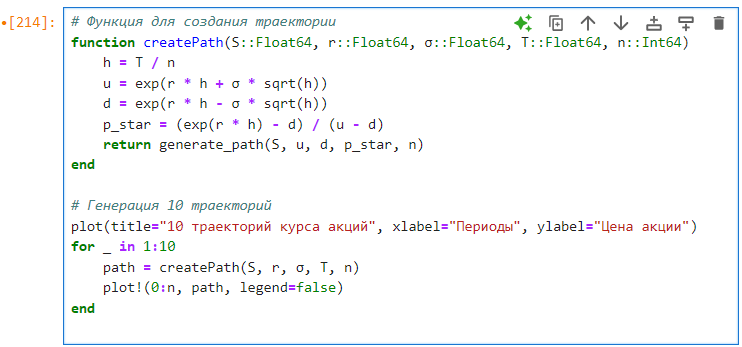


Рис. 33: Задание 3.2. Код

* Модель ценообразования биномиальных опционов (c) (рис. 34) (рис. 35)

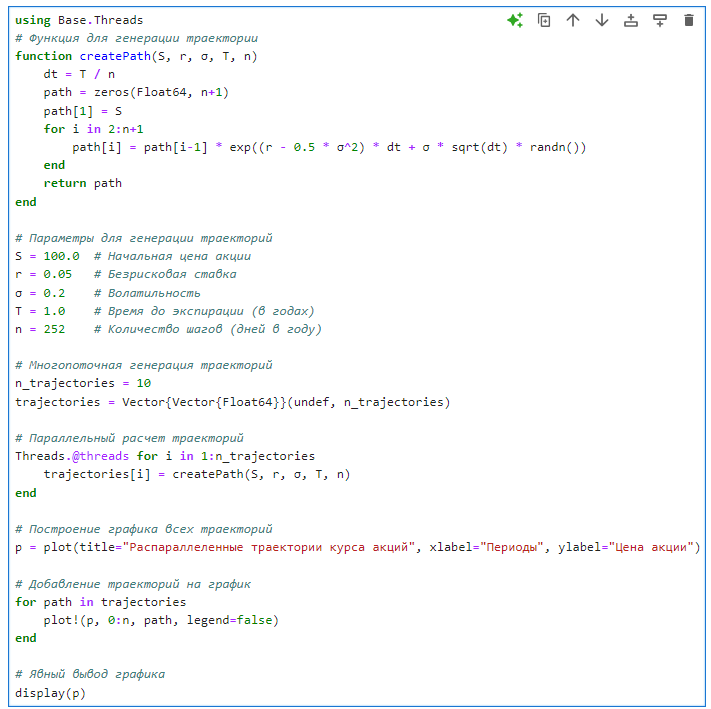


Рис. 34: Задание 3.3. Код

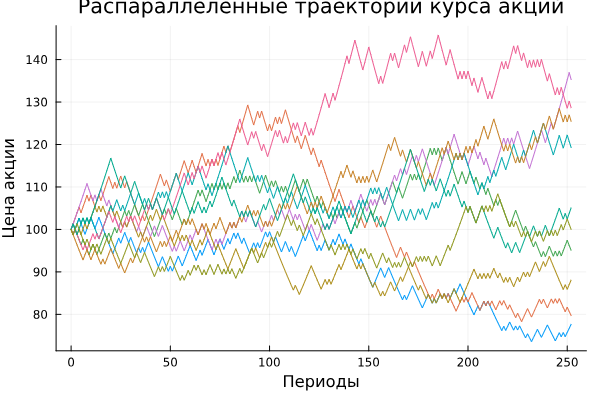


Рис. 35: Задание 3.3. График

* Модель ценообразования биномиальных опционов (d) (рис. 36)

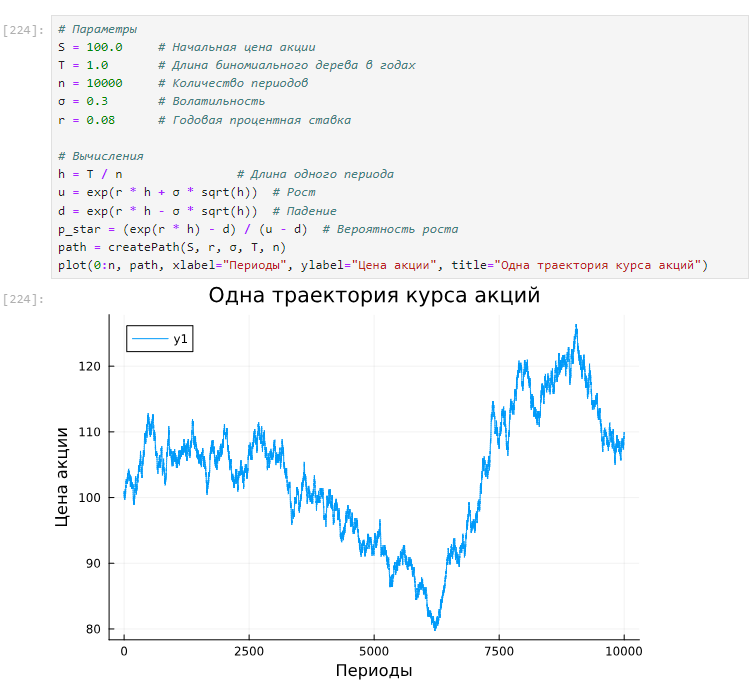


Рис. 36: Задание 3.4. Код и График

# 4 Вывод

Изучила специализированные пакеты Julia для обработки данных.