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 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
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

variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
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 featurres and the original date is dropped

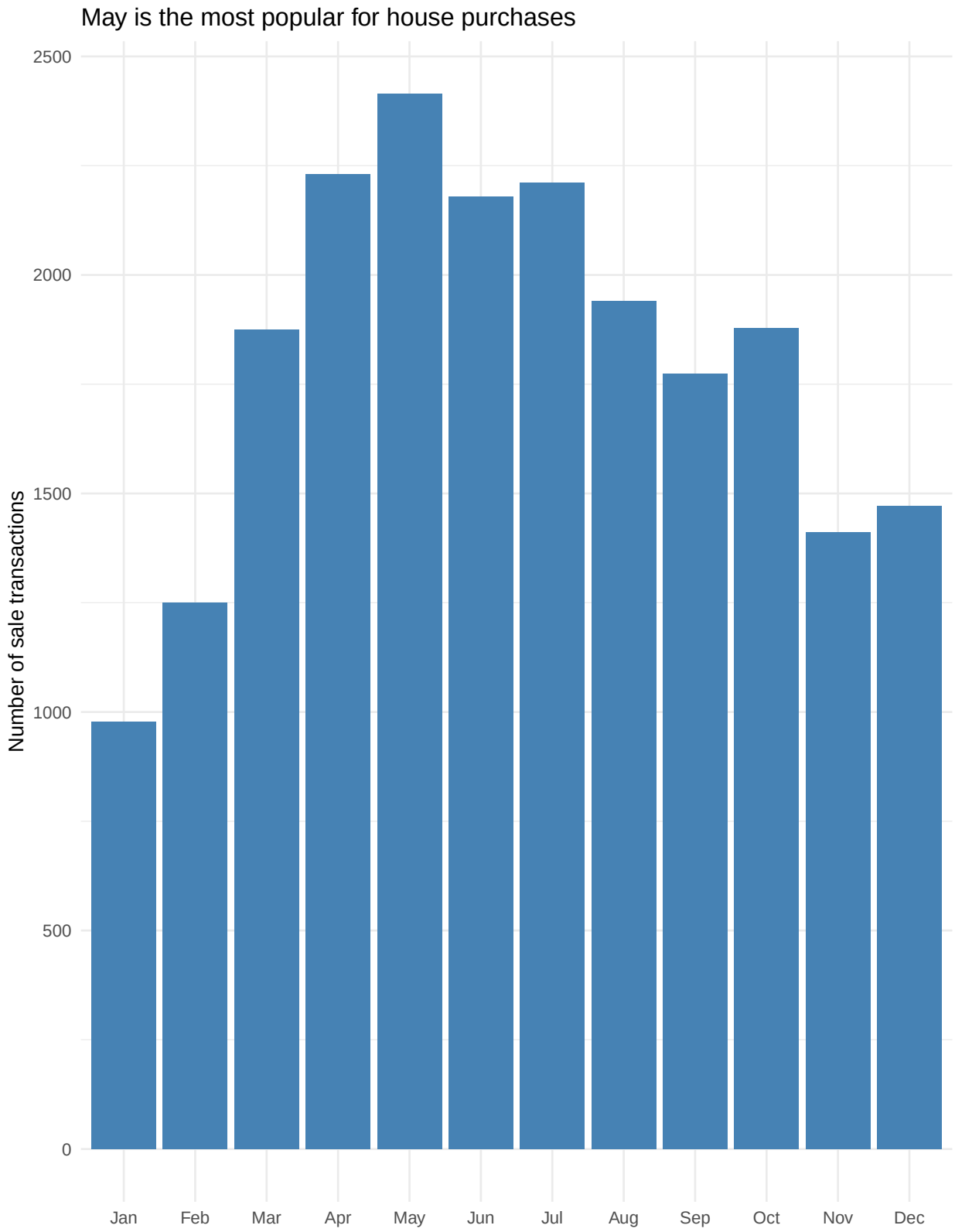
```
df <- df %>%
  mutate(year = year(date),
         month = month(date),
         age = year - yr_built) %>%
  select(-date)
glimpse(df)
```

```
## Rows: 21,613
## Columns: 19
## $ id           <chr> "7129300520", "6414100192", "5631500400", "2487200875...
## $ price        <dbl> 221900, 538000, 180000, 604000, 510000, 1225000, 2575...
## $ bedrooms     <dbl> 3, 3, 2, 4, 3, 4, 3, 3, 3, 3, 3, 2, 3, 3, 5, 4, 3, 4,...
## $ bathrooms    <dbl> 1.00, 2.25, 1.00, 3.00, 2.00, 4.50, 2.25, 1.50, 1.00,...
## $ sqft_living  <dbl> 1180, 2570, 770, 1960, 1680, 5420, 1715, 1060, 1780, ...
## $ sqft_lot     <dbl> 5650, 7242, 10000, 5000, 8080, 101930, 6819, 9711, 74...
## $ floors       <dbl> 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 2.0, 1.0...
## $ condition    <dbl> 3, 3, 3, 5, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 3, 3, 3, 4,...
## $ grade        <dbl> 7, 7, 6, 7, 8, 11, 7, 7, 7, 7, 8, 7, 7, 7, 7, 9, 7, 7...
## $ sqft_above   <dbl> 1180, 2170, 770, 1050, 1680, 3890, 1715, 1060, 1050, ...
## $ yr_built     <dbl> 1955, 1951, 1933, 1965, 1987, 2001, 1995, 1963, 1960,...
## $ zipcode      <dbl> 98178, 98125, 98028, 98136, 98074, 98053, 98003, 9819...
## $ lat          <dbl> 47.5112, 47.7210, 47.7379, 47.5208, 47.6168, 47.6561,...
## $ long         <dbl> -122.257, -122.319, -122.233, -122.393, -122.045, -12...
```

```
## $ sqft_living15 <dbl> 1340, 1690, 2720, 1360, 1800, 4760, 2238, 1650, 1780,...
## $ sqft_lot15    <dbl> 5650, 7639, 8062, 5000, 7503, 101930, 6819, 9711, 811...
## $ year          <dbl> 2014, 2014, 2015, 2014, 2015, 2014, 2014, 2015, 2015,...
## $ month         <dbl> 10, 12, 2, 12, 2, 5, 6, 1, 4, 3, 4, 5, 5, 10, 3, 1, 7...
## $ age           <dbl> 59, 63, 82, 49, 28, 13, 19, 52, 55, 12, 50, 72, 87, 3...
```

3.1 Is there a month when most sales occur?

```
## `summarise()` ungrouping output (override with `.groups` argument)
```


May is the most popular for house purchases
Number of sale transactions
2500
2000
1500
1000
500
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

3.2 Which are the 10 most expensive districts based on the number of properties sold over 3 million GBP?

3.3 Which are the 10 least expensive districts based on the number of properties under 100K GBP?

3.4 How are properties under 100k GBP clustered ?

3.5 Are the sales evenly spread across Greater London ?

3.6 Which type of property sold the most ?

3.7 Which year had the most sales

3.8 How often do properties change ownership ?

3.9 How are house prices distributed ?

4 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.

5 Results

6 Conclusion