STATISTICAL ANALYSIS - 2
Group Assignment

Group 6

Student Name	PG ID
Akshay Ramdev	12220032
Charanjeet Singh	12220064
Pooja Nilesh Doshi	12220028
Snigdha Debashis Bhattacharjee	12220067
Vinayak Dave	12220047

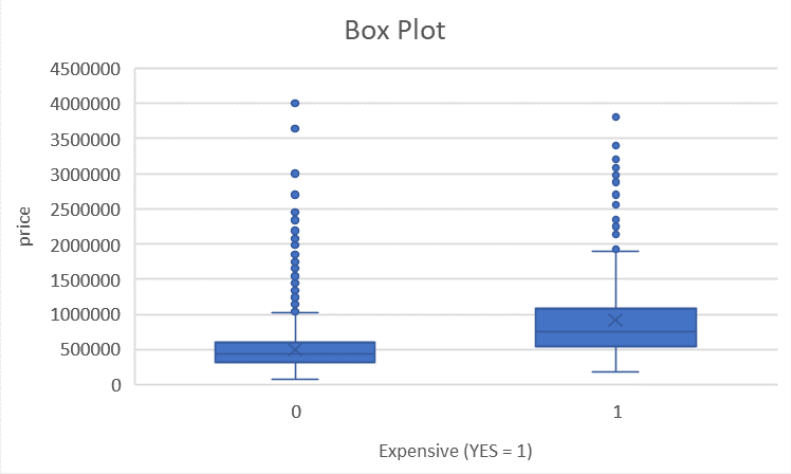
Note:
EDA is done in Python. Please refer to the SA2 Group Assignment _ Group-6.ipynb file. EDA is then continued in Excel.
Please refer to the Statistical Analysis 2_Group Assignment_Excel_Group 6.xlsx file, even for Q1-7.

Q1.

Edits 1:
"sqft_basement" indicate whether there is a basement or not. If there is a value present, then 1, otherwise 0. New variable "basement_present".

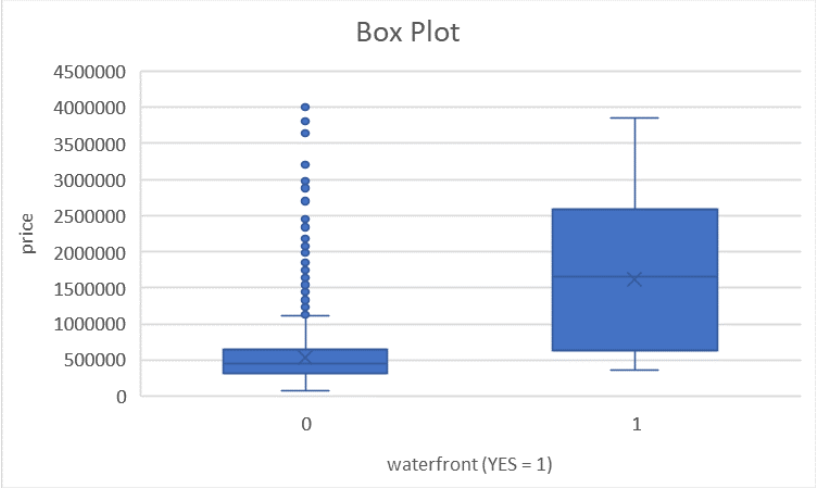
"yr_renovated" indicates whether the house has been renovated or not. 1 if the year of renovation is given, otherwise 0. New variable "renovated".

An additional column called "age" has been introduced, calculated as the difference between "yr_built" and "yr_sold." This new column provides information about the age of each house by subtracting the year it was built from the year it was sold. The dataset contains several houses with IDs 595000, 455000, 385195, 455000, and 597326, which originally had a calculated age of -1. However, an age of -1 is not a valid or possible value for a house. Therefore, to address this issue, the age values for these houses have been corrected and edited to 0 years. "yr_built" and "yr_sold" can be dropped.



The categorical variable "expensive" has been converted into a binary or dummy variable. The values have been recoded to 0 and 1. The value 0 represents "NO," indicating that the house is not considered expensive, while the value 1 represents "YES," indicating that the house is categorized as expensive. Also, as average price is different for each of the two "Expensive" categories, the "Expensive" variable will be

included in the model as a dummy variable. To ensure average is not different due to outliers, we also plotted the box plots to check for difference in median.



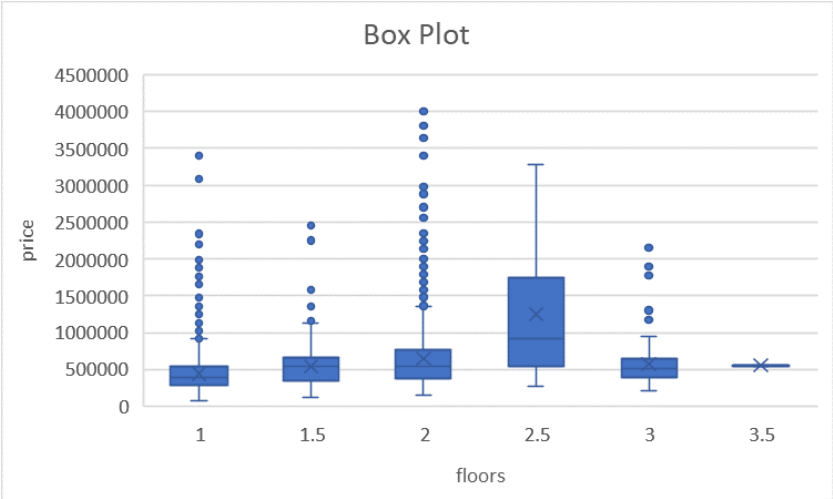
Similarly, the categorical variable "waterfront" has been transformed into a binary or dummy variable. It now represents whether a house has waterfront access or not. The value 0 corresponds to "NO," indicating that the house does not have a waterfront, while the value 1 corresponds to "YES," indicating that the house is accompanied by a waterfront. Also, average Price is different for each of the two 'Waterfront' categories, the 'Waterfront' variable will be included in the model as a dummy variable.

Dependent Variable "price" and Independent Variables "sqft_living", "sqft_above" and "sqft_basement".
The correlation matrix is as follows:

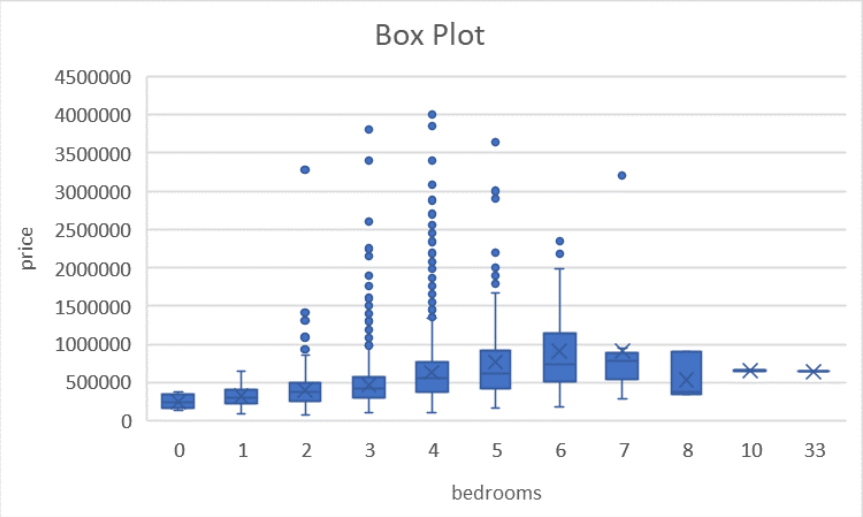
	price	sqft_living	sqft_above	sqft_basement
price	1.000			
sqft_living	0.698	1.000		
sqft_above	0.607	0.881	1.000	
sqft_basement	0.323	0.441	-0.036	1.000

First, we remove multi-collinearity, for which we identify highly correlated variables and consider removing one of them from the analysis. In this case, we observe a high correlation between “sqft_living” and “sqft_above” (correlation coefficient: 0.881).
Secondly, it is also determined that, sqft_living = sqft_above + sqft_basement, which ascertains “sqft_living” as the interaction variable.
Therefore, we drop the column “sqft_above”.

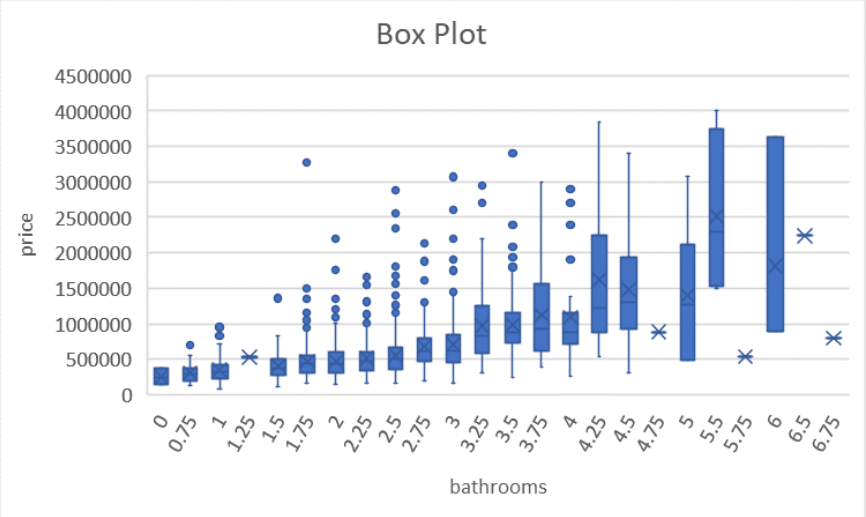
Edits 2:



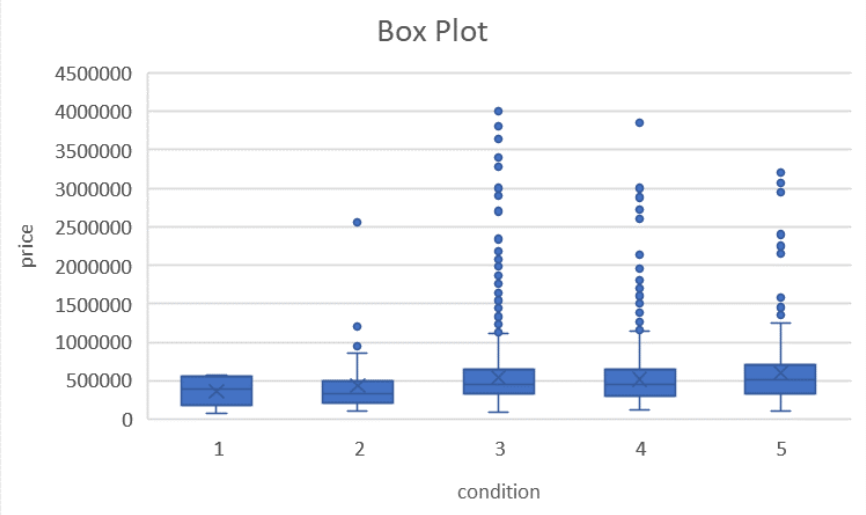
“floors” are grouped into “1-2 floor” which contain, 1.5 and 2 floors, “2.5floors”, and “3 and more floors” which contain 3 and 3.5 floors. Dummy variables introduced.



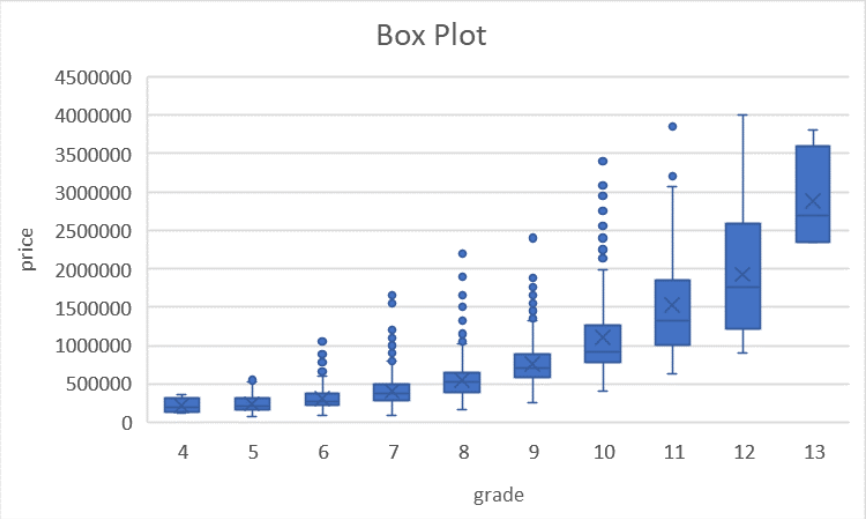
“bedrooms” range from 0 to 33. They have been grouped into ranges, “0-3 bedrooms”, “4-6 bedrooms”, “7 and more bedrooms”, and dummy variables introduced.



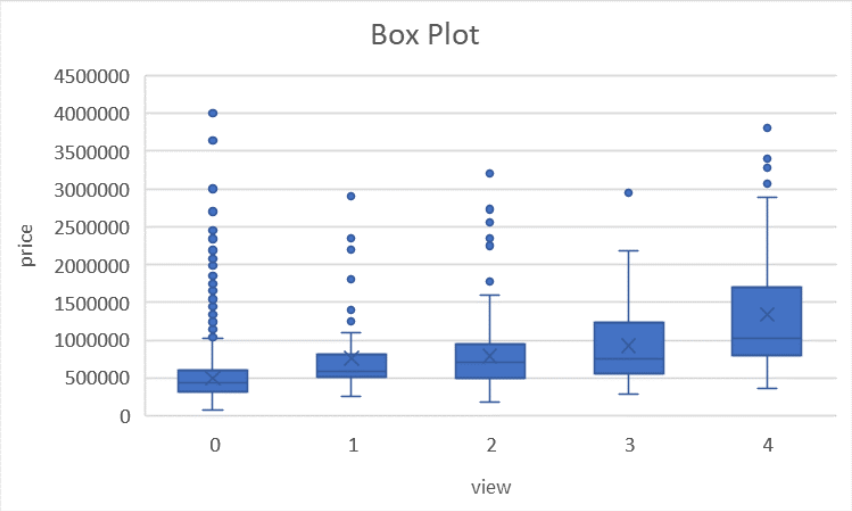
Similarly, “bathrooms” range from 0 to 6.75. They have been grouped into ranges, “0-3 bathrooms”, “3.25-4.5 bathrooms” and “4.75 and more bathrooms”. Dummy variables introduced.



As average price is different for each of the different “condition” categories, the variable will be included in the model as dummy variables. based on Box Plot, “condition” is grouped into “bad condition” which is 1-2, “ok condition” which is 3 and “good condition” which is 4-5.



Similarly, “grade” is grouped into “grade worse” which is 1-3, “grade bad” which is 4-6 and “grade ok” which is 7-9, and “grade good” which is 10 and above, dummy variables introduced. In our data there is no data for “grade worse” so this column is removed. Also, “grade” appears to have impact on “price”.



“view” is grouped into “view 0, 1 & 2”, and “view 3&4”.

From the Correlation Matrix:
“price” and “sqft_living” exhibit a strong positive correlation of 0.698. Given the high correlation, we will include “sqft_living” as a predictor in the model to capture its impact.

“bedrooms” (0.295) and “bathrooms” (0.527) are positively correlated, indicating that properties with more bedrooms tend to have more bathrooms as well. Both variables will be included in the model to capture their individual contributions to the price.

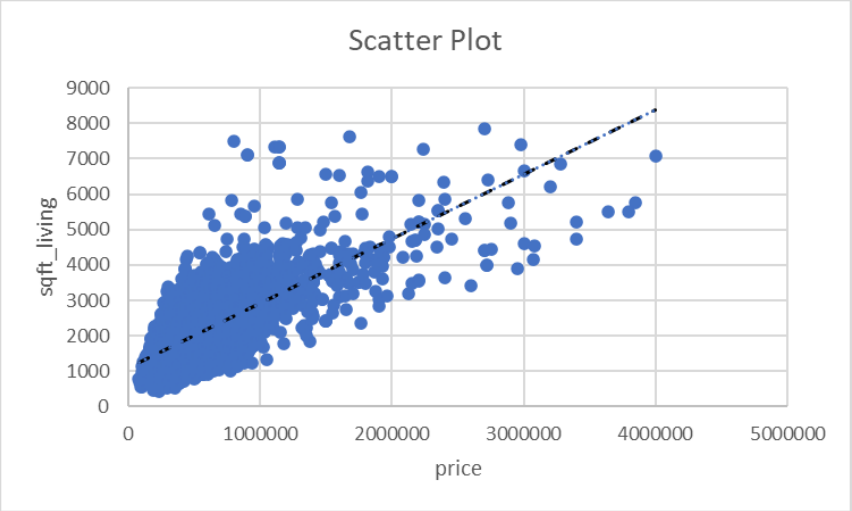
“Distance” (-0.275) exhibits a slight negative correlation with price. Despite the relatively weak association, we will still include this variable in the model to account for its potential impact on price.

We will not consider “zipcode” (-0.059), “lat” (0.313) and “long” (0.036) in our model owing to low correlation. Even though “lat” is higher, as a group as a whole we will not consider these variables.

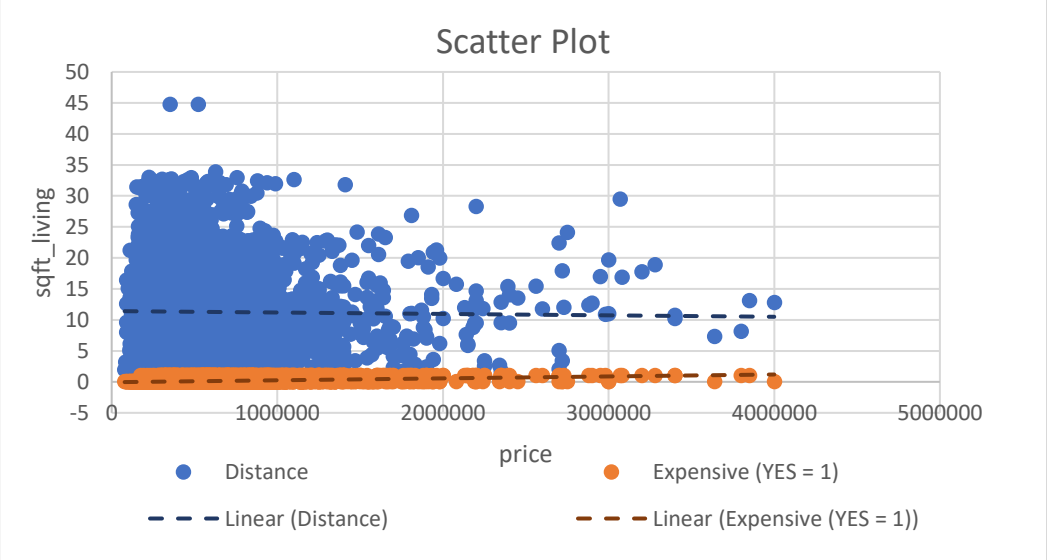
We will also not consider “sqft_lot” (0.088), “renovated” (0.089), “condition” (0.026) and “sqft_lot15” (0.085) since correlation is low.

Since “view” and “Expensive” are highly correlated (0.929), we will eventually drop one of them to take care of multi-collinearity. “sqft_living” and “sqft_living15” also have a high correlation (0.754), so we will not consider “sqft_living” in our model. Similarly, “sqft_living” and “grade” have a similar correlation (0.698 and 0.682) with “price” and together they have a correlation as well (0.763). Therefore we will retain “sqft_living” and drop “grade.
The “age” (-0.068) of the house displays a slight negative correlation with price. Since the relationship is not particularly strong, we will not include the “age” variable in the model to consider its potential influence on the price.

Q2.

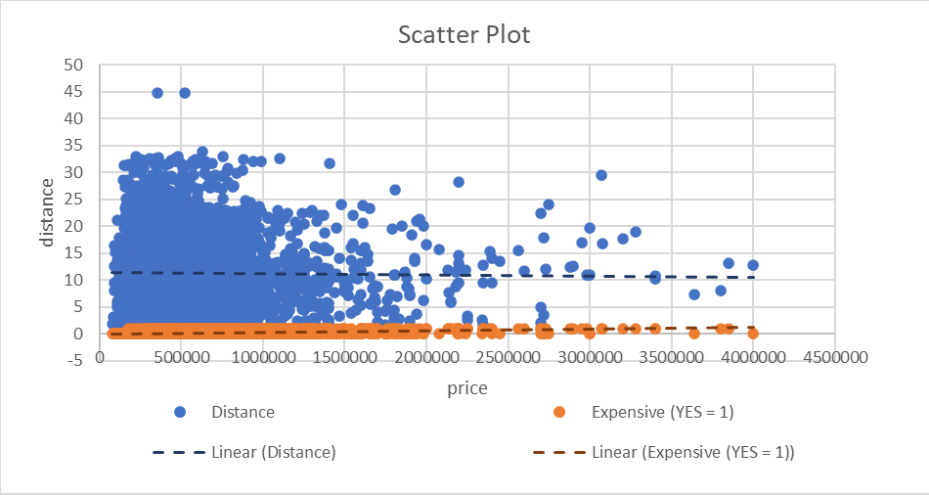


From the correlation matrix as well as Scatter Plot we can ascertain, that “price” and “sqft_living” exhibit a strong positive correlation of 0.698. However, it is worth noting that there are instances where properties with high “sqft_living” have relatively low prices. This observation suggests that “sqft_living” might be interacting with another variable, influencing the price. Given the high correlation between “price” and “sqft_living”, we will include “sqft_living” as a predictor in the model to capture its impact.



Interaction of "Expensive" & "Sqft_living": As the slopes of the fitted line are different, these variables have interaction effect.

Interaction Effect between “Distance” and “Expensive”:



There seem to be few points of “Expensive” that are pricier than Expensive-No which reflects a slight interaction effect.

Dependent Variable "price" and Independent Variables "condition" and "grade".
We create an interaction variable, by combining variables "condition" and "grade";
Interaction Variable: condition*grade, since higher-grade house in better condition is generally more desirable and may command a higher price.

The correlation matrix is as follows:

	price	condition	grade	condition*grade
price	1.000			
condition	0.026	1.000		
grade	0.682	-0.142	1.000	
condition*grade	0.495	0.733	0.557	1.000

The correlation coefficient between "condition*grade" and "price" is 0.495, indicating a moderate positive correlation between the interaction variable and the price of the house.
The correlation coefficient between "condition*grade" and "condition" is 0.733, indicating a strong positive correlation between the interaction variable and the condition of the house, therefore, we drop the column “condition” to avoid multi-collinearity.
The correlation coefficient between "condition*grade" and "grade" is 0.557, indicating a moderate positive correlation between the interaction variable and the grade of the house. Therefore, we drop the column “grade” to avoid multi-collinearity.
Based on these correlation coefficients, we can infer that the interaction variable "condition*grade" has some influence on the price of the house and is correlated with both the condition and grade variables.

Independent Variables "sqft_living" and "grade".
The interaction between square footage of living area and grade captures the combined effect of these two variables on the house price. Houses with larger living areas and higher grades (indicating better quality) tend to command higher prices.

Q3.

From Q1 and Q2, the following additional variables are created for our final dataset.

The following Dummy Variables have been introduced for the following:
For “floors”: “1-2 floor”, “2.5 floors”, and “3 and more floors
For “bedrooms”: “0-3 bedrooms”, “4-6 bedrooms”, and “7 and more bedrooms.

For “bathrooms”: “0-3 bathrooms”, “3.25-4.5 bathrooms” and “4.75 and more bathrooms
For “view”: “view 0”, “view 1&2”, “view 3”, and “view 4”.

Additional variables introduced are:
sqft_living * Expensive showing interaction effect.

Q4.

Regression with “price” as the Dependent Variable, and all others as Independent Variables. We will follow the Backward Stepwise calculation method.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.821
R Square	0.674
Adjusted R Square	0.673
Standard Error	2,03,065.198
Observations	5,000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	16	42,55,80,38,77,75,269.000	2,65,98,77,42,35,954.300	860.061	0.000
Residual	4,987	20,56,41,31,18,92,198.000	41,23,54,74,612.432		
Total	5,003	63,12,21,69,96,67,467.000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	6,16,126.807	70,016.326	8.800	0.000	4,78,864.016	7,53,389.599
sqft_living	252.184	4.712	53.514	0.000	242.946	261.423
sqft_living * Expensive	45.650	4.379	10.424	0.000	37.064	54.236
bedrooms 0-3	0.000	0.000	65,535.000	#NUM!	0.000	0.000
bedrooms 4-6	-44,653.182	7,050.691	-6.333	#NUM!	-58,475.638	-30,830.726
bedrooms 7 and more	-3,16,016.210	49,678.649	-6.361	0.000	-4,13,408.210	-2,18,624.211
bathrooms 0-3	-1,07,667.761	51,586.301	-2.087	0.037	-2,08,799.598	-6,535.923
bathrooms 3.25-4.5	-1,693.096	50,825.608	-0.033	0.973	-1,01,333.640	97,947.449
bathrooms 4.75 and more	0.000	0.000	65,535.000	#NUM!	0.000	0.000
floors 1-2	-2,37,062.513	36,444.582	-6.505	#NUM!	-3,08,509.922	-1,65,615.105
floors 2.5	0.000	0.000	65,535.000	#NUM!	0.000	0.000
floors 3 and more	-2,58,678.900	40,330.702	-6.414	#NUM!	-3,37,744.812	-1,79,612.987
basement_present (YES = 1)	-58,558.617	6,421.170	-9.120	0.000	-71,146.935	-45,970.299
waterfront (YES = 1)	6,43,217.489	36,297.944	17.720	0.000	5,72,057.556	7,14,377.422
view 0, 1&2	-33,316.140	19,078.923	-1.746	0.081	-70,719.220	4,086.940
view 3&4	0.000	0.000	65,535.000	#NUM!	0.000	0.000
Distance	-18,588.334	476.865	-38.980	#NUM!	-19,523.199	-17,653.468

The highlighted variables “bedrooms 0-3”, “bathrooms 4.75 and more”, “floors 2.5”, and “view 3&4” have coefficients as 0 and very hight T-Statistic. So we will remove these variables from our dataset.

Additional variables we will re-introduce are:
“age” = “yr_sold” – “yr_built”

We run the Regression again with “price” as the Dependent Variable, and all others as Independent Variables.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.822
R Square	0.675
Adjusted R Square	0.674
Standard Error	2,02,736.549
Observations	5,000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	13	42,62,86,58,67,53,972.000	3,27,91,27,59,04,151.700	797.800	0.000
Residual	4,986	20,49,35,11,29,13,495.000	41,10,21,08,486.461		
Total	4,999	63,12,21,69,96,67,467.000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5,65,350.243	70,968.234	7.966	0.000	4,26,221.287	7,04,479.199
sqft_living	258.288	4.930	52.392	0.000	248.623	267.952
sqft_living * Expensive	43.682	4.398	9.932	0.000	35.060	52.304
bedrooms 4-6	-46,915.327	7,060.404	-6.645	0.000	-60,756.824	-33,073.829
bedrooms 7 and more	-3,28,021.273	49,682.736	-6.602	0.000	-4,25,421.289	-2,30,621.256
bathrooms 0-3	-1,07,336.232	51,502.874	-2.084	0.037	-2,08,304.521	-6,367.944
bathrooms 3.25-4.5	3,003.249	50,755.997	0.059	0.953	-96,500.832	1,02,507.331
floors 1-2	-2,29,359.543	36,433.024	-6.295	0.000	-3,00,784.296	-1,57,934.789
floors 3 and more	-2,29,578.983	40,872.857	-5.617	0.000	-3,09,707.762	-1,49,450.204
basement_present (YES = 1)	-60,546.548	6,428.692	-9.418	0.000	-73,149.612	-47,943.484
waterfront (YES = 1)	6,38,940.111	36,253.887	17.624	0.000	5,67,866.545	7,10,013.676
view 0, 1&2	-31,528.928	19,052.924	-1.655	0.098	-68,881.041	5,823.184
Distance	-17,821.710	510.755	-34.893	0.000	-18,823.015	-16,820.405
age	487.853	117.695	4.145	0.000	257.119	718.587

“bathrooms 3.25-4.5” has a high p-value, so it is not significant. Running the regression model without this variable

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.822
R Square	0.675
Adjusted R Square	0.675

Standard Error	2,02,716.293
Observations	5,000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	12	42,62,86,44,28,50,232.000	3,55,23,87,02,37,519.400	864.456	0.000
Residual	4,987	20,49,35,25,68,17,235.000	41,09,38,95,491.726		
Total	4,999	63,12,21,69,96,67,467.000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5,68,522.212	46,500.571	12.226	0.000	4,77,360.643	6,59,683.780
sqft_living	258.245	4.877	52.948	0.000	248.684	267.807
sqft_living * Expensive	43.668	4.391	9.945	0.000	35.059	52.277
bedrooms 4-6	-46,887.940	7,044.512	-6.656	0.000	-60,698.282	-33,077.597
bedrooms 7 and more	-3,28,573.266	48,794.163	-6.734	0.000	-4,24,231.283	-2,32,915.248
bathrooms 0-3	-1,10,294.582	12,361.102	-8.923	0.000	-1,34,527.778	-86,061.387
floors 1-2	-2,29,462.393	36,387.900	-6.306	0.000	-3,00,798.680	-1,58,126.105
floors 3 and more	-2,29,679.993	40,833.111	-5.625	0.000	-3,09,730.849	-1,49,629.137
basement_present (YES = 1)	-60,537.838	6,426.364	-9.420	0.000	-73,136.338	-47,939.338
waterfront (YES = 1)	6,38,924.290	36,249.279	17.626	0.000	5,67,859.762	7,09,988.818
view 0, 1&2	-31,563.782	19,041.914	-1.658	0.097	-68,894.308	5,766.744
Distance	-17,821.610	510.701	-34.896	0.000	-18,822.809	-16,820.410
age	487.698	117.654	4.145	0.000	257.045	718.351

"view 0, 1 & 2" has a high p-value, so it is not significant. Running the regression model without this variable.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.822
R Square	0.675
Adjusted R Square	0.674
Standard Error	2,02,751.802
Observations	5,000

ANOVA

	df	***SS***	***MS***	***F***	***Significance F***
Regression	11	42,61,73,53,23,87,141.000	3,87,43,04,83,98,831.000	942.463	0.000
Residual	4,988	20,50,48,16,72,80,327.000	41,10,82,93,360.130		
Total	4,999	63,12,21,69,96,67,467.000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5,35,454.975	42,011.447	12.745	0.000	4,53,094.067	6,17,815.884
sqft_living	258.109	4.878	52.918	0.000	248.547	267.672
sqft_living * Expensive	47.910	3.569	13.423	0.000	40.912	54.907
bedrooms 4-6	-46,876.146	7,045.743	-6.653	0.000	-60,688.900	-33,063.392
bedrooms 7 and more	-3,32,280.581	48,751.415	-6.816	0.000	-4,27,854.790	-2,36,706.371
bathrooms 0-3	-1,09,847.019	12,360.317	-8.887	0.000	-1,34,078.675	-85,615.362
floors 1-2	-2,28,488.716	36,389.532	-6.279	0.000	-2,99,828.199	-1,57,149.234
floors 3 and more	-2,28,199.118	40,830.488	-5.589	0.000	-3,08,244.827	-1,48,153.409
basement_present (YES = 1)	-60,297.090	6,425.848	-9.384	0.000	-72,894.578	-47,699.602
waterfront (YES = 1)	6,56,639.754	34,644.032	18.954	0.000	5,88,722.218	7,24,557.290
Distance	-17,804.175	510.683	-34.863	0.000	-18,805.337	-16,803.012
age	491.980	117.646	4.182	0.000	261.342	722.618

Based on the above, the P-Value of the model is 0.000 (Significance F) which is less than 0.05, and therefore the model is statistically significant. Also the P-Values of all variables are also 0.000 or less that 0.05. This is the final regression model.

Regression Equation

price = **β0 + β1sqft_living + β2sqft_living * Expensive + β3bedrooms 4-6 + β4bedrooms 7 and more + β5bathrooms 0-3 + β6floors 1-2 + β7floors 3 and more + β9basement_present + β10waterfront + β11Distance + β12age +ε**

price = **5,35,454.975 + 258.109sqft_living + 47.910sqft_living*Expensive - 46,876.146bedrooms 4-6 - 3,32,280.581bedrooms 7 and more - 1,09,847.019bathrooms 0-3 - 2,28,488.716floors 1-2 - 2,28,199.118floors 3 and more - 60,297.090basement_present + 6,56,639.754waterfront - 17,804.175Distance + 491.980age +ε**

Q5.

At a significance of 0.10, we will rerun the models with the original variables again. We will follow the same Backward Stepwise calculation method.

(i) The highlighted variables “bedrooms 0-3”, “bathrooms 4.75 and more”, and “floors 2.5” have coefficients as 0 and very hight T-Statistic. So we will remove these variables from our dataset. Additional variables we will re-introduce are:
“age” = “yr_sold” – “yr_built”

(ii) “bathrooms 3.25-4.5” has a high p-value, so it is not significant. Running the regression model without this variable

(iii) “view 0, 1 & 2” has a high p-value, so it is not significant. Running the regression model without this variable.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.822
R Square	0.675
Adjusted R Square	0.674
Standard Error	2,02,751.802
Observations	5,000.000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	11.000	42,61,73,53,23,87,141.000	3,87,43,04,83,98,831.000	942.463	0.000
Residual	4,988.000	20,50,48,16,72,80,327.000	41,10,82,93,360.130		
Total	4,999.000	63,12,21,69,96,67,467.000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 90.0%*	*Upper 90.0%*
Intercept	5,35,454.975	42,011.447	12.745	0.000	4,66,339.458	6,04,570.493
sqft_living	258.109	4.878	52.918	0.000	250.085	266.134
sqft_living * Expensive	47.910	3.569	13.423	0.000	42.038	53.781
bedrooms 4-6	-46,876.146	7,045.743	-6.653	0.000	-58,467.514	-35,284.778
bedrooms 7 and more	-3,32,280.581	48,751.415	-6.816	0.000	-4,12,484.419	-2,52,076.742
bathrooms 0-3	-1,09,847.019	12,360.317	-8.887	0.000	-1,30,181.707	-89,512.330
floors 1-2	-2,28,488.716	36,389.532	-6.279	0.000	-2,88,355.288	-1,68,622.144
floors 3 and more	-2,28,199.118	40,830.488	-5.589	0.000	-2,95,371.770	-1,61,026.467
basement_present (YES = 1)	-60,297.090	6,425.848	-9.384	0.000	-70,868.633	-49,725.547
waterfront (YES = 1)	6,56,639.754	34,644.032	18.954	0.000	5,99,644.807	7,13,634.702
Distance	-17,804.175	510.683	-34.863	0.000	-18,644.329	-16,964.021
age	491.980	117.646	4.182	0.000	298.433	685.526

Based on the above, the P-Value of the model is 0.000 (Significance F) which is less than 0.05, and therefore the model is statistically significant. Also the P-Values of all variables are also 0.000 or less that 0.05. This is the final regression model.

	At Alpha = 0.05	**At Alpha = 0.10**
Multiple R	0.822	0.822
R^2	0.675	0.675
Adjusted R^2	0.674	0.674

Q.6

The summary of the results are as below :

Regression Statistics	
Multiple R	0.41974243
R Square	0.176183708
Adjusted R Square	0.175358902
Standard Error	322687.3046
Observations	5000

ANOVA

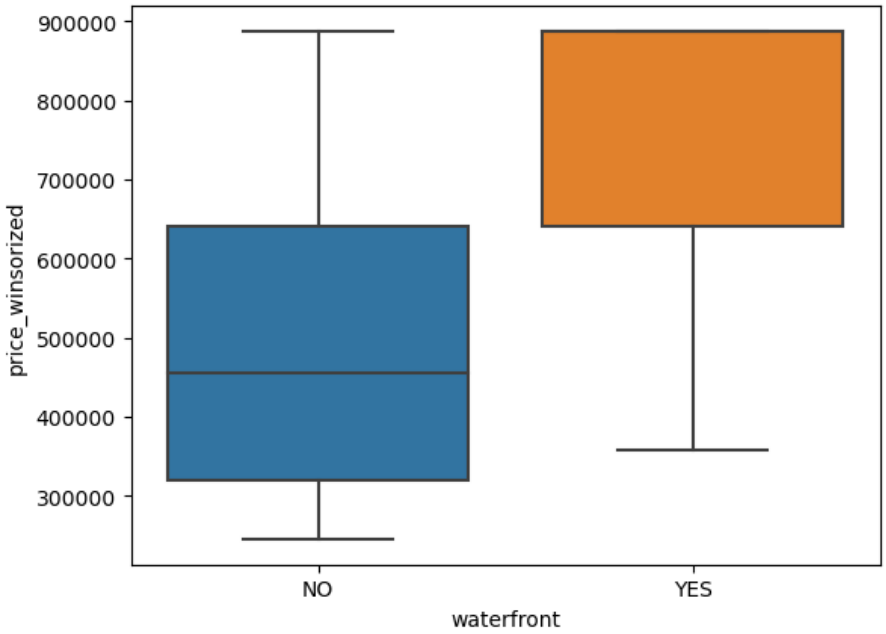
	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	1.11211E+14	2.22422E+13	213.6062241	4.6928E-207
Residual	4994	5.20011E+14	1.04127E+11		
Total	4999	6.31222E+14			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1173428.634	40579.3307	28.9169046	3.8961E-170	1093875.327	1252981.942	1093875.327	1252981.942
view 0	-677911.6059	40864.58569	-16.58922009	3.32082E-60	-758024.1384	-597799.0734	-758024.1384	-597799.0734
view 1	-415783.1979	55796.60298	-7.451765442	1.0794E-13	-525169.0412	-306397.3545	-525169.0412	-306397.3545
view 2	-385838.6367	45972.25526	-8.392858574	6.11074E-17	-475964.4444	-295712.829	-475964.4444	-295712.829
View 3	-258551.8712	48134.50555	-5.371445459	8.16504E-08	-352916.639	-164187.1035	-352916.639	-164187.1035
waterfront (YES = 1)	479637.4769	63742.40549	7.524621533	6.23505E-14	354674.3715	604600.5823	354674.3715	604600.5823

The above results indicate the following :

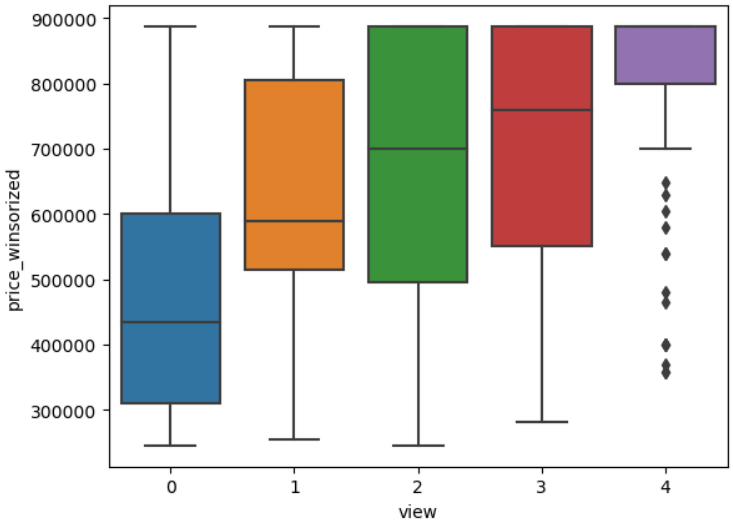
1) The views are significant in estimating the pricing of a plot.
2) However, the presence of 0 views affects the pricing adversely.
3) The waterfront adds to the cost of the property.
4) The two variables, Views and Waterfront are not **ALONE** sufficient to determine the price.
5) However, the 2 variables are significant to estimate the price of a property.

Also, as per the box-plot.



It is clearly visible that waterfront houses are expensive.

After performing the Two Sample Independent T Test also we can conclude that the Prices are dependent on the fact whether the house is Waterfront facing or not.



We can clearly see that the view Category 4 is costlier than the rest. It indicates that the different views have different prices.

Q7.

As per our initial model, there were some Independent Variables which showed significant correlation and therefore high occurrence of multicollinearity:

(i) "sqft_living" and "bedrooms": These variables have a correlation coefficient of 0.542, indicating a fairly positive correlation.

(ii) "Expensive" and "view": These variables have a correlation coefficient of 0.929, indicating a strong positive correlation. This suggests that expensive properties are more likely to have a good view.

(iii) "sqft_living" and "grade": These variables are highly correlated with a coefficient of 0.763. This suggests that the overall quality and construction of a house tend to be related to its living area.

(iv) "sqft_living" and "sqft_living15": These variables have a correlation coefficient of 0.754, indicating a strong positive correlation. This suggests that the living area of a house is also correlated with the average living area of the nearest 15 houses

(v) "sqft_lot" and "sqft_lot15": These variables are highly correlated with a coefficient of 0.748. This indicates that the size of a house's lot is related to the average size of the nearest 15 lots.

In our final data model on which we fitted the regression line, there was significantly lower occurrence of multicollinearity:

(i) "sqft_living" and "sqft_living * Expensive": These variables have a correlation coefficient of 0.435, indicating a moderate positive correlation. This suggests that the interaction term between "sqft_living" and the variable Expensive is moderately correlated with "sqft_living". This correlation might suggest that the effect of "sqft_living" on price differs for expensive and non-expensive properties.

(ii) "sqft_living" and "bathrooms 0-3": These variables have a correlation coefficient of -0.539, indicating a moderate negative correlation. This suggests that the living area and the number of bathrooms are moderately related. This correlation might suggest that larger living areas tend to have more bathrooms.

Note:
The following questions are done in Python. Please refer to the SA2_Q8-10_Group6.ipynb file.

Q8.

We have performed the Breusch-Pagan test to check for Heteroskedasticity. This test examines whether there is a relationship between the squared residuals and the independent variables. The test is based on the intuition that if there is heteroskedasticity, the squared residuals will exhibit a pattern or relationship with the independent variables.

Based on the test statistic and p-value, we can draw conclusions regarding the presence or absence of heteroskedasticity. If the p-value is below a chosen significance level (here 0.05), it indicates significant evidence against homoskedasticity, supporting the presence of heteroskedasticity in the data.

Test-statistic = 5.77644704971242e-200,
which is an extremely small value close to zero. This indicates a significant departure from the assumption of homoskedasticity. The closer the test statistic is to zero, the stronger the evidence against the null hypothesis.

P=Value = 1.1552894099424838e-203,
is an even smaller value close to zero. The p-value represents the probability of observing the given test statistic under the assumption that the null hypothesis is true (i.e., homoskedasticity). In this case, the extremely small p-value suggests that the likelihood of obtaining the observed test statistic by chance, assuming homoskedasticity, is virtually impossible.

Conclusion: Consequently, based on the extremely small test statistic and p-value, we can confidently conclude that there is compelling evidence to reject the null hypothesis of homoskedasticity. Thus, we can ascertain the presence of heteroskedasticity in the data.

Q9.

We have checked for normal distribution of errors by conduction the Jarque Bera Test and the Shapiro-Wilk Test

Jarque Bera Test: This is a statistical test used to assess the normality assumption of a set of data by examining the skewness and kurtosis of the data distribution. It calculates a test statistic and p-value, where a low p-value suggests a departure from normality.
Test-Statistic = 35879.687
P-Value = 0.0
This suggests a significant departure from normality. A higher test statistic and a lower p-value provide stronger evidence against the null hypothesis of normality.

Shapiro-Wilk Test: This is a statistical test used to assess the normality assumption of a dataset. It examines whether the data follows a normal distribution by calculating a test statistic and p-value. A low p-value indicates a significant departure from normality, suggesting that the data does not follow a normal distribution.
Test-Statistic = 0.879
P-Value = 0.0
This also indicates a significant deviation from normality. In this test, a test statistic closer to 1 and a higher p-value would support the normality assumption.

Based on these results, we can conclude that there is substantial evidence to reject the normality assumption about errors. This implies that the errors in the model do not follow a normal distribution.

Q10.

Regression Equation:

price = β0 + β1sqft_living + β2sqft_living * Expensive + β3bedrooms 4-6 + β4bedrooms 7 and more + β5bathrooms 0-3 + β6floors 1-2 + β7floors 3 and more + β9basement_present + β10waterfront + β11Distance + β12age +ε

price = 528,061.132324 + 281.5(sqft_living) - 59,614.62(bedrooms_4-6) - 354,167.07(bedrooms_7 and more) - 123,943.36(bathrooms_0-3) - 251,030.68(floors_1-2) - 240,274.33(floors_3 and more) - 52,041.64(basement_present) + 759,117.07(waterfront) - 17,904.38(Distance) + 729.39(age) +ε

Regression coefficients:

Intercept: The constant term in the regression model is 528,061.13. This represents the estimated baseline value of the price when all the independent variables are zero.

"sqft_living": For every unit increase in the square footage of living space, the predicted value of the price increases by 281.501191, assuming all other variables are held constant.

"bedrooms 4-6": Having 4-6 bedrooms is associated with a decrease of 59,614.62 in the predicted value of the price, compared to the reference category (likely the category with fewer bedrooms), assuming all other variables are held constant.

"bedrooms 7 and more": Having 7 or more bedrooms is associated with a decrease of 354,167.07 in the predicted value of the price, compared to the reference category, assuming all other variables are held constant.

"bathrooms 0-3": Having 0-3 bathrooms is associated with a decrease of 123,943.36 in the predicted value of the price, compared to the reference category, assuming all other variables are held constant.

"floors 1-2": Having 1-2 floors is associated with a decrease of 251,030.68 in the predicted value of the price, compared to the reference category, assuming all other variables are held constant.

"floors 3 and more": Having 3 or more floors is associated with a decrease of 240,274.33 in the predicted value of the price, compared to the reference category, assuming all other variables are held constant.

"basement_present": The presence of a basement (represented as 1) is associated with a decrease of 52,041.64 in the predicted value of the price, assuming all other variables are held constant.

"Waterfront": Having a waterfront property (represented as 1) is associated with an increase of 759,117.07 in the predicted value of the price, assuming all other variables are held constant.

"Distance": For every unit increase in distance from downtown area, the predicted value of the price decreases by 17,904.38, assuming all other variables are held constant.

"age": For every unit increase in age of the house, the predicted value of the price increases by 729.4, assuming all other variables are held constant.