Kings County Housing Sales

Final Project

Presented By:

Ahmed Raza Amaan

1. Provide a Table with descriptive statistics of all the numerical variables. What can you say about the variability in price and size of houses (sqft_living, sqft_lot, sqft_above, sqft_basement) in King County?

price		sqft_living		sqft_lot		sqft_above		sqft_basement	
Mean	540088.14	Mean	2079.90	Mean	15106.97	Mean	1788.39	Mean	291.51
Standard Error	2497.23	Standard Error	6.25	Standard Error	281.75	Standard Error	5.63	Standard Error	3.01
Median	450000.00	Median	1910.00	Median	7618.00	Median	1560.00	Median	0.00
Mode	450000.00	Mode	1300.00	Mode	5000.00	Mode	1300.00	Mode	0.00
Standard Deviation	367127.20	Standard Deviation	918.44	Standard Deviation	41420.51	Standard Deviation	828.09	Standard Deviation	442.58
Sample Variance	134782378397.25	Sample Variance	843533.68	Sample Variance	1715658774.18	Sample Variance	685734.67	Sample Variance	195872.67
Kurtosis	34.59	Kurtosis	5.24	Kurtosis	285.08	Kurtosis	3.40	Kurtosis	2.72
Skewness	4.02	Skewness	1.47	Skewness	13.06	Skewness	1.45	Skewness	1.58
Range	7625000.00	Range	13250.00	Range	1650839.00	Range	9120.00	Range	4820.00
Minimum	75000.00	Minimum	290.00	Minimum	520.00	Minimum	290.00	Minimum	0.00
Maximum	7700000.00	Maximum	13540.00	Maximum	1651359.00	Maximum	9410.00	Maximum	4820.00
Sum	11672925008.00	Sum	44952873.00	Sum	326506890.00	Sum	38652488.00	Sum	6300385.00
count	21613.00	count	21613.00	count	21613.00	count	21613.00	count	21613.00

Price:

• Prices fluctuate widely, from $75,000 to $7,700,000, with a median of $450,000.

• The data has a heavy-tailed distribution shown by its very high kurtosis of 34.59 and its sharp rightward skewness score of 4.02.

Sqft Living:

• Living spaces vary in size from 290 square feet to 13,540 square feet, with a median of 1,910 square feet.

• The data has a rather heavy-tailed distribution, as shown by the skewness value of 1.47 and the kurtosis value of 5.24.

Sqft Lot:

• Lot areas diUer greatly in size, from 520 square feet to 1,651,359 square feet, with a median of 7,618 square feet.

 With a skewness score of 13.06 and a very high kurtosis of 285.08, the data is strongly skewed to the right, suggesting an extremely heavy-tailed distribution.

Sqft above:

• Above-ground living spaces vary in size from 290 to 9,410 square feet, with a typical of 1,560 square feet.

• The data has a little heavy-tailed distribution indicated by a kurtosis value of 3.40 and a skewness value of 1.45, which is somewhat skewed to the right.

Sqft Basement:

• Basement spaces vary in size from 0 to 4,820 square feet, with a 0 square foot median.

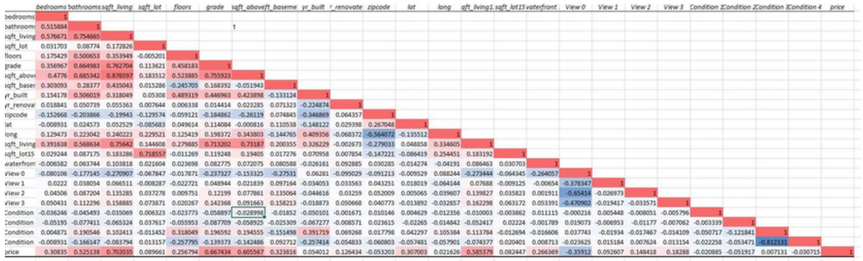
• A heavy-tailed distribution is indicated by the data's strong rightward skewness value of 1.58 and kurtosis value of 2.72.

2. Develop a regression model to predict the price of houses in King County. What are the variables aUecting the price? Be mindful of multicollinearity.

To solve this issue, we will use multiple regression analysis and conduct iterations. We will also exclude factors that are not statistically significant for determining the price. Specifically, we will remove variables with a p-value less than 0.05. Given that the dataset contains 19 variables, it is important to note that Excel can only accommodate 16 variables for regression analysis at a time. Therefore, we will first do the regression analysis using 16 variables. After identifying and eliminating the variables that do not significantly contribute to the analysis, we will then introduce more variables.

We have taken a dummy variable for the Year.

 Run regression taking Price on the Y-axis and the rest of variables on the X-axis



- Regression 1:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
	6091985.51							
intercept	26736.17	3125725.73	1.95	0.051	-34667.70	12218638.71	-34667.70	12218638.71
year dummy	-38884.03	3150.23	8.49	0.000	20561.50	32910.84	20561.50	32910.84
bedrooms	46557.67	2027.00	-19.18	0.000	-42857.10	-34910.97	-42857.10	-34910.97
bathrooms	166.77	3489.21	13.34	0.000	39718.56	53396.79	39718.56	53396.79
sqft_living	-0.26	4.69	35.58	0.000	157.58	175.95	157.58	175.95
sqft_lot	28807.48	0.04	-7.00	0.000	-0.33	-0.18	-0.33	-0.18
floors	578300.86	3830.67	7.52	0.000	21299.09	36315.88	21299.09	36315.88
waterfront	42893.91	18627.46	31.05	0.000	541789.66	614812.05	541789.66	614812.05
view	20211.36	2284.78	18.77	0.000	38415.57	47372.26	38415.57	47372.26
condition	120698.57	2523.27	8.01	0.000	15265.57	25157.15	15265.57	25157.15
grade		2251.72	53.60	0.000	116285.03	125112.10	116285.03	125112.10
sqft_above	0.00	4.58	-1.69	0.092	-16.70	1.26	-16.70	1.26
sqft_basement	-3597.30	0.00	65535.00	#NUM!	0.00	0.00	0.00	0.00
yr_built	10.44	74.52	-48.28	#NUM!	-3743.36	-3451.24	-3743.36	-3451.24
yr_renovated		3.92	2.66	0.008	2.76	18.12	2.76	18.12
zipcode	23.27	31.37	0.04	0.965	-60.10	62.86	-60.10	62.86
sqft_living15		3.64	6.39	0.000	16.13	30.41	16.13	30.41

Initially, we will exclude the variable "zip code" because to its elevated p-value.
Additionally, we will include the variable "sqft_lot_15" into the regression analysis.

Regression 2:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
	6184719.74462		44.74868	0.00000	5913817.71884	6455621.77039		
intercept	26652.34345	138210.11290	8.47005	0.00000	20484.66464	32820.02226	5913817.71884	6455621.77039
year dummy	-39511.90286	3146.65637	-19.52713	0.00000	-43477.98655	-35545.81917	20484.66464	32820.02226
bedrooms	45853.74237	2023.43586	13.15118	0.00000	39019.62183	52687.86291	-43477.98655	-35545.81917
bathrooms	167.39365	3486.66485	35.92125	0.00000	158.25968	176.52763	39019.62183	52687.86291
sqft_living		4.66002					158.25968	176.52763

sqft_lot	-0.00523	0.05120	-0.10212	0.91866	-0.10559	0.09513	-0.10559	0.09513
floors	27510.75015	3778.36534	7.28112	0.00000	20104.87511	34916.62519	20104.87511	34916.62519
waterfront	579670.81318	18606.11951	31.15485	0.00000	543201.44519	616140.18116	543201.44519	616140.18116
view	42887.67636	2269.82393	18.89472	0.00000	38438.65386	47336.69885	38438.65386	47336.69885
condition	20604.23397	2496.03805	8.25478	0.00000	15711.81510	25496.65284	15711.81510	25496.65284
grade	119983.48738	2245.68755	53.42840	0.00000	115581.77398	124385.20078	115581.77398	124385.20078
sqft_above	-8.32942	4.33709	-1.45912	0.15013	-15.42245	2.36361	-15.42245	2.36361
sqft_basement	0.00000	0.00000	65535.00000	#NUM!	0.00000	0.00000	0.00000	0.00000
yr_built	-3572.51728	70.87307	-50.40726	#NUM!	-3711.43372	-3433.60083	-3711.43372	-3433.60083
yr_renovated	10.97154	3.90861	2.80702	0.00500	3.31037	18.63272	3.31037	18.63272
sqft_living15	24.97683	3.59452	6.94859	0.00000	17.93132	32.02235	17.93132	32.02235
sqft_lot15	-0.54800	0.07823	-7.00467	0.00000	-0.70134	-0.39466	-0.70134	-0.39466

- The variable "sqft_lot" will be eliminated first because to its highest p-value.

Regression 3:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
intercept	6184509.4599	138191.6053	44.7531	0.0000	5913643.7110	6455375.2087	5913643.7110	6455375.2087
year dummy	26649.6554	3146.4742	8.4697	0.0000	20482.3337	32816.9771	20482.3337	32816.9771
bedrooms	-39506.9075	2022.7981	-19.5308	0.0000	-43471.7412	-35542.0739	-43471.7412	-35542.0739
bathrooms	45851.1413	3486.4919	13.1511	0.0000	39017.3597	52684.9229	39017.3597	52684.9229
sqft_living	167.3886	4.6596	35.9230	0.0000	158.2553	176.5218	158.2553	176.5218
floors	27524.9503	3775.7191	7.2900	0.0000	20124.2621	34925.6384	20124.2621	34925.6384
waterfront	579704.5573	18602.7587	31.1623	0.0000	543241.7767	616167.3378	543241.7767	616167.3378
view	42880.6848	2268.7391	18.9007	0.0000	38433.7886	47327.5809	38433.7886	47327.5809
condition	20607.1803	2495.8141	8.2567	0.0000	15715.2004	25499.1602	15715.2004	25499.1602
grade	119982.9150	2245.6291	53.4295	0.0000	115581.3162	124384.5139	115581.3162	124384.5139
sqft_above	0.0000		65535.00		0.0000	0.0000		2.3267
sqft_basement	-3572.4247	0.0000	00	#NUM!	-3711.3265	-3433.5228	0.0000	0.0000
yr_built	10.9747	70.8656	-50.4112	#NUM!	3.3139	18.6355	-3711.3265	
yr_renovated	24.9961	3.9084	2.8080		17.9604	32.0317	3.3139	-3433.5228
sqft_living15	-0.5536	3.5895	6.9637	0.0050	-0.6629	-0.4443	17.9604	18.6355
sqft_lot15		0.0558	-9.9251	0.0000			-0.6629	32.0317
				0.0000				-0.4443

- We are removing the variable “sqft_above” first because it has the highest p-value.

- Regression 4:

	Coefficients	Standard Error	t Stat	Pvalue	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%

intercept	6184509.4599	138191.6053	44.7531	0.0000	5913643.7110	6455375.2087	5913643.7110	6455375.2087
year dummy	26649.6554	3146.4742	8.4697	0.0000	20482.3337	32816.9771	20482.3337	32816.9771
bedrooms	-39506.9075	2022.7981	-19.5308	0.0000	-43471.7412	-35542.0739	-43471.7412	-35542.0739
bathrooms	45851.1413	3486.4919	13.1511	0.0000	39017.3597	52684.9229	39017.3597	52684.9229
sqft_living	160.8348	3.8840	41.4097	0.0000	153.2219	168.4477	153.2219	168.4477
floors	27524.9503	3775.7191	7.2900	0.0000	20124.2621	34925.6384	20124.2621	34925.6384
waterfront	579704.5573	18602.7587	31.1623	0.0000	543241.7767	616167.3378	543241.7767	616167.3378
view	42880.6848	2268.7391	18.9007	0.0000	38433.7886	47327.5809	38433.7886	47327.5809
condition	20607.1803	2495.8141	8.2567	0.0000	15715.2004	25499.1602	15715.2004	25499.1602
grade	119982.9150	2245.6291	53.4295	0.0000	115581.3162	124384.5139	115581.3162	124384.5139
sqft_basement	-3572.4247	70.8656	-50.4112	0.0000	-3711.3265	-3433.5228	-3711.3265	-3433.5228
yr_built	10.9747	3.9084	2.8080	0.0050	3.3139	18.6355	3.3139	18.6355
yr_renovated	24.9961	3.5895	6.9637	0.0000	17.9604	32.0317	17.9604	32.0317
sqft_living15	-0.5536	0.0558	-9.9251	0.0000	-0.6629	-0.4443	-0.6629	-0.4443
sqft_lot15								

- Now, we are removing the variable “sqft_basement” first as it has the highest p-value

- Regression 5:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
intercept	6205923.2142	137399.8938	45.1669	0.0000	5936609.2790	6475237.1493	5936609.2790	6475237.1493
year dummy	26612.7929	3146.4505	8.4580	0.0000	20445.5176	32780.0683	20445.5176	32780.0683
bedrooms	-39504.3814	2022.8485	-19.5291	0.0000	-43469.3138	-35539.4489	-43469.3138	-35539.4489
bathrooms	46680.4370	3439.1209	13.5734	0.0000	39939.5062	53421.3679	39939.5062	53421.3679
sqft_living	163.0538	3.5683	45.6950	0.0000	156.0596	170.0479	156.0596	170.0479
floors	25242.6274	3430.3400	7.3586	0.0000	18518.9077	31966.3470	18518.9077	31966.3470
waterfront	578740.0249	18591.2750	31.1297	0.0000	542299.7534	615180.2963	542299.7534	615180.2963
view	43477.3392	2230.9854	19.4880	0.0000	39104.4432	47850.2352	39104.4432	47850.2352
condition	20819.5345	2491.5560	8.3560	0.0000	15935.9008	25703.1681	15935.9008	25703.1681
grade	119668.4649	2235.1393	53.5396	0.0000	115287.4268	124049.5030	115287.4268	124049.5030
yr_built	-3581.9591	70.5602	-50.7646	0.0000	-3720.2624	-3443.6559	-3720.2624	-3443.6559
yr_renovated	10.9248	3.9083	2.7952	0.0052	3.2641	18.5854	3.2641	18.5854
sqft_living15	23.9884	3.5213	6.8123	0.0000	17.0863	30.8905	17.0863	30.8905
sqft_lot15	-0.5621	0.0555	-10.1342	0.0000	-0.6708	-0.4534	-0.6708	-0.4534

- All the above variables are significant as the p values are 0.05.

- R square on Regression 1

- R square on Regression 5

Regression	Statistics	Regression	Statistics
Multiple R	0.808760912	Multiple R	0.80922472
R Square	0.654094212	R Square	0.654844647
Adjusted R Square	0.653807664	Adjusted R Square	0.654636905
Standard Error	215996.263	Standard Error	215751.8466
Observations	21613	Observations	21613

Additionally, it is evident that the R square value has risen, indicating that the model has improved in accuracy after eliminating the variables that were not statistically significant.

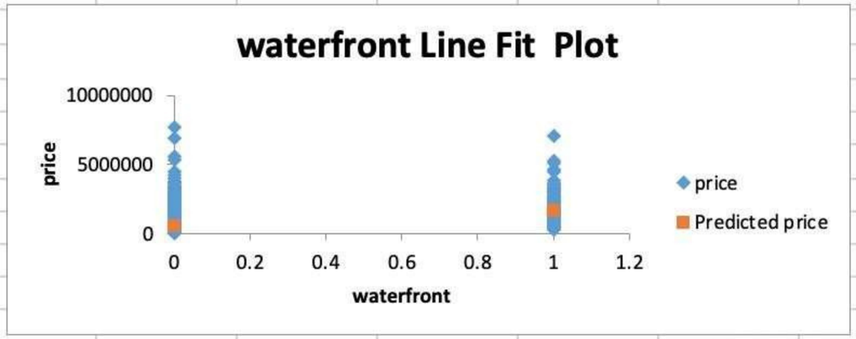
3. Test the following hypotheses and provide your conclusion (bonus)

a. Average price of houses with waterfront are higher than those without a waterfront.

For this, running regression on Price and waterfront, taking price on the Y axis and waterfront on the X axis.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	531563.600	2416.194	220.000	0.000	526827.681	536299.519	526827.681	536299.519
waterfront	1130312.425	27822.465	40.626	0.000	1075778.342	1184846.508	1075778.342	1184846.508

By the above regression, it is evident that the price of residences increases by 1130312.425 units when one unit of waterfront is added.
To get a more comprehensive clarification, see the graph provided below.



As we can see from the graph above, “0” denotes the houses with no waterfront and “1” with houses with waterfront.

The predicted price for houses with a waterfront is higher than the houses without a waterfront.

b. Older houses have lower prices. (Create the “age” variable with respect to 2014 and 2015 using yr_built data)

For above question, A new variable, named "age," was formed by subtracting the year from the year_built, resulting in the determination of the house's "Age."

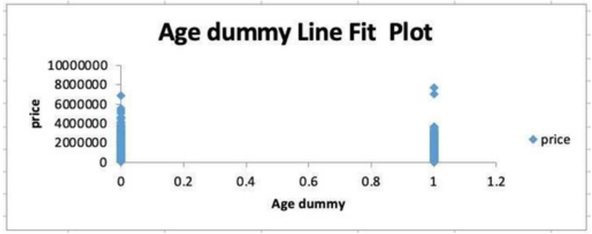
price	Year	yr_built	Age
221900	2014	1955	59
538000	2014	1951	63
180000	2015	1933	82
604000	2014	1965	49
510000	2015	1987	28
1.23E+06	2014	2001	13
257500	2014	1995	19
291850	2015	1963	52
229500	2015	1960	55

Subsequently, we computed the Mean of the column labeled "Age," yielding an average value of 43. Subsequently, we generated a dummy variable in which buildings with an age beyond 43 were assigned a value of 1 (representing old), whereas dwellings with an age below 43 were assigned a value of 0 (representing new).

Upon running regression, we got the following results:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	571124.4609	3426.3261	166.6871	0.0000	564408.6090	577840.3128	564408.6090	577840.3128
Age dummy	-65615.5694	4981.9233	-13.1707	0.0000	-75380.5066	-55850.6322	-75380.5066	-55850.6322

The data above indicates that for each additional year of the "Age dummy" variable, the price of the property decreases by $65615.5694.
For more clarification, we can see the graph below:



The prices of old houses (represented by dummy variable 1) are much lower than the prices of new houses (represented by dummy variable 0).