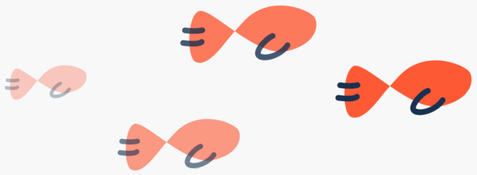
EDA Project_Kings County

Hassan B.



Contents

1. Purpose

- Charles C. is our stake holder
- Interested in Buying and selling houses
- Average investment budget
- Willing to invest in big returns only!
- Needs advise on:
    - -Should he renovate ?
    - - Waiting period?

2. Importing Libraries

Libraries for:

- For plots
- For map plot
- For regression model

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from scipy import stats
import seaborn as sns
```

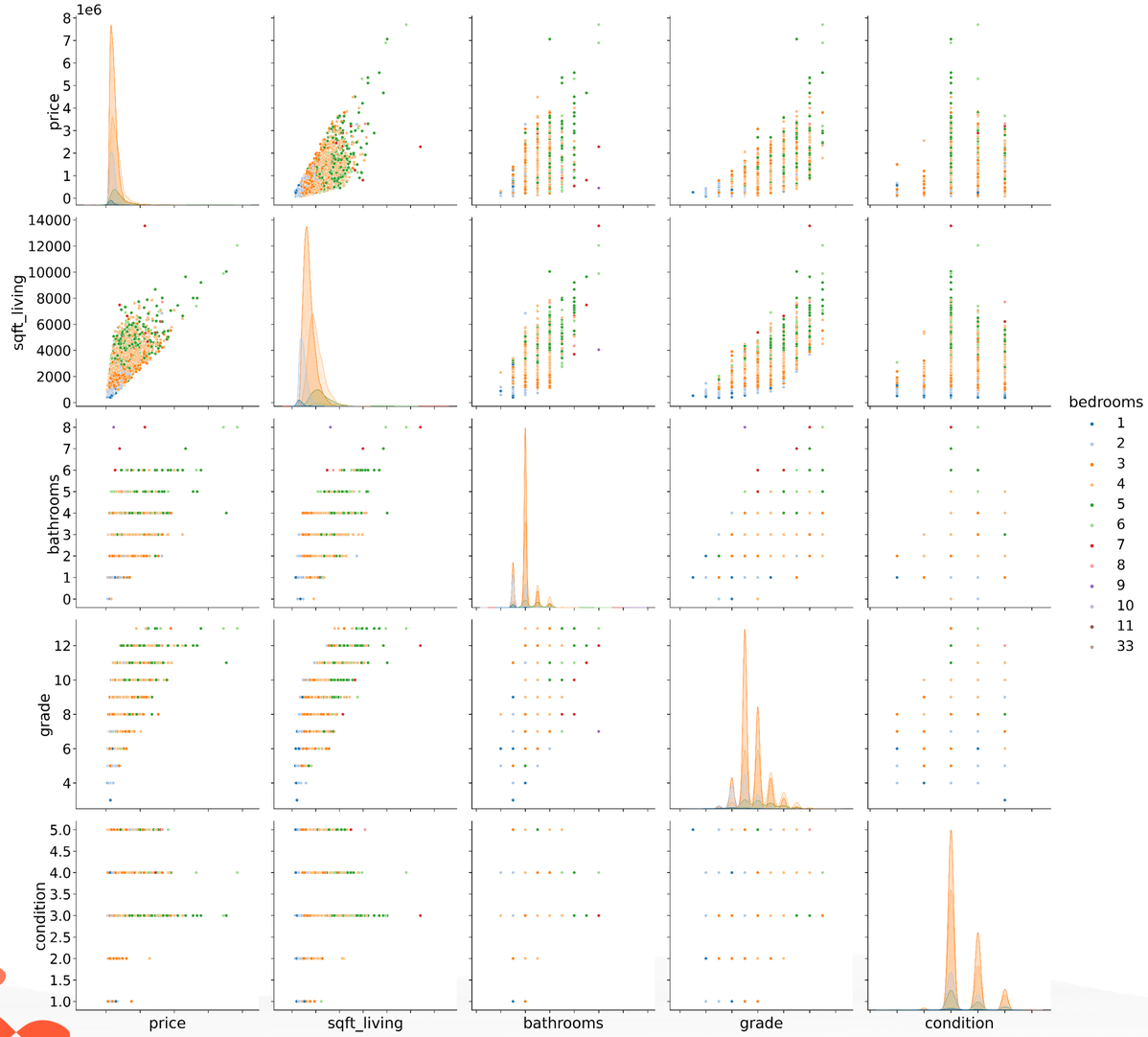
```
import folium
from folium import plugins
from folium.plugins import HeatMap
```

```
import statsmodels.api as sms
import statsmodels.formula.api as smf
```

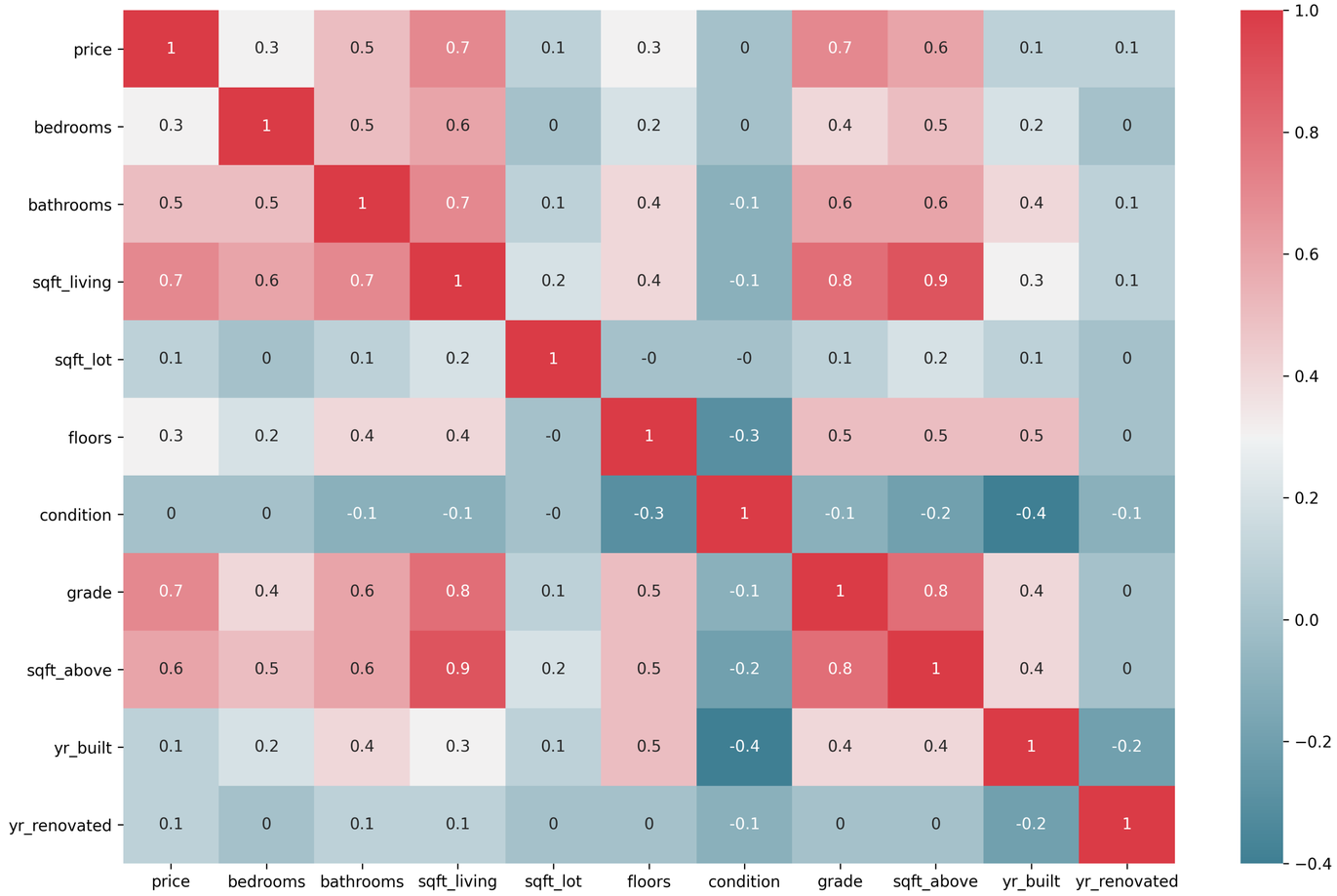
Overview of the data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21597 entries, 0 to 21596
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   id             21597 non-null  int64  
 1   date           21597 non-null  datetime64[ns]
 2   price          21597 non-null  float64
 3   bedrooms       21597 non-null  int64  
 4   bathrooms      21597 non-null  float64
 5   sqft_living    21597 non-null  int64  
 6   sqft_lot       21597 non-null  int64  
 7   floors         21597 non-null  float64
 8   waterfront     19221 non-null  float64
 9   view           21534 non-null  float64
 10  condition      21597 non-null  int64  
 11  grade          21597 non-null  int64  
 12  sqft_above     21597 non-null  int64  
 13  sqft_basement  21143 non-null  float64
 14  yr_built       21597 non-null  int64  
 15  yr_renovated   17755 non-null  float64
 16  zipcode        21597 non-null  int64  
 17  lat            21597 non-null  float64
 18  long           21597 non-null  float64
 19  sqft_living15  21597 non-null  int64  
 20  sqft_lot15     21597 non-null  int64  
dtypes: datetime64[ns](1), float64(9), int64(11)
memory usage: 3.5 MB
None
```

4. Analyzing the Data with plots



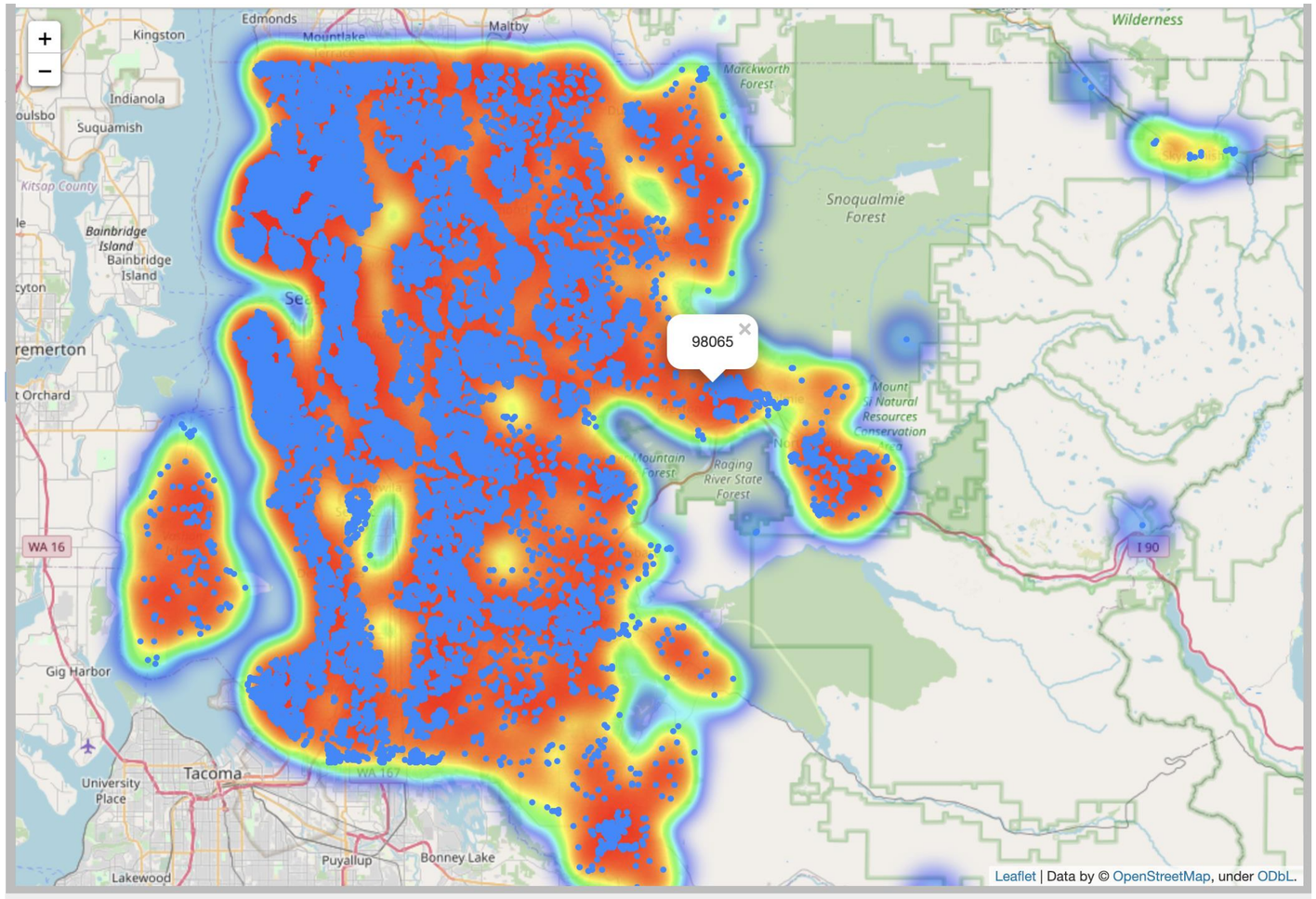
4. Analyzing the Data with plots



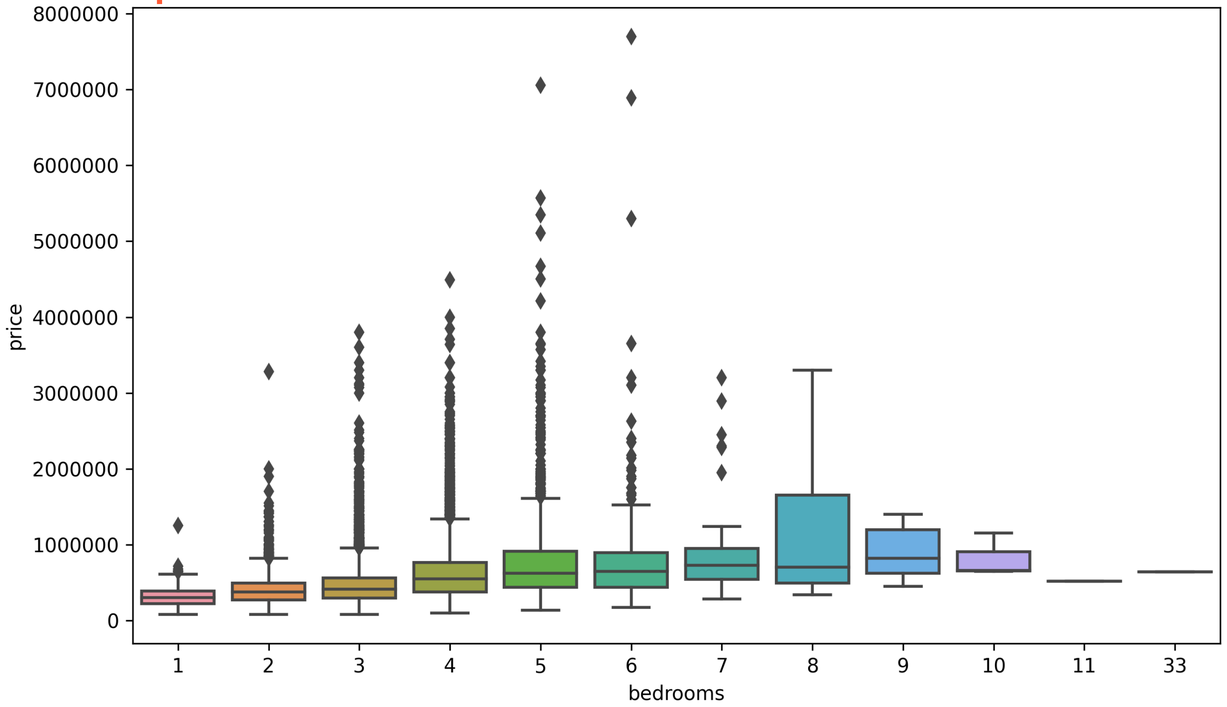
<u>Somethings we have learned from the data are:</u>

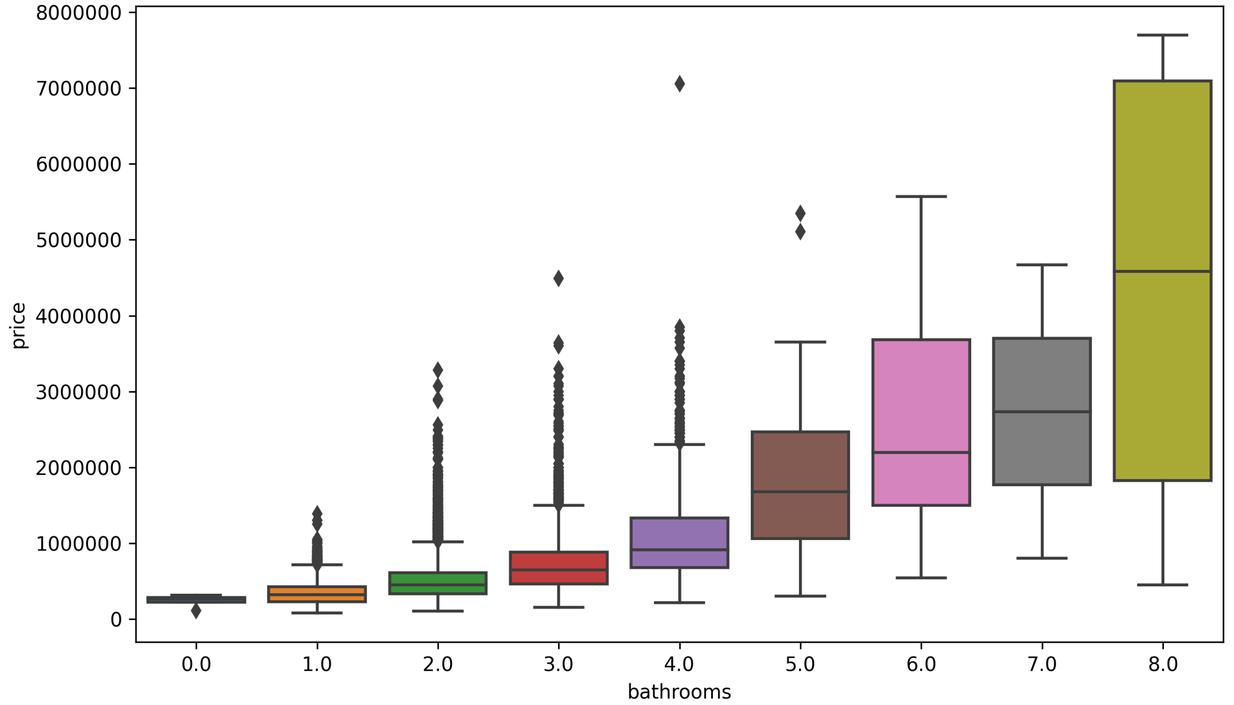
- One house has a room of 33 rooms which my stakeholder will not be interested.
- Condition is ranging from 1 - 5 scale
- View ranges from 0 - 4 scale
- Grade from scale 3 - 13
- Houses are built from 1900 - 2015 with mean around 1970.

4. Analyzing the Data with plots



4. Analyzing the Data with plots

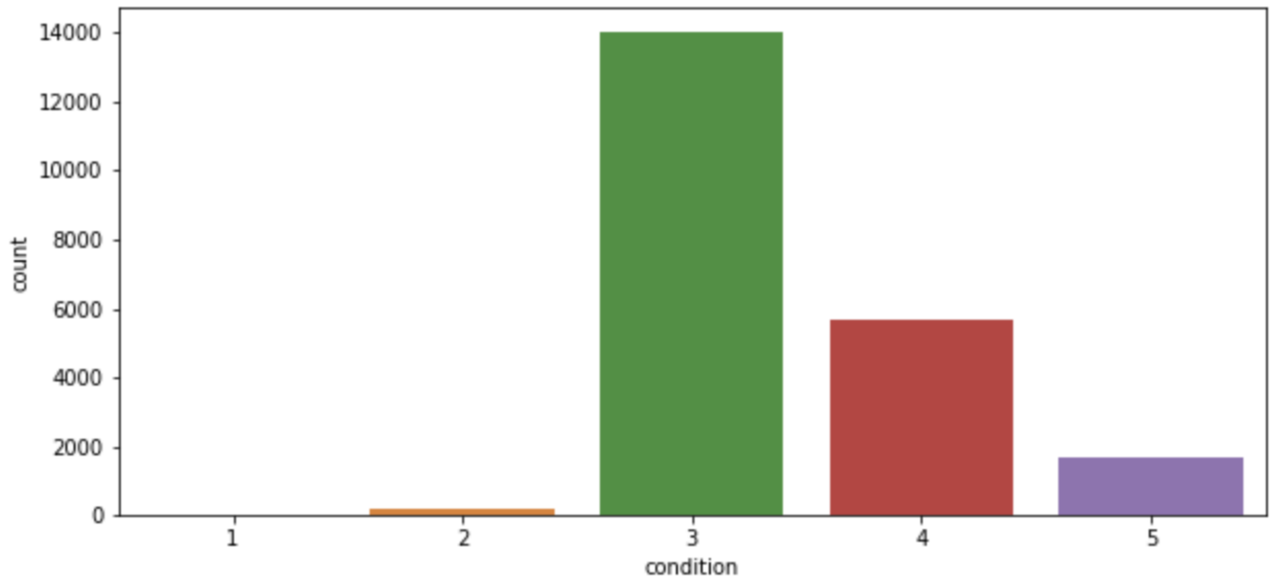


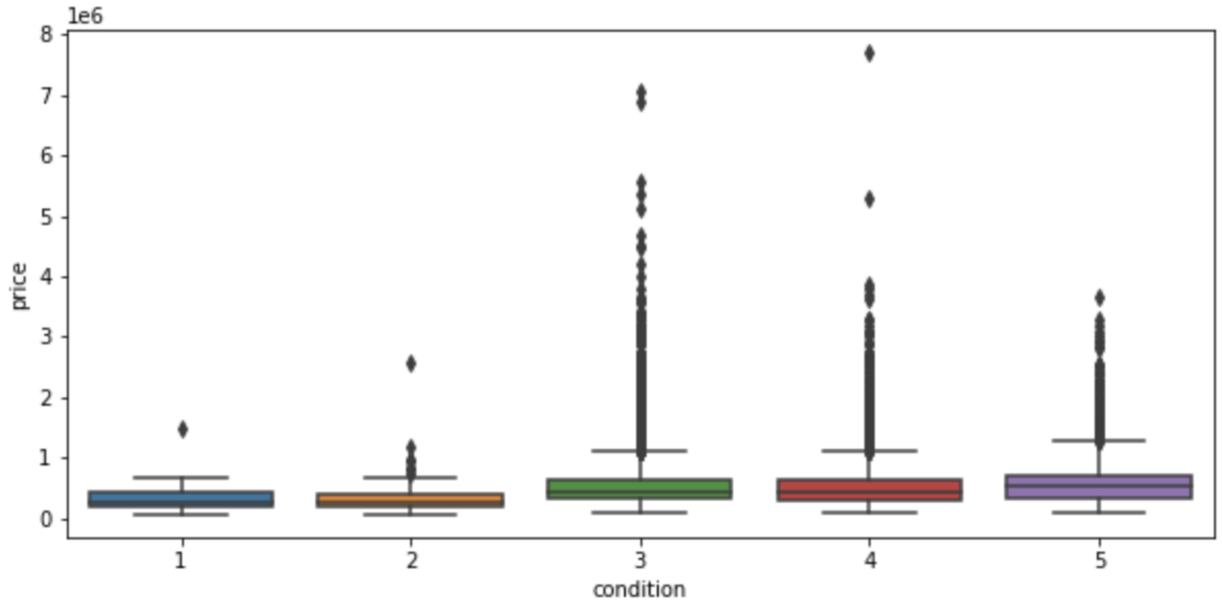


bedrooms	1	2	3	4	5	6	7	8	9	10	11	33
bathrooms												
0.0	1	2	0	1	0	0	0	0	0	0	0	0
1.0	167	1587	1800	327	43	6	1	0	0	0	0	0
2.0	28	1129	7186	4709	695	90	7	1	0	1	0	1
3.0	0	41	656	1219	506	88	7	6	2	1	1	0
4.0	0	1	182	601	321	72	17	4	3	0	0	0
5.0	0	0	0	19	22	12	2	1	0	1	0	0
6.0	0	0	0	6	13	2	2	1	0	0	0	0
7.0	0	0	0	0	1	0	1	0	0	0	0	0
8.0	0	0	0	0	0	2	1	0	1	0	0	0

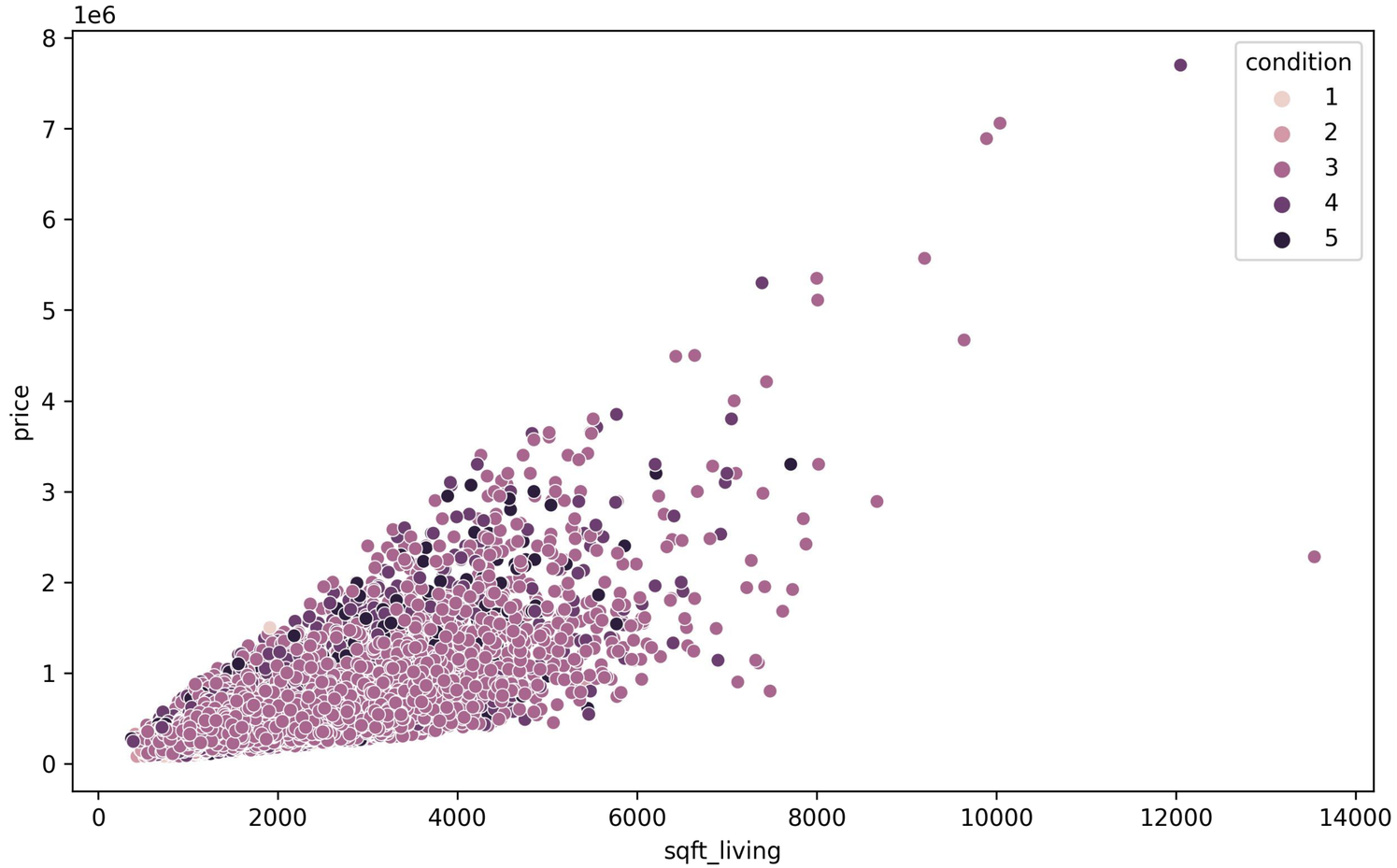
4. Analyzing the Data with plots

condition	1	2	3	4	5
bedrooms					
1	4	10	123	47	12
2	12	51	1779	718	200
3	8	69	6308	2711	728
4	4	36	4580	1682	580
5	0	1	1031	418	151
6	1	3	158	87	23
7	0	0	25	9	4
8	0	0	8	3	2
9	0	0	6	0	0
10	0	0	1	2	0
11	0	0	1	0	0
33	0	0	0	0	1





4. Analyzing the Data with plots



5. Regression modelling

- 1 explanatory variable and response variable

```
# Create an OLS model
X = df.sqft_living
y = df.price
X = sms.add_constant(X)

model = sms.OLS(y, X).fit()
# Our model needs an intercept so we add a column of 1s:

# return the output of the model
model.summary()
```

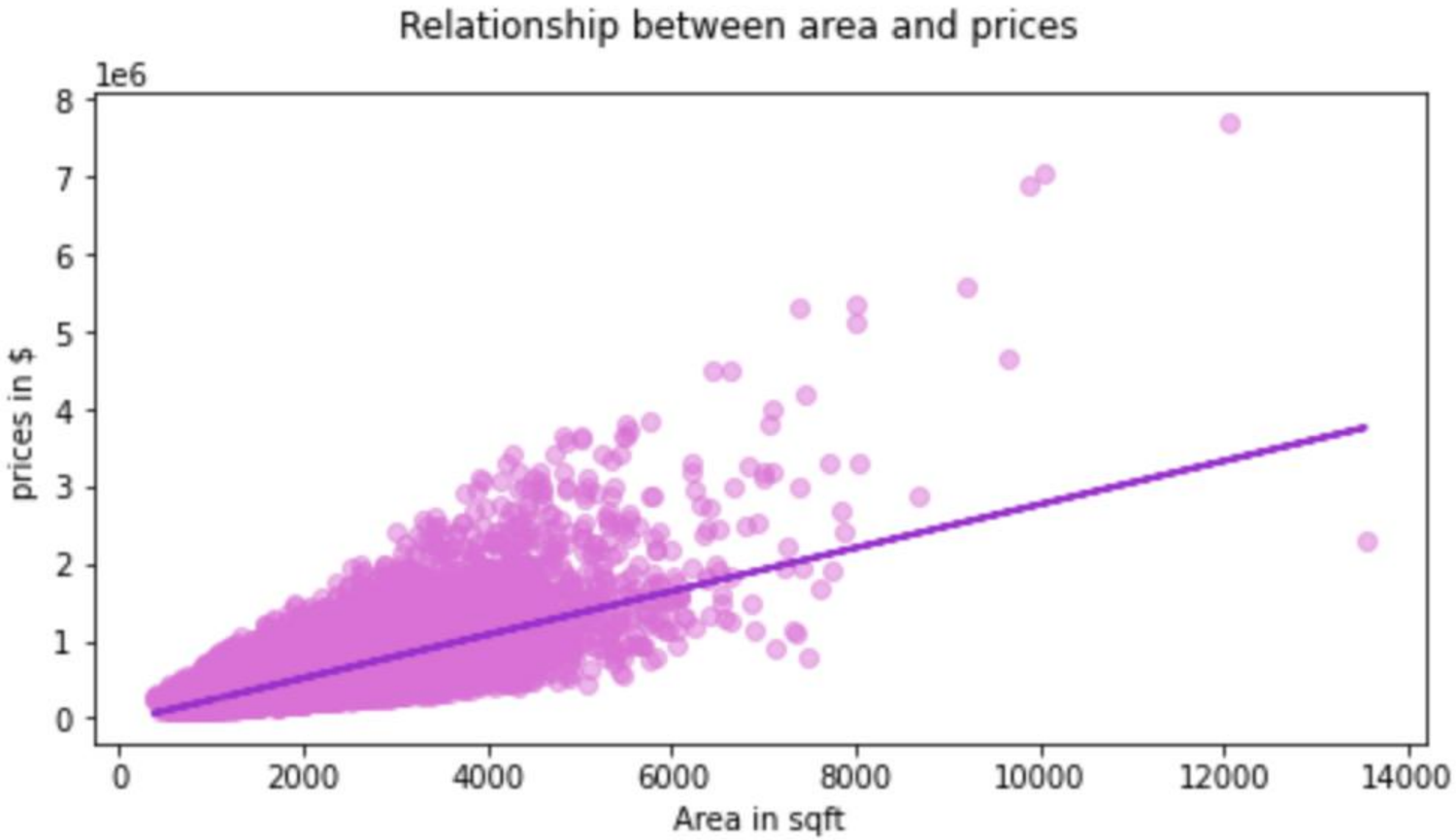
OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.493
Model:	OLS	**Adj. R-squared:**	0.493
Method:	Least Squares	**F-statistic:**	2.097e+04
Date:	Thu, 18 Feb 2021	**Prob (F-statistic):**	0.00
Time:	17:03:28	**Log-Likelihood:**	-3.0006e+05
No. Observations:	21597	**AIC:**	6.001e+05
Df Residuals:	21595	**BIC:**	6.001e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.399e+04	4410.023	-9.975	0.000	-5.26e+04	-3.53e+04
sqft_living	280.8630	1.939	144.819	0.000	277.062	284.664

Omnibus:	14801.942	**Durbin-Watson:**	1.982
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	542662.604
Skew:	2.820	**Prob(JB):**	0.00
Kurtosis:	26.901	**Cond. No.**	5.63e+03

5. Regression modelling



5. Regression modelling

```
R squared for each possible explanatory variable:
bedrooms                = 0.0953
bathrooms               = 0.27
sqft_living             = 0.493
sqft_lot                = 0.00808
floors                  = 0.0659
condition               = 0.0013
grade                   = 0.446
sqft_above              = 0.366
yr_built                = 0.00291
yr_renovated            = 0.0139
rooms                   = 0.218
```

5. Regression modelling

- Multiple explanatory variable and response variable

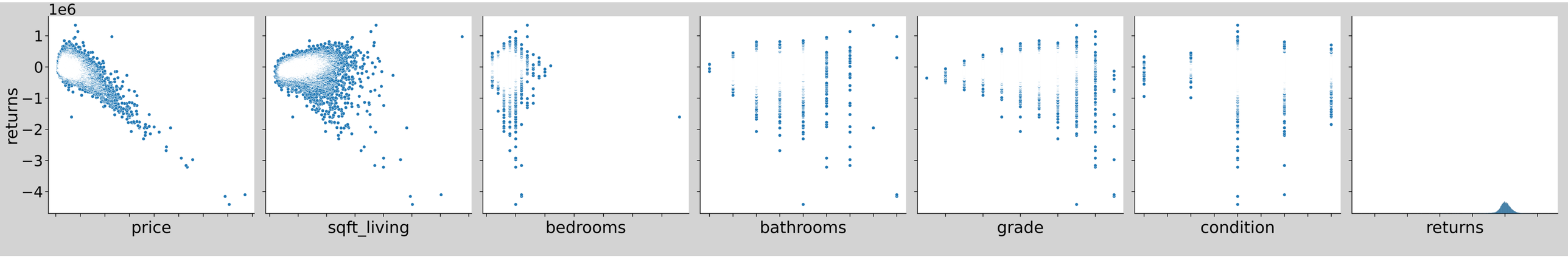
OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.621
Model:	OLS	**Adj. R-squared:**	0.620
Method:	Least Squares	**F-statistic:**	3531.
Date:	Thu, 18 Feb 2021	**Prob (F-statistic):**	0.00
Time:	17:03:28	**Log-Likelihood:**	-2.9693e+05
No. Observations:	21597	**AIC:**	5.939e+05
Df Residuals:	21586	**BIC:**	5.940e+05
Df Model:	10		
Covariance Type:	nonrobust		

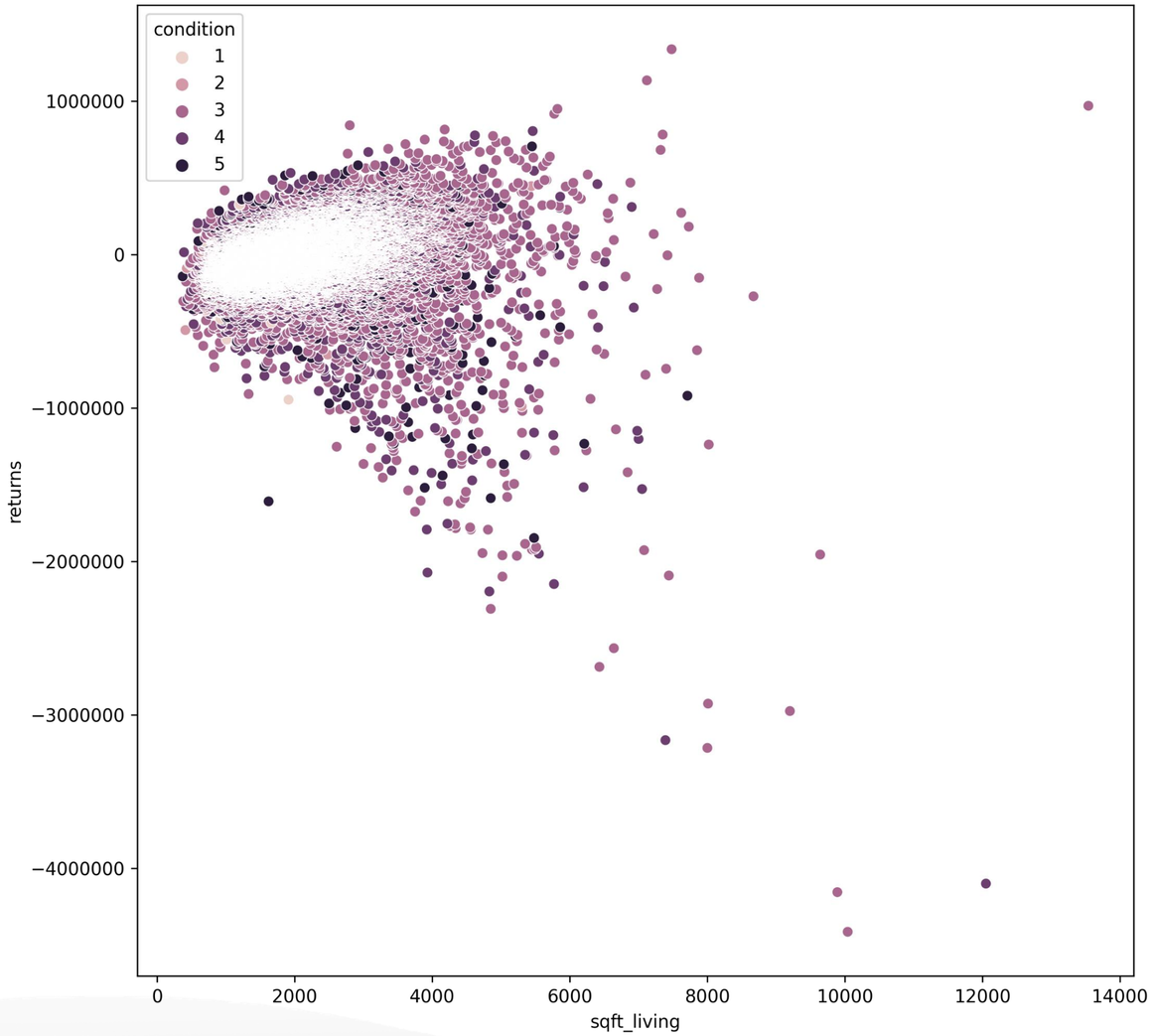
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.492e+06	1.36e+05	47.731	0.000	6.23e+06	6.76e+06
bedrooms	-4.886e+04	2104.006	-23.221	0.000	-5.3e+04	-4.47e+04
bathrooms	5.035e+04	3070.238	16.399	0.000	4.43e+04	5.64e+04
sqft_living	200.3544	4.607	43.489	0.000	191.324	209.385
floors	3.346e+04	3790.859	8.827	0.000	2.6e+04	4.09e+04
condition	2.067e+04	2607.407	7.929	0.000	1.56e+04	2.58e+04
grade	1.32e+05	2251.955	58.629	0.000	1.28e+05	1.36e+05
yr_built	-3755.7136	69.785	-53.818	0.000	-3892.498	-3618.929
yr_renovated	23.7525	4.441	5.348	0.000	15.047	32.457
sqft_lot	-0.2338	0.038	-6.106	0.000	-0.309	-0.159
sqft_above	-19.3217	4.594	-4.206	0.000	-28.325	-10.318

Omnibus:	17301.994	**Durbin-Watson:**	1.984
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1218066.285
Skew:	3.350	**Prob(JB):**	0.00
Kurtosis:	39.176	**Cond. No.**	3.89e+06

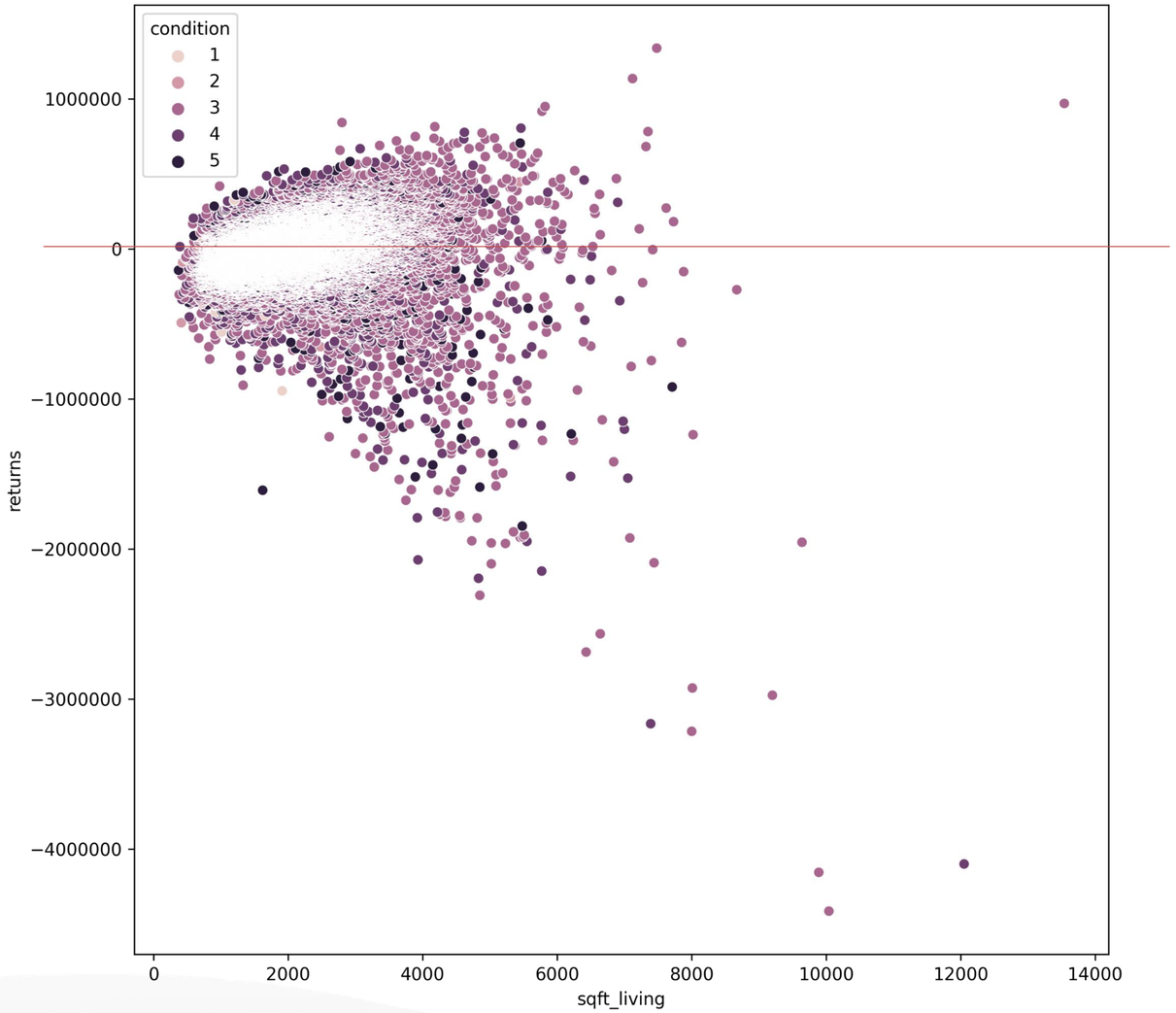
5. Regression modelling – Plotting the returns



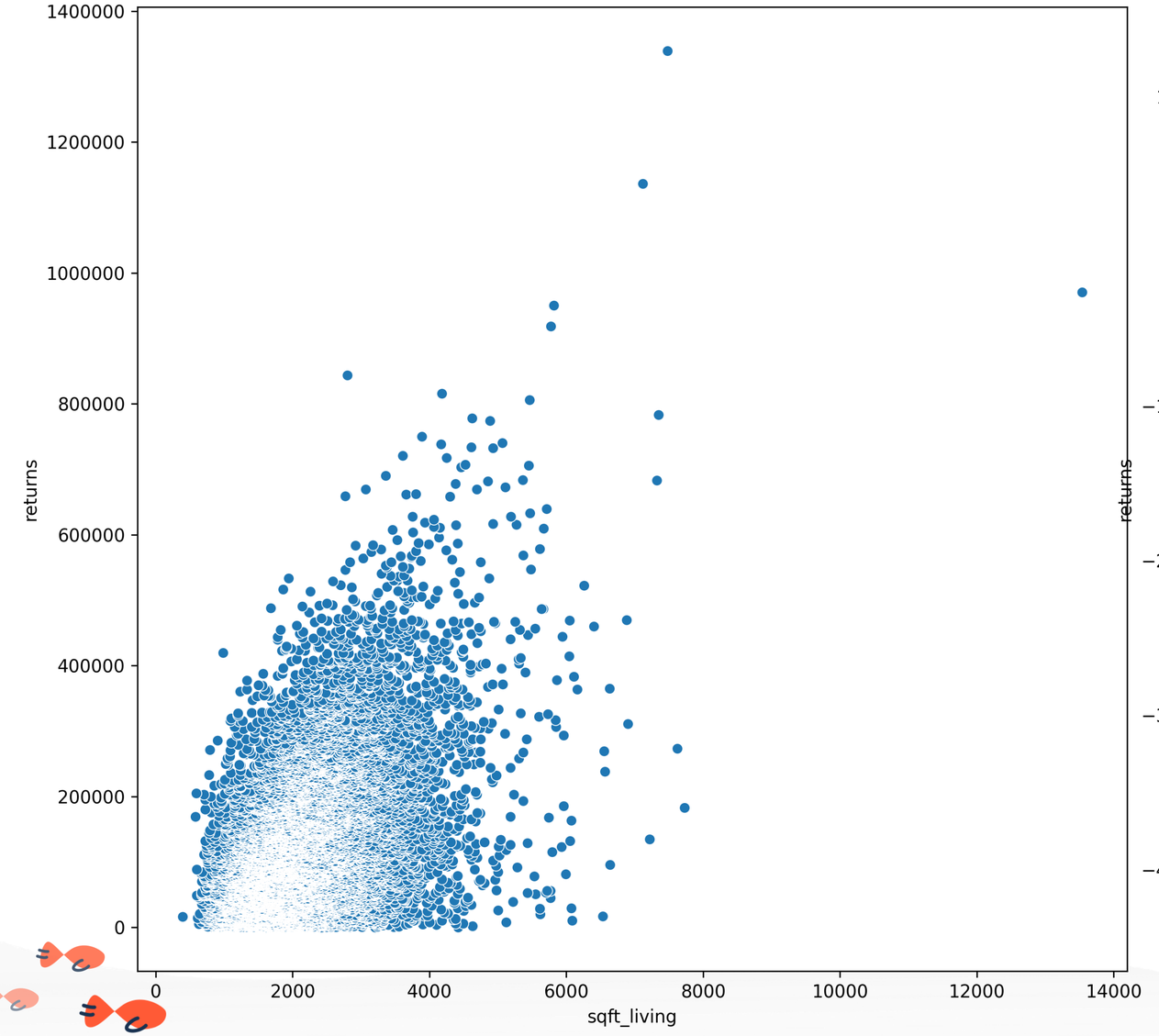
5. Regression modelling – Plotting the returns

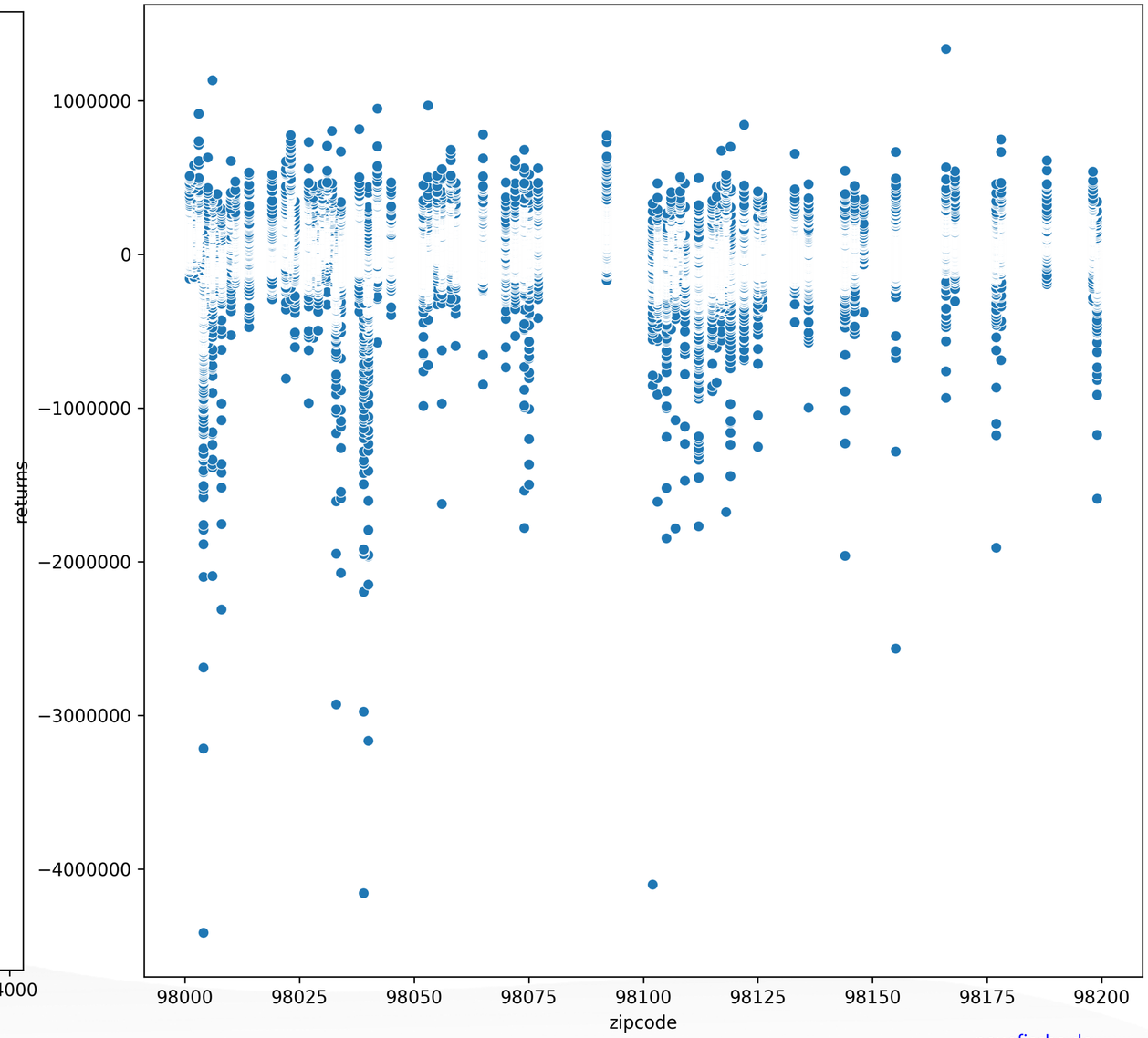


5. Regression modelling – Plotting the returns



5. Regression modelling – Plotting the returns





6. Open Questions

Some questions to answer:

Impact of having basement on the price. This can be done by categorizing the basement for same living area and compare it with price trend.

Thank you. Any questions?