Contents

Active Group Members

Farrukh Nadeem

Iris Kontogiorgi

Matheus Pereira da Silva Florencio

Cynthia Anyanwu

Dataset overview:

We started our project by getting familiar with overall picture of our data. The dataset represents houses data with 21 columns and 21597 records.

Types of Data

Variable	Type of Variable	Explanation
Id	Nominal	They are unique identifiers randomly assigned without any order in mind
Date	Ordinal	Each date is either bigger or smaller than the other
Price	Ratio	They cannot be less than zero and have equal spacing.
Bedrooms	Ratio	They cannot be less than zero and have equal spacing.
Bathrooms	Ratio	They cannot be less than zero and have equal spacing.
sqft_living	Ratio	They cannot be less than zero and have equal spacing.
sqft_lot	Ratio	They cannot be less than zero and have equal spacing.
Floors	Ratio	They cannot be less than zero and have equal spacing.
Waterfront	Nominal	They only represent existence of waterfront with no inherent order.
View	Ordinal	They have an order with 0 being lowest and 4 highest
Condition	Ordinal	They have an order with 1 being lowest and 5 highest

Grade	Ordinal	They have an order with 1 being lowest and 13 highest
sqft_above	Ratio	They cannot be less than zero and have equal spacing.
sqft_basement	Ratio	They cannot be less than zero and have equal spacing.
yr_built	Ordinal	Each year is either bigger or smaller than the other
yr_renovated	Ordinal	Each year is either bigger or smaller than the other
Zipcode	Nominal	They are unique identifiers randomly assigned without any order in mind
Lat	Interval	They can be less than zero and have equal spacing.
Long	Interval	They can be less than zero and have equal spacing.
sqft_living15	Ratio	They cannot be less than zero and have equal spacing.
sqft_lot15	Ratio	They cannot be less than zero and have equal spacing.

Project Questions

What are the ideal locations (Zip codes) to invest in considering profit margins on resale in mind?

Does the ratio of bedroom to lot area have any impact on price? If yes, what is the suggested ratio?

Does the ratio of bedroom to floors have any impact on price? If yes, what is the suggested ratio?

Does the ratio of bedroom to bathrooms have any impact on price? If yes, what is the suggested ratio?

Data Cleaning:

Irrelevant Data

We considered three columns: **ID, sqft_living15, and sqft_lot15** to be irrelevant initially, but kept them for following reasons.

ID column: Though this column seemed irrelevant at first, but on analysis we saw that few houses were listed multiple times on different dates with different prices. This led us to keep this column as we thought it might prove to be useful in future analysis.

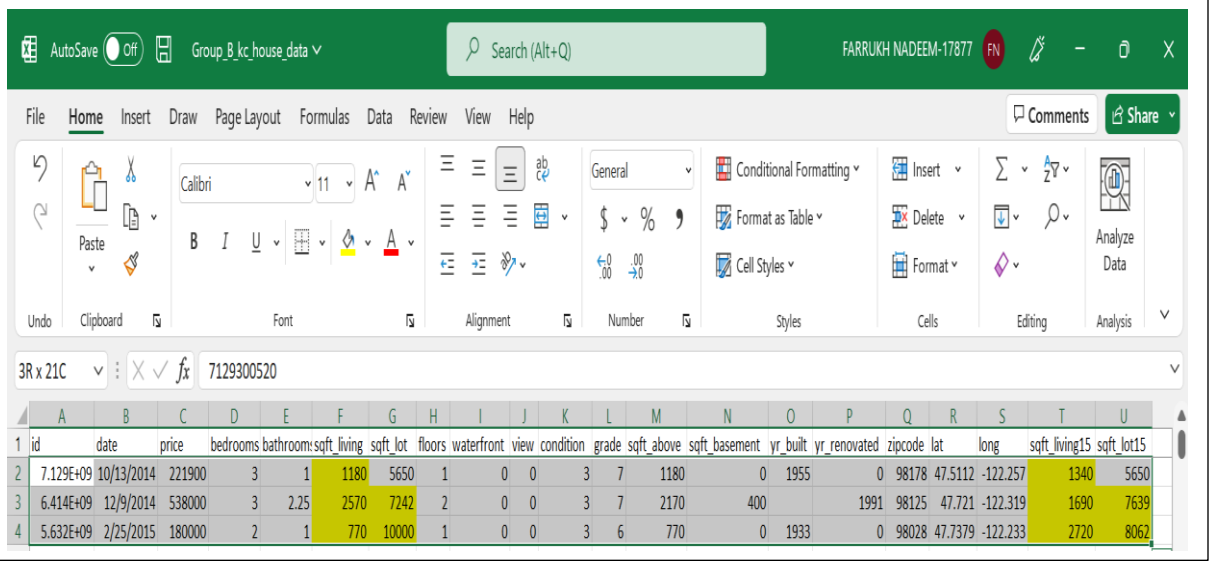
Figure 1

	id	date	price	bedrooms	...	lat	long	sqft_living15	sqft_lot15
17588	795000620	9/24/2014	115000.0	3	...	47.5045	-122.33	1070.0	6250.0
17589	795000620	12/15/2014	124000.0	3	...	47.5045	-122.33	1070.0	6250.0
17590	795000620	3/11/2015	157000.0	3	...	47.5045	-122.33	1070.0	6250.0

[3 rows x 21 columns]

sqft_living15 and sqft_lot15 columns: These two columns looked like sqft_living and sqft_lot, but still we kept them as in many cases they represented different values. Here is an example:

Figure 2



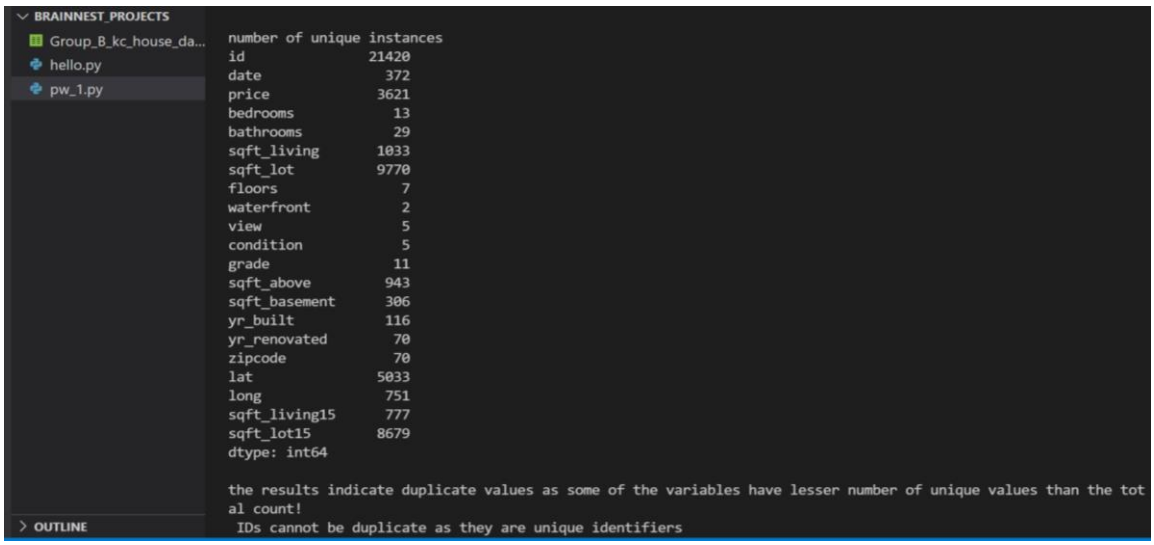
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
2	7.129E+09	10/13/2014	221900	3	1	1180	5650	1	0	0	3	7	1180	0	1955	0	98178	47.5112	-122.257	1340	5650
3	6.414E+09	12/9/2014	538000	3	2.25	2570	7242	2	0	0	3	7	2170	400		1991	98125	47.721	-122.319	1690	7639
4	5.632E+09	2/25/2015	180000	2	1	770	10000	1	0	0	3	6	770	0	1933	0	98028	47.7379	-122.233	2720	8062

Duplicate Data

Before getting to completely duplicated rows, we first checked how many unique values each column had.

Below is the image showing that result.

Figure 3



ID column: Figure 3 shows that there are lesser number of IDs than the total rows indicating that there are duplicate IDs. But as indicated already by **figure 1**, we knew that even though some IDs were duplicated but they still had different values in some other columns.

Latitude and Longitude: In our opinion the combination of latitude and longitude for each house must be unique unless they are apartments in the same building. We saw additional 177 houses with **different IDs** but **duplicated combination of latitude and longitude along** with other similar features. But we still decided to keep them knowing that they might be different houses (as indicated by different IDs) in the same building with similar features.

One such example can be seen below:

Figure 4

```
number of combinations of latitude and longitude with duplicate instances
note True means similar combinations were detected already

False    21420
True       177
dtype: int64
              id       date     price  bedrooms  ...      lat     long  sqft_living15  sqft_lot15
14004  9136103136  3/13/2015  580000.0         2  ...  47.6652 -122.338         1490.0      4013.0
17852  7732410360  8/22/2014  752888.0         3  ...  47.6599 -122.146         2630.0      9000.0
19568  1025049266  9/30/2014  555000.0         2  ...  47.6647 -122.284         1160.0      1327.0
21209  3629700080   1/8/2015  635000.0         3  ...  47.5446 -122.017         2290.0      1407.0
21371  2895800750  4/17/2015  274800.0         3  ...  47.5171 -122.347         1410.0      1899.0
21482  1972200555  7/14/2014       NaN         3  ...  47.6536 -122.354         1570.0      1335.0

[6 rows x 21 columns]
              id       date     price  bedrooms  bathrooms  sqft_living  sqft_lot  floors      lat     long
211   1025049114  7/17/2014  625504.0         3       2.25       1270.0    1566.0     2.0  47.6647 -122.284
13459 1025049115  6/25/2014  594000.0         3       2.25       1270.0    1406.0     2.0  47.6647 -122.284
18297 1025049268  7/21/2014  549900.0         2       1.75       1140.0     936.0     2.0  47.6647 -122.284
19568 1025049266  9/30/2014  555000.0         2       2.25       1160.0     954.0     2.0  47.6647 -122.284
              id      lat     long  waterfront  view  condition  grade  sqft_above  sqft_basement  yr_built
211   1025049114  47.6647 -122.284           0   0.0        3.0    8.0      1060.0          210.0    2014.0
13459 1025049115  47.6647 -122.284           0   0.0        3.0    8.0      1060.0          210.0    2014.0
18297 1025049268  47.6647 -122.284           0   0.0        3.0    8.0       940.0          200.0    2014.0
19568 1025049266  47.6647 -122.284           0   0.0        3.0    8.0       960.0          200.0    2014.0
PS D:\Brainnest\PW_1\Brainnest_projects>
```

These analyses still did not rule out the possibility of entire row having exact duplicate, so we checked to see whether that was the case, but we saw no duplication.

Structural Errors

We did not identify any structural errors. However, we found it odd to see Bathrooms and floors to not have discrete values.

Figure 5

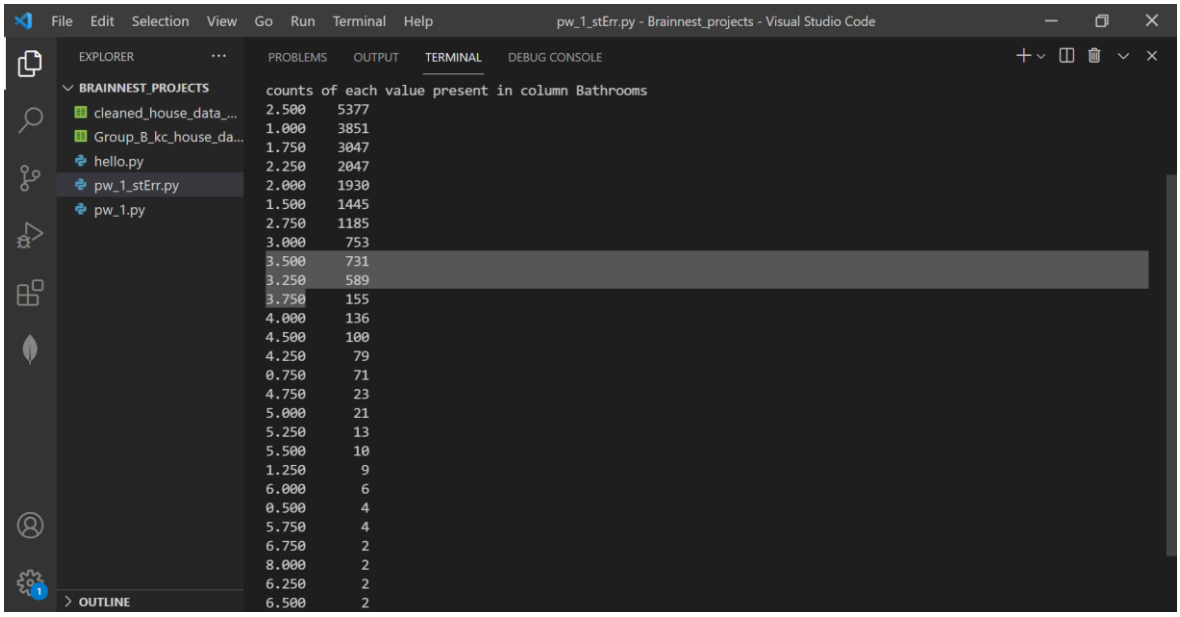
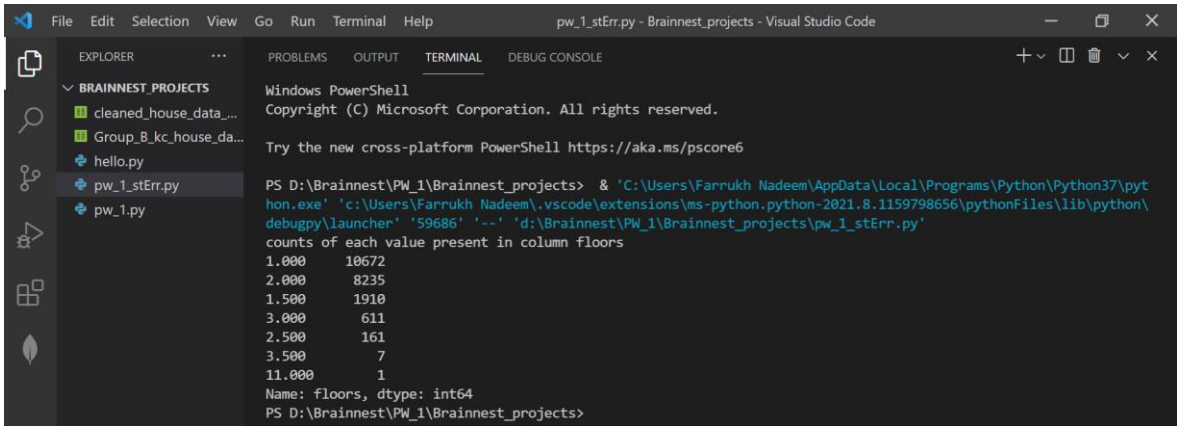


Figure 6

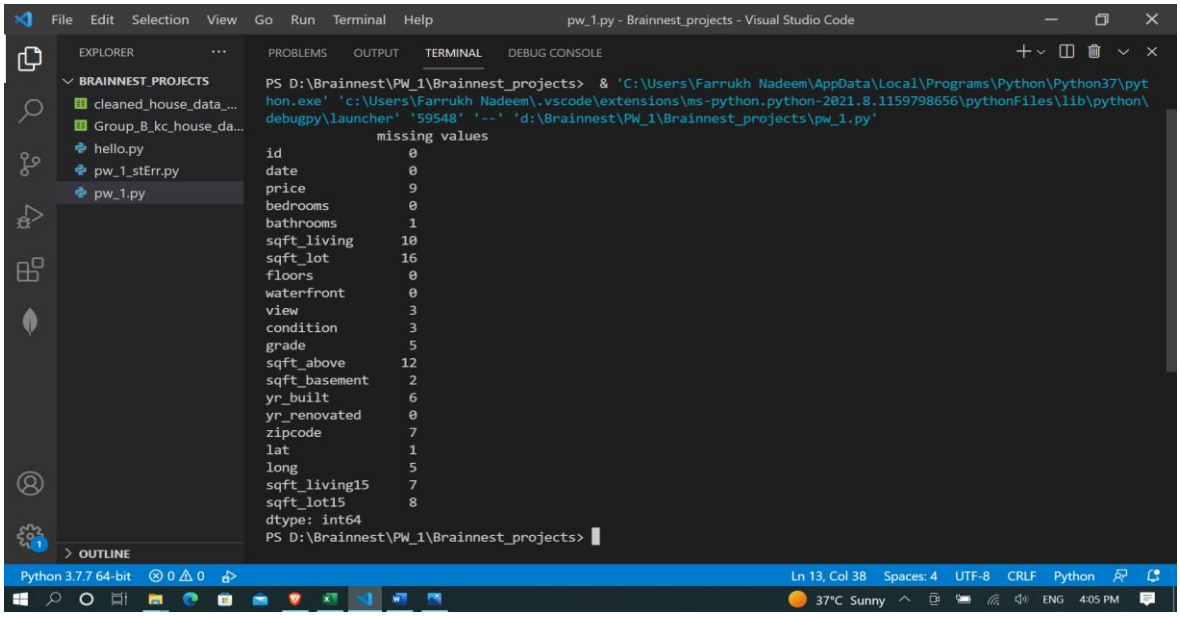


We still did not correct them as their high frequency indicates that it cannot be by mistake and might actually mean something important for real estate agents.

Missing Data

We checked to see how many missing values were present in each column, **figure 7** shows the result.

Figure 7



The percentage of missing values for any column was very low. And except for the columns: zipcode, lat, and long, all other columns were missing completely at random (MCAR) and hence, were imputed with median.

For columns: zipcode, lat, and long, they were missing at random as the missing data could be explained/derived from the other column. Hence, for zipcodes we entered the values of lat, and long on the following website to find their respective zip code.

https://www.freemaptools.com/convert-us-zip-code-to-lat-lng.htm

And since missing lat, and long were very few for us to manually check them and replace them we imputed missing values in these columns to the best of our knowledge.

Data Outliers

During analyses we saw outliers in bedrooms and floors columns.

Figure 8

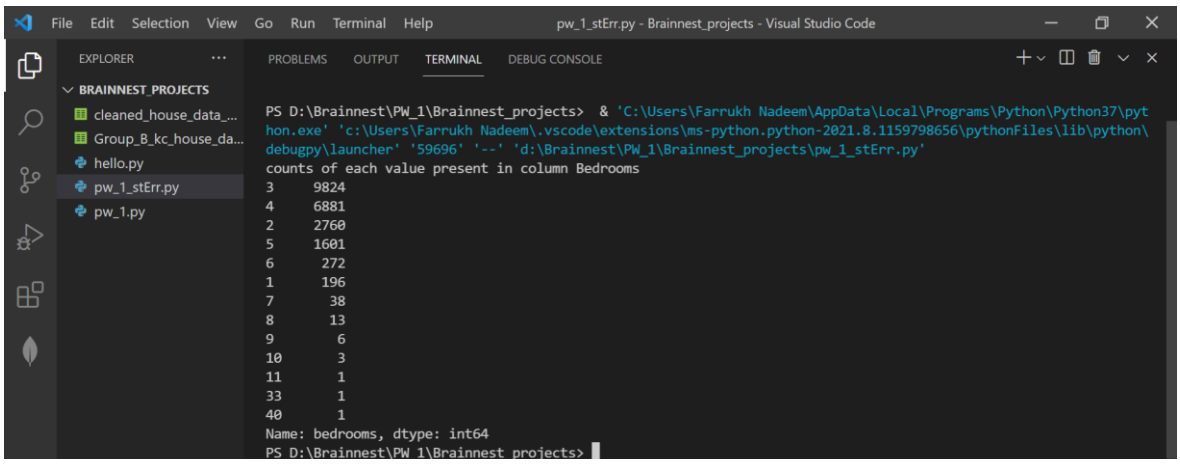
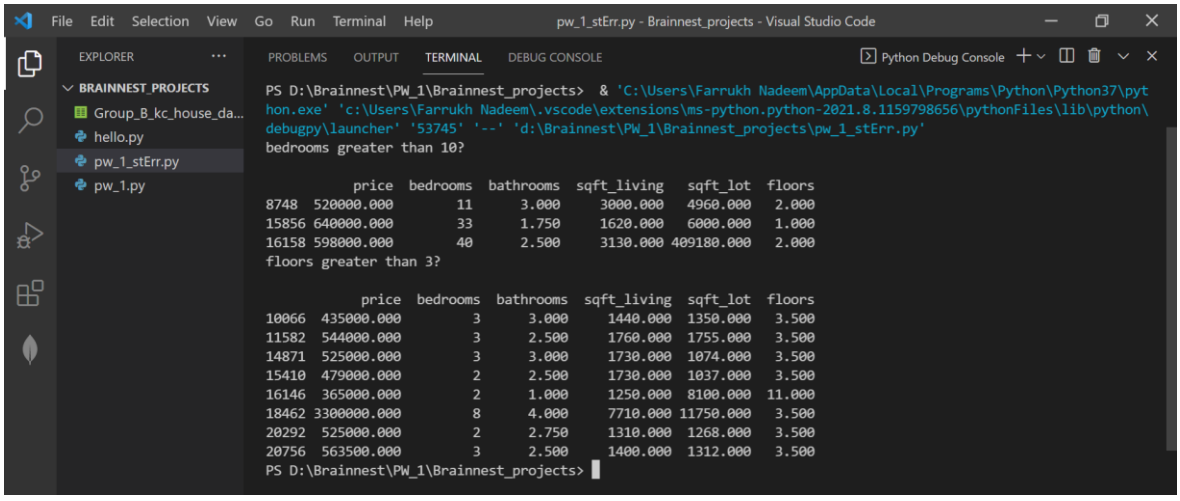


Figure 9



We believe having 33 or 40 bedrooms in 2 floors with such low number of bathrooms and small sqft_living must be an error. We also found that odd to have only 2 bedrooms in 11 floors! Hence, we decided to replace these values that were more reasonable.

Descriptive Statistics

Figure 10

```
debugpy\launcher' '53639' '--' 'd:\Brainnest\PW_1\Brainnest_projects\pw_1_stErr.py'
              price  bedrooms  bathrooms  sqft_living     sqft_lot     floors
count   21588.000 21597.000  21596.000    21587.000    21581.000 21597.000
mean    540382.754     3.375      2.116     2080.185    15113.467     1.495
std     367405.169     0.959      0.769      917.968    41511.996     0.544
min      78000.000     1.000      0.500      370.000      520.000     1.000
25%     322000.000     3.000      1.750     1430.000     5040.000     1.000
50%     450000.000     3.000      2.250     1910.000     7617.000     1.500
75%     645000.000     4.000      2.500     2550.000    10677.000     2.000
max    7700000.000    40.000      8.000    13540.000 1651359.000    11.000
         waterfront      view  condition      grade  sqft_above  sqft_basement  yr_built  yr_renovated
count   21597.000 21594.000  21594.000 21592.000   21585.000      21595.000 21591.000      21597.000
mean        0.008     0.234      3.410     7.658    1789.258        291.710  1970.998         84.465
std         0.087     0.766      0.651     1.173     834.846        442.664    29.377        401.821
min         0.000     0.000      1.000     3.000     370.000          0.000  1900.000          0.000
25%         0.000     0.000      3.000     7.000    1190.000          0.000  1951.000          0.000
50%         0.000     0.000      3.000     7.000    1560.000          0.000  1975.000          0.000
75%         0.000     0.000      4.000     8.000    2210.000        560.000  1997.000          0.000
max         1.000     4.000      5.000    13.000   18000.000       4820.000  2015.000       2015.000
            zipcode       lat      long  sqft_living15  sqft_lot15
count 21590.000 21596.000 21592.000      21590.000    21589.000
mean  98077.950    47.560  -122.214       1986.586    12761.042
std      53.513     0.139     0.141        685.296    27279.087
min   98001.000    47.156  -122.519        399.000      651.000
25%   98033.000    47.471  -122.328       1490.000     5100.000
50%   98065.000    47.572  -122.231       1840.000     7620.000
75%   98118.000    47.678  -122.125       2360.000    10083.000
max   98199.000    47.778  -121.315       6210.000   871200.000
```

Feature Engineering

Based on the Project Questions/Objectives we felt a need to create some columns that could best serve our needs.

Figure 11

Feature Engineering

Column 1	Column 2
Price	sqft_lot
sqft_lot	bedrooms
bedrooms	floors
bedrooms	bathrooms

Calculation
Price/sqft_lot
sqft_lot/bedrooms
bedrooms/floors
bedrooms/bathrooms

Calculated field name	Intended outcome
pricelot	priceper unit lot size
lotbedroom	lot size to bedroom ratio
bedroomfloor	bedroom to floor ratio
bedroomsbathrooms	bedrooms to bathrooms ratio

From figure 11, on the left most group we have two columns, by diving Variables listed in Column 1 by the variables in Column 2 we obtain results that are listed in the right most group's first column i.e., Calculated Field Name.

We believe that comparing price per square foot for different lots is a more relevant measure than the price itself. Thus, we created for ourselves a column named **pricelot**.

We also created the other three columns: lotbedroom, bedroomfloor, and bedroomsbathrooms to facilitate ourselves in trying to answer project questions/obejectives 2, 3 ,and 4 respectively.

Normality Tests

We checked the normality of each variable (Column) using following methods:

Frequency Histogram
comparing mean, median, and mode
Skewness and kurtosis
Z-scores (Skewness and kurtosis)
Q-Q Plots
P-P Plots
Box Plots

Here is an example of the tests applied on one of these columns.

Figure 12

Price

Visual and statistical tests like Histogram, Box Plots, Q-Q plots and P-P plots, showed that the price is not normally distributed with skewness of-4.023(SE= 0.017) and kurtosis of-34.536(SE= 0.033).

Test	Normally Distributed
Histogram	No
QQ Plot	N0
PP Plot	No
Box Plot	No
Skewness	No
Z score (Skewness)	No (236.64)
Kurtosis	No
Z score (Kurtosis)	No (1046)
Final conclusion	No (8/8)

For the sake of keeping report neat, the following table shows the results of Normality Tests for only the columns that are required to answer our project questions.

Table 1: Normality Test

Variable	Normality Test	Value
Price	Not Normal	100%
Bedrooms	Not Normal	75%
Bathrooms	Not Normal	75%
Sqft_lot	Not Normal	100%
Floors	Approximately Normal	50%
Zipcode	Categorical Variable. Hence, did not test.	

Conclusion on Normality Tests

For Floors we are 50% sure it is Normally distributed

For other variables we are at least 75% sure they are not normally distributed.

So, we suggest going for non-parametric tests in our future analysis.

Hypothesis Testing

As can be seen in Table 2, for every Project Objective (PO.) we created its hypothesis H0 (null) and H1 (alternate). Since our Normality Test results suggested us to use non-parametric tests, and for all of our tests we had one dependent variable and one independent variable with more than two groups Kruskal-Wallis test was our best option.

Here is one example of the application of Kruskal-Wallis Test for one of our Hypothesis:

Figure 13

Nonparametric Tests

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of pricelot is the same across categories of zipcode.	Independent-Samples Kruskal-Wallis Test	<.001	Reject the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

Independent-Samples Kruskal-Wallis Test

pricelot across zipcode

Independent-Samples Kruskal-Wallis Test Summary

Total N	21597
Test Statistic	12075.477[a]
Degree Of Freedom	69
Asymptotic Sig.(2-sided test)	<.001

a. The test statistic is adjusted for ties.

Independent-Samples Kruskal-Wallis Test

Like this test, for all our hypothesis the p-value was less than 0.05, therefore, we rejected the null hypothesis in favor of alternate hypothesis.

Following is the summary of our results on non-parametric tests:

Table 2: Hypothesis Test

PO.	H0	H1	Applied Test	Result
1	The cost of land per unit area is same across all neighborhoods (i.e., zip codes).	The cost of land per unit area is affected by the neighborhood.	Kruskal-Wallis Test	Accept H1
2	The mean price is not affected by different ratios of lot area to bedroom.	The mean price is affected by lot area to bedroom ratio.		Accept H1
3	The mean price is not affected by different ratios of bedroom to floors.	The mean price is affected by bedroom to floor ratio.		Accept H1
4	The mean price is not affected by different ratios of bedroom to bathrooms.	The mean price is affected by bedroom to bathroom ratio.		Accept H1

Accepting H1 for all tests indicate that our questions were indeed valid and there is a need for in-depth analysis to detect any meaningful patterns.

Pattern Recognition

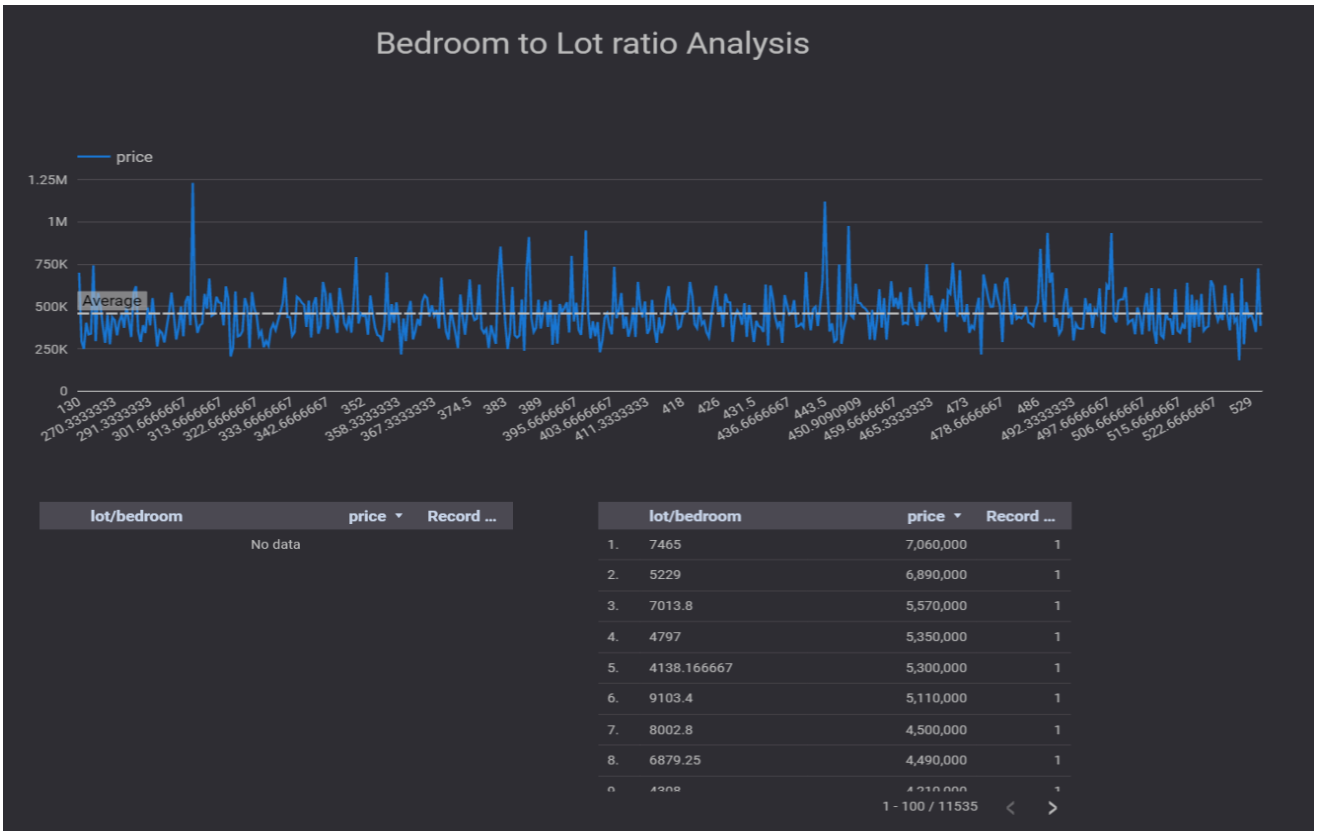
To find patterns between certain variables we first set for ourselves a threshold of at least 2.5% representation for any group in a factor.

PO2, PO3, and PO4

When trying to perform in-depth analysis for PO2, PO3, and PO4 most of the groups failed to meet the threshold. Thus, we decided not to investigate any further as deducing any rules from such small proportion of data would be meaningless and misleading in our opinion.

Below is the example of PO2 where square feet lot to bedroom ratio was to be investigated:

Figure 14



As can be seen from Figure 14, on the bottom left there is an empty table (this is due to the application of filter that restricts groups falling behind the threshold). This can be explained by the line chart on top and table on bottom-right which highlight the fact that there are too many groups within this factor thus rendering it impossible for any group to occupy a larger share of the sample.

Similar scenarios were observed for PO3 and PO4 where bedroom to floors, and bedroom to bathroom ratios were to be investigated respectively. Hence, we decided to end our quest to find any meaningful patterns for these objectives/questions.

PO1: Impact of Location on prices

Unlike the rest of the questions, the number of groups in PO1 were few. This meant that the groups had enough representation from their overall samples, hence, deducing any rules from this analysis was possible.

We conducted in-depth studies using different graphs for price/lot against zipcodes and then selected 7 locations (zipcodes) of interest. We then divided these 7 locations into three groups (G1, G2, and G3) based on their average price per square foot of land, with G1 requiring minimum investment and G3 the highest for a lot of same size.

Figure 15

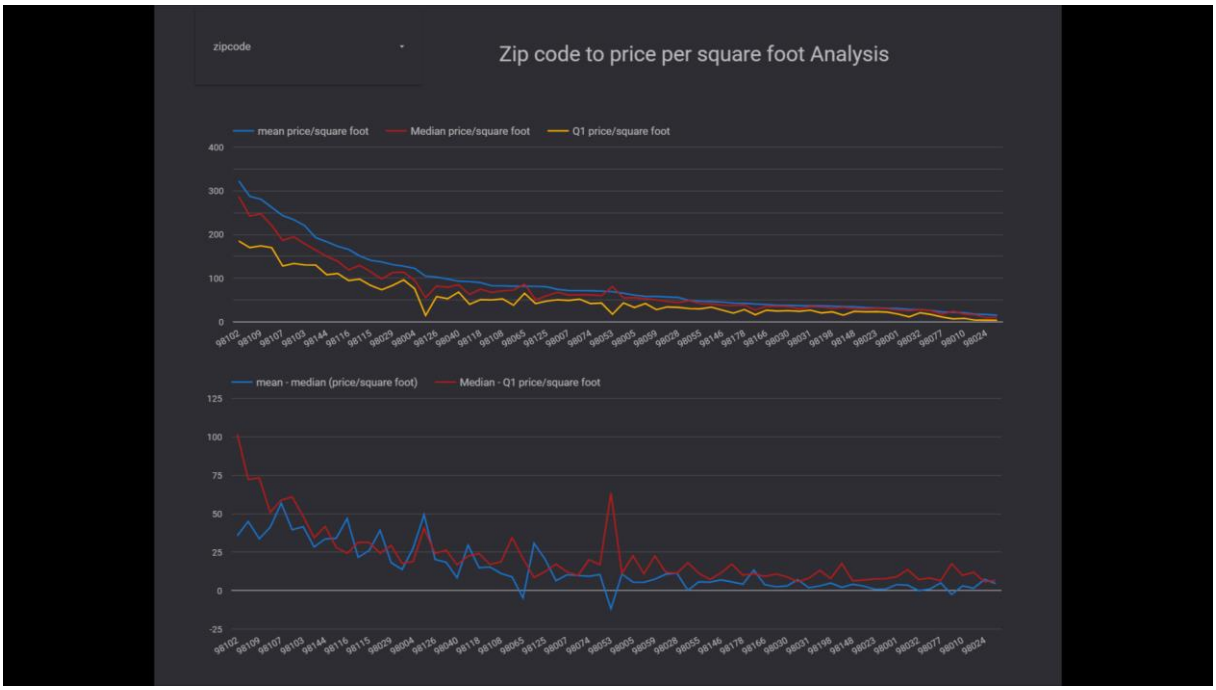


Figure 16

zipcode: 98027, 98053, 98022, 98... (7)

Zip code to price per square foot Analysis

	zipcode	price/lot	mean - median (price/square foot)	Median - Q1 price/square foot	profit% (price/square feet) (se...	Record Count
1.	98102	322.56	35.7	101.74	35.47	104
2.	98119	287.43	45.02	72.33	29.84	184
3.	98109	281.08	33.69	73.31	29.63	109
4.	98027	104.82	49.43	40.53	73.17	412
5.	98053	69.19	-11.78	63.36	78.25	403
6.	98022	21.99	-2.46	17.53	71.71	233
7.	98014	17.77	1.6	11.97	74.02	124

1 - 7 / 7

By studying carefully charts from Figures 15, and 16 we decided that our focus should be to try to buy property that is priced under the lower quartile of their respective groups, and then flip them (sell them) at a price equal to or greater than the median price of its respective group.

Though finding property at a price below lower quartile will be hard but it will be a little easy to sell them for median prices as they are mostly closer to average prices and yet cheaper than 30-40% of the houses in same location.

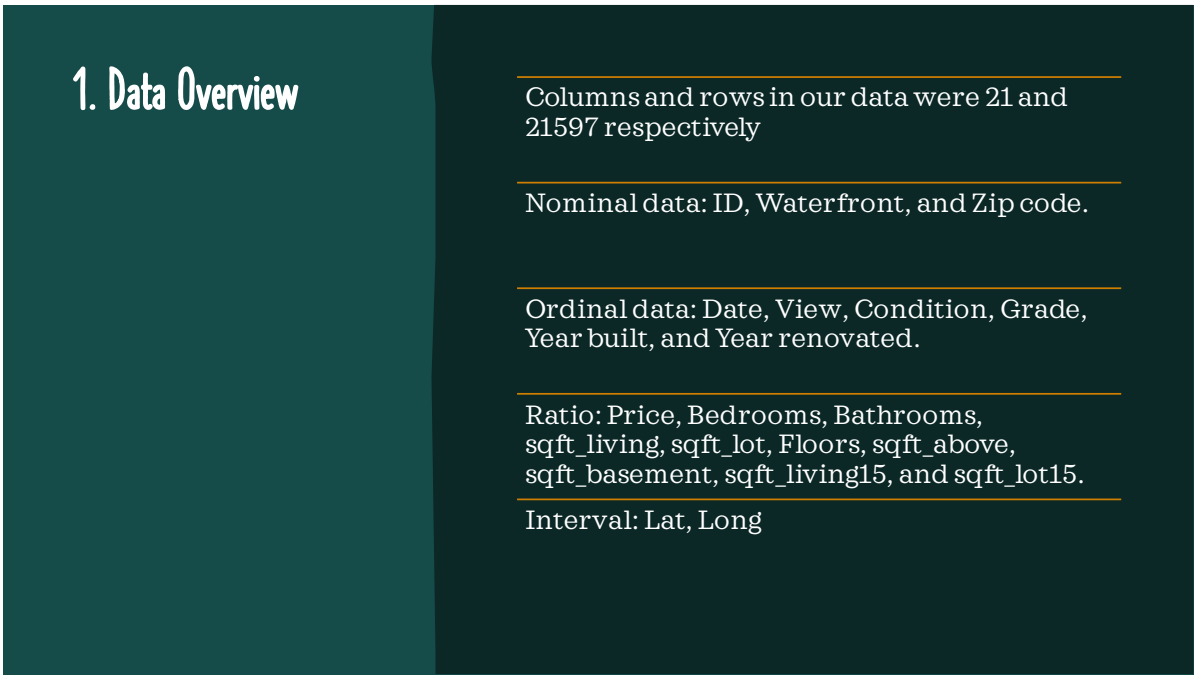
Findings

We came to following conclusions/suggestions:

- For G1 (98022, 98014), if investors can buy property under **$6.9/sqft** and **$4.2/sqft** respectively and flip them for at least **$24.4/sqft** and **$16.2/sqft**, they can expect profit margin between **71-74%**.
- For G2 (98027, 98053), if investors can buy property under **$14.8/sqft** and **$17.6/sqft** respectively and flip them for at least **$55.4/sqft** and **$81/sqft**, they can expect profit margin between **73-78%**.
- For G3 (98102, 98109), if investors can buy property under **$185/sqft** and **$174/sqft** respectively and flip them for at least **$287/sqft** and **$247/sqft**, they can expect profit margin between **29-35%**.

Appendix A: Project Presentations

Project Week 1



2.1 Irrelevant Data

- Considered ID column
- Multiple entries with different values in other variables
- Left them untouched

```
                 id        date     price  bedrooms  ...      lat     long  sqft_living15  sqft_lot15
17588  795000620   9/24/2014  115000.0         3  ...  47.5045 -122.33         1070.0      6250.0
17589  795000620  12/15/2014  124000.0         3  ...  47.5045 -122.33         1070.0      6250.0
17590  795000620   3/11/2015  157000.0         3  ...  47.5045 -122.33         1070.0      6250.0

[3 rows x 21 columns]

the reults above show that though same house has been listed thrice, they were all listed on different dates wi
th different prices!
Hence, I conclude that there are no duplicates in my opinion!
PS D:\Brainnest\PW_1\Brainnest_projects>
```

OUTLINE

Python 3.7.7 64-bit ⊗ 0 ⚠ 0 Ln 18, Col 38 Spaces: 4 UTF-8 CRLF Python

2.1 Irrelevant Data

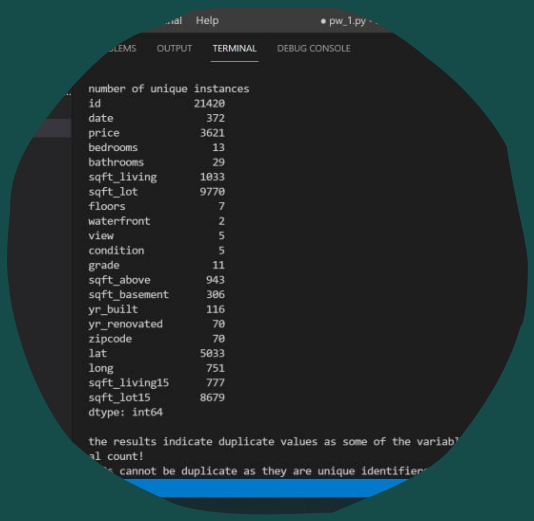
- Considered sqft_living15 and sqft_lot15 columns
- Different values for same records
- Hence, kept them.

3R x 21C fx 7129300520

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
2	7.129E+09	10/13/2014	221900	3	1	1180	5650	1	0	0	3	7	1180	0	1955	0	98178	47.5112	-122.257	1340	5650
3	6.414E+09	12/9/2014	538000	3	2.25	2570	7242	2	0	0	3	7	2170	400		1991	98125	47.721	-122.319	1690	7639
4	5.632E+09	2/25/2015	180000	2	1	770	10000	1	0	0	3	6	770	0	1933	0	98028	47.7379	-122.233	2720	8062
5	2.487E+09	12/9/2014	604000	4	3	1960	5000	1	0	0	5	7	1050	910	1965	0	98136	47.5208	-122.393	1360	5000
6	1.954E+09	2/18/2015	510000	3	2	1680	8080	1	0	0	3	8	1680	0	1987	0	98074	47.6168	-122.045		7503
7	7.238E+09	5/12/2014	1.23E+06	4	4.5		101930	1	0	0	3	11	3890	1530		0	98053	47.6561	-122.005	4760	101930
8	1.321E+09	6/27/2014	257500	3	2.25	1715	6819	2	0	0	3	7	1715	0	1995	0		47.3097	-122.327	2238	
9	2.008E+09	1/15/2015	291850	3	1.5	1060	9711	1	0	0	3	7	1060	0	1963	0	98198	47.4095	-122.315	1650	9711
10	2.415E+09	4/15/2015	229500	3	1	1780	7470	1	0	0	3	7	1050	730	1960	0	98146	47.5123	-122.337	1780	8113
11	3.794E+09	3/12/2015	323000	3	2.5	1890	6560	2	0	0	3	7		0	2003	0	98038	47.3684	-122.031		7570
12	1.737E+09	4/3/2015	662500	3	2.5	3560		1	0	0	3	8	1860	1700	1965	0	98007	47.6007			8925
13	9.213E+09	5/27/2014	468000	2	1	1160	6000	1	0	0	4	7	860	300	1942	0	98115	47.69	-122.292	1330	6000
14	114101516	5/28/2014	310000	3	1	1430	19901	1.5	0	0	4	7	1430	0	1927	0	98028	47.7558	-122.229	1780	12697
15	6.055E+09	10/7/2014	400000	3	1.75	1370	9680	1	0	0	4	7	1370	0	1977	0	98074	47.6127	-122.045	1370	10208
16	1.175E+09	3/12/2015	530000	5	2	1810	4850	1.5	0	0	3	7	1810	0	1900	0	98107	47.67	-122.394	1360	4850
17	9.297E+09	1/24/2015	650000	4	3	2950	5000	2	0	3	3	9	1980	970		0	98126	47.5714	-122.375	2140	4000
18	1.876E+09	7/31/2014	395000	3	2		14040	2	0	0	3	7	1890	0	1994	0	98019	47.7277	-121.962	1890	14018
19	6.865E+09	5/29/2014	485000	4	1	1600	4300	1.5	0	0	4	7	1600	0	1916	0	98103	47.6648	-122.343	1610	4300
20	16000397	12/5/2014	189000	2	1	1200	9850	1	0	0	4	7	1200	0	1921	0	98002	47.3089	-122.21	1060	5095
21	7.983E+09	4/24/2015	230000	3	1	1250		1	0	0	4	7	1250	0	1969	0	98003	47.3343	-122.306	1280	8850
22	6.301E+09	5/14/2014	385000	4	1.75	1620	4980	1	0	0	4	7	860	760	1947	0	98133	47.7025	-122.341	1400	4980

2.2 Duplicate Data

First checked unique values in each column



2.2 Duplicate Data

- Considered ID column
- Multiple entries with different values in other variables
- Therefore, did not remove them

```
              id        date     price  bedrooms  ...      lat     long  sqft_living15  sqft_lot15
17588  795000620   9/24/2014  115000.0         3  ...  47.5045 -122.33         1070.0      6250.0
17589  795000620  12/15/2014  124000.0         3  ...  47.5045 -122.33         1070.0      6250.0
17590  795000620   3/11/2015  157000.0         3  ...  47.5045 -122.33         1070.0      6250.0

[3 rows x 21 columns]

the reults above show that though same house has been listed thrice, they were all listed on different dates wi
th different prices!
Hence, I conclude that there are no duplicates in my opinion!
PS D:\Brainnest\PW_1\Brainnest_projects>
```

2.2 Duplicate Data

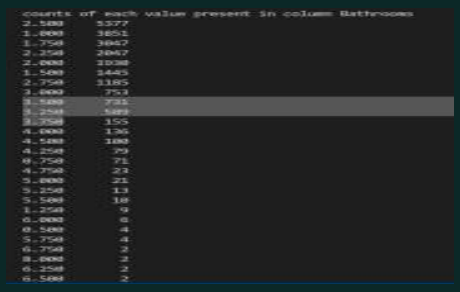
- Considered combination of latitude and longitude
- Houses could be apartments in same building
- Hence kept them

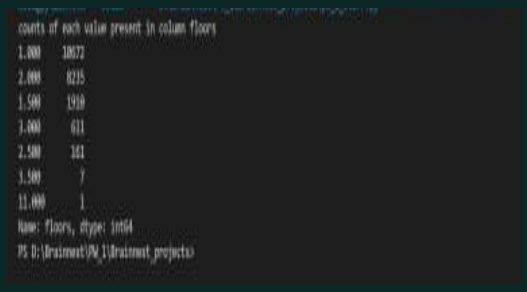
```
             id       date     price  bedrooms  ...      lat     long  sqft_living15  sqft_lot15
14004  9136103136  3/13/2015  580000.0         2  ...  47.6652 -122.338         1490.0      4013.0
17852  7732410360  8/22/2014  752888.0         3  ...  47.6599 -122.146         2630.0      9000.0
19568  1025049266  9/30/2014  555000.0         2  ...  47.6647 -122.284         1160.0      1327.0
21209  3629700080   1/8/2015  635000.0         3  ...  47.5446 -122.017         2290.0      1407.0
21371  2895800750  4/17/2015  274800.0         3  ...  47.5171 -122.347         1410.0      1899.0
21482  1972200555  7/14/2014       NaN         3  ...  47.6536 -122.354         1570.0      1335.0

[6 rows x 21 columns]
             id       date     price  bedrooms  bathrooms  sqft_living  sqft_lot  floors      lat     long
211    1025049114  7/17/2014  625504.0         3       2.25       1270.0    1566.0     2.0  47.6647 -122.284
13459  1025049115  6/25/2014  594000.0         3       2.25       1270.0    1406.0     2.0  47.6647 -122.284
18297  1025049268  7/21/2014  549900.0         2       1.75       1140.0     936.0     2.0  47.6647 -122.284
19568  1025049266  9/30/2014  555000.0         2       2.25       1160.0     954.0     2.0  47.6647 -122.284
             id      lat     long  waterfront  view  condition  grade  sqft_above  sqft_basement  yr_built
211    1025049114  47.6647 -122.284           0   0.0        3.0    8.0      1060.0          210.0    2014.0
13459  1025049115  47.6647 -122.284           0   0.0        3.0    8.0      1060.0          210.0    2014.0
18297  1025049268  47.6647 -122.284           0   0.0        3.0    8.0       940.0          200.0    2014.0
19568  1025049266  47.6647 -122.284           0   0.0        3.0    8.0       960.0          200.0    2014.0
PS D:\Brainnest\PW_1\Brainnest_projects>
```

2.3 Structural errors

- Odd to see bathrooms and floors not in discrete numbers
- Still left it as it is





2.4 Missing Data

2.5 Data Outliers

33 OR 40 BEDROOMS?

11 FLOORS

```
counts of each value present in column floors
1.000     10672
2.000      8235
1.500      1910
3.000       611
2.500       161
3.500         7
11.000        1
Name: floors, dtype: int64
counts of each value present in column Bedrooms
3     9824
4     6881
2     2760
5     1601
6      272
1      196
7       38
8       13
9        6
10       3
11       1
33       1
40       1
Name: bedrooms, dtype: int64
```

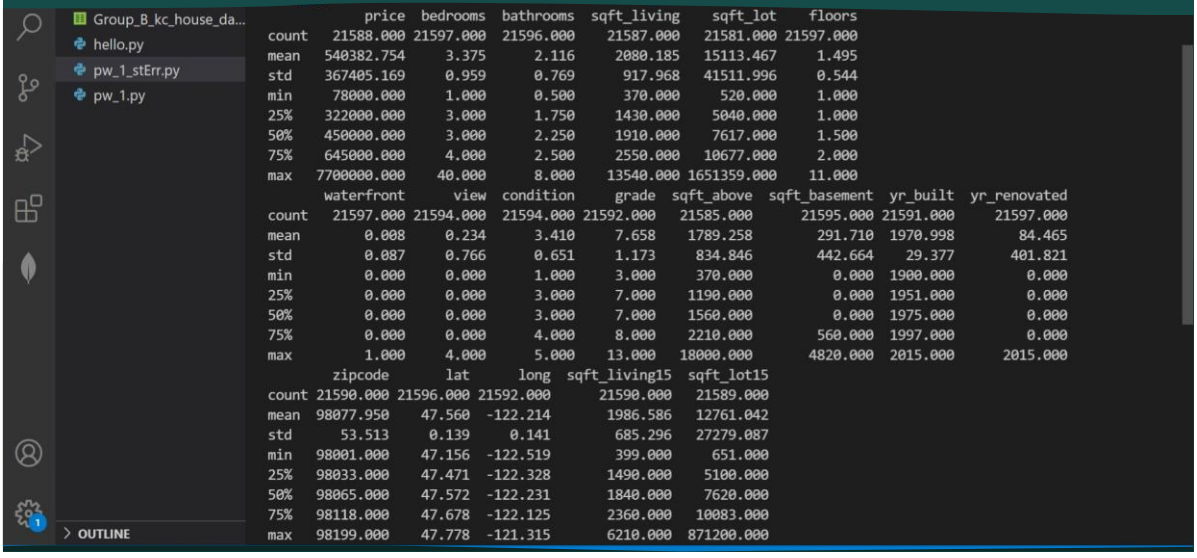
TOO FEW BATHROOMS FOR THAT MANY BEDROOMS!

11 FLOORS FOR JUST 2 BEDROOMS?

```
            price  bedrooms  bathrooms  sqft_living  sqft_lot  floors
8748   520000.000        11      3.000     3000.000  4960.000   2.000
15856  640000.000        33      1.750     1620.000  6000.000   1.000
16158  598000.000        40      2.500     3130.000 409180.000  2.000
floors greater than 3?

            price  bedrooms  bathrooms  sqft_living  sqft_lot  floors
10066  435000.000         3      3.000     1440.000  1350.000   3.500
11582  544000.000         3      2.500     1760.000  1755.000   3.500
14871  525000.000         3      3.000     1730.000  1074.000   3.500
15410  479000.000         2      2.500     1730.000  1037.000   3.500
16146  365000.000         2      1.000     1250.000  8100.000  11.000
18462 3300000.000         8      4.000     7710.000 11750.000   3.500
20292  525000.000         2      2.750     1310.000  1268.000   3.500
20756  563500.000         3      2.500     1400.000  1312.000   3.500
PS D:\Brainnest\PW_1\Brainnest_projects>
```

3. Descriptive Statistics



	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors
count	21588.000	21597.000	21596.000	21587.000	21581.000	21597.000
mean	540382.754	3.375	2.116	2080.185	15113.467	1.495
std	367405.169	0.959	0.769	917.968	41511.996	0.544
min	78000.000	1.000	0.500	370.000	520.000	1.000
25%	322000.000	3.000	1.750	1430.000	5040.000	1.000
50%	450000.000	3.000	2.250	1910.000	7617.000	1.500
75%	645000.000	4.000	2.500	2550.000	10677.000	2.000
max	7700000.000	40.000	8.000	13540.000	1651359.000	11.000

	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated
count	21597.000	21594.000	21594.000	21592.000	21585.000	21595.000	21591.000	21597.000
mean	0.008	0.234	3.410	7.658	1789.258	291.710	1970.998	84.465
std	0.087	0.766	0.651	1.173	834.846	442.664	29.377	401.821
min	0.000	0.000	1.000	3.000	370.000	0.000	1900.000	0.000
25%	0.000	0.000	3.000	7.000	1190.000	0.000	1951.000	0.000
50%	0.000	0.000	3.000	7.000	1560.000	0.000	1975.000	0.000
75%	0.000	0.000	4.000	8.000	2210.000	560.000	1997.000	0.000
max	1.000	4.000	5.000	13.000	18000.000	4820.000	2015.000	2015.000

	zipcode	lat	long	sqft_living15	sqft_lot15
count	21590.000	21596.000	21592.000	21590.000	21589.000
mean	98077.950	47.560	-122.214	1986.586	12761.042
std	53.513	0.139	0.141	685.296	27279.087
min	98001.000	47.156	-122.519	399.000	651.000
25%	98033.000	47.471	-122.328	1490.000	5100.000
50%	98065.000	47.572	-122.231	1840.000	7620.000
75%	98118.000	47.678	-122.125	2360.000	10083.000
max	98199.000	47.778	-121.315	6210.000	871200.000

4. Project Questions

- What are the ideal locations (Zip codes) to invest in considering profit margins on resale in mind?
- Does the ratio of bedroom to lot area have any impact on price? If yes, what is the suggested ratio?
- Does the ratio of bedroom to floors have any impact on price? If yes, what is the suggested ratio?
- Does the ratio of bedroom to bathrooms have any impact on price? If yes, what is the suggested ratio?

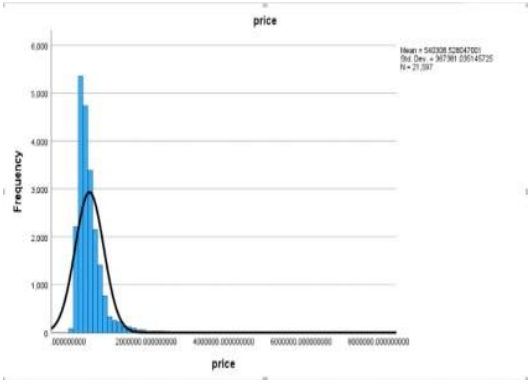
Project week 2

Skewness and Kurtosis

	Skewness			Kurtosis		
variable name	Statistic	Std. Error	Stat/std. error	Statistic	Std. Error	Stat/std. error
id	0.243	0.017	14.29411765	-1.261	0.033	-38.21212121
price	4.023	0.017	236.6470588	34.536	0.033	1046.545455
bedrooms	0.549	0.017	32.29411765	1.797	0.033	54.45454545
bathrooms	0.52	0.017	30.58823529	1.279	0.033	38.75757576
sqft_living	1.471	0.017	86.52941176	5.253	0.033	159.1818182
sqft_lot	13.075	0.017	769.1176471	285.571	0.033	8653.666667
floors	0.615	0.017	36.17647059	-0.491	0.033	-14.87878788
waterfront	11.275	0.017	663.2352941	125.139	0.033	3792.090909
view	3.395	0.017	199.7058824	10.89	0.033	330
condition	1.028	0.017	60.47058824	0.536	0.033	16.24242424
grade	0.76	0.017	44.70588235	1.245	0.033	37.72727273
sqft_above	1.745	0.017	102.6470588	9.77	0.033	296.0606061
sqft_basement	1.577	0.017	92.76470588	2.713	0.033	82.21212121
yr_built	-20.444	0.017	-1202.588235	1062.416	0.033	32194.42424
yr_renovated	4.548	0.017	267.5294118	18.684	0.033	566.1818182
zipcode	-84.549	0.017	-4973.470588	7162.595	0.033	217048.3333
lat	-77.329	0.017	-4548.764706	6358.629	0.033	192685.7273
long	83.655	0.017	4920.882353	7062.384	0.033	214011.6364
sqft_living15	1.103	0.017	64.88235294	1.595	0.033	48.33333333
sqft_lot15	9.525	0.017	560.2941176	151.406	0.033	4588.060606

Price

Histogram

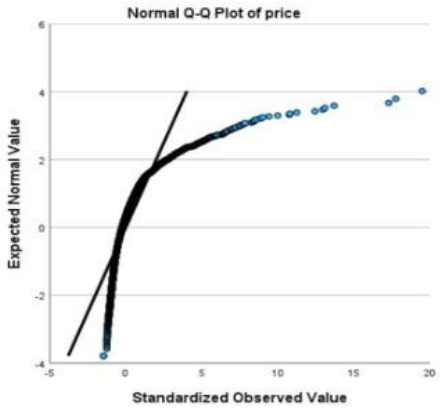


- We can see that the tail is skewing more to the right
- The distribution isn't symmetric
- The median is going also to the right.
- If Median=M, Mean=x, and Mode=m, then $x \neq M \neq m$

Price

We can see that the distribution is not normal, many of the values are far away from the straight line.

Q-Q Plot

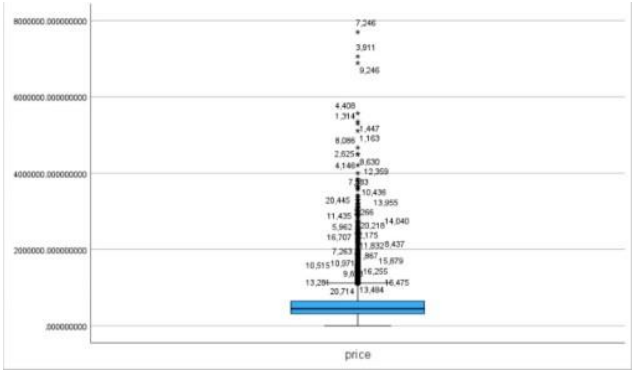


P-P Plot

Price

Box Plot

- We can see that the whiskers of the box plot are almost equals but the boxes aren't.
- The distribution is skewed up (positive)
- Non symmetric distribution
- There are outliers. There are two different types. The stars are the Extreme outliers, and the circles are the mild outliers



Price

Visual and statistical tests like Histogram, Box Plots, Q-Q plots and P-P plots, showed that the price is not normally distributed with skewness of-4.023(SE= 0.017) and kurtosis of-34.536(SE= 0.033).

Test	Normally Distributed
Histogram	No
QQ Plot	N0
PP Plot	No
Box Plot	No
Skewness	No
Z score (Skewness)	No (236.64)
Kurtosis	No
Z score (Kurtosis)	No (1046)
Final conclusion	No (8/8)

Bathrooms

Visual and statistical tests like Histogram, Box Plots, Q-Q plots and P-P plots, showed that the Bathrooms is not normally distributed with skewness of 0.52(SE=0.017) and kurtosis of 1.279 (SE=0.033).

Test	Normally Distributed
Histogram	No
QQ Plot	No
PP Plot	Yes
Box Plot	No
Skewness	Close to 0
Z score (Skewness)	30.58
Kurtosis	Not that close to 0
Z score (Kurtosis)	38.75
Final conclusion	Not normal(6/8)

Bedroom

Visual and statistical tests like Histogram, Box Plots, Q-Q plots and P-P plots, showed that the Bedroom is not normally distributed with skewness of 0.549(SE=0.017) and kurtosis of 1.797(SE=0.033).

Test	Normally Distributed
Histogram	No
QQ Plot	No
PP Plot	Yes
Box Plot	No
Skewness	Is close to 0
Z score (Skewness)	(32.29) Big
Kurtosis	1.797
Z score (Kurtosis)	(54.45)
Final conclusion	Non normal (6/8)

Square Feet Living

Visual and statistical tests like Histogram, Box Plots, Q-Q plots and P-P plots, showed that the Square Feet Living is not normally distributed with skewness of 1.471(SE=0.017) and kurtosis of 5.253 (SE=0.033).

Test	Normally Distributed
Histogram	No
QQ Plot	No
PP Plot	No
Box Plot	No
Skewness	No
Z score (Skewness)	No(86.52)
Kurtosis	No
Z score (Kurtosis)	No(159.18)
Final conclusion	No

Square Feet Lot

Visual and statistical tests like Histogram, Box Plots, Q-Q plots and P-P plots, showed that the Square Feet Lot is not normally distributed with skewness of 13.075(SE=0.017) and kurtosis of 285.571 (SE=0.033).

Test	Normally Distributed
Histogram	No
QQ Plot	No
PP Plot	No
Box Plot	No
Skewness	No
Z score (Skewness)	No
Kurtosis	No
Z score (Kurtosis)	No
Final conclusion	No

Floors

Visual and statistical tests like Histogram, Box Plots, Q-Q plots and P-P plots, showed that the Floors is approximately normally distributed with skewness of 0.615 (SE=0.017) and kurtosis of -0.491(SE=0.033).

Test	Normally Distributed
Histogram	No
QQ Plot	Yes
PP Plot	Yes
Box Plot	No
Skewness	Close to zero
Z score (Skewness)	36.1764
Kurtosis	Close to zero
Z score (Kurtosis)	-14.8787
Conclusion	Approximatelynormally distributed (4/8)

Square Feet Above

Visual and statistical tests like Histogram, Box Plots, Q-Q plots and P-P plots, showed that the Square Feet Above variable is not normally distributed with skewness of -1.745 (SE=0.017) and kurtosis of 9.77 (SE=0.033).

Test	Normally Distributed
Histogram	No
QQ Plot	No
PP Plot	No
Box Plot	No
Skewness	Not close to 0
Z score (Skewness)	No (102.64)
Kurtosis	Not close to 0
Z score (Kurtosis)	No (296.06)
Final conclusion	No (8/8)

Square Feet Basement

Visual and statistical tests like Histogram, Box Plots, Q-Q plots and P-P plots, showed that the Square Feet Basement is not normally distributed with skewness of 1.577(SE=0.017) and kurtosis of 2.713(SE=0.033).

Test	Normally Distributed
Histogram	No
QQ Plot	No
PP Plot	No
Box Plot	No
Skewness	Not close to 0
Z score (Skewness)	No (92.76)
Kurtosis	Not close to 0
Z score (Kurtosis)	No (82.21)
Final conclusion	No (8/8)

Year Built

Visual and statistical tests like Histogram, Box Plots, Q-Q plots and P-P plots, showed that the year built is not normally distributed with skewness of -20.44 (SE=0.017) and kurtosis of 1062 (SE=0.033).

Test	Normally Distributed
Histogram	No
QQ Plot	Yes
PP Plot	Yes
Box Plot	No
Skewness	No
Z score (Skewness)	No
Kurtosis	No
Z score (Kurtosis)	No
Final conclusion	NO

Year Renovated

Visual and statistical tests like Histogram, Box Plots, Q-Q plots and P-P plots, showed that the year renovated is or not normally distributed with skewness of 4.548 (SE=0.017) and kurtosis of 18.684 (SE=0.033).

Test	Normally Distributed
Histogram	No
QQ Plot	No
PP Plot	No
Box Plot	No
Skewness	No
Z score (Skewness)	No (4.548)
Kurtosis	No
Z score (Kurtosis)	No (18.684)
Final conclusion	No 8/8

Sqft_living15

Visual and statistical tests like Histogram, Box Plots, Q-Q plots and P-P plots, showed that the {Column Name} is {approximately normally distributed or not normally distributed} with skewness of 1.103 (SE = 0.017) and kurtosis of 1.595 (SE=0.033).

Test	Normally Distributed
Histogram	No
QQ Plot	No
PP Plot	No
Box Plot	No
Skewness	No
Z score (Skewness)	Yes (1.103)
Kurtosis	No
Z score (Kurtosis)	Yes (1.595)
Final conclusion	No 6/8

Sqft_lot15_1

Visual and statistical tests like Histogram, Box Plots, Q-Q plots and P-P plots, showed that the {Column Name} is {approximately normally distributed or not normally distributed} with skewness of 9.525 (SE=0.017) and kurtosis of 151.406 (SE= 0.033).

Test	Normally Distributed
Histogram	No
QQ Plot	No
PP Plot	No
Box Plot	No
Skewness	No
Z score (Skewness)	No (9.525)
Kurtosis	No
Z score (Kurtosis)	No (151.406)
Final conclusion	No (8/8)

Overall Picture

Variable	Remarks	Value
Id	Unique identifiers no need to test normality	
Date	Considered them to be categorical so no need to test normality	
Price	Not Normal	100%
Bedrooms	Not Normal	75%
Bathrooms	Not Normal	75%
sqft_living	Not Normal	100%
sqft_lot	Not Normal	100%
Floors	Approximately normal	50%
Waterfront	Categorical variable. Hence, no need to test normality.	

Overall Picture Continued…

Variable	Remarks	Value
View	Categorical variable. Hence, no need to test normality.	
Condition	Categorical variable. Hence, no need to test normality.	
Grade	Categorical variable. Hence, no need to test normality.	
sqft_above	Not Normal	100%
sqft_basement	Not Normal	100%
yr_built	Not Normal	100%
yr_renovated	Not Normal	100%
Zipcode	Categorical variable. Hence, no need to test normality.	
Lat	Unique identifiers no need to test normality	

Overall Picture Continued…

Variable	Remarks	Value
Long	Unique identifiers no need to test normality	
sqft_living15	Not Normal	75%
sqft_lot15	Not Normal	100%

- For Floors we are 50% sure it is Normally distributed
- For other variables we are at least 75% sure they are not normally distributed.
- So we suggest going for non-parametric tests in our future analysis.

Final Presentation

Project Objectives

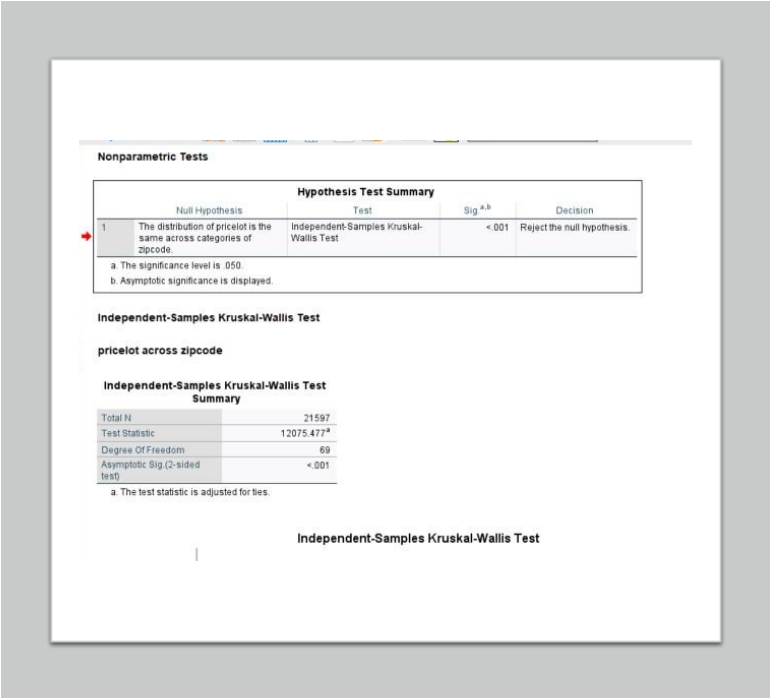
PO.	Objective
1	What are the ideal locations (Zip codes) to invest in considering profi margins on resale in mind?
2	Does the ratio of bedroom to lot area have any impact on price? If ye what is the suggested ratio?
3	Does the ratio of bedroom to floors have any impact on price? If yes what is the suggested ratio?
4	Does the ratio of bedroom to bathrooms have any impact on price? yes, what is the suggested ratio?

Hypothesis Testing

PO.	H0	H1	Applied Test	Result
1	The cost of land per unit area is same across all neighborhoods (i.e., zip codes).	The cost of land per unit area is affected by the neighborhood.	Kruskal-Wallis Test	Accept H1
2	The price is not affected by different ratios of lot area to bedroom.	The price is affected by lot area to bedroom ratio.		Accept H1
3	The price is not affected by different ratios of bedroom to floors.	The price is affected by bedroom to floor ratio.		Accept H1
4	The price is not affected by different ratios of bedroom to bathrooms.	The price is affected by bedroom to bathroom ratio.		Accept H1

Hypothesis Test for PO1

- The p-value is less than 0.05, hence we reject null hypotheses and conclude that:
- Cost of land per square foot is indeed affected by the location (i.e., Zip codes).



Nonparametric Tests

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of pricelot is the same across categories of zipcode.	Independent-Samples Kruskal-Wallis Test	<.001	Reject the null hypothesis.

a. The significance level is .050.
b. Asymptotic significance is displayed.

Independent-Samples Kruskal-Wallis Test

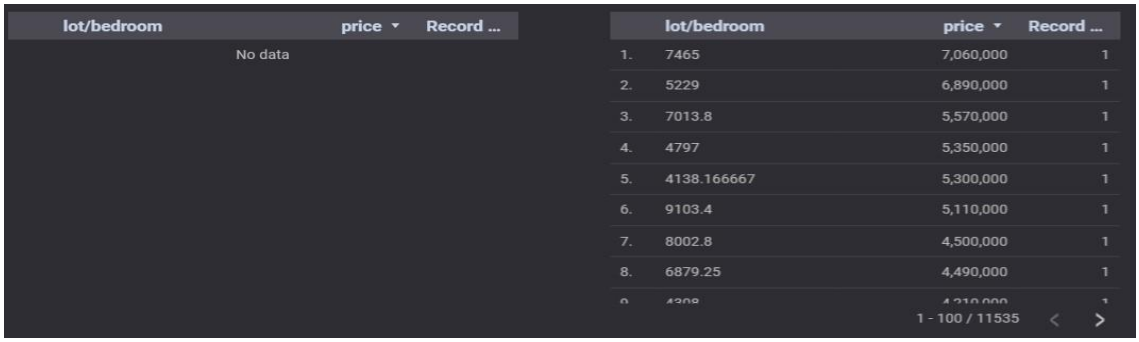
pricelot across zipcode

Independent-Samples Kruskal-Wallis Test Summary

Total N	21597
Test Statistic	12075.477[a]
Degree Of Freedom	69
Asymptotic Sig.(2-sided test)	<.001

a. The test statistic is adjusted for ties.

Independent-Samples Kruskal-Wallis Test



lot/bedroom	price	Record ...
No data		

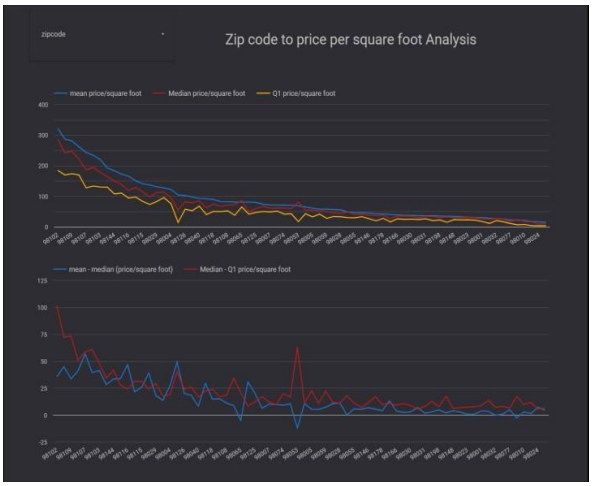
	lot/bedroom	price	Record ...
1.	7465	7,060,000	1
2.	5229	6,890,000	1
3.	7013.8	5,570,000	1
4.	4797	5,350,000	1
5.	4138.166667	5,300,000	1
6.	9103.4	5,110,000	1
7.	8002.8	4,500,000	1
8.	6879.25	4,490,000	1

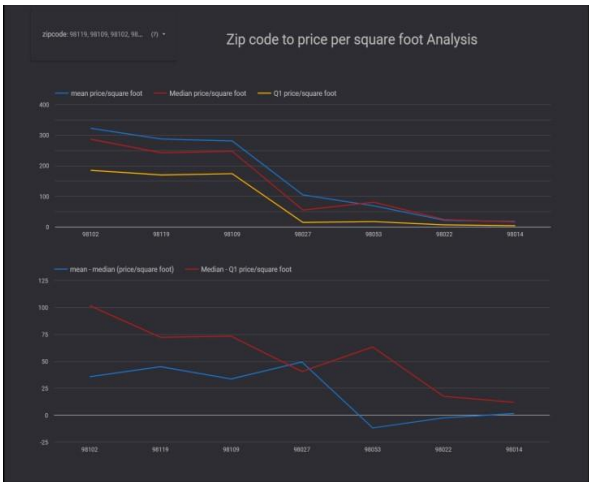
1 - 100 / 11535

PO2
PO3
PO4

- Results from hypotheses testing did indicate possible existence of patterns.
- The group sizes within respective samples did not meet the minimum threshold of at least 2.5% representation of the total sample size.
- Thus, detecting any pattern from such small groups is meaningless in our opinion.

PO1: What are the ideal locations (Zip codes) to invest in considering profit margins on resale in mind?





zipcode: 98027, 98053, 98022, 98... (7)

Zip code to price per square foot Analysis

	zipcode	mean price/square fo...	Median price/square foot	Q1 price/square foot	Median - Q1 price/square fo...	profit% (price/square feet) (selling at me...
1.	98102	322.56	286.86	185.12	101.74	35.47
2.	98119	287.43	242.41	170.08	72.33	29.84
3.	98109	281.08	247.39	174.08	73.31	29.63
4.	98027	104.82	55.39	14.86	40.53	73.17
5.	98053	69.19	80.97	17.61	63.36	78.25
6.	98022	21.99	24.45	6.92	17.53	71.71
7.	98014	17.77	16.17	4.2	11.97	74.02

1 - 7 / 7

PO1

- We divided our zones of interest into three groups G1, G2, and G3; with G1 requiring minimum investment and G3 the highest for same property size.
- Buy property at a price per square foot that lies within first quartile of its respective zip-code
- Try to sell them at a price equal to or greater than Median to earn desired profit.

PO1: Report/Insights

- For G1 (98022, 98014), if investors can buy property under **$6.9/sqft** and **$4.2/sqft** respectively and flip them for at least **$24.4/sqft** and **$16.2/sqft**, they can expect profit margin between **71-74%**.
- For G2 (98027, 98053), if investors can buy property under **$14.8/sqft** and **$17.6/sqft** respectively and flip them for at least **$55.4/sqft** and **$81/sqft**, they can expect profit margin between **73-78%**.
- For G3 (98102, 98109), if investors can buy property under **$185/sqft** and **$174/sqft** respectively and flip them for at least **$287/sqft** and **$247/sqft**, they can expect profit margin between **29-35%**.

Appendix B: Tools Used

Throughout the project we used following tools:

Python (Data cleaning and EDA)

SPSS (Statistical analysis, and some visualizations)

Google Data Studio (Data Visualization)