MA678 Final Project

Tong Sun

2022-12-11

Abstract

This report focuses on the relationship between housing factors and prices, examining which factors have a large impact on prices and how they affect. In the report, I use multilevel model with different level of floors and view to examine the correlation. The square footage of the apartments interior living space has the greatest impact on the price of the house.

Introduction

The data set I choose contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. From https://www.kaggle.com/datasets/harlfoxem/housesalesprediction

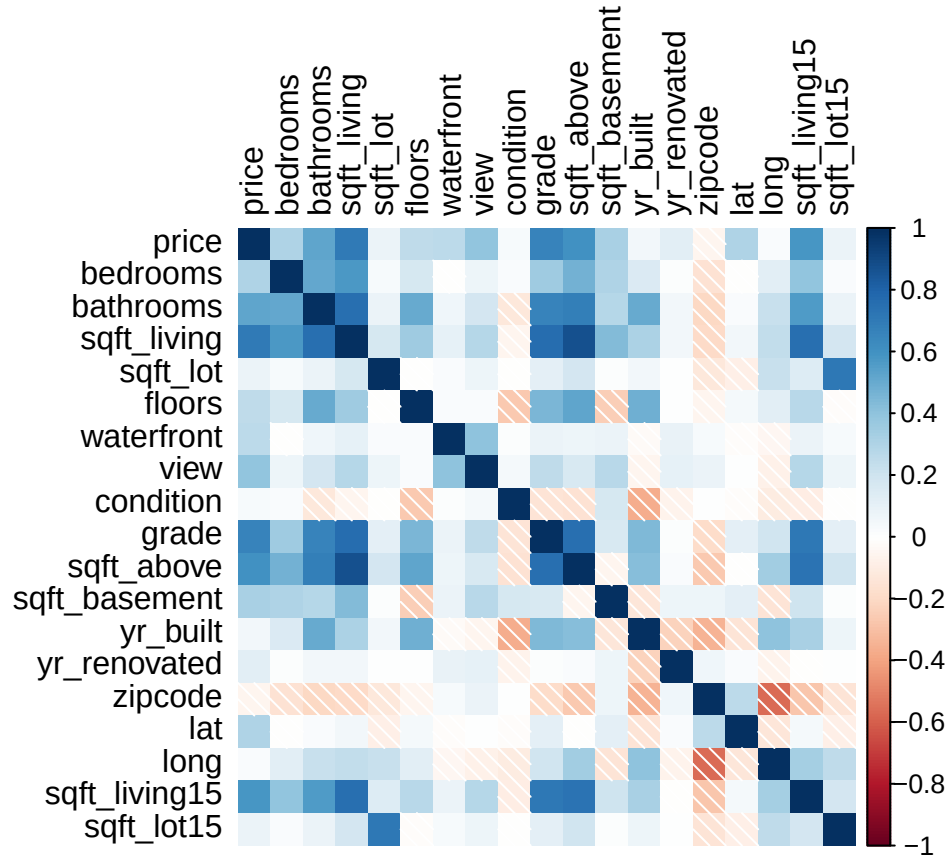
The explanation of column names is listed below:

column names	explanation
id	Unique ID for each home sold
date	Date of the home sale
price	Price of each home sold
bedrooms	Number of bedrooms
bathrooms	Number of bathrooms, where .5 accounts for a room with a toilet but no shower
sqft_living	Square footage of the apartments interior living space
sqft_lot	Square footage of the land space
floors	Number of floors
waterfront	A dummy variable for whether the apartment was overlooking the waterfront or not
view	An index from 0 to 4 of how good the view of the property was
condition	An index from 1 to 5 on the condition of the apartment
grade	An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design
sqft_above	The square footage of the interior housing space that is above ground level
sqft_basement	The square footage of the interior housing space that is below ground level
yr_built	The year the house was initially built
yr_renovated	The year of the house's last renovation
zipcode	What zipcode area the house is in

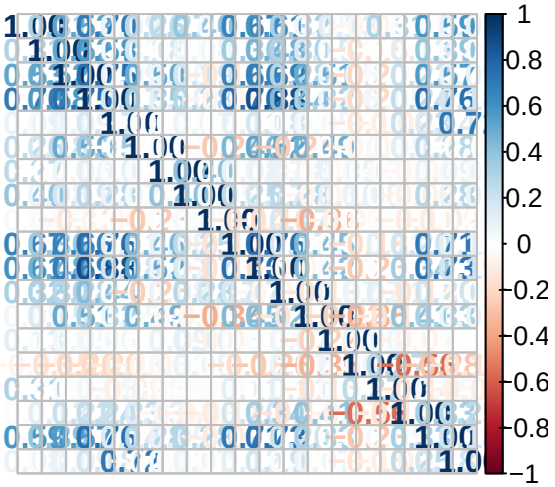
column names	explanation
lat	Latitude
long	Longitude
sqft_living15	The square footage of interior housing living space for the nearest 15 neighbors
sqft_lot15	The square footage of the land lots of the nearest 15 neighbors

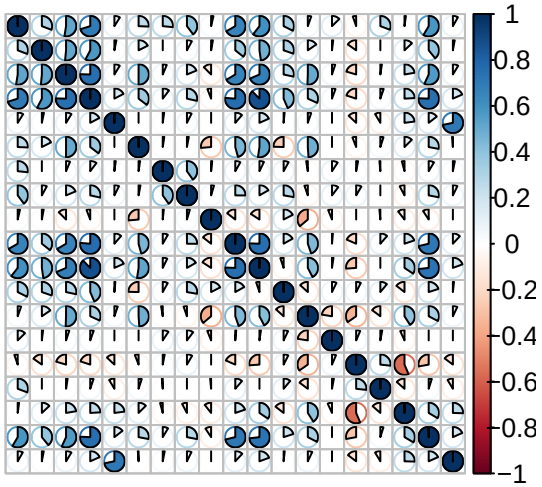
Method

Correlation



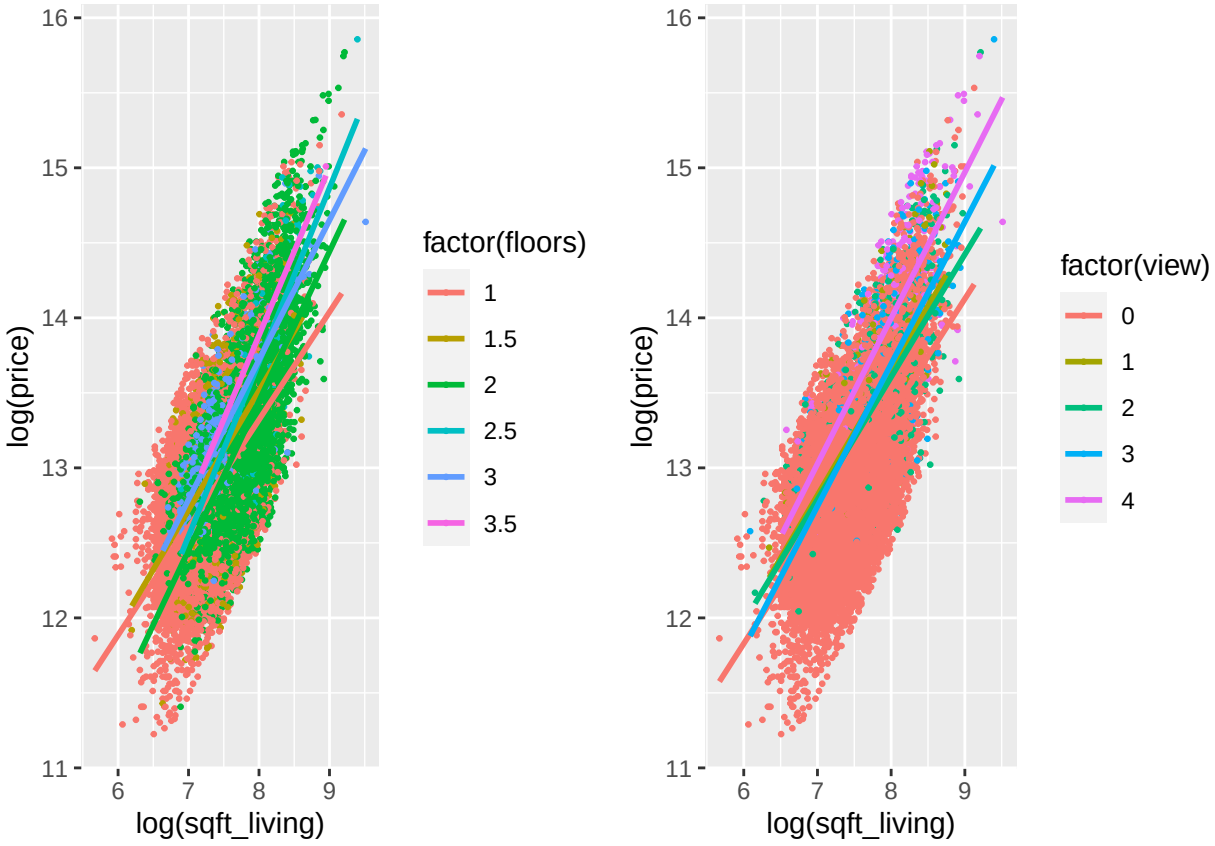
method = 'number' **method = 'pie'**

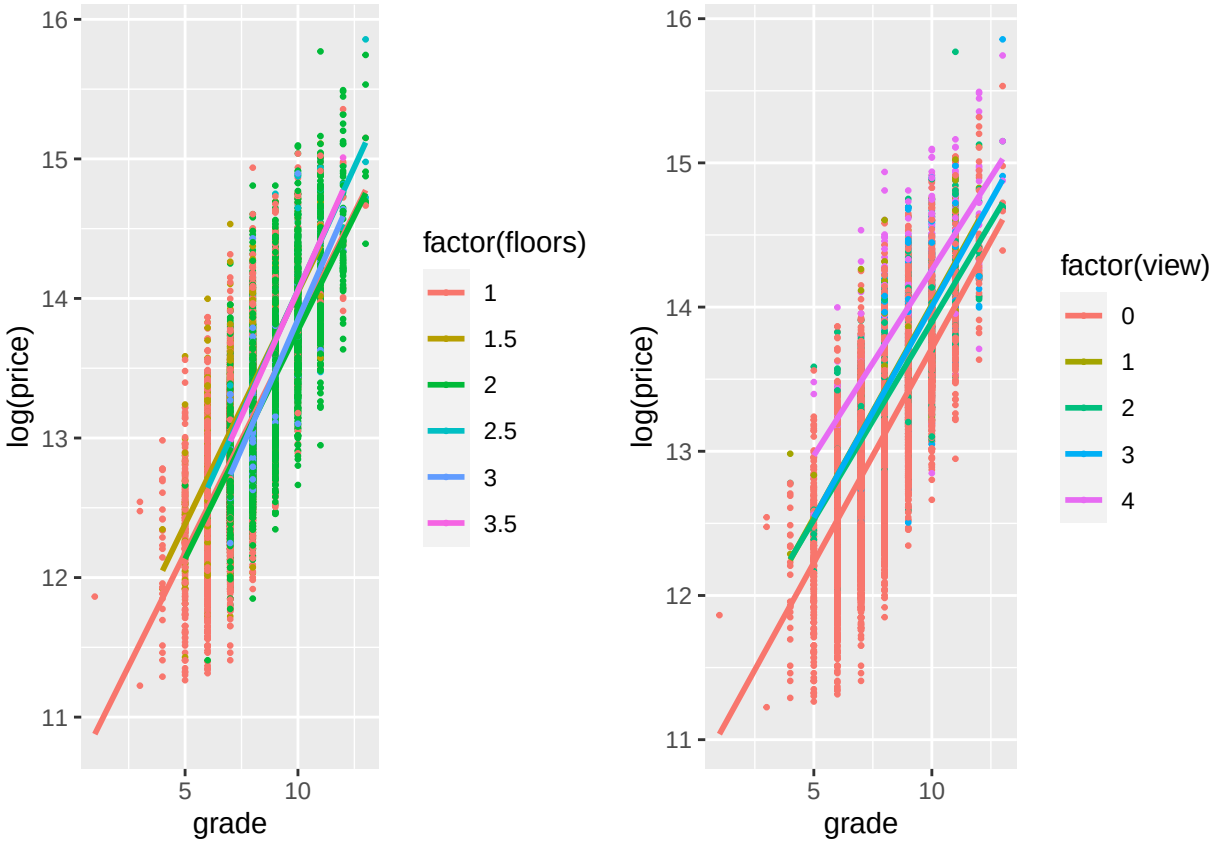




First I plot the correlation plot between the 19 variable (There are total 21 variable and I remove the id and date). From the plot, some factor have high correlation with price such as sqft_living, grade and sqft_above. Some have little to do with price like condition, yr_built and latitude.

Then I choose to plot the relation between living space and grade with price at different floors and view level to see their change. From the plots, we can see that as the living space and grade of the house increases, the price of increases as well, which is the same correlation as in the previous analysis.





Model Fitting

Then we can choose some factors fit our model to analyze.

```
model <- lmer(log(price) ~ bedrooms + bathrooms + log(sqft_living) + waterfront + grade + log(sqft_above
```

Result

Random effects

Groups	Name	Variance	Std.Dev.
floors	(Intercept)	0.01460	0.1208
view	(Intercept)	0.01317	0.1148
Residual		0.11070	0.3327

Fixed effects

	Estimate	Std. Error	df	t value	Pr(>
(Intercept)	8.274e+00	1.074e-01	3.997e+01	77.037	< 2e-16 ***

	Estimate	Std. Error	df	t value	Pr(>
bedrooms	-2.167e-02	3.238e-03	2.160e+04	-6.691	2.27e-11 ***
bathrooms	1.681e-02	5.098e-03	2.157e+04	3.297	0.000979 ***
log(sqft_living)	4.926e-01	1.455e-02	2.160e+04	33.845	< 2e-16 ***
waterfront	3.649e-01	3.234e-02	1.367e+04	11.283	< 2e-16 ***
grade	1.947e-01	3.334e-03	2.160e+04	58.397	< 2e-16 ***
log(sqft_above)	-2.313e-01	1.313e-02	2.158e+04	-17.617	< 2e-16 ***
log(sqft_living15)	2.086e-01	1.131e-02	2.160e+04	18.451	< 2e-16 ***

Correlation of Fixed Effects

	(Intr)	bedrms	bthrms	lg(sqft_l)	wtrfrn	grade	lg(sqft_b)
bedrooms	0.203						
bathrooms	0.217	-0.165					
lg(sqft_lv)	-0.214	-0.325	-0.384				
waterfront	-0.028	0.042	-0.006	-0.004			
grade	0.279	0.167	-0.157	-0.177	0.022		
lg(sqft_bv)	-0.148	-0.025	0.111	-0.577	-0.018	-0.165	
lg(sqft_15)	-0.360	0.041	0.029	-0.220	0.015	-0.250	-0.171

VIF

VIF equal to 1 = variables are not correlated VIF between 1 and 5 = variables are moderately correlated VIF greater than 5 = variables are highly correlated

bedrooms	bathrooms	log(sqft_living)	waterfront	grade	log(sqft_above)	log(sqft_living15)
1.643611	1.994794	5.469905	1.002886	2.059336	3.590107	2.057776

Formula with fixex Effects

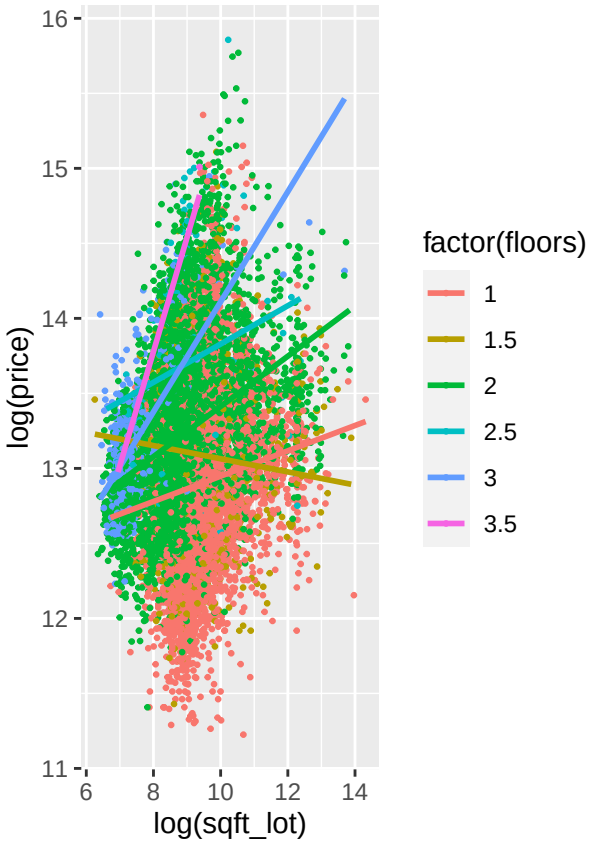
$log(price) = 8.274 - 0.02167 * bedrooms + 0.01681 * bathrooms + 0.4926 * log(sqft_living) + 0.3649 * waterfront + 0.1947 * grade - 0.2$

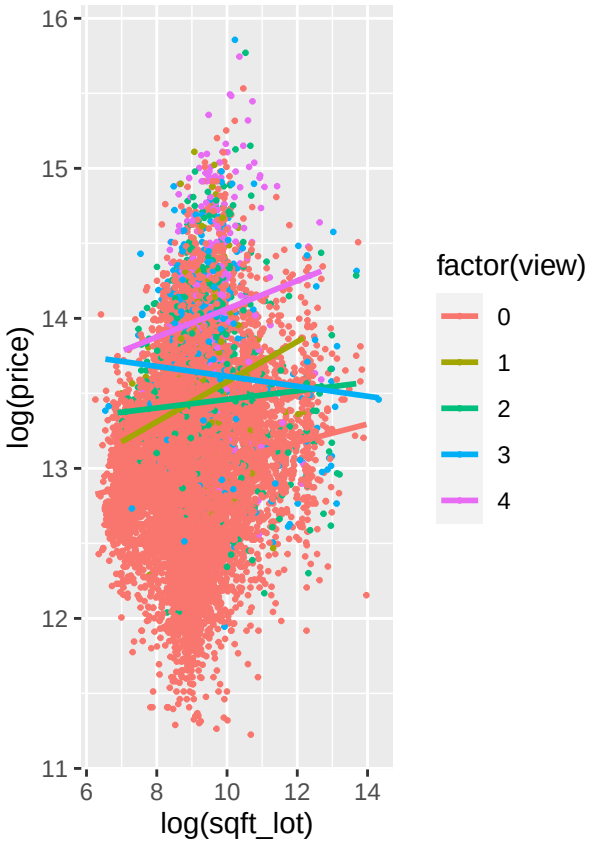
From the formula, we can see bedrooms and the interior housing space that is above ground level have negative effect on price and others have positive effect. Among the variables, living space are highly correlated with price.

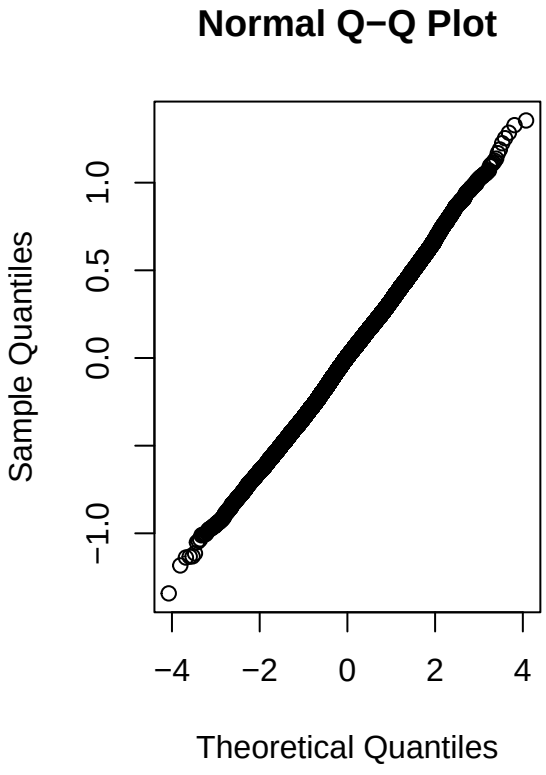
Discussion

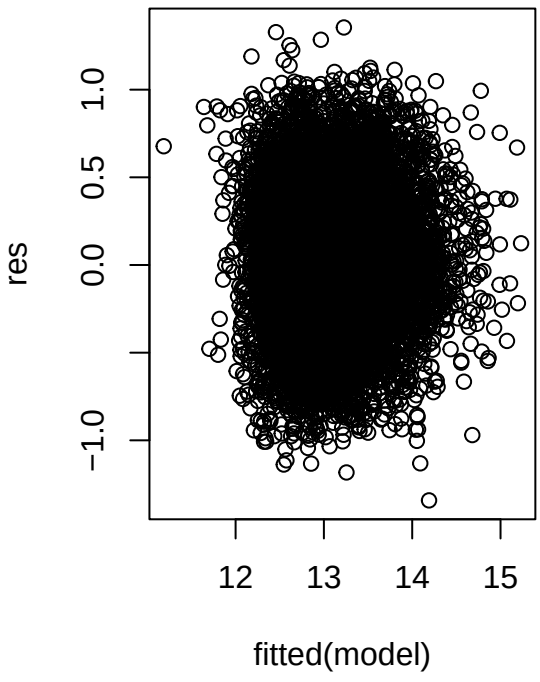
From the analyze, the greatest impact on the price of house is the living space of the house, followed by the number of rooms and the level of construction and design also have a relative impact. But like the condition of the apartment and the year the house was initially built are nothing to influence.

Appendix


log(price)
log(sqft_lot)
factor(floors)
1
1.5
2
2.5
3
3.5


log(price)
log(sqft_lot)
factor(view)
0
1
2
3
4


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles


res
fitted(model)