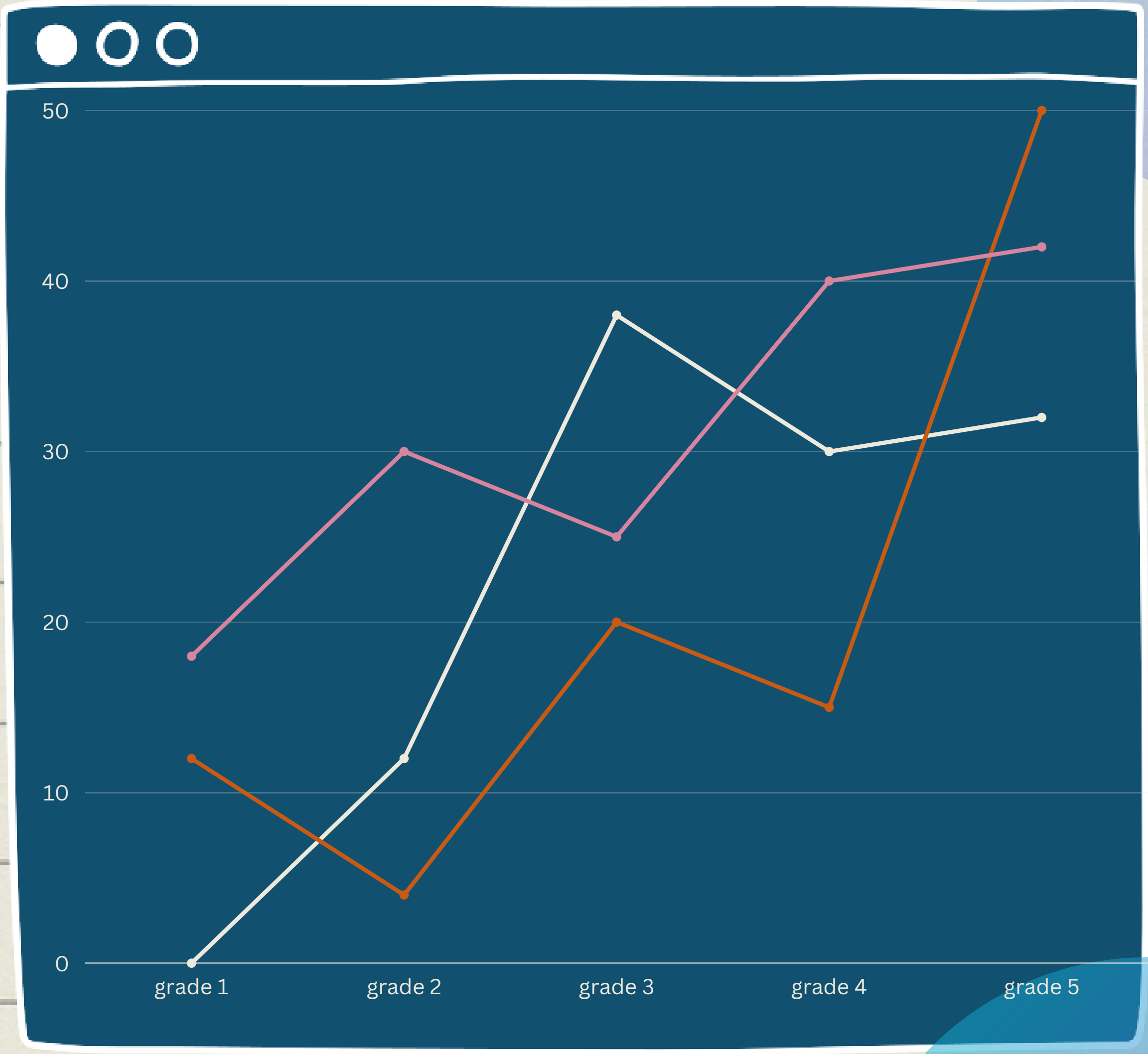
PHASE 2 PROJECT

Student :Joy Chepchumba
Student pace : Part-time
project review time : 26/03/2023
Instructor Name: Samuel Jane.



Introduction

This is the Phase two project presentation on the analysis of the Kings County House Sales Dataset. We are analyzing this data for a real estate company that would like to know which features they should invest in when selling houses so that they can sell houses at the best price.

Business Understanding

Our stake holders are a Real Estate agent or A real estate company. When selling a house, a real estate agent or company, would like to sell a house at the most reasonable price in the market, that would be both good to the customers and to them earning them a profit.

In order to sell the house at such a price, the house has to have some features that are pleasing to the customers which makes them want to buy the house. We are then going to conduct thorough analysis on a house sale dataset and find out the features that are related to a good house sale price, then advice the agent or company on the features they should focus on when they want to sell houses , we also would like to advice them on a good price to sell their houses at so based on our findings.

Data Understanding

For this analysis, we shall be using the Kings county house sale dataset which is sourced from canvas. We shall use the Price variable as our target against the various feature's that have been provided, which are 20 feature's in total. Our data also contains 21,597 sales information

My approach to the business problem

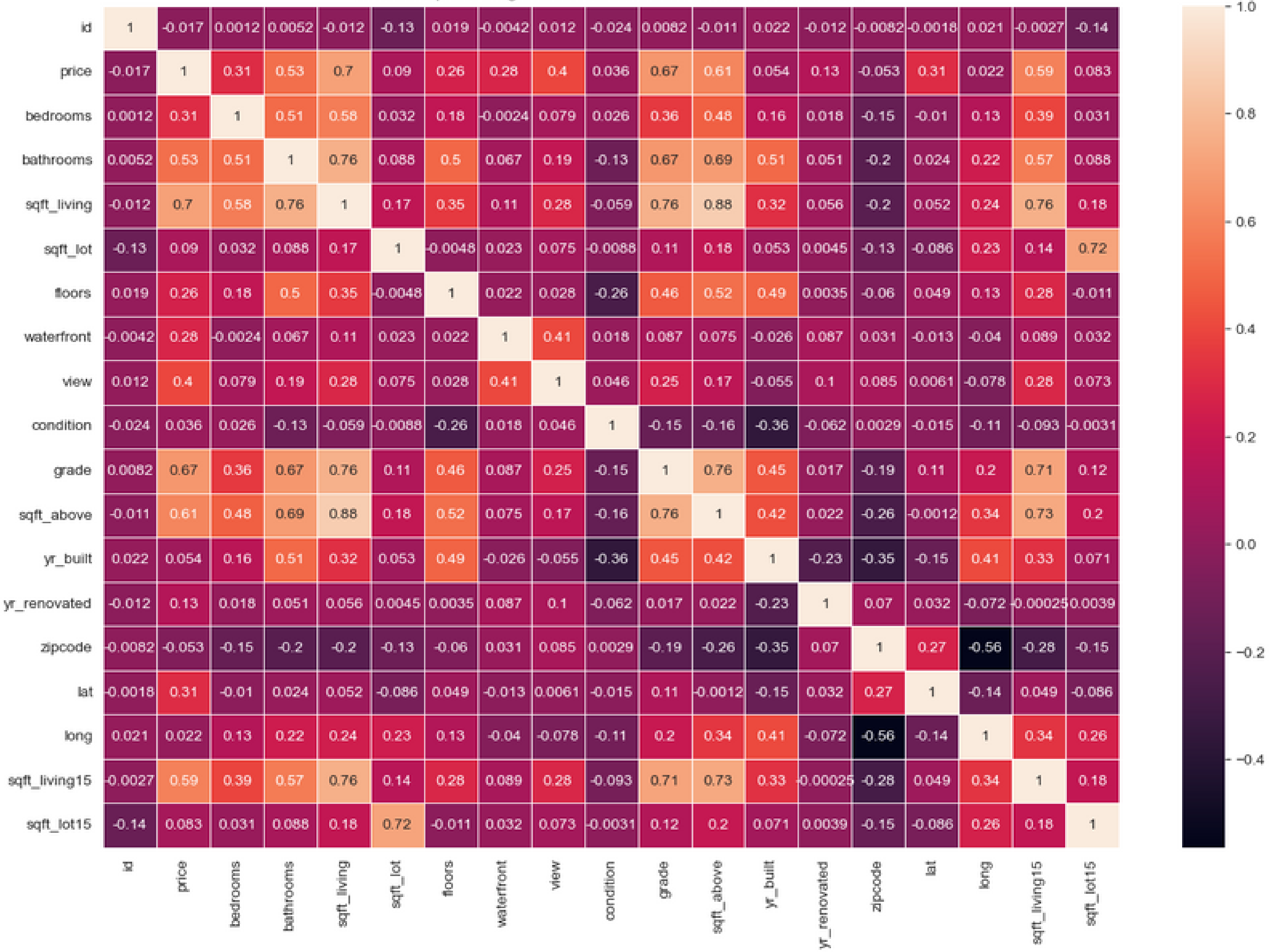
NEXT

1)Lets find the features that have a high correlation with our target variable

A correlation is a relationship or an association between two variables. When we find a variable that has a high correlation with our target variable it means there is a strong relationship between the two and most likely that is a feature that customers like hence why it has a high correlation with our variable. To do this, we can plot a heat map. We should then do further analysis to show the relationship between this variables. From my analysis I found out that the variables that have a high correlation with target variable (price) include:

- sqft_living : footage of the home.
- sqft_lot : footage of the lot.
- sqft_above : footage of the house apart from the basement.
- yr_built : The year the house was built.
- sqft_living15 : The square footage of interior housing living space for the nearest 15 neighbors
- grade :overall grade given to the housing unit, based on King County grading system,

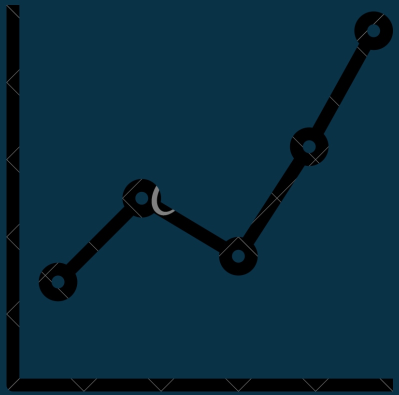
A heatmap showing the correlations between various variables

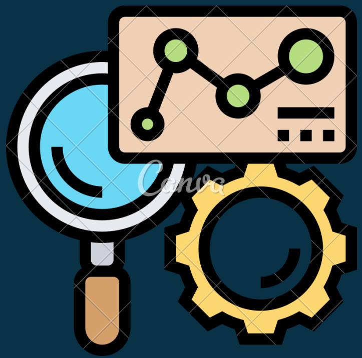


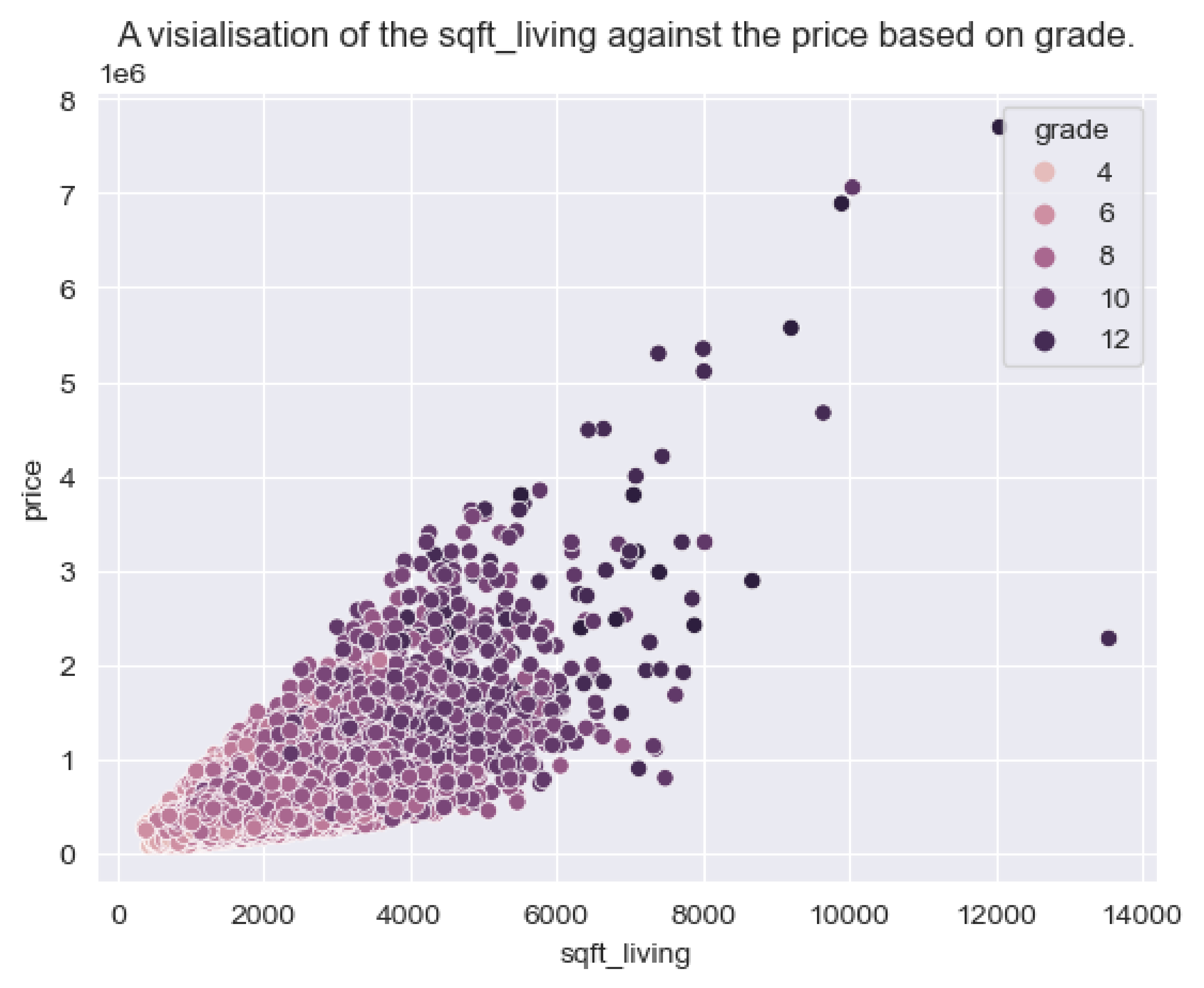
	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
id	1	-0.017	0.0012	0.0052	-0.012	-0.13	0.019	-0.0042	0.012	-0.024	0.0082	-0.011	0.022	-0.012	-0.0082	-0.0018	0.021	-0.0027	-0.14
price	-0.017	1	0.31	0.53	0.7	0.09	0.26	0.28	0.4	0.036	0.67	0.61	0.054	0.13	-0.053	0.31	0.022	0.59	0.083
bedrooms	0.0012	0.31	1	0.51	0.58	0.032	0.18	-0.0024	0.079	0.026	0.36	0.48	0.16	0.018	-0.15	-0.01	0.13	0.39	0.031
bathrooms	0.0052	0.53	0.51	1	0.76	0.088	0.5	0.067	0.19	-0.13	0.67	0.69	0.51	0.051	-0.2	0.024	0.22	0.57	0.088
sqft_living	-0.012	0.7	0.58	0.76	1	0.17	0.35	0.11	0.28	-0.059	0.76	0.88	0.32	0.056	-0.2	0.052	0.24	0.76	0.18
sqft_lot	-0.13	0.09	0.032	0.088	0.17	1	-0.0048	0.023	0.075	-0.0088	0.11	0.18	0.053	0.0045	-0.13	-0.086	0.23	0.14	0.72
floors	0.019	0.26	0.18	0.5	0.35	-0.0048	1	0.022	0.028	-0.26	0.46	0.52	0.49	0.0035	-0.06	0.049	0.13	0.28	-0.011
waterfront	-0.0042	0.28	-0.0024	0.067	0.11	0.023	0.022	1	0.41	0.018	0.087	0.075	-0.026	0.087	0.031	-0.013	-0.04	0.089	0.032
view	0.012	0.4	0.079	0.19	0.28	0.075	0.028	0.41	1	0.046	0.25	0.17	-0.055	0.1	0.085	0.0061	-0.078	0.28	0.073
condition	-0.024	0.036	0.026	-0.13	-0.059	-0.0088	-0.26	0.018	0.046	1	-0.15	-0.16	-0.36	-0.062	0.0029	-0.015	-0.11	-0.093	-0.0031
grade	0.0082	0.67	0.36	0.67	0.76	0.11	0.46	0.087	0.25	-0.15	1	0.76	0.45	0.017	-0.19	0.11	0.2	0.71	0.12
sqft_above	-0.011	0.61	0.48	0.69	0.88	0.18	0.52	0.075	0.17	-0.16	0.76	1	0.42	0.022	-0.26	-0.0012	0.34	0.73	0.2
yr_built	0.022	0.054	0.16	0.51	0.32	0.053	0.49	-0.026	-0.055	-0.36	0.45	0.42	1	-0.23	-0.35	-0.15	0.41	0.33	0.071
yr_renovated	-0.012	0.13	0.018	0.051	0.056	0.0045	0.0035	0.087	0.1	-0.062	0.017	0.022	-0.23	1	0.07	0.032	-0.072	-0.00025	0.0039
zipcode	-0.0082	-0.053	-0.15	-0.2	-0.2	-0.13	-0.06	0.031	0.085	0.0029	-0.19	-0.26	-0.35	0.07	1	0.27	-0.56	-0.28	-0.15
lat	-0.0018	0.31	-0.01	0.024	0.052	-0.086	0.049	-0.013	0.0061	-0.015	0.11	-0.0012	-0.15	0.032	0.27	1	-0.14	0.049	-0.086
long	0.021	0.022	0.13	0.22	0.24	0.23	0.13	-0.04	-0.078	-0.11	0.2	0.34	0.41	-0.072	-0.56	-0.14	1	0.34	0.26
sqft_living15	-0.0027	0.59	0.39	0.57	0.76	0.14	0.28	0.089	0.28	-0.093	0.71	0.73	0.33	-0.00025	-0.28	0.049	0.34	1	0.18
sqft_lot15	-0.14	0.083	0.031	0.088	0.18	0.72	-0.011	0.032	0.073	-0.0031	0.12	0.2	0.071	0.0039	-0.15	-0.086	0.26	0.18	1

2) A relationship between price sqft_living and grade

Previously, we had seen that there is a correlation between price and grade , price and sqft_living. I decided to do further analysis on this I created a linear model between price and target variable ,sqft_living and grade . I found out that there was a relationship between the sqft_living but the relationship wasn't that strong. I then created an interaction between grade and sqft_living and saw that there was a good relationship.Upon this discovery, I found out that as the sqft_living increase, the grade increases and so does the price, this can be proven in a visualization between the variables, as shown in the next slide.



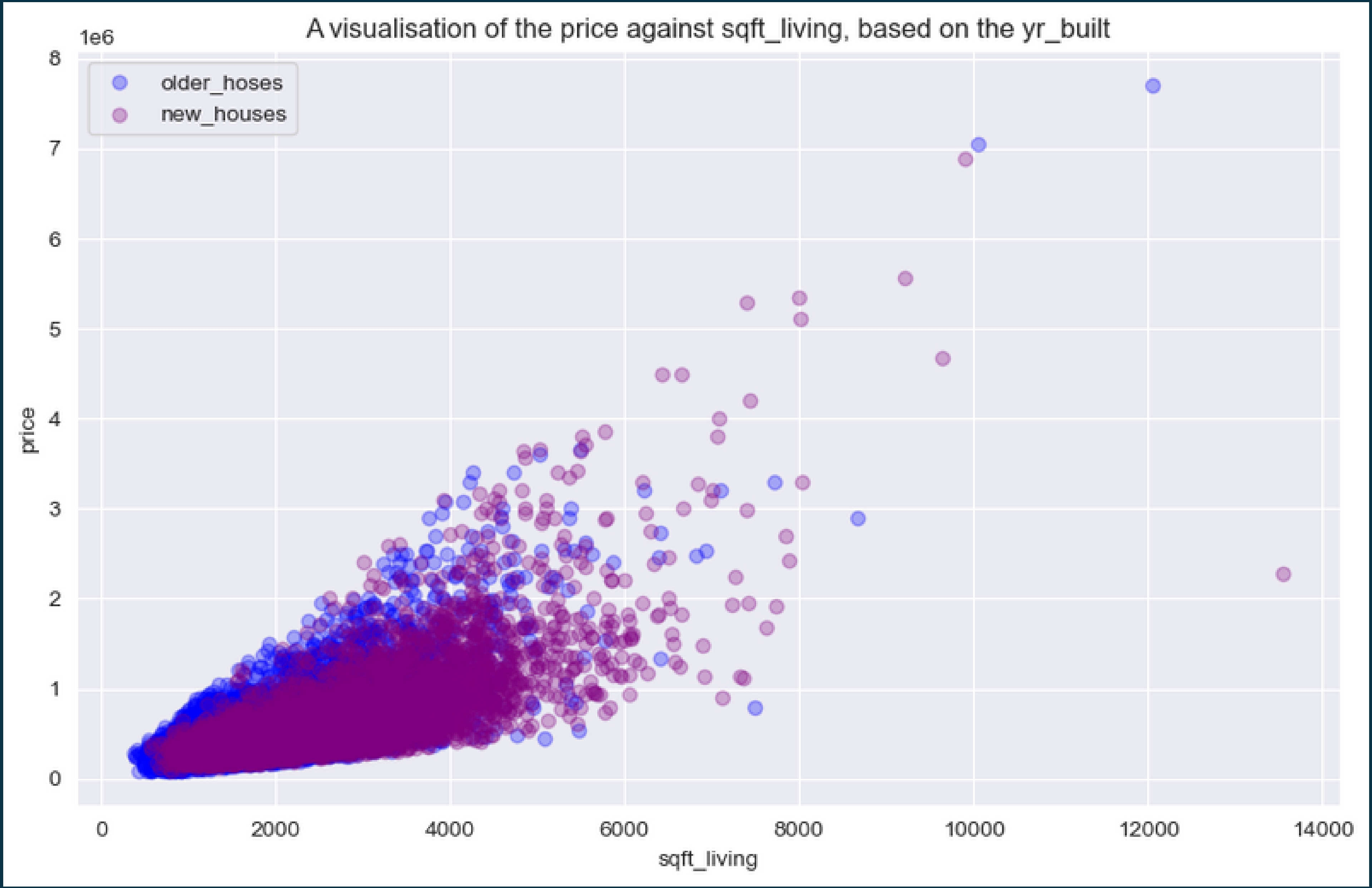



A visialisation of the sqft_living against the price based on grade.
1e6
8
7
6
5
4
3
2
1
0
price
0
2000
4000
6000
8000
10000
12000
14000
sqft_living
grade
4
6
8
10
12

3)Finding out the relationship between price, sqft_living and the year built.

We go on deeper again and try to find if there is a relationship between the sqft_living, the yr_built and the price .Here we classify houses that are built past 1969(mean) as newly built houses and houses built before 1969 as old houses.We then use this to create an interaction between the sqft_living of the new houses and those of the old houses. We then find out that there is an interaction between the sqft_living, price and the year built, where most houses that were recently built have higher chances of selling at a good price unlike the old houses ,however we do have some old houses selling at relatively high prices, this could be as by chance or due some other factors hence more analysis needs to be done. In the next slide is a visualization that show my findings.

NEXT

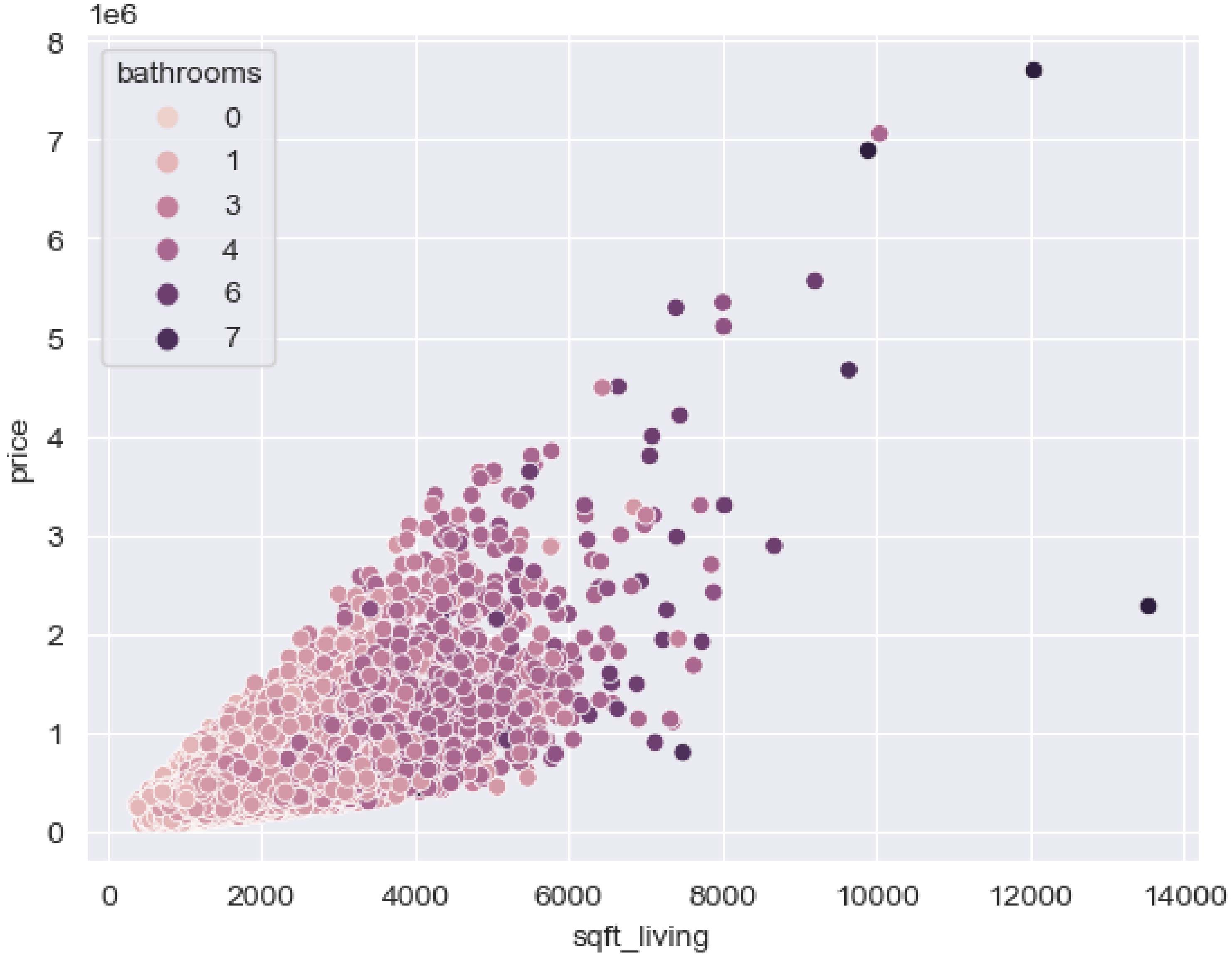

A visualisation of the price against sqft_living, based on the yr_built
1e6
8
7
6
5
4
3
2
1
0
price
older_hoses
new_houses
0
2000
4000
6000
8000
10000
12000
14000
sqft_living

4)Finding a relationship between sqft_living , price and the number of bathrooms in a house

I wanted to find out, does the price of a house increase based on the number of bathrooms, in a house? There was no linear relationship between price and the number of bathrooms , hence we couldn't be able to form a linear regression model on bathrooms alone, I discovered that there is however an interaction between sqft_living and the number of bathrooms in a house. A linear model between price and the interaction shows that as the sqft_living and number of bathrooms increases higher chances that the price of our house increases. This can clearly be seen in a visualization provided in the next slide.

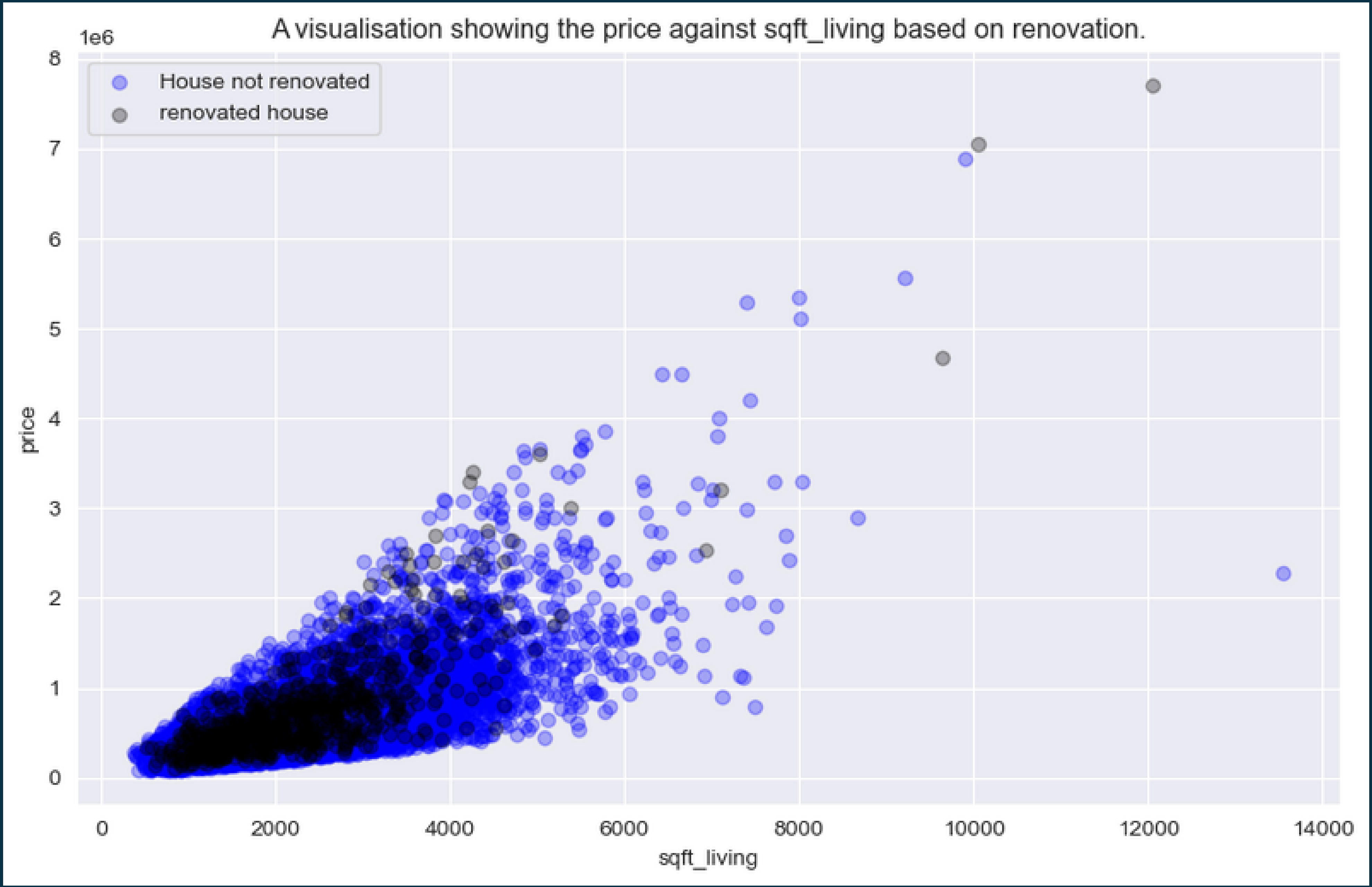
NEXT

A visualisation of the sqft_living against the prices based on the number of bathrooms in a house.



5)Does the renovation of a house affect its price.

Here we are seeking to find out if the renovation of a house has an effect on price. I categorized these into two where houses that had a year of renovation were encoded as a one(meaning renovation was done) and the rest as zero(meaning no renovation was done).I did analysis and found out that renovation doesn't really have much effect on the price. This could be because the most houses could be newly built, or that we don't really have much information about this renovation.



A visualisation showing the price against sqft_living based on renovation.

6) Predicting a value that can be used as a good sale price when selling a house based on the features our model found relevant.

Our problem was finding relevant features that an estate agent or company should look into when selling a house so that they can get the most reasonable price of sale. Based on our data our model recognized that the agent should look into the following features.

- number of bedrooms
- number of bathrooms
- number of floors
- House condition
- house grade
- sqft_lot
- sqft_living
- views
- yr_built.
- If the house has a waterfront.

From these features our predicted reasonable sale price would be in the range of $1,316,044.93 - $673648.12

7)Conclusions

From my analysis I can conclude that :

1)There is a high correlation between sqft_living and the target price.

2)There is a linear relationship between sqft_living and price.

3)The price of a house increase as the sqft_living and grade increase.

4)The price of a house increase as the sqft_living and the number of bedrooms increase.

5)Renovation does not have much of an effect on the house price.It should be notted that maybe this is because some of this houses were newly built hence more research is needed on this.

6)The best features to be considered when selling a house include:

a)sqft_living

b)yr_built

c)waterfront

d)views

e)grade

f)floors

g)yr_built

h)floors

i)bathrooms

7)The range of sale price for houses with these features is between $1,316,044.94-$976,256.16

8)Recommendations

1)I would recommend that more focus on the sale of houses with a good sqft_living as this has quite much of an impact on the house price.
2)I would recommend that the company or agent should look into the following features and seek to satisfy them for a good sale:

a)sqft_living
b)yr_built
c)waterfront
d)views
e)grade
f)floors
g)yr_built
h)floors
i)bathrooms

3)I would recommend that more research be done on the renovated houses and their sale price. 4)I would recommend that a sales agent should target at selling a house at a price range between 1,316,044.93 *to* 4,976,256.16 .

Thank you