King County, USA, Housing Prices

Nigel Brown

2020/10/21

Contents

1 Introduction

The goal of this project is to predict house prices from the House Sales in King County, USA dataset downloaded from kaggle.

The project followed the stages of:

1. Data Exploration
2. Cleaning the data to ready it for modeling
3. Modeling: Linear, randomForest and xgboost
4. Evaluating the models performance and finalizing the results

2 Initial data exploration

The dataset consists of house prices in King County, Washington, USA from observed sales between May 2014 and May 2015. The data consists of 21613 rows of data, each row observes a single sale. There are 21 features in the dataset. The features are:

Variable	Description
id	Unique ID for each sale
date	Date of the observed sale
price	Price of each house sold in USD
bedrooms	Number of bedrooms
bathrooms	Number of bathrooms, where .5 is a room with a toilet but no with bath or shower
sqft living	Square footage of the interior living space of the house
sqft lot	Square footage of the land area the house resides on
floors	Number of floors
waterfront	Does the house overlook the waterfront front
view	An index from 0 to 4 of how good the view of the property is
condition	An index from 1 to 5 on the condition of the house when sold
grade	An index from 1 to 13, where 1-3 have poor construction and design, 4 - 6 have below average construction and design, 7 has an average level of construction and design, 8 -11 have above average construction and design and 11 -13 have high quality construction and design
sqrt above	Square footage of the interior housing space above ground level
sqrt basement	Square footage of the interior housing space below ground level
yr_built	the year the house was built
yr_renovated	the year of the house's last renovation
zipcode	the area zip code where the house is situated
lat	Latitude
long	Longitude
sqft_living15	The square footage of the interior living space for the closest 15 neighbors
sqft_lot15	The square footage of land for the closest 15 neighbors' houses

Price will be utilized as the outcome column for our models.

The next step is to analyze the data for missing values. For this analysis the df_status function from the funModelling package is utilized.

Table 2: Dataset status

variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
id	0	0.00	0	0	0	0	character	21436
date	0	0.00	0	0	0	0	POSIXct/POSIXt	372
price	0	0.00	0	0	0	0	numeric	4028
bedrooms	13	0.06	0	0	0	0	numeric	13
bathrooms	10	0.05	0	0	0	0	numeric	30
sqft_living	0	0.00	0	0	0	0	numeric	1038
sqft_lot	0	0.00	0	0	0	0	numeric	9782
floors	0	0.00	0	0	0	0	numeric	6
waterfront	21450	99.25	0	0	0	0	numeric	2
view	19489	90.17	0	0	0	0	numeric	5
condition	0	0.00	0	0	0	0	numeric	5
grade	0	0.00	0	0	0	0	numeric	12
sqft_above	0	0.00	0	0	0	0	numeric	946
sqft_basement	13126	60.73	0	0	0	0	numeric	306
yr_built	0	0.00	0	0	0	0	numeric	116
yr_renovated	20699	95.77	0	0	0	0	numeric	70
zipcode	0	0.00	0	0	0	0	numeric	70
lat	0	0.00	0	0	0	0	numeric	5034
long	0	0.00	0	0	0	0	numeric	752
sqft_living15	0	0.00	0	0	0	0	numeric	777
sqft_lot15	0	0.00	0	0	0	0	numeric	8689

- **q_zeros:** quantity of zeros (p_zeros: in percent)
- **q_inf:** quantity of infinite values (p_inf: in percent)
- **q_na:** quantity of NA (p_na: in percent)
- **type:** the variable type
- **unique:** quantity of unique values

3 Data Preprocessing

As is shown in the table above there are no instances of missing data or values being infinite, however there are a number of variables where the percentage of zeros is greater than 60%, these may not be useful for modeling and they may dramatically bias the model. Therefore For this project the decision is made to remove these variables from the dataset. The features removed are: waterfront, view, sqft_basement, yr_renovated.

Once the predominately zero columns are removed, there are 17 left in the dataset. The date feature is split into year and month features and the original date feature is dropped. It is also decided to convert all variables of type double except bathrooms, lat and long to integer type.

3.1 Data Anomilies

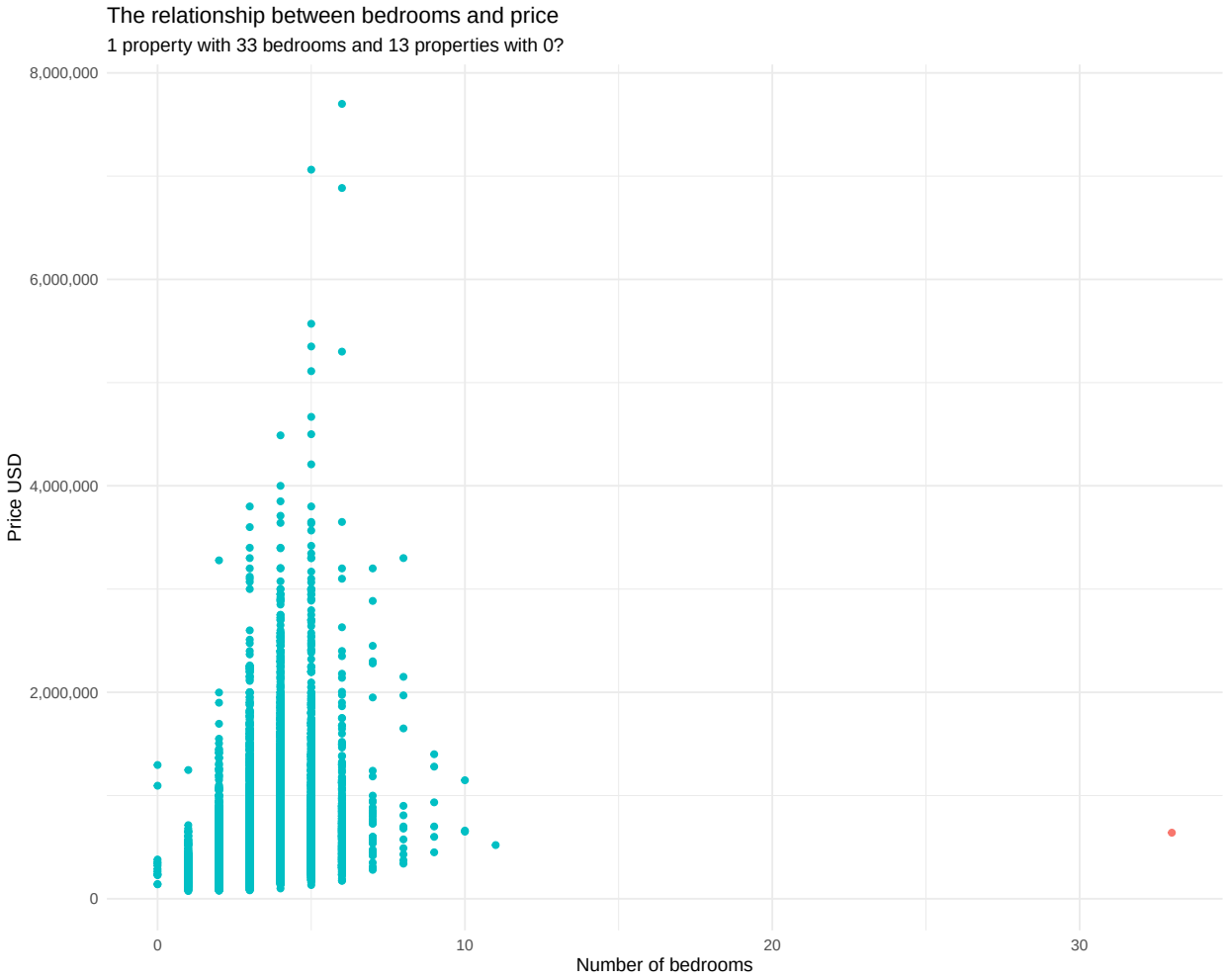


Table 3: Anomalous Data

id	price	bedrooms	bathrooms	floors	sqft_living
6306400140	1095000	0	0.00	3	3064
3918400017	380000	0	0.00	3	1470
1453602309	288000	0	1.50	3	1430
6896300380	228000	0	1.00	1	390
2954400190	1295650	0	0.00	2	4810
2569500210	339950	0	2.50	2	2290
2310060040	240000	0	2.50	2	1810
3374500520	355000	0	0.00	2	2460
7849202190	235000	0	0.00	2	1470
7849202299	320000	0	2.50	2	1490
9543000205	139950	0	0.00	1	844
2402100895	640000	33	1.75	1	1620
1222029077	265000	0	0.75	1	384
3980300371	142000	0	0.00	1	290

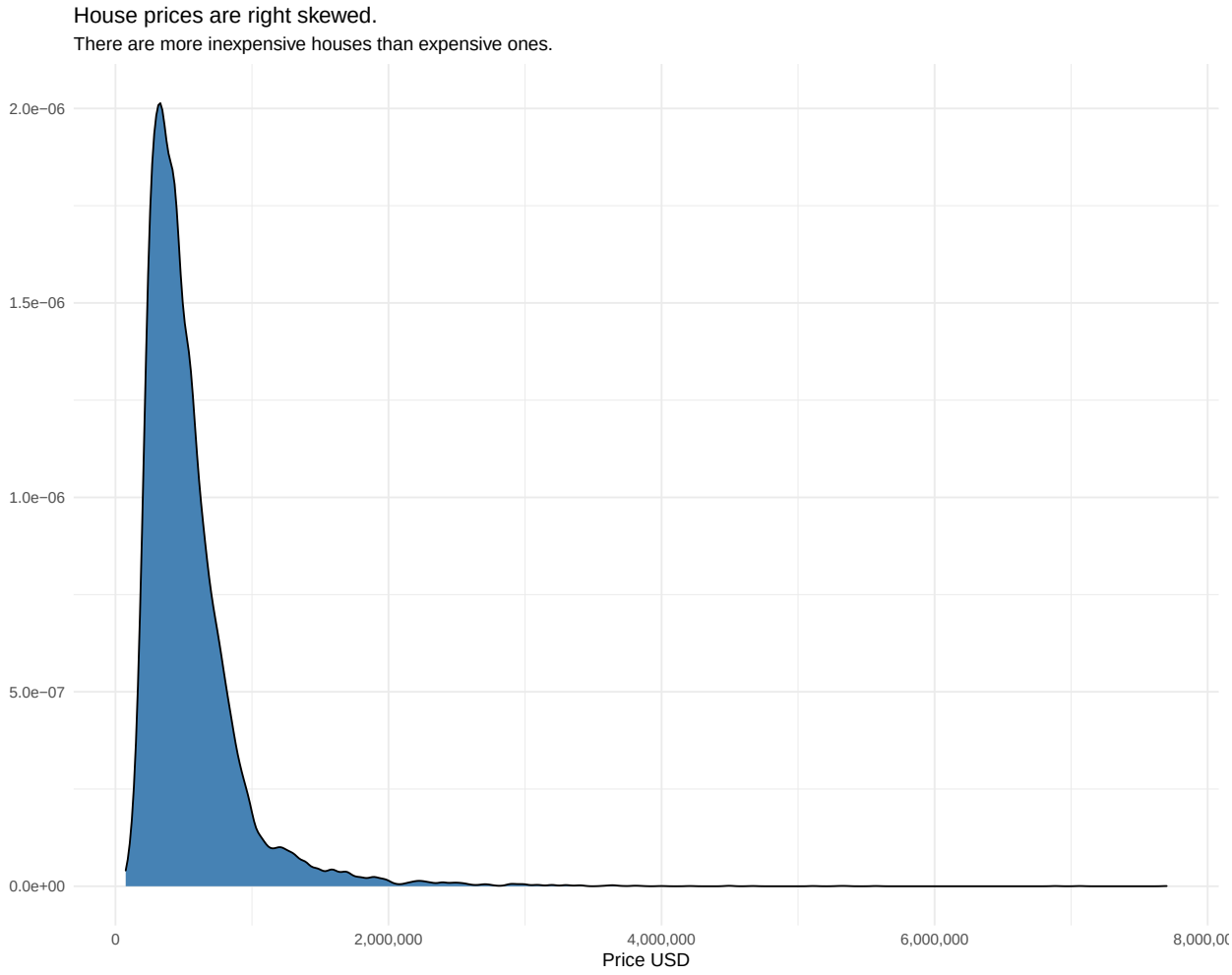
Exploring the data it is found that a single property has been listed with 33 bedrooms. This property is

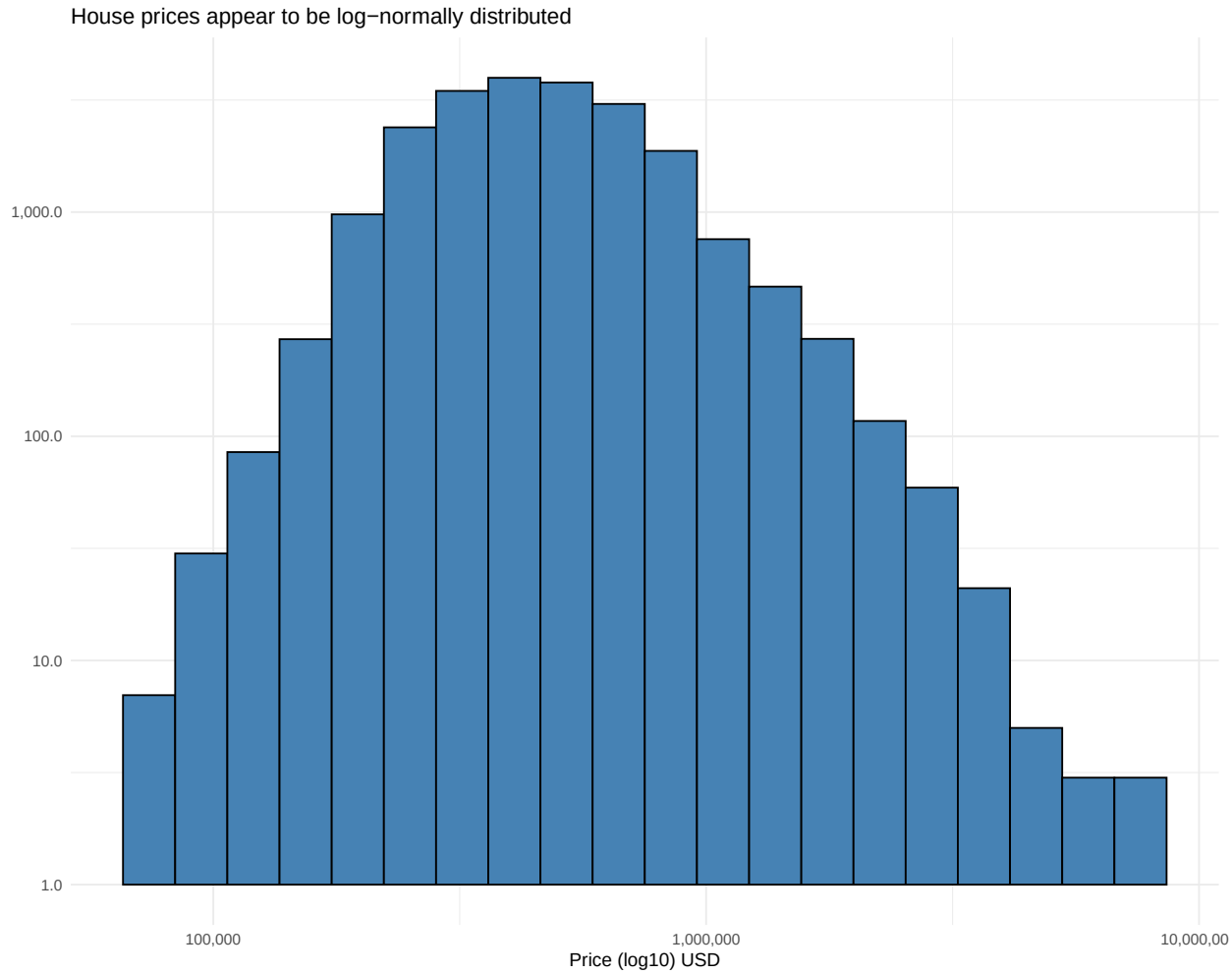
re-entered as a 3 bedroom property. Also 13 properties are listed with zero bedrooms, as these properties range in size and there is no obvious method of reincorporating these properties in the data with a bedroom count that is explainable, these properties are dropped from the data.

4 Exploritory Data Analysis

Now that the data is cleaned it consists of 21600rows of data and 18 columns. An exploratory data analysis is now performed to visualize relationships and patterns in the data.

4.1 How are house prices distributed ?



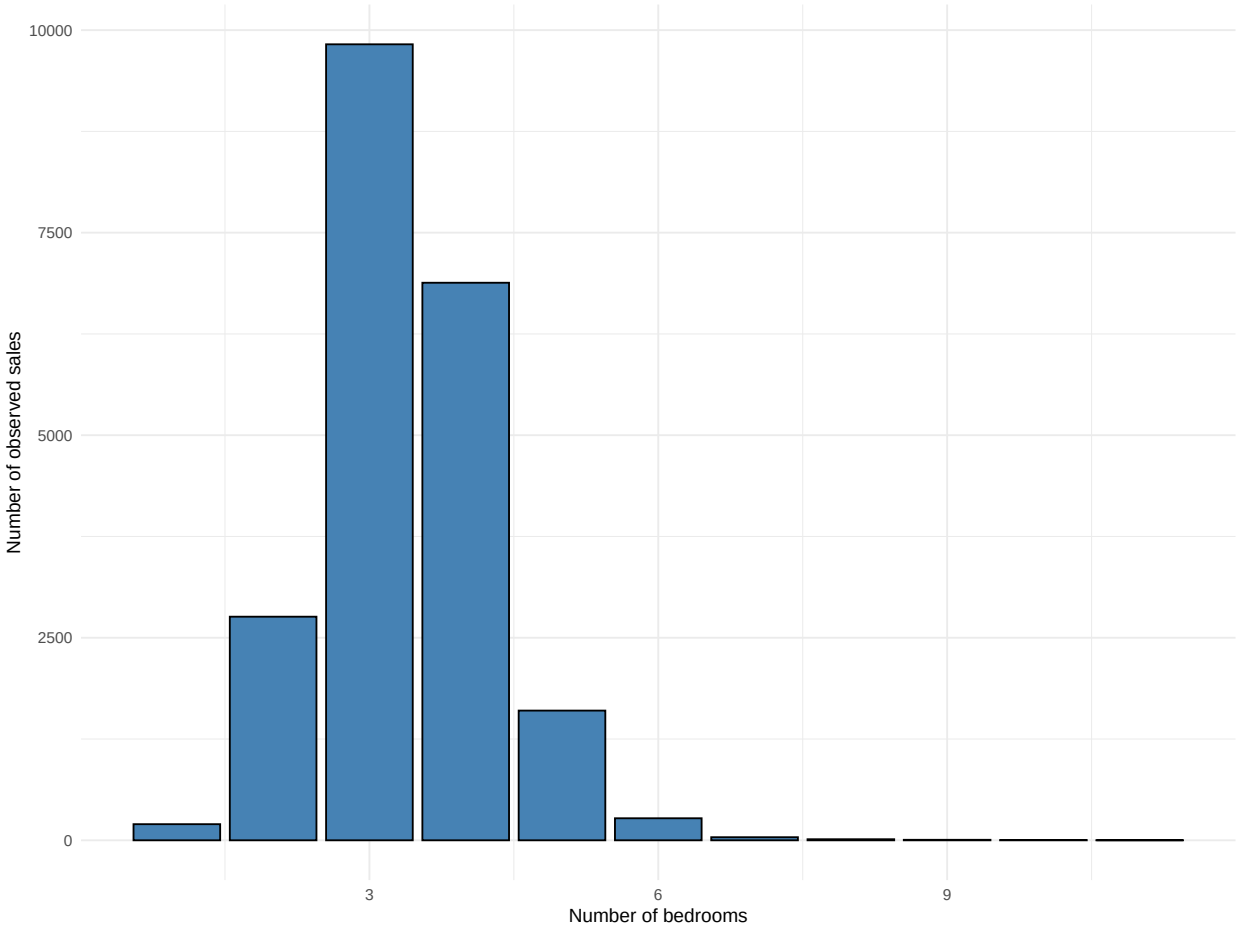


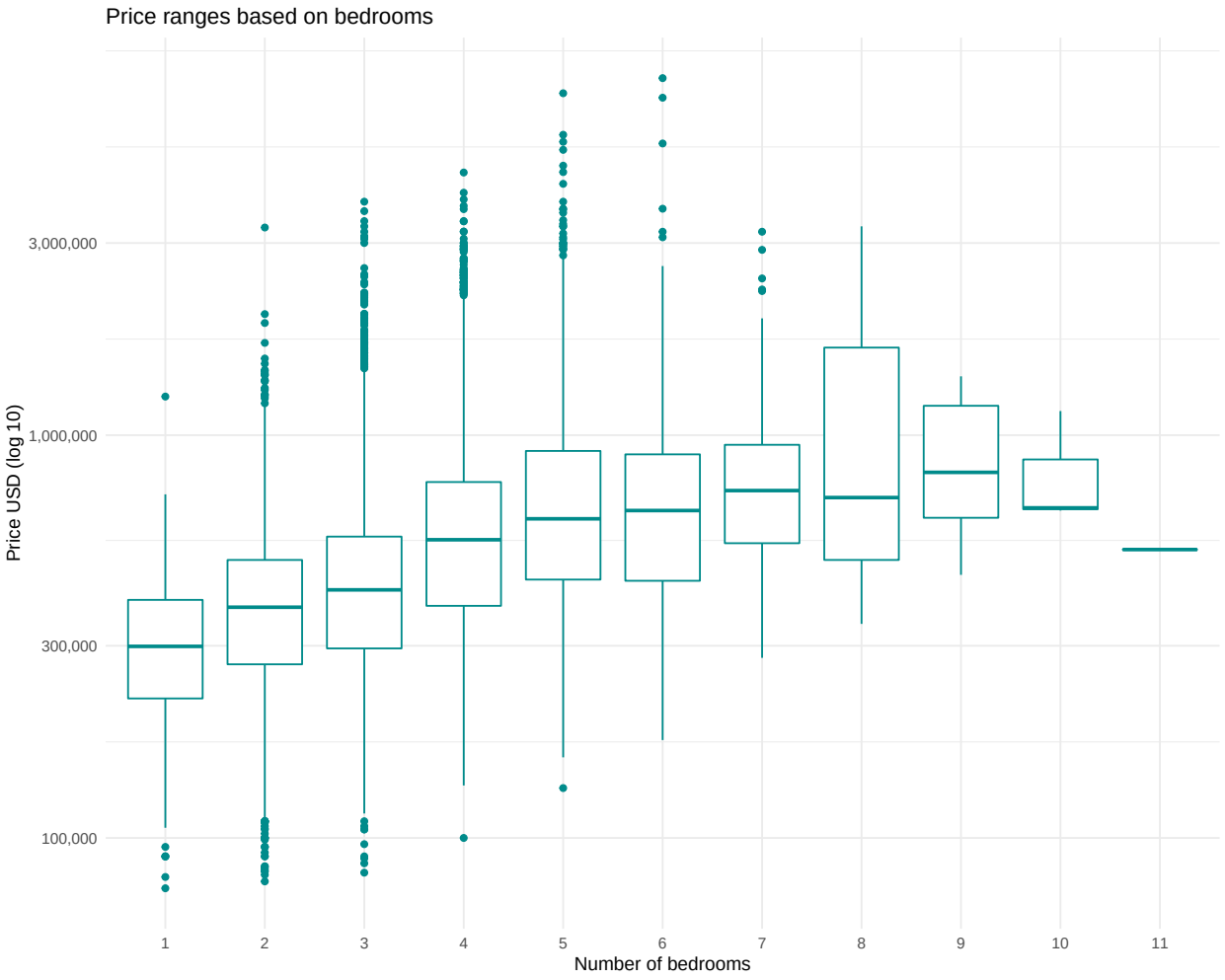
A strong argument can be made that the price should be log-transformed. The advantages of doing this are that no houses would be predicted with negative sale prices and that errors in predicting expensive houses will not have an undue influence on the model. Also, from a statistical perspective, a logarithmic transform may also stabilize the variance in a way that makes inference more legitimate. When the models are built a final pre-processing step of transforming the prices into logs will be performed.

How many houses of each bedroom count were sold?

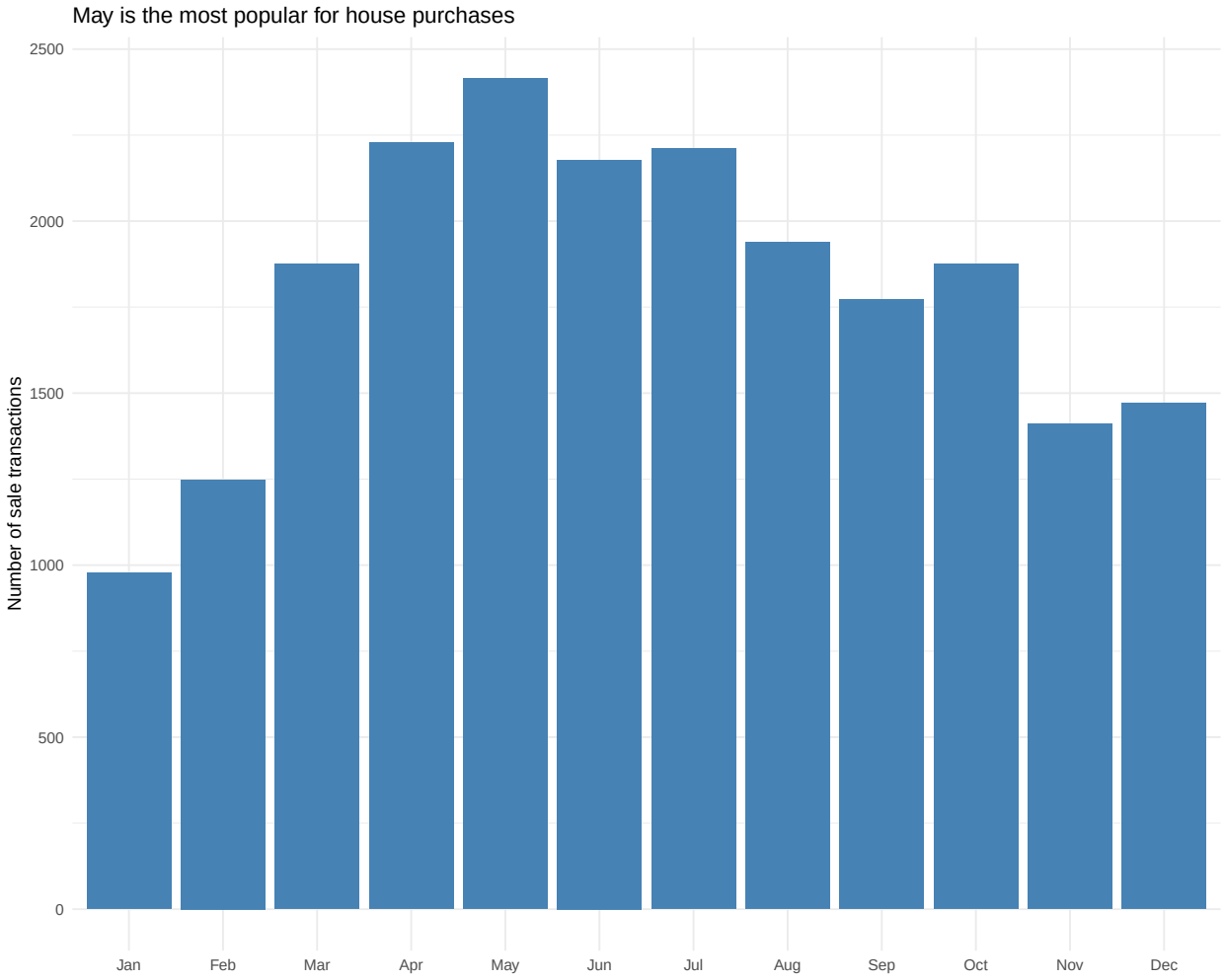
4 bedroom houses sell more often than other house sizes

3 bedroom houses are the next most frequently sold




Price ranges based on bedrooms
Price USD (log 10)
3,000,000
1,000,000
300,000
100,000
1
2
3
4
5
6
7
8
9
10
11
Number of bedrooms

4.2 Is there a month when most sales occur?



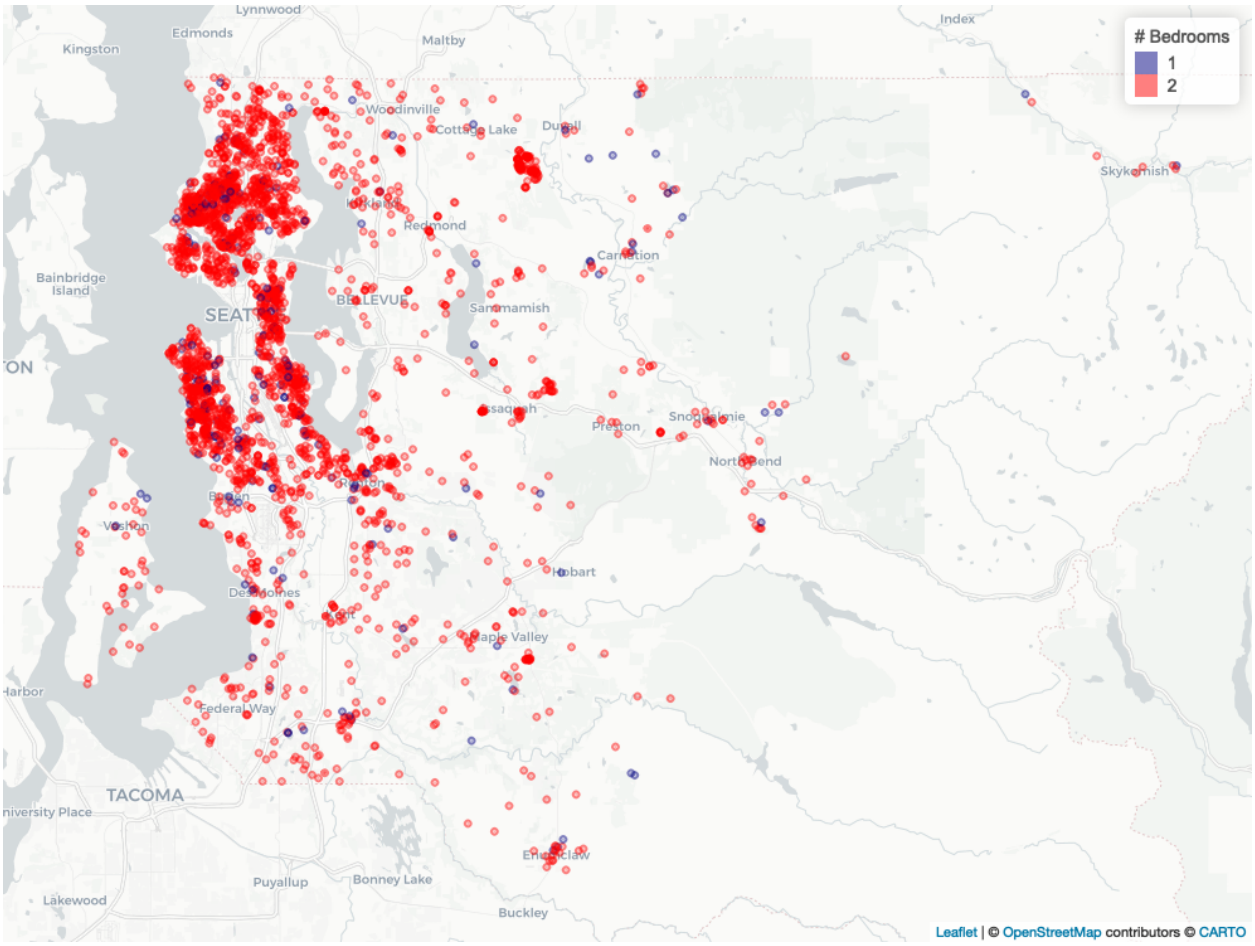


Figure 1: Location 1 and 2 bedroom houses sold

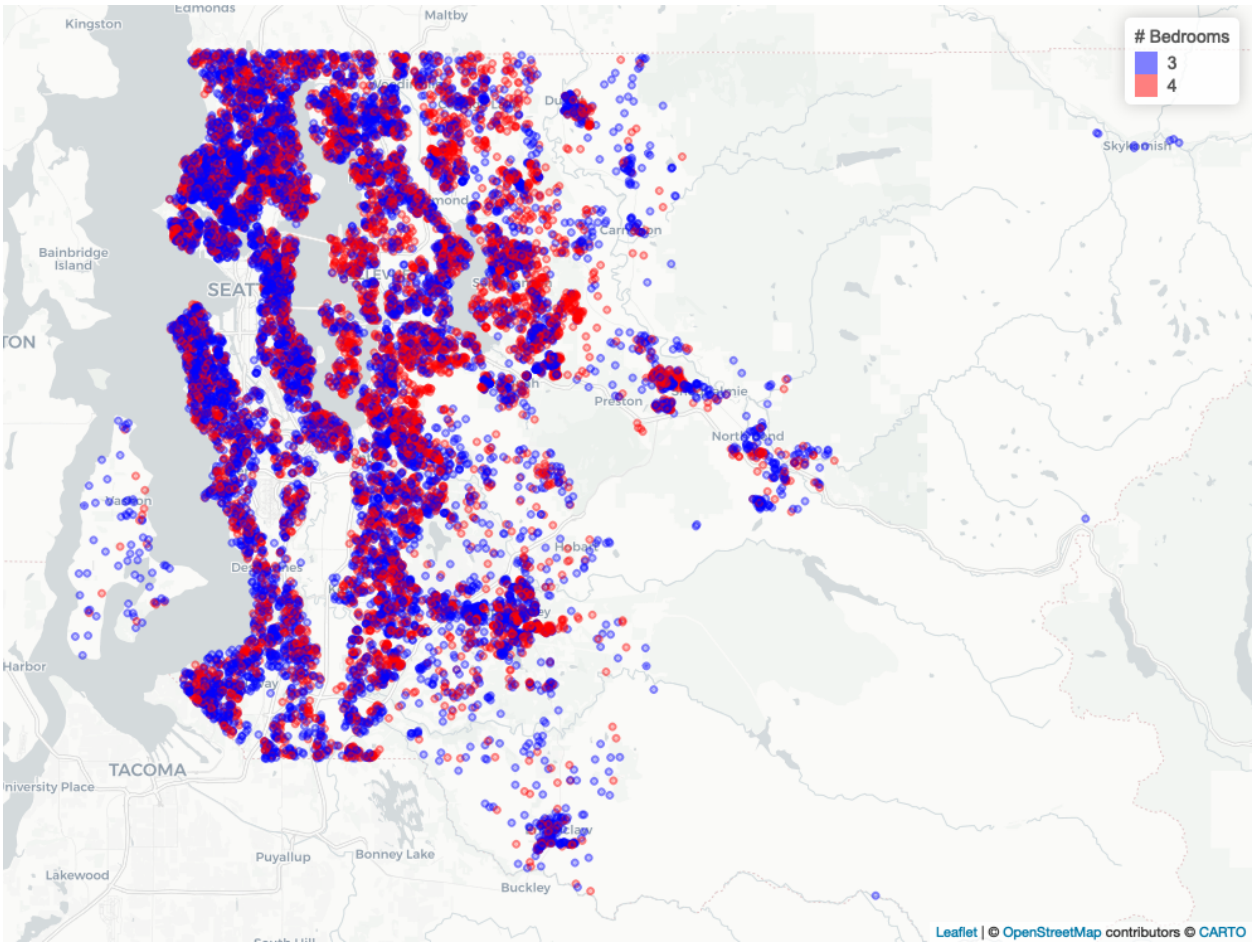


Figure 2: Location 3 and 4 bedroom houses sold

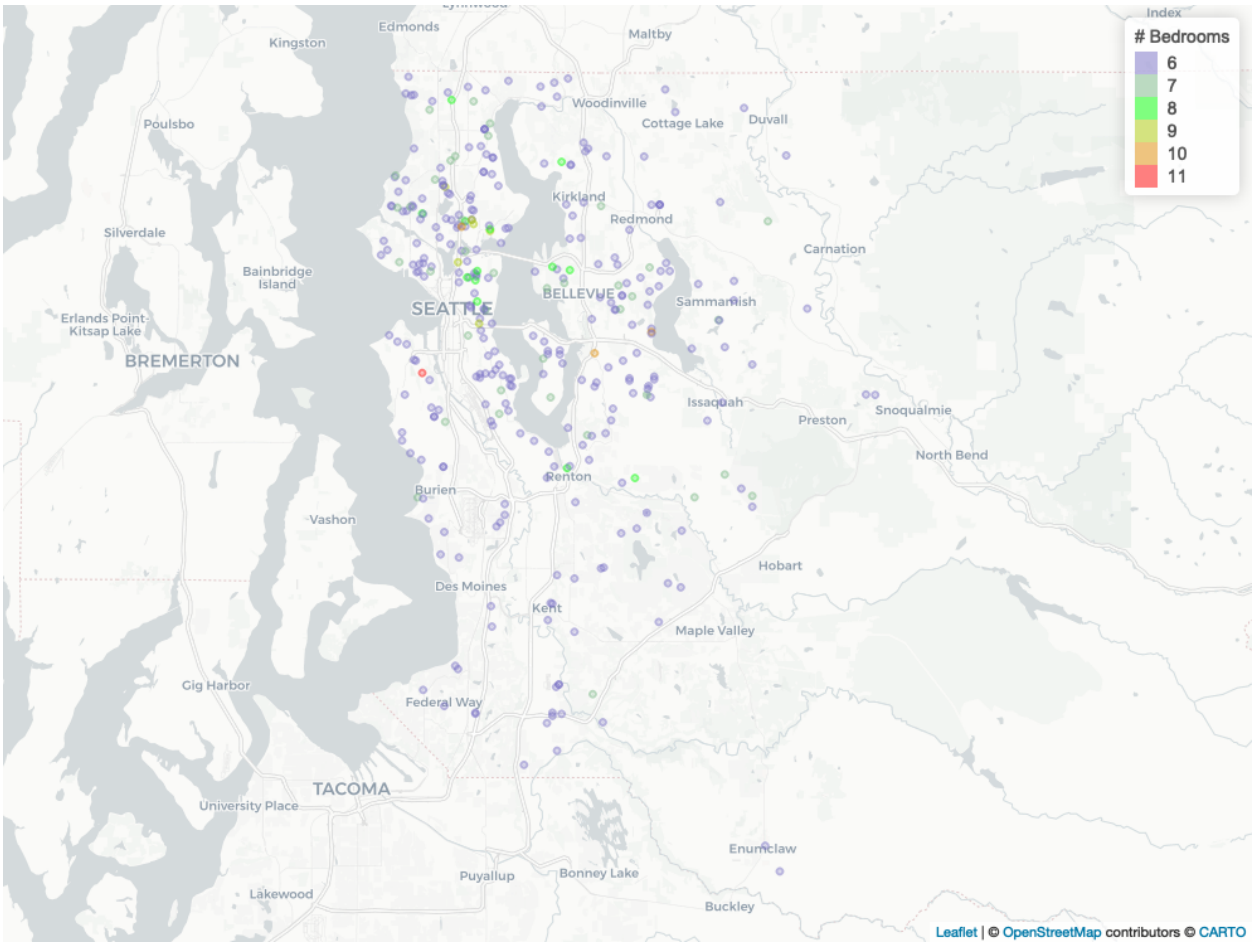
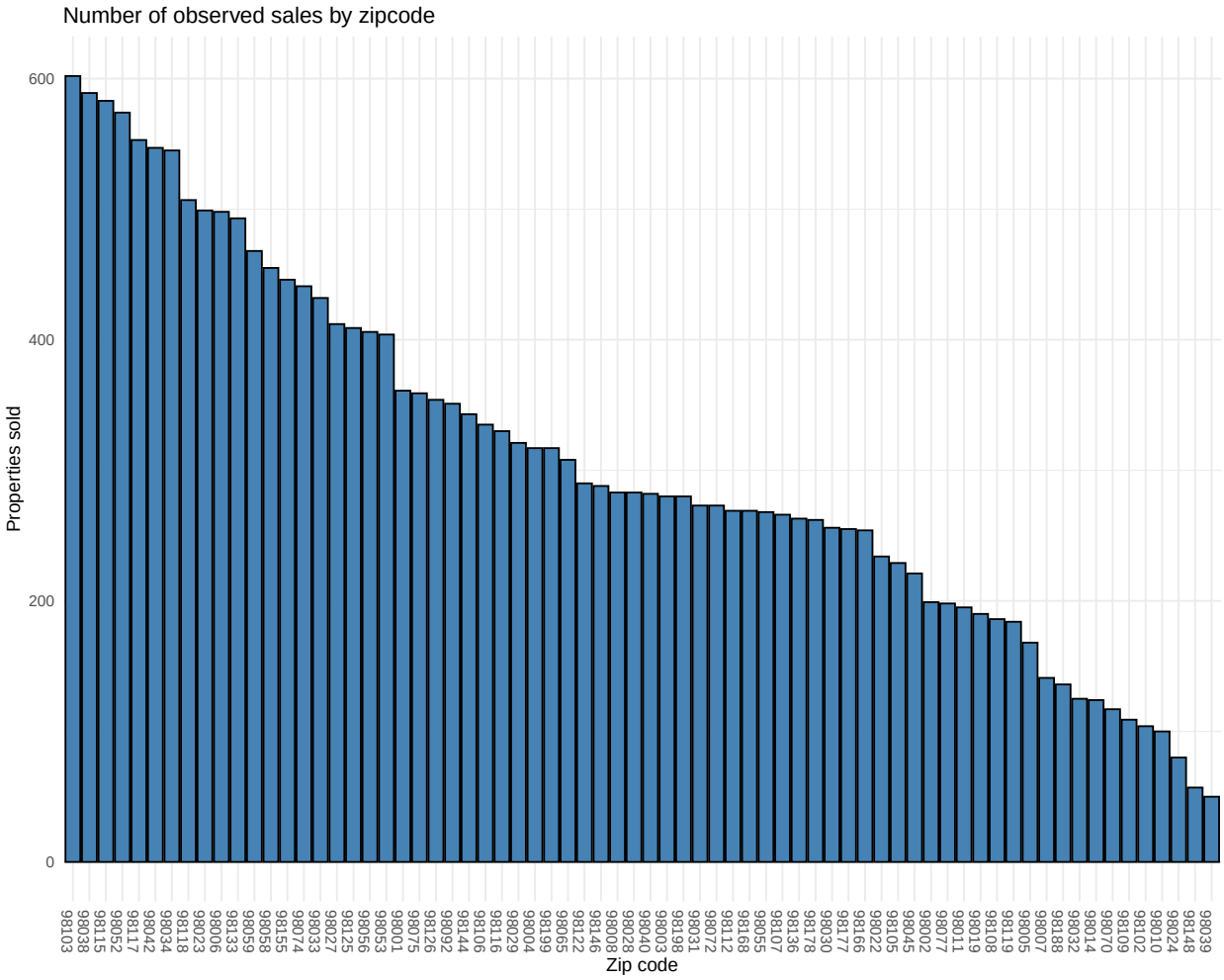


Figure 3: Location 5 or more bedroom houses sold

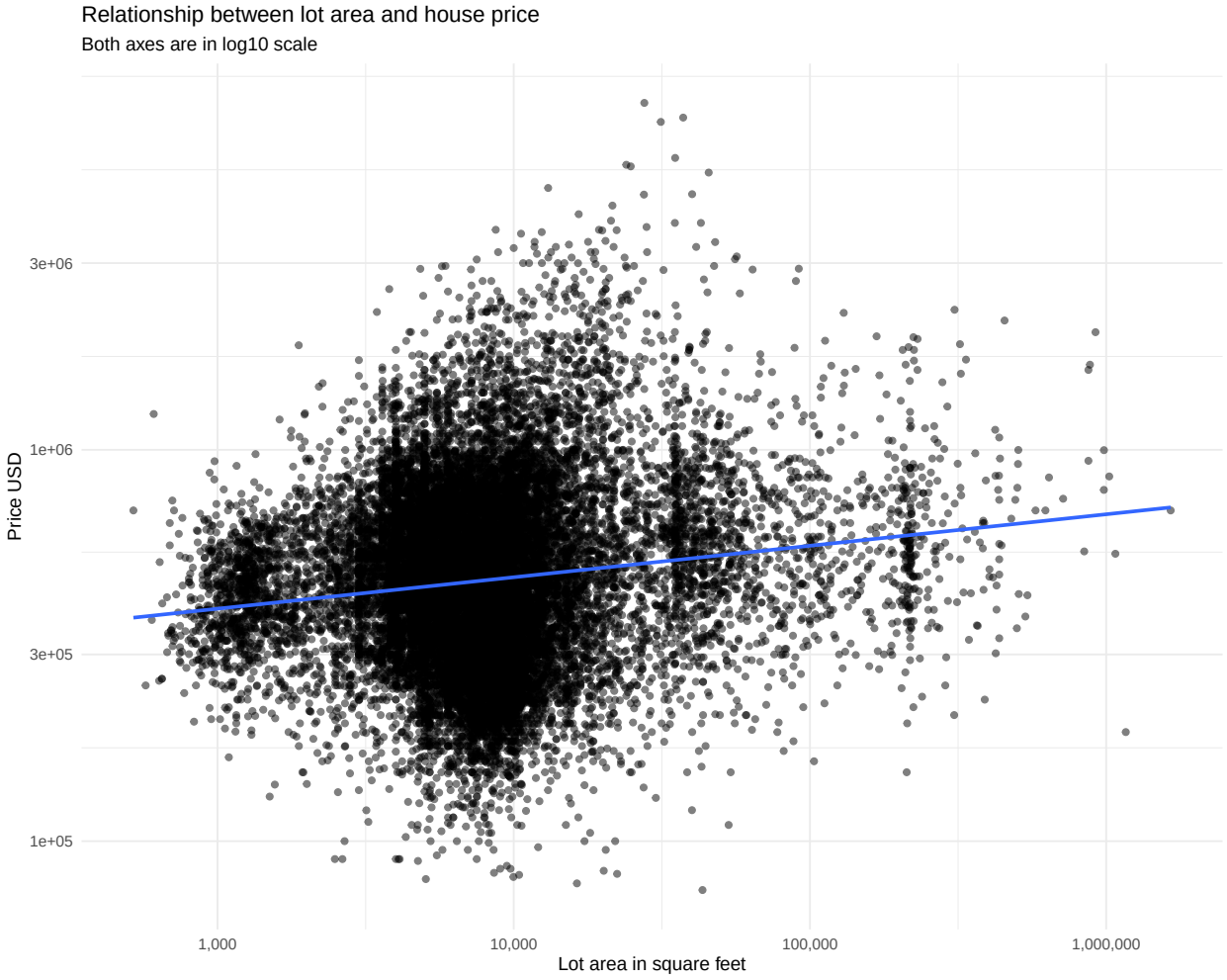
4.3 Spatial analysis plots

4.3.1 Where are the sales located within King County?

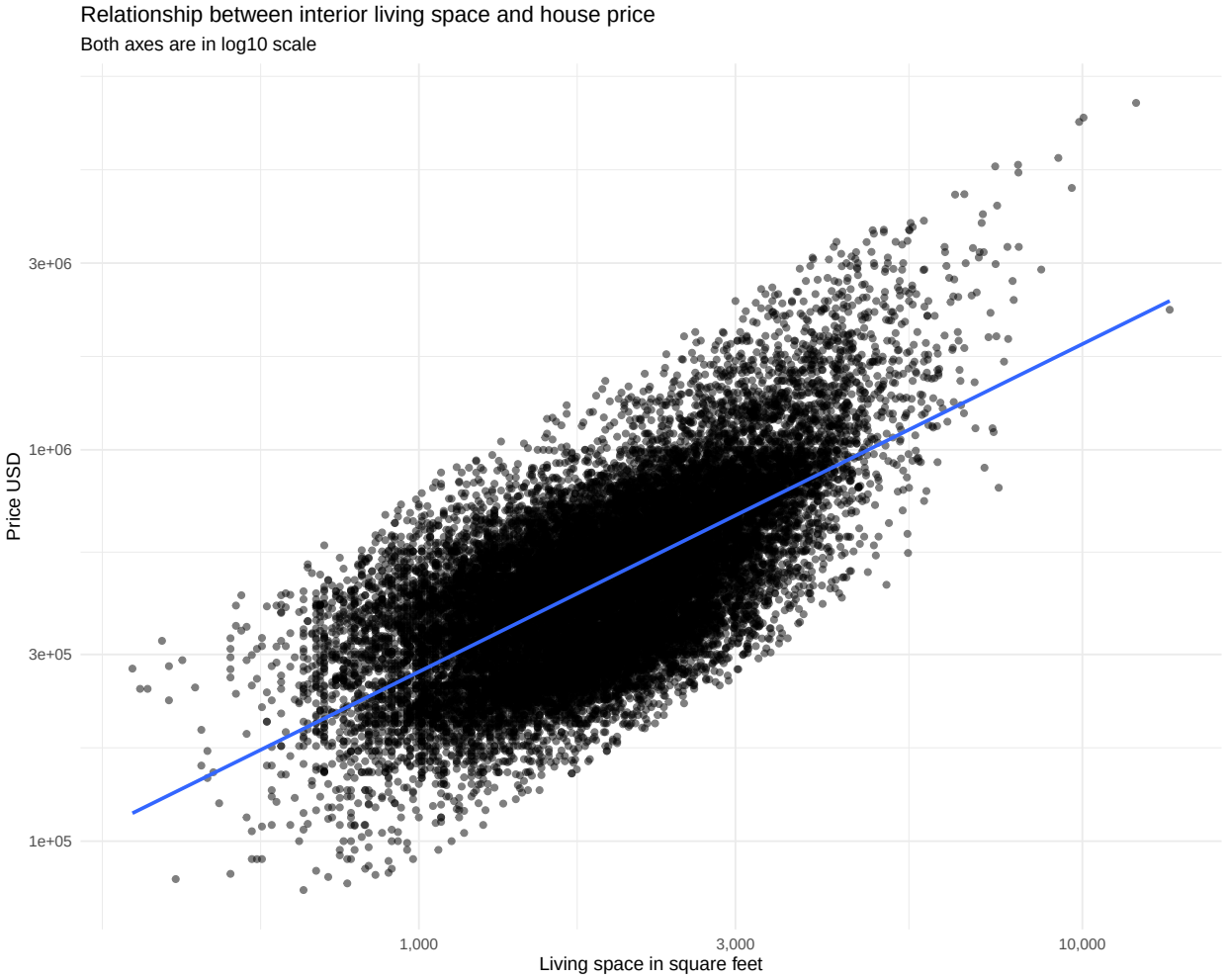
4.4 Are the sales evenly spread across the county ?



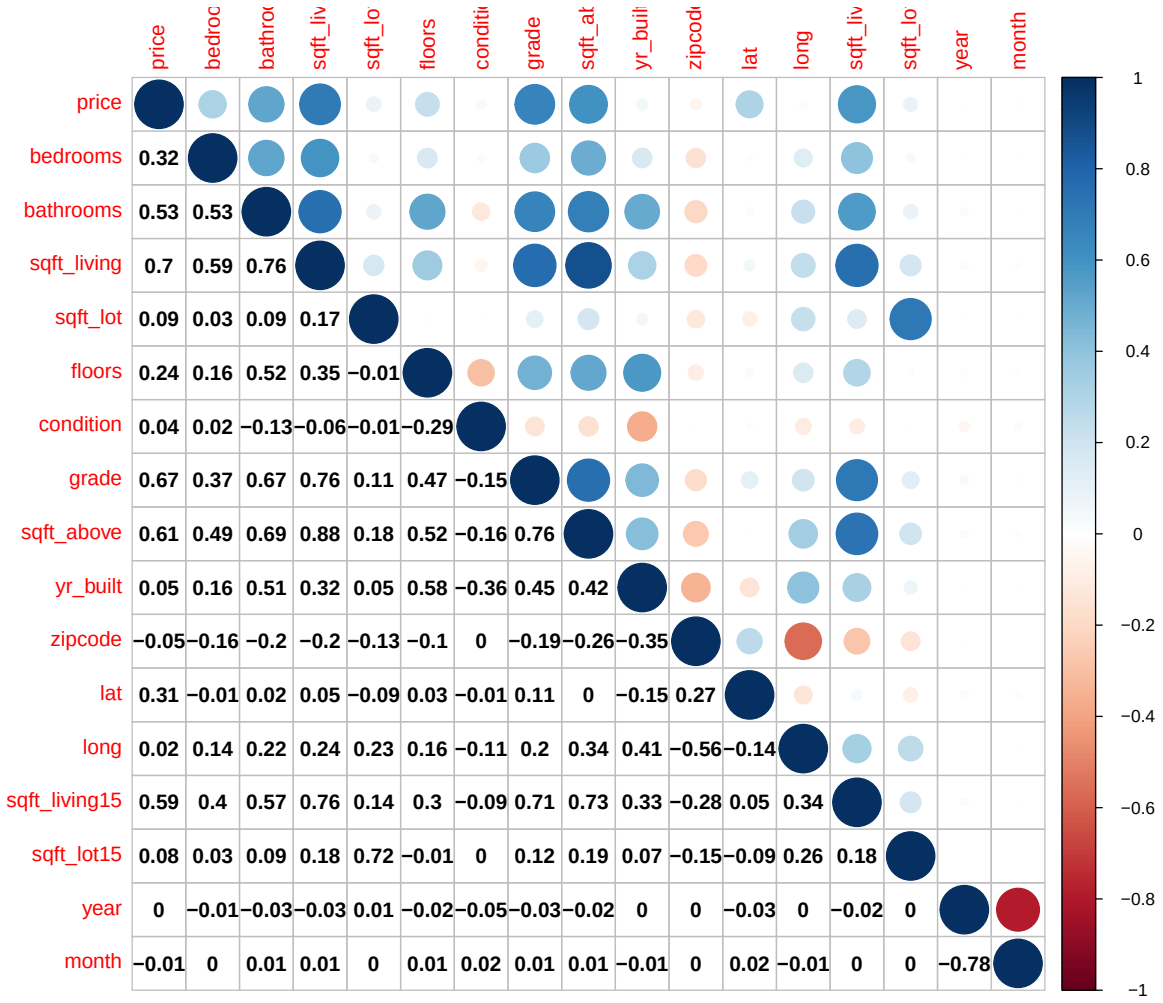
4.5 Does price increase as the lot area gets larger ?



4.6 Does price increase as the interior area gets larger ?



4.7 How are the features correlated ?



The matrix indicates some correlation between price and:

- **sqft_ features**, logically this makes sense as the more space a property has normally equates to a higher price
- **grade**, again this makes sense as the better condition aproperty is in demands a higher price.
- **bathrooms**: the number of bathrooms correlates with the price, this is probably due to the space needed to have multiple bathrooms in a property.

5 Methods/Analysis

The first step is to split the data into training and testing datasets. The split will be an 80/20 split with 80% of the data in the training set and the remaining 20% in the test set. The split of the data will be stratified by district, to ensure that the number of data points in the training data is equivalent to the proportions in the original data set.

6 Results

7 Conclusion