Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

«Уральский федеральный университет имени первого Президента России Б.Н. Ельцина» (УрФУ)

Институт радиоэлектроники и информационных технологий – РтФ

Школа бакалавриата

Оценка работы

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Члены комиссии

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

РАЗРАБОТКА СИСТЕМЫ ПРОГНОЗИРОВАНИЯ СТОИМОСТИ ДОМА

Пояснительная записка к курсовому проекту по модулю

«Методы анализа Big Data»

Преподаватель Мирвода С.Г.

Студенты Куштанов Д.И.

Изотов И.Н.

Группа РИ-450004

Екатеринбург

2019

ОГЛАВЛЕНИЕ

[ВВЕДЕНИЕ 3](#_Toc534931151)

[1 ОПИСАНИЕ НАБОРА ДАННЫХ 4](#_Toc534931152)

[2 ТРЕНИРОВКА ЛИНЕЙНОЙ МОДЕЛИ 7](#_Toc534931153)

[3 ПРОЕКТИРОВАНИЕ ПРИЛОЖЕНИЯ 11](#_Toc534931154)

[4 ИНСТРУКЦИЯ ПОЛЬЗОВАТЕЛЯ 12](#_Toc534931155)

[ЗАКЛЮЧНИЕ 13](#_Toc534931156)

[БИБЛИОГРАФИЧЕСКИЙ СПИСОК 14](#_Toc534931157)

# ВВЕДЕНИЕ

В текущее время объемы информации растут по экспоненте. Для того, чтобы быстрее реагировать на изменения рынка, получить конкурентные преимущества и повысить эффективность производства, нужно получить, обработать и проанализировать огромное количество данных. Для этого используются специальные программные средства, и одним из таких решений может являться язык программирования для статистической обработки данных и работы с графикой R.

Целью данной работы является разработка системы прогнозирования стоимости домов, с использованием открытого набора данных.

Гипотезой служит предположение о том, что технические и геолокационные характеристики дома вносят влияние на формирование его стоимости.

Для достижения поставленной цели необходимо выполнить следующие задачи:

1. изучить набор данных;
2. провести анализ данных;
3. натренировать линейную модель;
4. написать пользовательское приложение.

# 1 ОПИСАНИЕ НАБОРА ДАННЫХ

Для данный работы использовался набор данных стоимостей домов США в определенном регионе [1]. Пример данных изображен на рисунке 1.

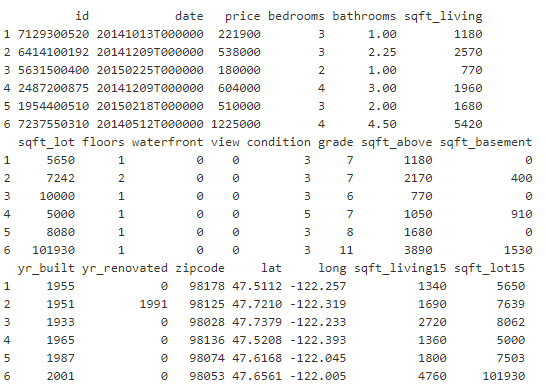


Рисунок 1 – Пример данных

В данном наборе данных колонки имеют следующие значения:

* id – номер дома;
* date – дата продажи дома;
* price – цена;
* bedrooms – число спален;
* bathrooms – число ванных комнат;
* sqft\_living – площадь дома;
* sqft\_lot – площадь участка;
* floors – число этажей;
* waterfront – есть ли выход на набережную;
* view – был ли осмотрен;
* condition – состояние дома;
* grade – оценка дома основанная на King County;
* sqft\_above – площадь дома отдельно от подвала;
* sqft\_basement – площадь подвала;
* yr\_built – год постройи;
* yr\_renovated – год ремонта дома;
* zipcode – почтовый zip код;
* lat – координата широты;
* long – координата долготы;
* sqft\_living15 – жилая площадь на 2015 год;
* sqft\_lot15lot – площадь участка на 2015 год.

На рисунках ниже (рисунки 2, 3, 4, 5) отображены некоторые зависимости атрибутов от цены.

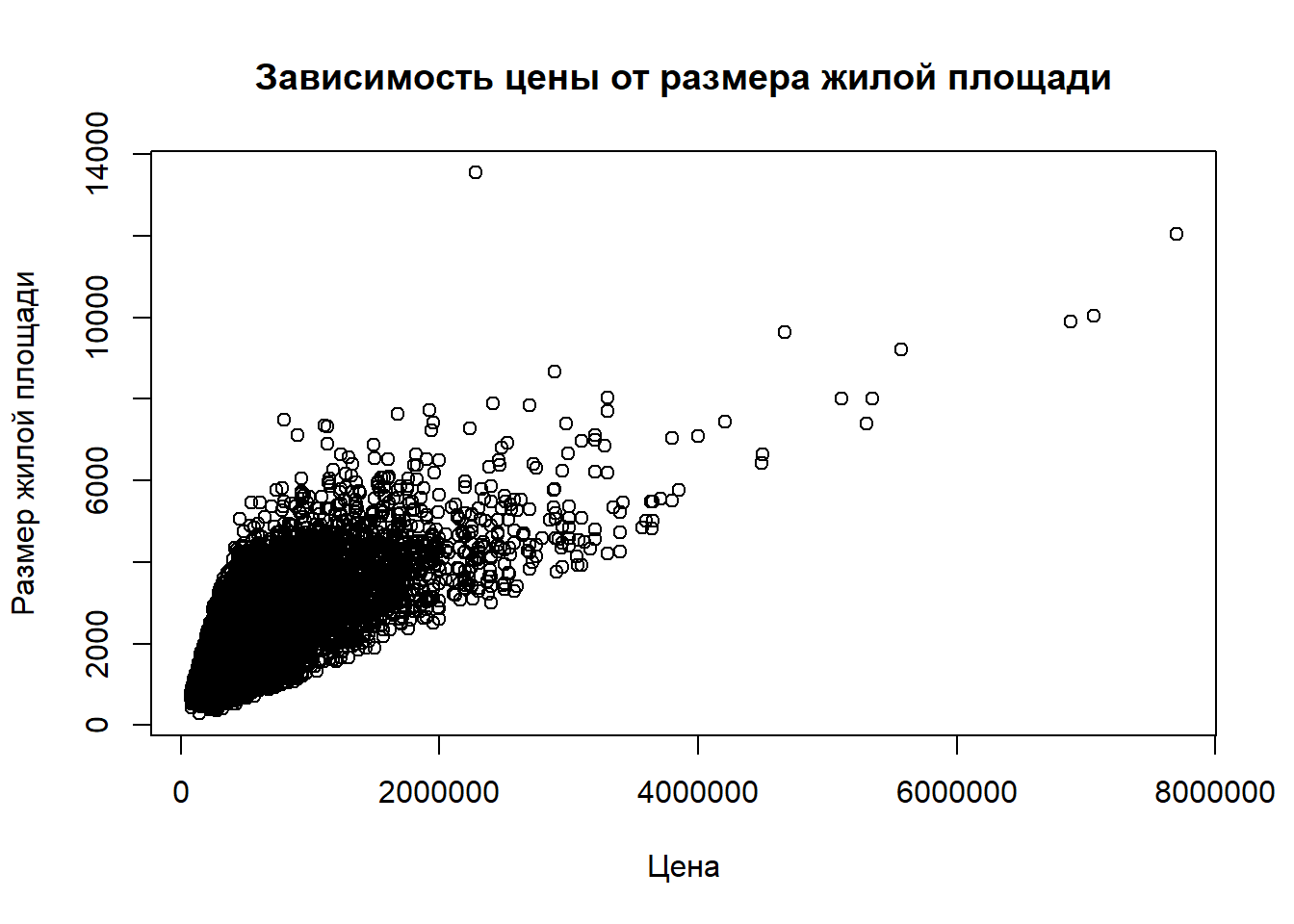


Рисунок 2 – Зависимость цены от размера жилой площади

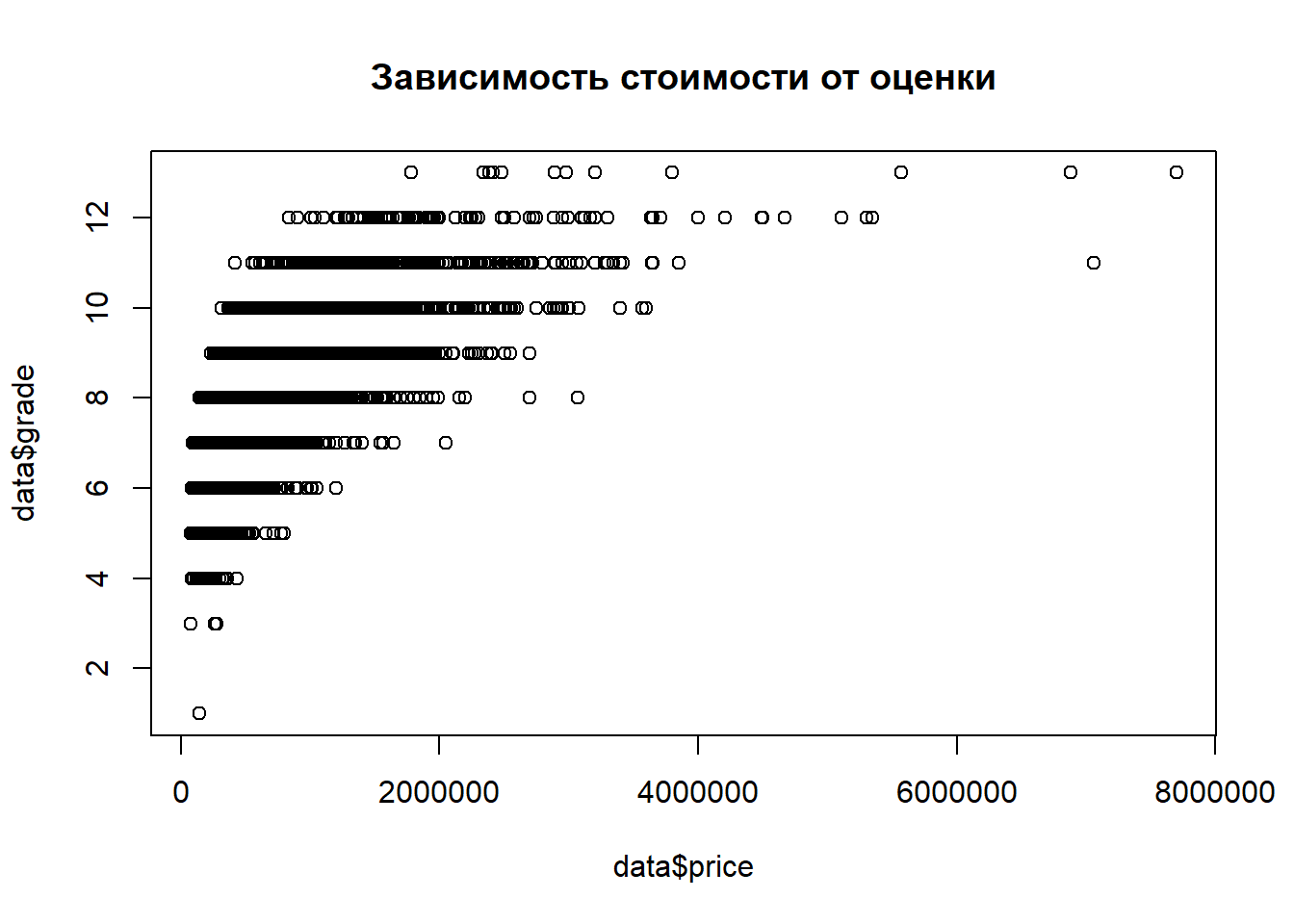


Рисунок 3 – Зависимость цены от оценки

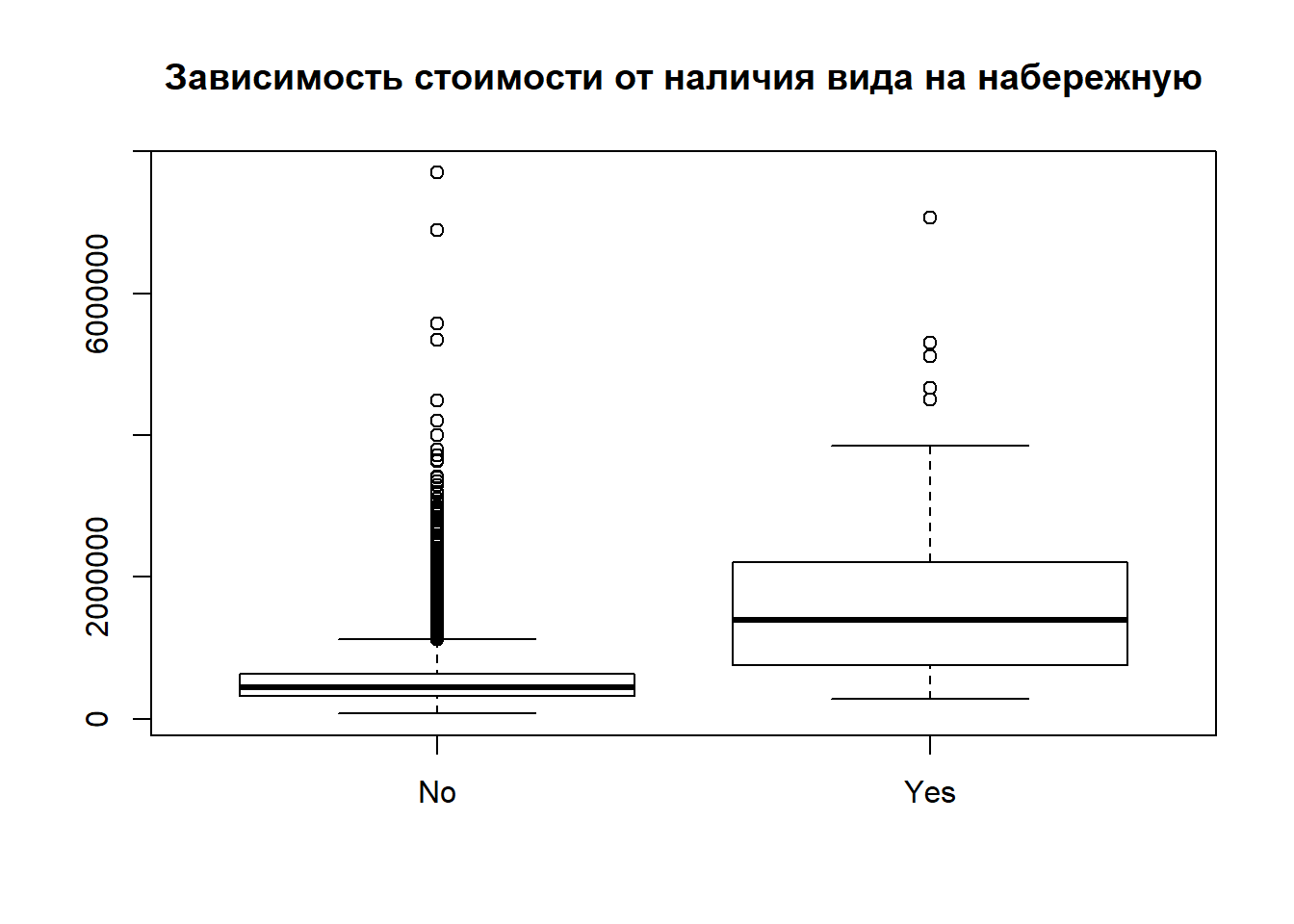


Рисунок 4 – Зависимость цены от наличия вида на набережную.

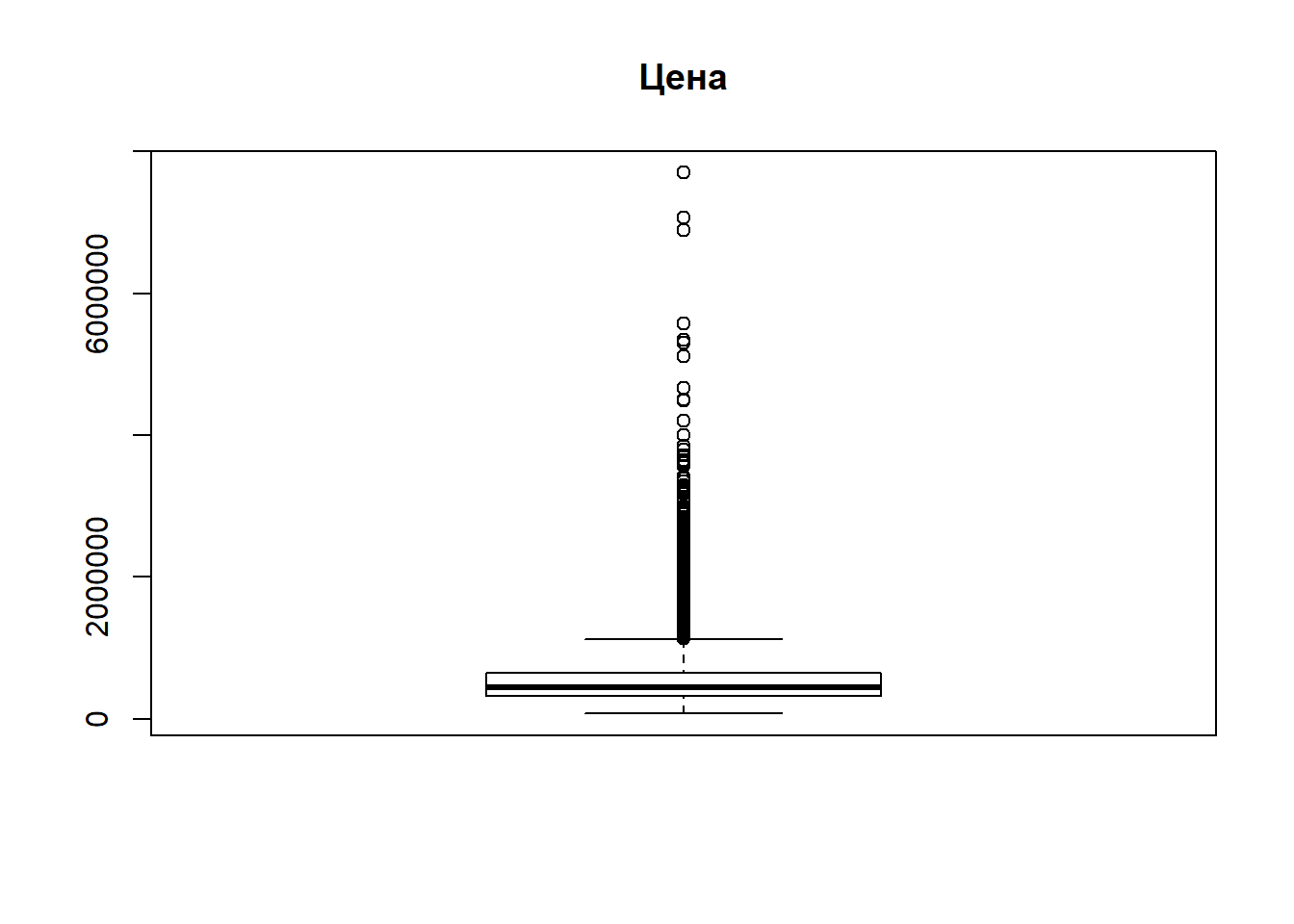


Рисунок 5 – График стоимости

# 2 ТРЕНИРОВКА ЛИНЕЙНОЙ МОДЕЛИ

Для тренировки модели была проведена чистка данных (результат на рисунке 6) и построена линейная модель на основе всех данных, чтобы определить наиболее значимые параметры (рисунок 7).



Рисунок 6 – Очищенная зависимость стоимости от вида на набережную

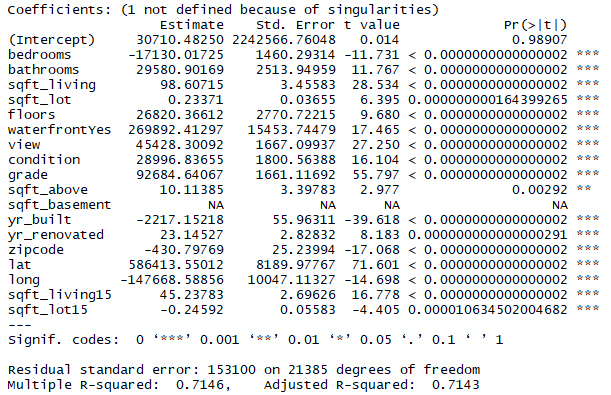


Рисунок 7 – Зависимость влияния параметров на цену

Для построения более точной модели были убраны наименее значимые столбцы и создан дополнительный avgZipCodePrice являющийся средней ценой по почтовому индексу. Результаты показаны на рисунках 8, 9, 10.

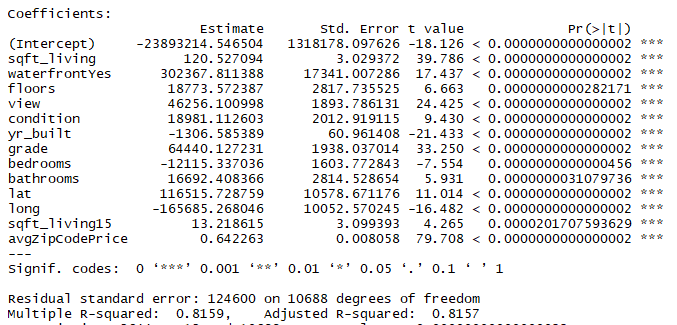


Рисунок 8 – Коэффициенты новой модели

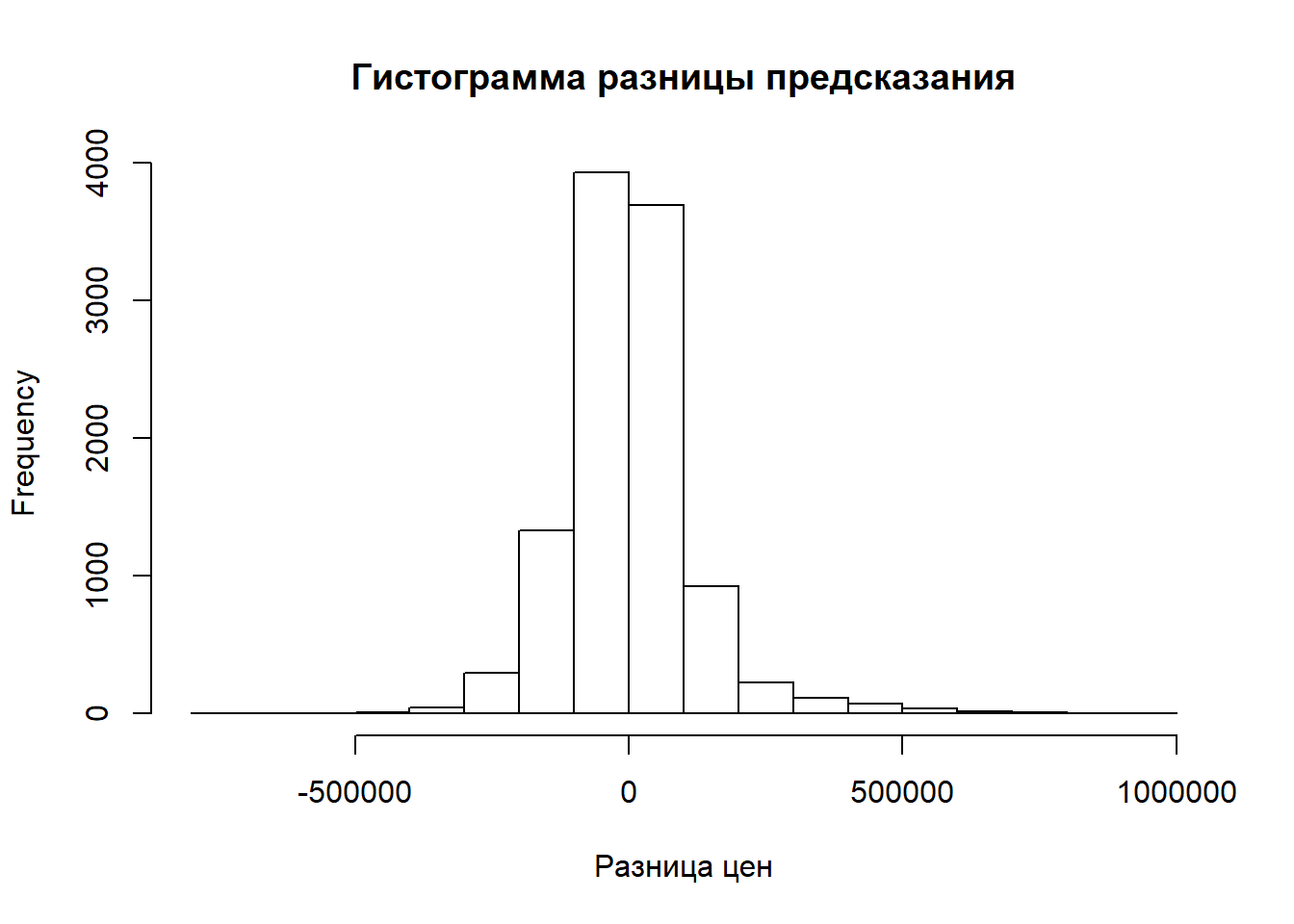


Рисунок 9 – Гистограмма зависимости предсказанных значений от проверяемых

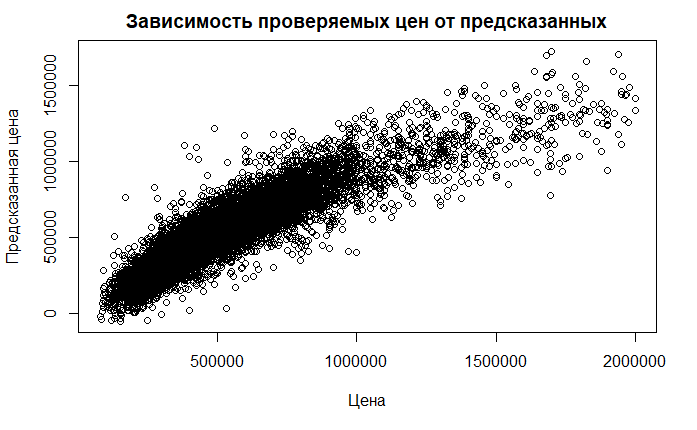


Рисунок 10 – Зависимость предсказанных значений от проверяемых

Для приложения было принято решение использовать меньшее число параметров, для этого была создана дополнительная модель. Ее результат показан на рисунках 11, 12, 13.

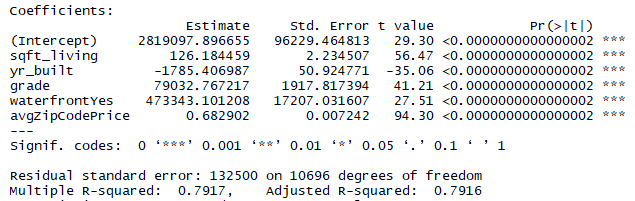


Рисунок 11 – Коэффициенты модели с меньшим числом параметров



Рисунок 12 – Гистограмма зависимости предсказанных значений от проверяемых модели с меньшим числом параметров



Рисунок 13 – Зависимость предсказанных значений от проверяемых модели с меньшим числом параметров

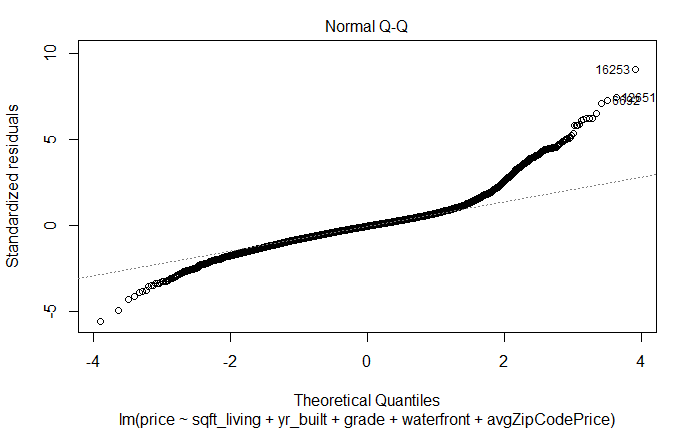


Рисунок 14 – Проверка зависимости при помощи графика Q-Q

# 3 ПРОЕКТИРОВАНИЕ ПРИЛОЖЕНИЯ

Приложение было написано на языке Python. Его интерфейс изображен на рисунке 12.

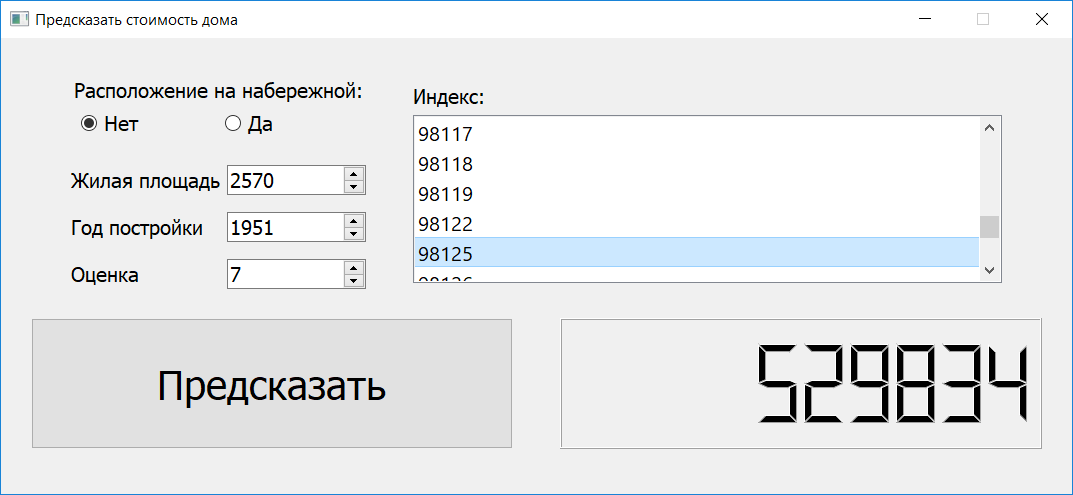


Рисунок 15 – Макет приложения

Результат вызова функции из R с теми же параметрами, что представлены на изображение макета, показан на рисунке 13. Результаты вывода у них совпадают.



Рисунок 16 – Вызов функции предсказания в R

Данное приложение получает значения почтовых индексов, а при нажатии кнопки предсказать вызывает функции R для расчета цены.

# 4 ИНСТРУКЦИЯ ПОЛЬЗОВАТЕЛЯ

1. Запустить приложение;
2. Ввести требуемые параметры в окна ввода (возможные вводимые параметры ограничены граничными значениями). Выбрать необходимый почтовый индекс;
3. Нажить на кнопку «Предсказать»;
4. Получить значение в окне справа от нажимаемой кнопки.

# ЗАКЛЮЧНИЕ

В результате выполнения данной работы был изучен открытый набор данных стоимости домов, путем изучения представленной информации на сайте источника и отображением графиков и параметров при помощи функция языка программирования R в программе Rstudio. После была построена линейная модель по этим данным для предсказания стоимости домов, используя для этого язык Rmarkdown в котором были построены возможные зависимости между ценой и другими параметрами, проведена очистка от выбросов и изучены значения наиболее влияющие факторы на цену дома. Значение для оптимального числа параметров в линейной модели составило 0,82.

Для отображения возможного применения данных расчетов было реализовано приложение языке Python, в котором была реализована функциональность в виде обращения к функциям языка R для применения и расчета линейной модели. В данном случае использовалась линейная модель с меньшим числом параметров, в котором .

# БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. House Sales in King County, USA. Predict house price using regression – https://www.kaggle.com/harlfoxem/housesalesprediction

2.