# Постановка задачи

## Формулировка прикладной задачи

Задачу, которую мы с ставим перед собой, мы обозначили следующим образом: «Определить какие показатели жилого дома и в какой степени влияют на его стоимость».

Для этого мы проанализируем следующие переменные:

* количество спален,
* количество ванн,
* площадь дома,
* площадь участка,
* количество этажей,
* наличие вида на берег,
* количество раз, которое покупатели осматривали дом,
* состояние,
* общая оценка по системе Округа Кинг,
* площадь подвала,
* год постройки,
* год реновации,
* цена.

Будут взяты данные 21613 продаж жилых домов в 2015 году в Округе Кинг, штат [Вашингтон](https://ru.wikipedia.org/wiki/%D0%92%D0%B0%D1%88%D0%B8%D0%BD%D0%B3%D1%82%D0%BE%D0%BD_(%D1%88%D1%82%D0%B0%D1%82)), [США](https://ru.wikipedia.org/wiki/%D0%A1%D0%A8%D0%90).

## Потенциальные потребители решения

Используя построенную модель и найденные зависимости между показателями жилого дома и его ценой, агенты на данном рынке недвижимости смогут рационализировать свои действия купли-продажи. Главным потребителем результатов нашего исследования будут риэлтерские агентства и обычные люди, которые хотят оптимально для себя продать или купить жилой дом или просто рационально оценить цену того или иного жилого дома. Также возможным потребителем исследований могут стать налоговые службы, которым необходимо рационально обозначить сумму налога для конкретного жилого дома.

## Основные гипотезы

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Характеристика объекта/явления | Название переменной | Шкала измерения | Роль: зависимая/независимая |
| 1 | Количество спален | bedrooms | количественная | независимая |
| 2 | Количество ванн | bathrooms | количественная | независимая |
| 3 | Площадь дома без подвала | sqft\_above | количественная | независимая |
| 4 | Количество этажей | floors | количественная | независимая |
| 5 | Наличие вида на берег | waterfront | бинарная номинальная | независимая |
| 6 | Общая оценка по системе Округа Кинг | grade | порядковая | независимая |
| 7 | Площадь подвала | sqft\_basement | количественная | независимая |
| 8 | Год постройки | yr\_built | порядковая | независимая |
| 9 | Цена | price | количественная | зависимая |

1. Первая, самая очевидная, гипотеза заключается в наличии прямой зависимости между ценой дома и его количественных (площадь дома и участка, количество спален, этажей и т.д.), а также порядковых переменных (состояние, год постройки и реновации) и номинальной переменной (наличие вида на берег).
2. Потенциальные покупатели обращают больше внимания на площадь участка и дома, его состояние, чем на количество ванн, спален, этажей, площадь подвала.
3. Существует зависимость между ценой дома и его общей оценкой по системе Округа Кинг, то есть покупатели обращают внимание на рейтинг жилья.

## Основные источники данных

Единственный источник данных (более 20 000 наблюдений)

https://www.kaggle.com/harlfoxem/housesalesprediction/data

## Публикации

<https://www.kaggle.com/arthurtok/feature-ranking-rfe-random-forest-linear-models>

<https://www.kaggle.com/harlfoxem/house-price-prediction-part-1>

<https://www.kaggle.com/auygur/step-by-step-house-price-prediction-r-2-0-77/output>

# Предварительный анализ собранных данных

## Анализ особенностей данных: потенциальные ошибки и пропущенные значения, группы и выбросы

##### Анализ количественных переменных

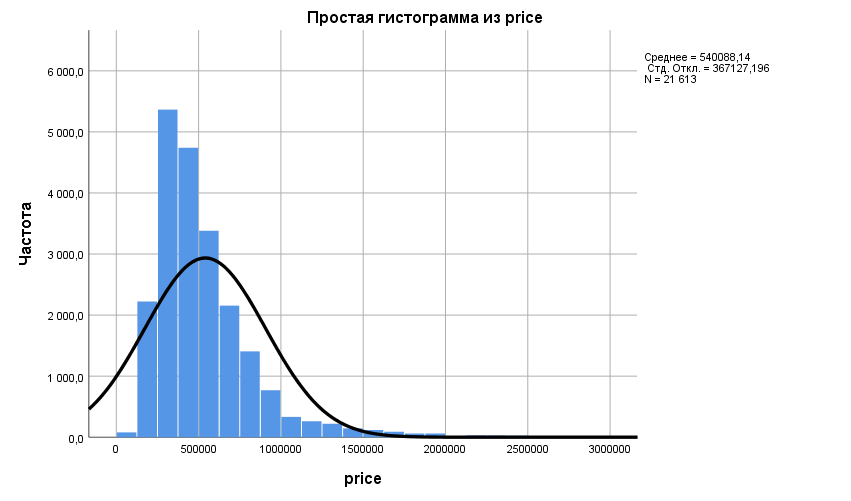
1. **Цена дома**

* **Форма гистограммы**

Правый хвост, очевидно, тяжелее, чем левый.

С помощью инструментов IBM SPSS Statistics произведем подсчет критерия эксцесса, позволяющего проверить «плосковершинность» или «узковершинность» эмпирического распределения, выраженную в числовой форме, и коэффициента асимметрии, меры отклонения распределения частоты от симметричного распределения. Распределение является нормальным, при условии, что показатели эксцесса и асимметрии находятся в диапазоне [-1;1]. Если показатели или асимметрии, или эксцесса не находятся в данном диапазоне, распределение нельзя назвать нормальным.

В таблице указаны значения, выходящие за рамки диапазона (Ex = 34,586; As = 4,024). Следовательно, принимая за основную гипотезу, что данные взяты из нормально распределенной генеральной совокупности, а за альтернативную, соответственно, что данные взяты не из нормально распределенной генеральной совокупности, мы отвергаем основную гипотезу. В данном случае перед нами «узковершинность».



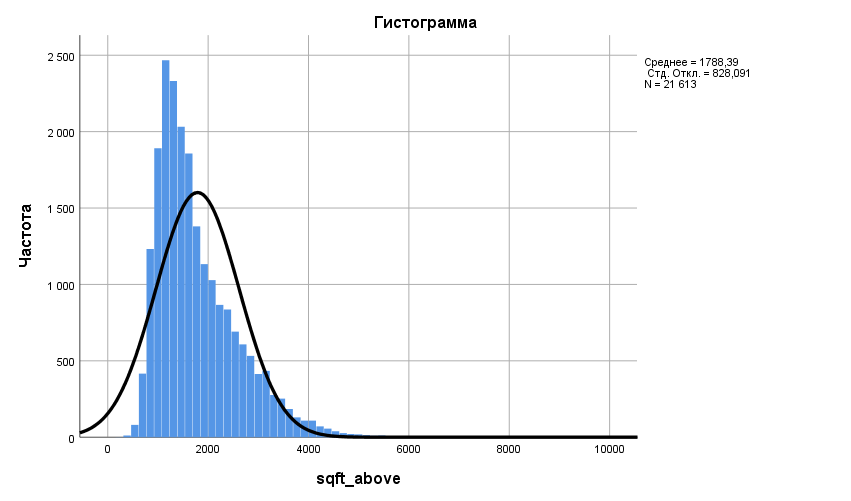
|  |  |  |
| --- | --- | --- |
| **Статистика** | | |
| price | | |
| N | Валидные | 21613 |
| Пропущенные | 0 |
| Среднее | | 540088,14 |
| Медиана | | 450000,00 |
| Стандартная отклонения | | 367127,196 |
| Дисперсия | | 134782378397,247 |
| Асимметрия | | 4,024 |
| Стандартная Ошибка асимметрии | | ,017 |
| Эксцесс | | 34,586 |
| Стандартная ошибка эксцесса | | ,033 |
| Диапазон | | 7625000 |
| Процентили | 25 | 321725,00 |
| 50 | 450000,00 |
| 75 | 645000,00 |

1. **Площадь без подвала**

* **Форма гистограммы**

Правый хвост, очевидно, тяжелее, чем левый.

С помощью инструментов IBM SPSS Statistics произведем подсчет критерия эксцесса и асимметрии. В таблице 2 указаны значения как эксцесса, так и асимметрии, не входящие в диапазон [-1;1] (Ex = 3,402; As = 1,447). Так, принимая за основную гипотезу, что данные взяты из нормально распределенной генеральной совокупности, а за альтернативную, соответственно, что данные взяты не из нормально распределенной генеральной совокупности, мы принимаем альтернативную гипотезу.

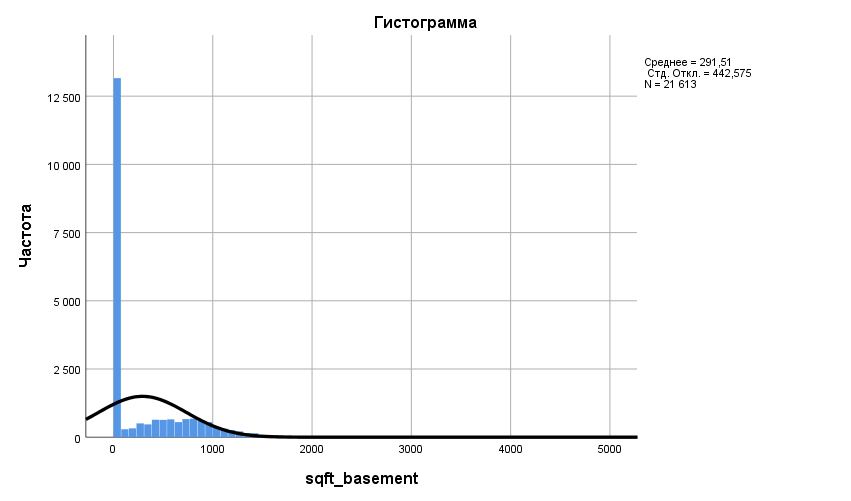


|  |  |  |
| --- | --- | --- |
| **Статистика** | | |
| sqft\_above | | |
| N | Валидные | 21613 |
| Пропущенные | 0 |
| Среднее | | 1788,39 |
| Медиана | | 1560,00 |
| Стандартная отклонения | | 828,091 |
| Дисперсия | | 685734,667 |
| Асимметрия | | 1,447 |
| Стандартная Ошибка асимметрии | | ,017 |
| Эксцесс | | 3,402 |
| Стандартная ошибка эксцесса | | ,033 |
| Диапазон | | 9120 |
| Процентили | 25 | 1190,00 |
| 50 | 1560,00 |
| 75 | 2210,00 |

1. **Площадь подвала**

* **Форма гистограммы**

С помощью инструментов IBM SPSS Statistics произведем подсчет критерия эксцесса и асимметрии. В таблице указаны значения как эксцесса, так и асимметрии, не входящие в диапазон [-1;1] (Ex = 2,716; As = 1,578). Так, принимая за основную гипотезу, что данные взяты из нормально распределенной генеральной совокупности, а за альтернативную, соответственно, что данные взяты не из нормально распределенной генеральной совокупности, мы принимаем альтернативную гипотезу.



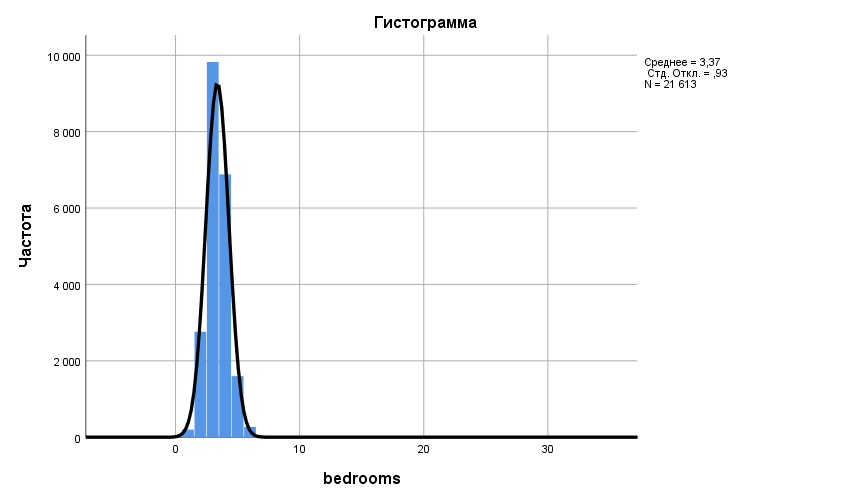
|  |  |  |
| --- | --- | --- |
| **Статистика** | | |
| sqft\_basement | | |
| N | Валидные | 21613 |
| Пропущенные | 0 |
| Среднее | | 291,51 |
| Медиана | | ,00 |
| Стандартная отклонения | | 442,575 |
| Дисперсия | | 195872,668 |
| Асимметрия | | 1,578 |
| Стандартная Ошибка асимметрии | | ,017 |
| Эксцесс | | 2,716 |
| Стандартная ошибка эксцесса | | ,033 |
| Диапазон | | 4820 |
| Процентили | 25 | ,00 |
| 50 | ,00 |
| 75 | 560,00 |

1. **Количество спален**

* **Форма гистограммы**

Полученная форма гистограммы схожа с формой нормальной кривой.

С помощью инструментов IBM SPSS Statistics произведем подсчет критерия эксцесса и асимметрии. В таблице указаны значения как эксцесса, так и асимметрии, не входящие в диапазон [-1;1] (Ex = 49,064; As = 1,974). Так, принимая за основную гипотезу, что данные взяты из нормально распределенной генеральной совокупности, а за альтернативную, соответственно, что данные взяты не из нормально распределенной генеральной совокупности, мы принимаем альтернативную гипотезу.



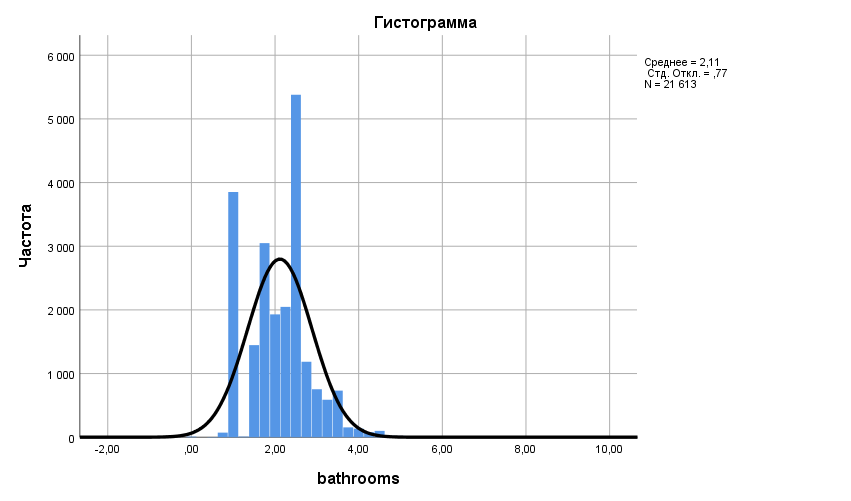
|  |  |  |
| --- | --- | --- |
| **Статистика** | | |
| bedrooms | | |
| N | Валидные | 21613 |
| Пропущенные | 0 |
| Среднее | | 3,37 |
| Медиана | | 3,00 |
| Стандартная отклонения | | ,930 |
| Дисперсия | | ,865 |
| Асимметрия | | 1,974 |
| Стандартная Ошибка асимметрии | | ,017 |
| Эксцесс | | 49,064 |
| Стандартная ошибка эксцесса | | ,033 |
| Диапазон | | 33 |
| Процентили | 25 | 3,00 |
| 50 | 3,00 |
| 75 | 4,00 |

1. **Количество ванн**

* **Форма гистограммы**

Полученная форма гистограммы схожа с формой нормальной кривой.

С помощью инструментов IBM SPSS Statistics произведем подсчет критерия эксцесса и асимметрии. В таблице указаны значения как эксцесса, так и асимметрии, не входящие в диапазон [-1;1] (Ex = 1,280; As = 0,511). Так, принимая за основную гипотезу, что данные взяты из нормально распределенной генеральной совокупности, а за альтернативную, соответственно, что данные взяты не из нормально распределенной генеральной совокупности, мы принимаем альтернативную гипотезу.



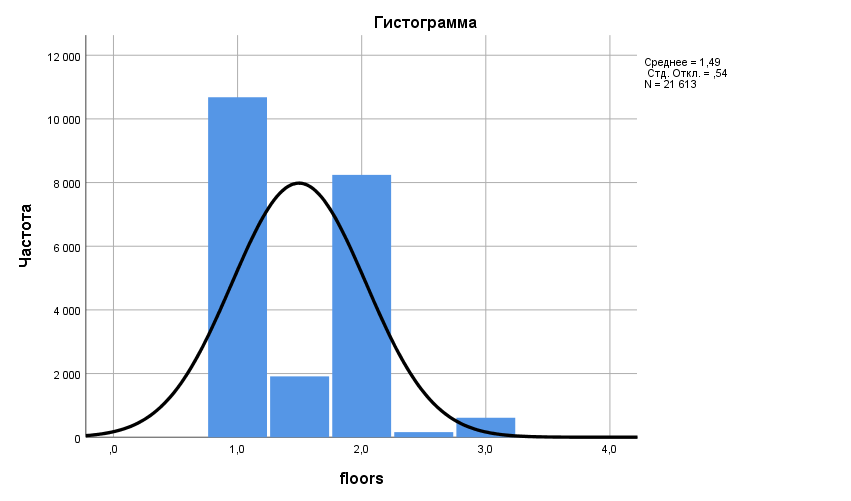
|  |  |  |
| --- | --- | --- |
| **Статистика** | | |
| bathrooms | | |
| N | Валидные | 21613 |
| Пропущенные | 0 |
| Среднее | | 2,1148 |
| Медиана | | 2,2500 |
| Стандартная отклонения | | ,77016 |
| Дисперсия | | ,593 |
| Асимметрия | | ,511 |
| Стандартная Ошибка асимметрии | | ,017 |
| Эксцесс | | 1,280 |
| Стандартная ошибка эксцесса | | ,033 |
| Диапазон | | 8,00 |
| Процентили | 25 | 1,7500 |
| 50 | 2,2500 |
| 75 | 2,5000 |

1. **Количество этажей**

* **Форма гистограммы**

Полученная форма гистограммы схожа с формой нормальной кривой.

С помощью инструментов IBM SPSS Statistics произведем подсчет критерия эксцесса и асимметрии. В таблице указаны значения как эксцесса, так и асимметрии, входящие в диапазон [-1;1] (Ex = -0,485; As = 0,616). Так, принимая за основную гипотезу, что данные взяты из нормально распределенной генеральной совокупности, а за альтернативную, соответственно, что данные взяты не из нормально распределенной генеральной совокупности, мы принимаем основную гипотезу.

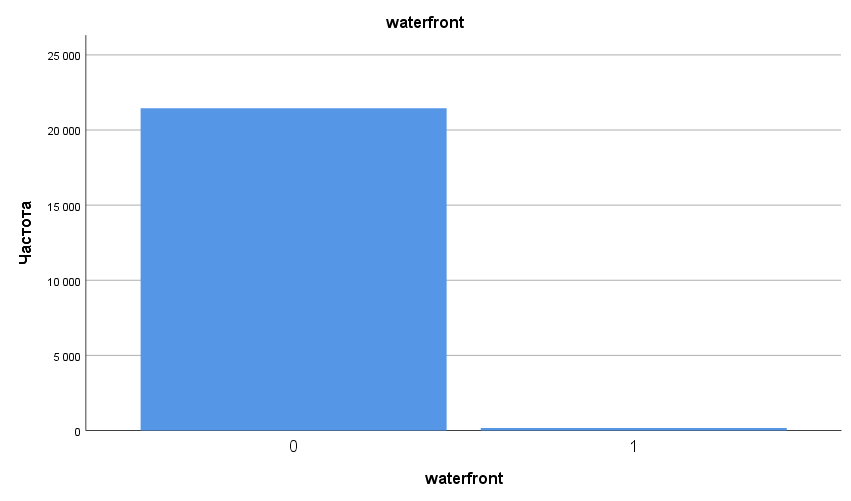


|  |  |  |
| --- | --- | --- |
| **Статистика** | | |
| floors | | |
| N | Валидные | 21613 |
| Пропущенные | 0 |
| Среднее | | 1,494 |
| Медиана | | 1,500 |
| Стандартная отклонения | | ,5400 |
| Дисперсия | | ,292 |
| Асимметрия | | ,616 |
| Стандартная Ошибка асимметрии | | ,017 |
| Эксцесс | | -,485 |
| Стандартная ошибка эксцесса | | ,033 |
| Диапазон | | 2,5 |
| Процентили | 25 | 1,000 |
| 50 | 1,500 |
| 75 | 2,000 |

### Анализ качественных переменных.

1. Наличие вида на берег

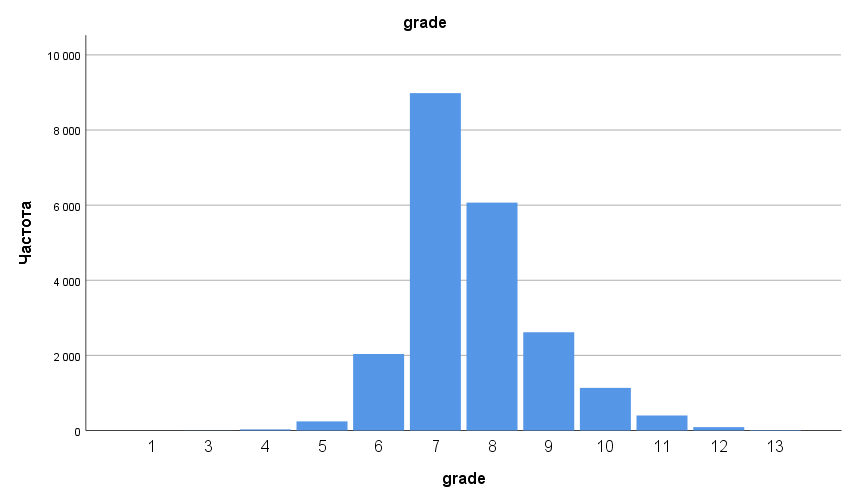
Из гистограммы выше видим значительную асимметрию влево. При этом преобладают дома без вида на берег.



1. Оценка дома по системе Округа Кинг

Из гистограммы можно сделать вывод об отсутствии асимметрии и эксцесса, а распределение данных кажется нормальным.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **waterfront** | | | | | |
|  | | Частота | Проценты | Валидный процент | Накопленный процент |
| Валидные | 0 | 21450 | 99,2 | 99,2 | 99,2 |
| 1 | 163 | ,8 | ,8 | 100,0 |
| Всего | 21613 | 100,0 | 100,0 |  |

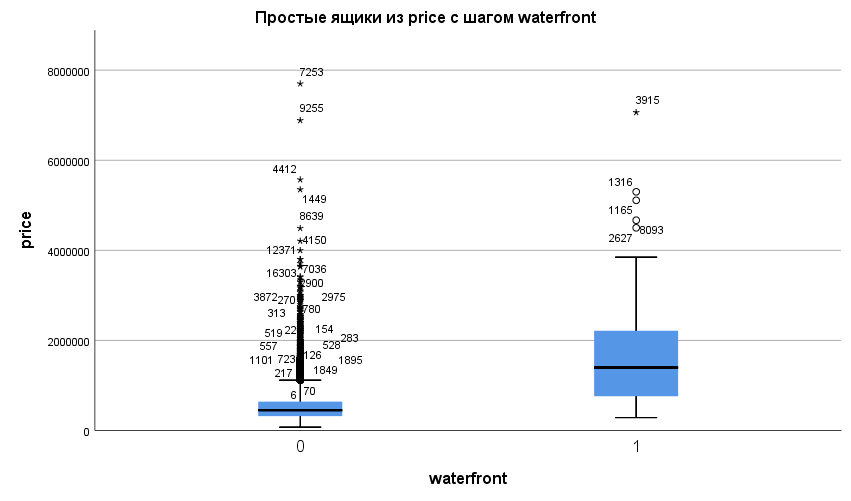


|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **grade** | | | | | |
|  | | Частота | Проценты | Валидный процент | Накопленный процент |
| Валидные | 1 | 1 | ,0 | ,0 | ,0 |
| 3 | 3 | ,0 | ,0 | ,0 |
| 4 | 29 | ,1 | ,1 | ,2 |
| 5 | 242 | 1,1 | 1,1 | 1,3 |
| 6 | 2038 | 9,4 | 9,4 | 10,7 |
| 7 | 8981 | 41,6 | 41,6 | 52,3 |
| 8 | 6068 | 28,1 | 28,1 | 80,3 |
| 9 | 2615 | 12,1 | 12,1 | 92,4 |
| 10 | 1134 | 5,2 | 5,2 | 97,7 |
| 11 | 399 | 1,8 | 1,8 | 99,5 |
| 12 | 90 | ,4 | ,4 | 99,9 |
| 13 | 13 | ,1 | ,1 | 100,0 |
| Всего | 21613 | 100,0 | 100,0 |  |

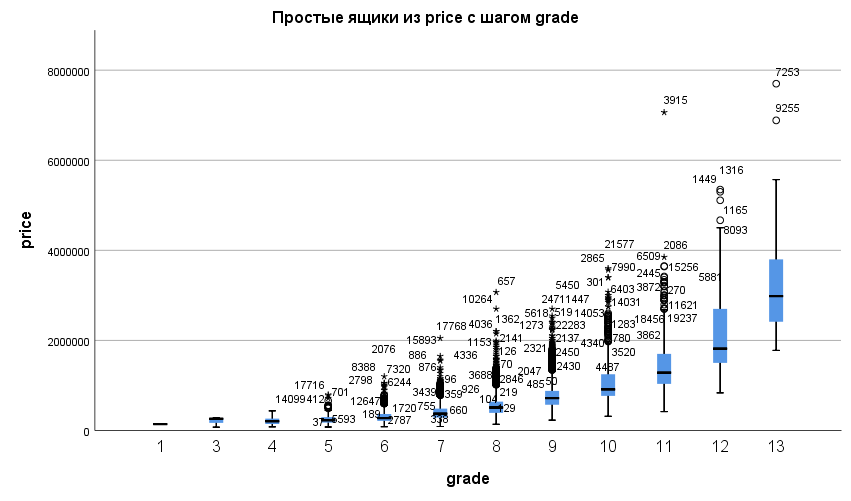
## 2.2 Анализ статистической связи.

### Графический анализ пары «числовая зависимая переменная – качественная независимая переменная».

Проанализируем влияние наличие вида на берега на цену дома. 0 – вид отсутствует, 1 – вид есть. Из диаграммы Бокса-Вискера видно, что дома с видом на берег стоят значительно больше.



Проанализируем влияние оценки по системе Округа Кинг на цену. Из диаграммы Бокса-Вискера видно что дома с большей оценкой стоят больше.



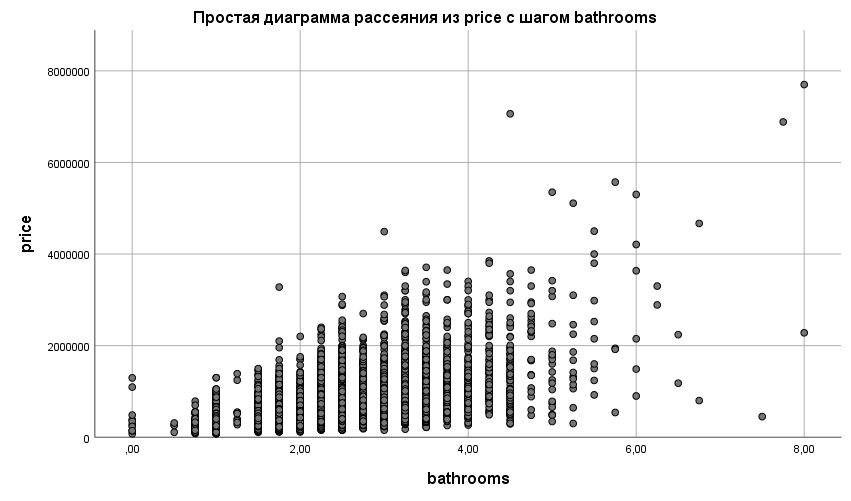
### Графический анализ пары «числовая зависимая переменная – числовая независимая переменная».



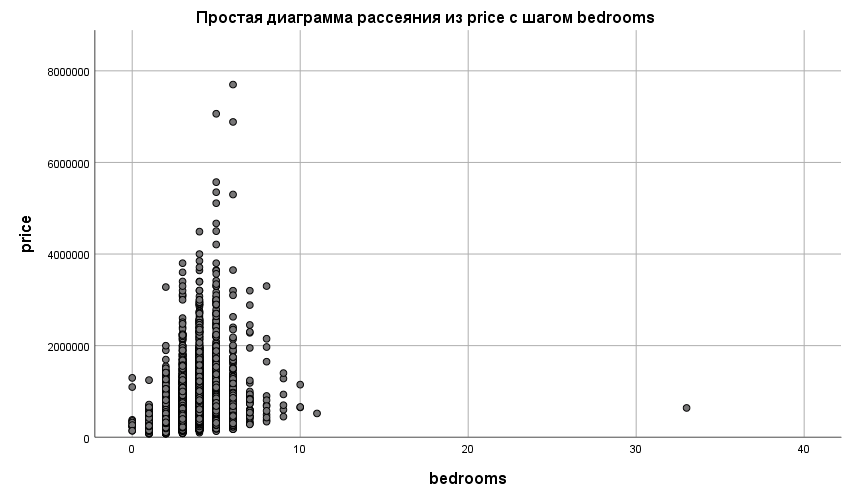
Очевидно, что площадь дома без подвала напрямую влияет на цену дома. Этот простой и логичный вывод подтверждает соответствующая диаграмма рассеивания. В данном случае зависимость прямая и, вероятнее всего, линейная.



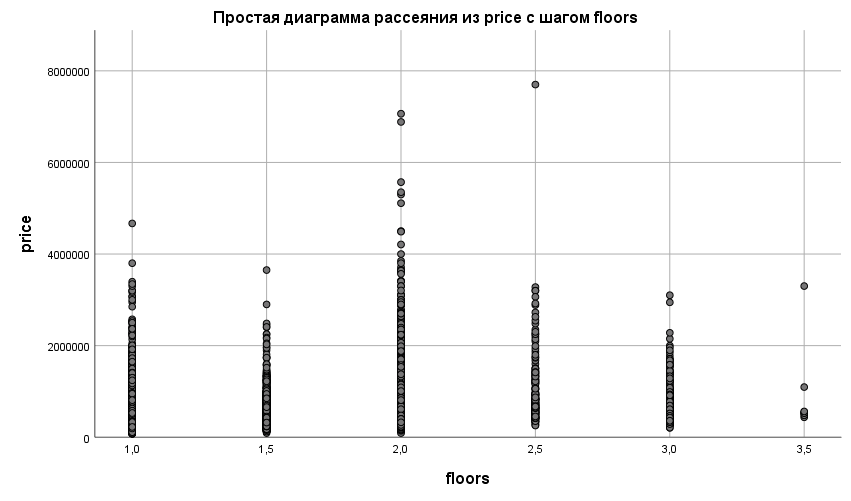
На данном графике зависимость между площадью подвала и ценой дома также прямая и, вероятнее всего, линейная.



На данном графике зависимость между количеством ванн и ценой дома также прямая и, вероятнее всего, линейная.



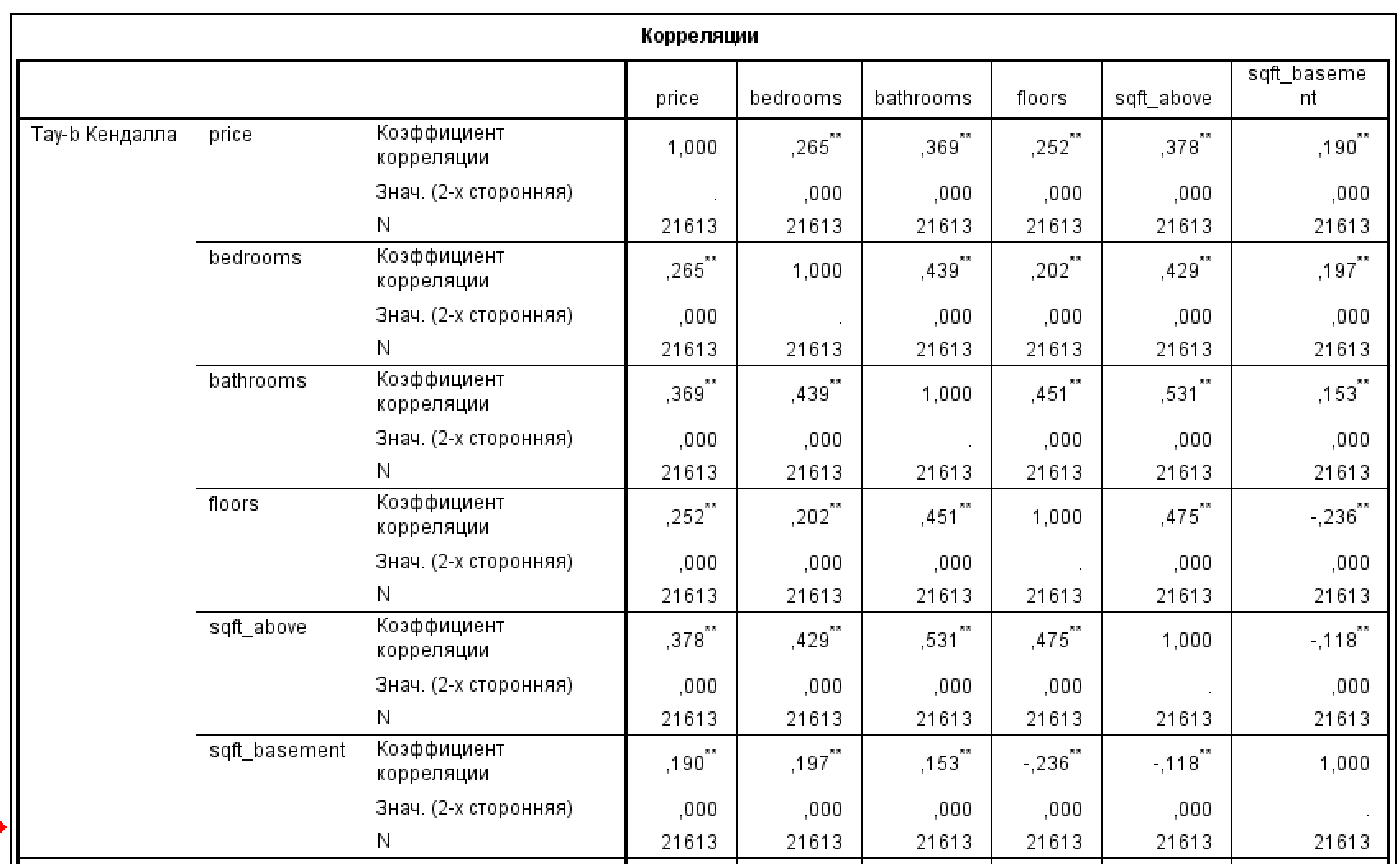
На данном графике зависимость между количеством спален и ценой дома также прямая и, вероятнее всего, линейная.

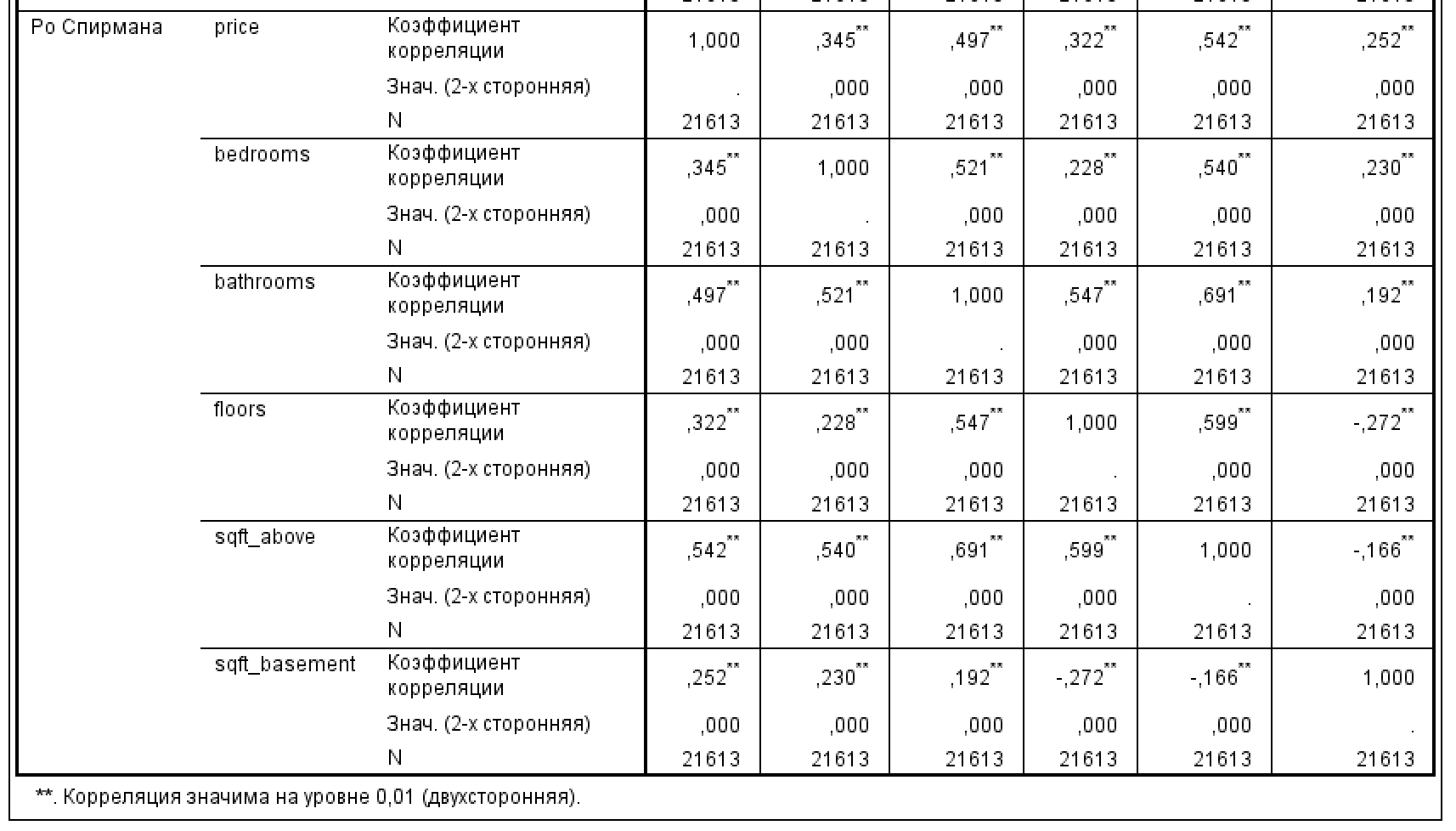


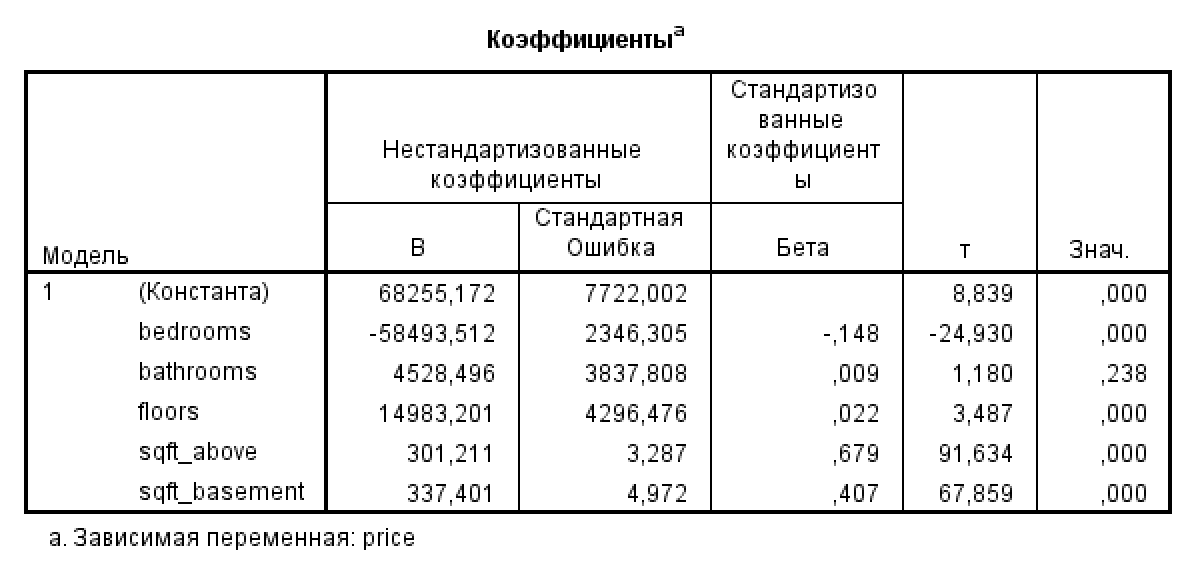
На данном графике зависимость между количеством этажей и ценой дома также прямая и, вероятнее всего, линейная.

Для формальной проверки гипотезы о наличии связи посчитаем коэффициенты корреляции Пирсона и Спирмена, а также тау Кендала и приведем результаты проверки их значимости.









Коэффициенты корреляции Пирсона и Спирмена, а также тау Кендала и проверка значимости подтверждают наличие связи между зависимой и независимыми переменными.

### Анализ наличия корреляции между независимыми переменными.



По таблице статистики коллинеарности можно увидеть, что между независимыми переменными корреляция наблюдается, но т.к. значения VIF < 3, то корреляция не сильно вредит исследованию.

### Предварительная проверка гипотез

1. Гипотеза о том, что существует зависимость между ценой дома и его количественных (площадь дома, количество спален, ванн, этажей и т.д.) и номинальной переменной (наличие вида на берег) подтвердилась.
2. Гипотеза о том, что потенциальные покупатели обращают больше внимания на площадь дома, чем на количество ванн, спален, этажей, площадь подвала, не подтвердилась.
3. Гипотеза о том, что существует зависимость между ценой дома и его общей оценкой по системе Округа Кинг, то есть покупатели обращают внимание на рейтинг жилья, подтвердилась.