MBA 810

King County Housing Sales

Regression Analysis

- Mrinal Mishra

06 MAY 2024

Que-1 Part I. Descriptive Statistics

Statistics Charts:

	Price	Bedrooms	Bathrooms	Sqft_living	Sqft_lot	Floors	Waterfront	View	Condition
Mean	$5,40,088.14	3.37	2.11	2079.90	15106.97	1.49	0.01	0.23	3.41
Standard Error	$2,497.23	0.01	0.01	6.25	281.75	0.00	0.00	0.01	0.00
Median	$4,50,000.00	3.00	2.25	1910.00	7618.00	1.50	0.00	0.00	3.00
Mode	$4,50,000.00	3.00	2.50	1300.00	5000.00	1.00	0.00	0.00	3.00
Standard Deviation	$3,67,127.20	0.93	0.77	918.44	41420.51	0.54	0.09	0.77	0.65
Sample Variance	$1,34,78,23,78,397.25	0.87	0.59	843533.68	1715658774.18	0.29	0.01	0.59	0.42
Kurtosis	$34.59	49.06	1.28	5.24	285.08	-0.48	127.63	10.89	0.53
Skewness	$4.02	1.97	0.51	1.47	13.06	0.62	11.39	3.40	1.03
Range	$76,25,000.00	33.00	8.00	13250.00	1650839.00	2.50	1.00	4.00	4.00
Minimum	$75,000.00	0.00	0.00	290.00	520.00	1.00	0.00	0.00	1.00
Maximum	$77,00,000.00	33.00	8.00	13540.00	1651359.00	3.50	1.00	4.00	5.00
Sum	$11,67,29,25,008.00	72854.00	45706.25	44952873.00	326506890.00	32296.50	163.00	5064.00	73688.00
Count	$21,613.00	21613.00	21613.00	21613.00	21613.00	21613.00	21613.00	21613.00	21613.00

	Grade	Sqft_above	Sqft_basement	Yr_built	Lat	Long	Sqft_living15	Sqft_lot15
Mean	7.66	1788.39	468.85	1971.01	47.56	-122.21	1986.55	12768.46
Standard Error	0.01	5.63	2.28	0.20	0.00	0.00	4.66	185.73
Median	7.00	1560.00	292.00	1975.00	47.57	-122.23	1840.00	7620.00
Mode	7.00	1300.00	292.00	2014.00	47.68	-122.29	1540.00	5000.00
Standard Deviation	1.18	828.09	335.87	29.37	0.14	0.14	685.39	27304.18
Sample Variance	1.38	685734.67	112811.92	862.80	0.02	0.02	469761.24	745518225.34
Kurtosis	1.19	3.40	7.33	-0.66	-0.68	1.05	1.60	150.76
Skewness	0.77	1.45	2.28	-0.47	-0.49	0.89	1.11	9.51
Range	12.00	9120.00	4810.00	115.00	0.62	1.20	5811.00	870549.00
Minimum	1.00	290.00	10.00	1900.00	47.16	-122.52	399.00	651.00
Maximum	13.00	9410.00	4820.00	2015.00	47.78	-121.32	6210.00	871200.00
Sum	165488.00	38652488.00	10133177.00	42599334.00	1027915.42	-2641408.94	42935359.00	275964632.00
Count	21613.00	21613.00	21613.00	21613.00	21613.00	21613.00	21613.00	21613.00

Variability in Price: In King County, the average home price is about $540,088, with a broad range indicated by a high standard deviation of $367,127 and a price range (difference between most and least expensive homes) of $7,625,000. The distribution of home prices is highly peaked and right skewed, showing that most homes are relatively affordable, but there are also many high-priced outliers.

Variability in Size of Houses: In King County, there is a wide range in house sizes, from the living area's average of 2,079.90 sqft to lots as large as 1,650,839 sqft. Living spaces both above and below ground show similar variability, with above-ground spaces averaging 1,788.39 sqft and basements ranging up to 4,820 sqft. This variance in house and lot sizes reflects the diverse housing options available in the area.

Que-2 Part I. Variables Sorting

1. Missing Data

Replaced missing data for Basement with mean value which is 292.

There was missing data for YRen too so converted it into binary numeric as below:

- **YRen = 1** implies house was renovated
- **YRen = 0** implies house was not renovated

2. Error in Data

Initially divided Date of Sale in 4 categories based on season as below:

- Feb-April: Spring (S1)
- May-July: Summer (S2)
- Aug- Oct: Fall (S3)
- Nov-Jan: Winter (S4)

We got error when we were applying **If-And** formula on **Date (DD/MM/YYYY)**. We then got the Month removed out separately as new field from date and applied formula only on Month part to remove the error.

3. Final list of converted Categorical Variables in Numeric Variables List:

Zip code:

- ZPC1: Zip codes from 98001 to 98050
- ZPC2: Zip codes from 98051 to 980100
- ZPC3: Zip codes from 98101 to 98150

Grades:

- Gr1: Lower grades from 1-3
- Gr2: Below Average grades from 4-6
- Gr3: Average grade 7
- Gr4: Above average grade 8-9

Months:

- MQ1: Months from January to March
- MQ2: Months from April to June
- MQ3: Months from July to September

Year Built:

- YB1: House was built in year from 1900 to year 1925.
- YB2: House was built in year from 1926 to year 1950.
- YB3: House was built in year from 1951 to year 1975.
- YB4: House was built in year from 1976 to year 2000.

Year of Sale

- Yr2014 = 1 implies house was sold in year 2014
- Yr2014 = 0 implies house was sold in year 2015

Que-2 Part II. Correlation Matrix

4. Correlation Matrix Analysis of Dataset 1 (Includes all 16 converted numerical variables)

	ZPC1	ZPC2	ZPC3	Yr2014	YB1	YB2	YB3	YB4	YRen	Gr1	Gr2	Gr3	Gr4	MQ1	MQ2	MQ3	price
ZPC1	1																
ZPC2	-0.40045	1															
ZPC3	-0.51757	-0.35198	1														
Yr2014	-0.00392	0.00072	0.002886	1													
YB1	-0.20093	-0.14532	0.380092	0.008934	1												
YB2	-0.22807	-0.15912	0.268202	0.005485	-0.13399	1											
YB3	0.112125	-0.0733	-0.14096	-0.01657	-0.19786	-0.24531	1										
YB4	0.20994	0.18468	-0.30328	0.005769	-0.20704	-0.2567	-0.37904	1									
YRen	-0.03796	-0.05585	0.074493	0.023688	0.163396	0.117604	0.003623	-0.10481	1								
Gr1	0.01773	-0.0071	-0.00918	-0.00515	0.006992	0.004186	0.007209	-0.00857	-0.00286	1							
Gr2	-0.09327	-0.0691	0.087888	-0.02124	0.18316	0.305573	-0.0244	-0.17862	0.004729	-0.00471	1						
Gr3	-0.00534	-0.12092	0.086842	-0.01296	0.047483	0.041249	0.213545	-0.08444	-0.02043	-0.01147	-0.29162	1					
Gr4	0.049203	0.14926	-0.11859	0.024729	-0.1505	-0.21576	-0.17087	0.169192	0.017475	-0.01241	-0.31551	-0.76922	1				
MQ1	0.014751	-0.00552	-0.01703	-0.70088	-0.01439	-0.00355	0.023764	-0.00974	-0.01613	0.01076	0.020495	0.014855	-0.02622	1			
MQ2	-0.02502	0.014365	0.01738	-0.14323	0.004808	0.006751	-0.01618	0.008108	0.002659	-0.00924	-0.01003	-0.00728	0.012388	-0.32885	1		
MQ3	0.012485	0.001979	-0.00999	0.424445	0.002137	-0.01278	-0.00199	0.006515	0.007955	-0.00074	-0.00738	-0.01496	0.019324	-0.29749	-0.4175	1	
price	0.010289	0.00202	0.041316	-0.00358	0.051939	-0.03698	-0.11199	0.008739	0.126092	-0.01298	-0.23068	-0.3158	0.311373	-0.01507	0.030736	-0.00449	1

Key Insights on Housing Prices:

- **Zip Codes:** Houses in the zip codes from 98101 to 98150 (ZPC3) tend to be slightly more expensive.
- **Year Built:** Older houses built between 1951 and 1975 (YB3) are usually cheaper, but the oldest houses, built from 1900 to 1925 (YB1), might be more valuable because they could be considered historic.
- **Renovation:** Houses that have been renovated generally sell for more.
- **House Grades:** High-grade houses (Gr4) are priced much higher. In contrast, houses with average (Gr3) or below-average grades (Gr2) tend to sell for less.
- **Sale Timing:** The month of the sale, from January to September, doesn't really impact the price.

Que-2 Part III. Regression Model

5. Regression Model Analysis

We performed 10 iterations of regression making sure we use all the variables at least once. Since all variables had significant variable below 0.05 and we were satisfied that our model is fit to explain, we stopped the regression here. And the analysis of our final regression model is:

Model Performance: The model has a good fit with an R Square of 0.6793, meaning it explains about 67.93% of the variability in house prices based on the included variables.

Significant Variables:

- **Location (Latitude and Longitude):** Latitude positively affects house prices significantly, indicating that houses located at higher latitudes within the dataset tend to be more expensive. Longitude negatively affects house prices, suggesting that houses located further east within the region are cheaper.
- **House Characteristics:** Square footage of the living area (sqft_living) and the number of bathrooms (bathrooms) positively influence house prices. More space and more bathrooms increase the house's value.
- **Property Features:** Houses with a waterfront view (waterfront) and those with better views (view) and condition (condition) fetch higher prices.
- **Grades:** The grades variable, which reflects the construction and design quality, significantly impacts prices. Higher grades generally increase prices, but the specific grade coefficients (Gr1 to Gr4) are negative, which may require further exploration to understand this context within the data's coding.
- **Square Footage of Neighbouring Properties:** Larger living spaces in neighbouring properties (sqft_living15) also positively affect a house's price, suggesting that more affluent neighbourhoods (with larger homes) have higher property values.

Negative Influences:

- **Year Sold:** Houses sold in 2014 (Yr2014) tend to have lower prices compared to those sold in 2015, indicated by a negative coefficient.
- **Number of Bedrooms:** Surprisingly, more bedrooms (bedrooms) have a negative impact on price. This could suggest that within the dataset, larger numbers of smaller or less desirable rooms could reduce a property's appeal.

Intercept and P-Values: The model's intercept and each variable's P-value are significant, indicating robust statistical validity for the model's estimates.

Que-3 Bonus One

3A. In conclusion, the hypothesis that "average price of houses with waterfront are higher than those without a waterfront" is supported by the regression analysis, with waterfront houses being more expensive by over a million dollars on average compared to non-waterfront houses.

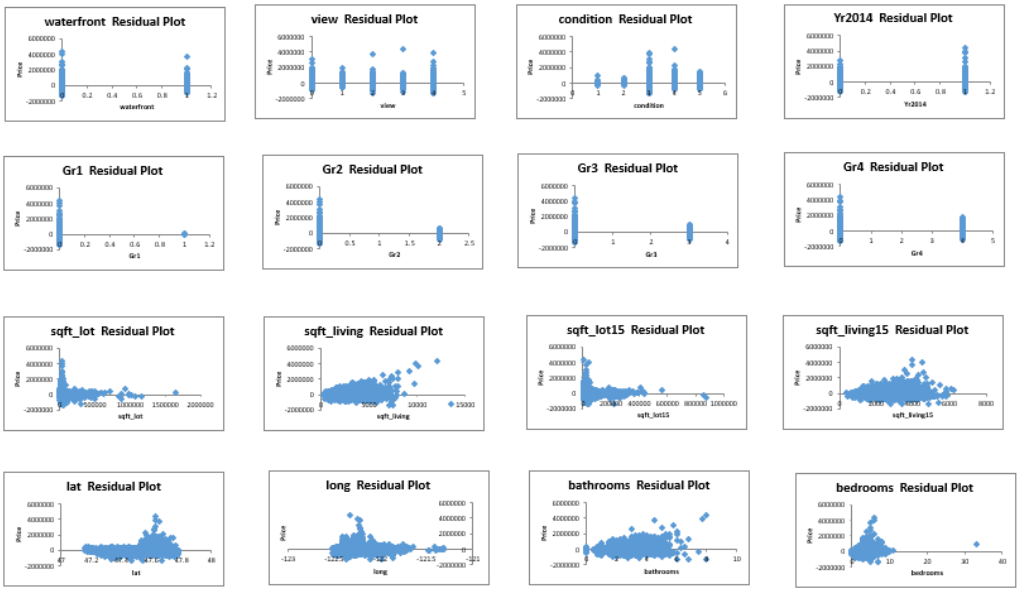
3B. In conclusion, the analysis supports the hypothesis that older houses have a lower price, with each additional year of age reducing the price by approximately $674. However, given the very low R-squared value (0.0029), age alone does not strongly predict the price of the houses. Other variables not included in the model are likely to have a significant impact on the house prices.

Also, no pattern between Age and Price so below Residual Plot seems good.

Appendix

Que-2 Part IV. Residual Plots

Final



Regression Model for Question 2:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8242
R Square	0.6793
Adjusted R Square	0.6790
Standard Error	207994.2789
Observations	21613.0000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	16.0000	1978638815189400.0000	123664925949338.0000	2858.5366	0.0000
Residual	21596.0000	934277946731895.0000	43261620056.1166		
Total	21612.0000	2912916761921300.0000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-60889746.3979	1409958.8716	-43.1855	0.0000	-63653369.8952	-58126122.9005	-63653369.8952	-58126122.9005
bedrooms	-28949.0236	1955.8417	-14.8013	0.0000	-32782.6177	-25115.4295	-32782.6177	-25115.4295
bathrooms	19896.3005	3015.4674	6.5981	0.0000	13985.7618	25806.8393	13985.7618	25806.8393
sqft_living	196.6601	3.3272	59.1064	0.0000	190.1385	203.1817	190.1385	203.1817
sqft_lot	0.1905	0.0495	3.8527	0.0001	0.0936	0.2875	0.0936	0.2875
waterfront	574939.0732	17904.1833	32.1120	0.0000	539845.5519	610032.5945	539845.5519	610032.5945
view	54693.0829	2153.8221	25.3935	0.0000	50471.4325	58914.7332	50471.4325	58914.7332
condition	50001.6620	2234.5759	22.3764	0.0000	45621.7284	54381.5957	45621.7284	54381.5957
lat	665656.2681	10466.5621	63.5984	0.0000	645141.0335	686171.5027	645141.0335	686171.5027
long	-242141.6293	11363.8963	-21.3080	0.0000	-264415.7051	-219867.5535	-264415.7051	-219867.5535
sqft_living15	55.5320	3.4276	16.2014	0.0000	48.8136	62.2504	48.8136	62.2504
sqft_lot15	-0.4392	0.0757	-5.8042	0.0000	-0.5876	-0.2909	-0.5876	-0.2909
Gr1	-247971.1798	105014.5911	-2.3613	0.0182	-453807.5325	-42134.8272	-453807.5325	-42134.8272
Gr2	-237845.6905	6592.6737	-36.0773	0.0000	-250767.8176	-224923.5633	-250767.8176	-224923.5633
Gr3	-162686.5456	3971.8922	-40.9595	0.0000	-170471.7475	-154901.3436	-170471.7475	-154901.3436
Gr4	-109295.8393	2695.2120	-40.5519	0.0000	-114578.6539	-104013.0247	-114578.6539	-104013.0247
Yr2014	-27144.0145	3032.9769	-8.9496	0.0000	-33088.8731	-21199.1559	-33088.8731	-21199.1559