Business Analytics and Stats Fall 2024

King County Housing Sales

Group

Aakash Janardhan Jadhav(aj68160n@pace.edu)

Varun Shah(vs68126n@pace.edu)

Ganai Izaan Hamid(ig79632n@pace.edu)

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	540088.1418	Mean	2079.899736	Mean	15106.96757	Mean	1788.390691	Mean	291.5090455
Standard Error	2497.232803	Standard Error	6.247319071	Standard Error	281.7461116	Standard Error	5.632750647	Standard Error	3.010435961
Median	450000	Median	1910	Median	7618	Median	1560	Median	0
Mode	450000	Mode	1300	Mode	5000	Mode	1300	Mode	0
Standard Deviation	367127.1965	Standard Deviation	918.440897	Standard Deviation	41420.51152	Standard Deviation	828.0909777	Standard Deviation	442.5750427
Sample Variance	1.34782E+11	Sample Variance	843533.6814	Sample Variance	1715658774	Sample Variance	685734.6673	Sample Variance	195872.6684
Kurtosis	34.58554043	Kurtosis	5.24309299	Kurtosis	285.0778197	Kurtosis	3.402303621	Kurtosis	2.715574211
Skewness	4.024069145	Skewness	1.471555427	Skewness	13.06001896	Skewness	1.446664473	Skewness	1.577965056
Range	7625000	Range	13250	Range	1650839	Range	9120	Range	4820
Minimum	75000	Minimum	290	Minimum	520	Minimum	290	Minimum	0
Maximum	7700000	Maximum	13540	Maximum	1651359	Maximum	9410	Maximum	4820
Sum	11672925008	Sum	44952873	Sum	326506890	Sum	38652488	Sum	6300385
Count	21613	Count	21613	Count	21613	Count	21613	Count	21613
Confidence Level(95.0%)	4894.760483	Confidence Level(95.0%)	12.24520616	Confidence Level(95.0%)	552.2431596	Confidence Level(95.0%)	11.04060672	Confidence Level(95.0%)	5.900676524

Price:

The mean house price is approximately $540,088. The wide range ($75,000 to $7,700,000) indicates significant variability, supported by a high standard deviation of $367,127.

The distribution is highly skewed to the right (skewness = 4.02) with a pronounced kurtosis (34.59), indicating the presence of outliers or very expensive properties.

Size Metrics (sqft):

Living Area (sqft_living): Mean size is 2,080 sqft with moderate variability (SD = 918 sqft). The range is wide, from 290 sqft to 13,540 sqft.

Lot Area (sqft_lot): Substantial variability with a mean of 15,107 sqft but an extremely high range of up to 1,651,359 sqft. Skewness (13.06) and kurtosis (285.08) suggest extreme values.

Above-ground Area (sqft_above): Mean of 1,788 sqft and less skewed than total living area, indicating a relatively consistent above-ground size.

Basement Area (sqft_basement): Mean is 292 sqft, with the majority of homes having no basement (median = 0).

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

Regression models were developed to understand factors influencing housing prices while addressing multicollinearity and variable significance. As the dataset has 19 variables, and Excel can only take 16 variables at a time to run a regression, we will be running the regression with 16 variables and after removing the insignificant ones, we will be adding more variables.

Regression 1: Initial Model

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	9470712.217	3084190.311	3.070728867	0.00	3425471.508	15515952.93	3425471.508	15515952.93
bedrooms	-39120.04388	2028.527386	-19.2849474	0.00	-43096.10732	-35143.98044	-43096.10732	-35143.98044
bathrooms	45807.76846	3490.447617	13.12375188	0.00	38966.23343	52649.30348	38966.23343	52649.30348
sqft_living	172.4147544	4.60658279	37.42790747	0.00	163.3855121	181.4439968	163.3855121	181.4439968
sqft_lot	-0.261303318	0.036705043	-7.11900325	0.00	-0.333247912	-0.189358725	-0.333247912	-0.189358725
floors	25743.01243	3804.011182	6.767333532	0.00	18286.86967	33199.15519	18286.86967	33199.15519
waterfront	574488.0937	18635.03912	30.82838141	0.00	537962.0412	611014.1461	537962.0412	611014.1461
view	45057.60821	2261.61677	19.92274235	0.00	40624.67237	49490.54405	40624.67237	49490.54405
condition	19542.12224	2523.413161	7.744321279	0.00	14596.04614	24488.19833	14596.04614	24488.19833
grade	124723.4952	2163.75382	57.64218372	0.00	120482.378	128964.6124	120482.378	128964.6124
sqft_above	-2.687621076	4.516388542	-0.5950819	**0.55**	-11.54007606	6.164833903	-11.54007606	6.164833903
sqft_basement	0	0	65535	#NUM!	0	0	0	0
yr_built	-3597.451338	74.58524768	-48.2327464	#NUM!	-3743.64393	-3451.258746	-3743.64393	-3451.258746
yr_renovated	9.18934065	3.917311324	2.345828526	0.02	1.511121247	16.86756005	1.511121247	16.86756005
zipcode	-32.75551924	30.9363901	-1.05880224	**0.29**	-93.3931278	27.88208933	-93.3931278	27.88208933
Year_dummy	-26651.48189	3153.098657	-8.4524732	0.00	-32831.78805	-20471.17574	-32831.78805	-20471.17574

Regression Statistics	
Multiple R	0.80835678
R Square	**0.653440684**
Adjusted R Square	0.653169741
Standard Error	216195.205
Observations	21613

- **R Square**: 0.6534
- **Significant Variables**: bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, grade, yr_renovated, Year_dummy
- **Removed Variable**: sqft_above (p-value > 0.05)

Regression 2: Added sqft_living15

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	6118721.675	3125715.088	1.957542995	0.05	-7910.679564	12245354.03	-7910.679564	12245354.03
bedrooms	-38884.03283	2026.997501	-19.18306896	0.00	-42857.09759	-34910.96807	-42857.09759	-34910.96807
bathrooms	46557.67388	3489.211229	13.34332341	0.00	39718.56225	53396.78551	39718.56225	53396.78551
sqft_living	159.0476416	3.893846275	40.84589641	0.00	151.4154154	166.6798678	151.4154154	166.6798678
sqft_lot	-0.256570652	0.03667875	-6.99507613	0.00	-0.328463711	-0.184677593	-0.328463711	-0.184677593
floors	28807.48158	3830.669888	7.520220333	0.00	21299.08577	36315.87739	21299.08577	36315.87739
waterfront	578300.8552	18627.45673	31.04561528	0.00	541789.6647	614812.0457	541789.6647	614812.0457
view	42893.91174	2284.783469	18.77373165	0.00	38415.56744	47372.25603	38415.56744	47372.25603
condition	20211.35654	2523.267085	8.009994923	0.00	15265.56675	25157.14632	15265.56675	25157.14632
grade	120698.5663	2251.719793	53.60283581	0.00	116285.0293	125112.1034	116285.0293	125112.1034
sqft_basement	7.718600913	4.580452811	1.68511744	0.09	-1.259424786	16.69662661	-1.259424786	16.69662661
yr_built	-3597.299055	74.51661844	-48.2751248	0.00	-3743.357128	-3451.240981	-3743.357128	-3451.240981
yr_renovated	10.43928236	3.918595295	2.664036875	0.01	2.758546264	18.12001846	2.758546264	18.12001846
zipcode	1.382001922	31.36654378	0.044059745	**0.96**	-60.09873978	62.86274362	-60.09873978	62.86274362
Year_dummy	-26736.16959	3150.225085	-8.487066437	0.00	-32910.84335	-20561.49583	-32910.84335	-20561.49583
sqft_living15	23.27291434	3.643349821	6.387779236	0.00	16.13167969	30.41414898	16.13167969	30.41414898

Regression Statistics	
Multiple R	0.808760912
R Square	**0.654094212**
Adjusted R Square	0.653853967
Standard Error	215996.263
Observations	21613

- **R Square**: 0.6541
- **Significant Variables**: bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, grade, yr_renovated, Year_dummy, sqft_living15
- **Removed Variable**: zipcode (p-value > 0.05)

Regression 3:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	6211372.088	138217.44	44.93913422	0.00	5940455.701	6482288.475	5940455.701	6482288.475
bedrooms	-39511.90286	2023.435863	-19.52713381	0.00	-43477.98655	-35545.81917	-43477.98655	-35545.81917
bathrooms	45853.74237	3486.66485	13.15117579	0.00	39019.62183	52687.86291	39019.62183	52687.86291
sqft_living	160.8642334	3.894788344	41.3024327	0.00	153.2301607	168.4983062	153.2301607	168.4983062
sqft_lot	-0.005228564	0.051201605	-0.102117189	**0.92**	-0.10558749	0.095130362	-0.10558749	0.095130362
floors	27510.75015	3778.365342	7.281124946	0.00	20104.87511	34916.62519	20104.87511	34916.62519
waterfront	579670.8132	18606.11951	31.15484736	0.00	543201.4452	616140.1812	543201.4452	616140.1812
view	42887.67636	2269.82393	18.89471504	0.00	38438.65386	47336.69885	38438.65386	47336.69885
condition	20604.23397	2496.038049	8.254775593	0.00	15711.8151	25496.65284	15711.8151	25496.65284
grade	119983.4874	2245.687548	53.42839767	0.00	115581.774	124385.2008	115581.774	124385.2008
sqft_basement	6.529420664	4.537088264	1.439121367	**0.15**	-2.363607321	15.42244865	-2.363607321	15.42244865
yr_built	-3572.517275	70.87306697	-50.40726228	0.00	-3711.43372	-3433.600831	-3711.43372	-3433.600831
yr_renovated	10.97154245	3.908614582	2.807015687	0.01	3.310369282	18.63271562	3.310369282	18.63271562
sqft_lot15	-0.54799844	0.078233307	-7.004669221	0.00	-0.701341499	-0.394655382	-0.701341499	-0.394655382
Year_dummy	-26652.34345	3146.656367	-8.470052126	0.00	-32820.02226	-20484.66464	-32820.02226	-20484.66464
sqft_living15	24.97683259	3.594515552	6.948594942	0.00	17.93131671	32.02234846	17.93131671	32.02234846

Regression Statistics	
Multiple R	0.809245482
R Square	**0.65487825**
Adjusted R Square	0.654638549
Standard Error	215751.3332
Observations	21613

- **R Square**: 0.6549
- **Significant Variables**: bedrooms, bathrooms, sqft_living, floors, waterfront, view, condition, grade, yr_renovated, Year_dummy, sqft_living15
- **Removed Variable**: sqft_lot (p-value > 0.05)

Regression 4: Final Model

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	6232536.007	137409.6103	45.35735159	**0.00**	5963203.027	6501868.987	5963203.027	6501868.987
bedrooms	-39504.38136	2022.848534	-19.5290852	**0.00**	-43469.31382	-35539.4489	-43469.31382	-35539.4489
bathrooms	46680.43704	3439.120883	13.57336326	**0.00**	39939.50622	53421.36786	39939.50622	53421.36786
sqft_living	163.0537659	3.568308126	45.69497927	**0.00**	156.0596185	170.0479132	156.0596185	170.0479132
floors	25242.62736	3430.339992	7.358637168	**0.00**	18518.90774	31966.34698	18518.90774	31966.34698
waterfront	578740.0249	18591.27499	31.1296576	**0.00**	542299.7534	615180.2963	542299.7534	615180.2963
view	43477.33918	2230.985351	19.48795369	**0.00**	39104.4432	47850.23517	39104.4432	47850.23517
condition	20819.53448	2491.555971	8.35603724	**0.00**	15935.90084	25703.16812	15935.90084	25703.16812
grade	119668.4649	2235.139331	53.53959963	**0.00**	115287.4268	124049.503	115287.4268	124049.503
yr_built	-3581.959122	70.5602288	-50.76456217	**0.00**	-3720.262379	-3443.655864	-3720.262379	-3443.655864
yr_renovated	10.92476503	3.908348947	2.795237882	**0.01**	3.26411257	18.58541749	3.26411257	18.58541749
sqft_lot15	-0.562118596	0.055467433	-10.1342096	**0.00**	-0.67083886	-0.453398332	-0.67083886	-0.453398332
Year_dummy	-26612.79292	3146.450549	-8.458036286	**0.00**	-32780.06828	-20445.51756	-32780.06828	-20445.51756
sqft_living15	23.98839679	3.521338227	6.812295566	**0.00**	17.08631391	30.89047967	17.08631391	30.89047967

Regression Statistics	
Multiple R	0.80922472
R Square	**0.654844647**
Adjusted R Square	0.654636905
Standard Error	215751.8466
Observations	21613

- **R Square**: 0.6548
- **Significant Variables**: bedrooms, bathrooms, sqft_living, floors, waterfront, view, condition, grade, yr_renovated, Year_dummy, sqft_living15
- **Removed Variable**: sqft_lot

Conclusion

The final regression model explains approximately 65.48% of the variance in house prices in King County. Key factors affecting house prices include the number of bathrooms, living area square footage, number of floors, waterfront presence, view quality, house condition, construction grade, and recent renovations. These insights can help homeowners, buyers, and real estate professionals make informed decisions.

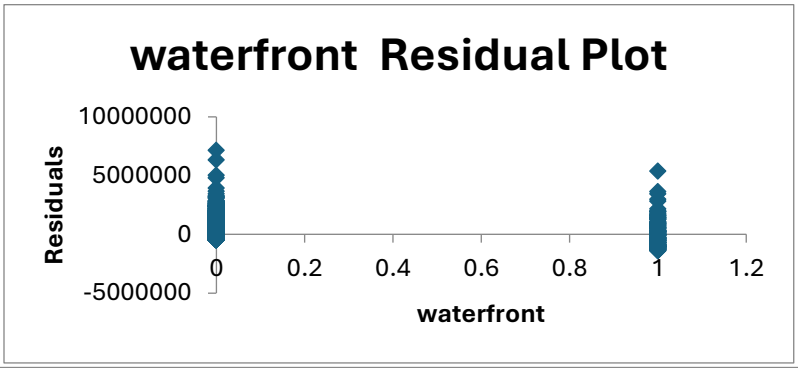
3. Test the following hypotheses and provide your conclusion

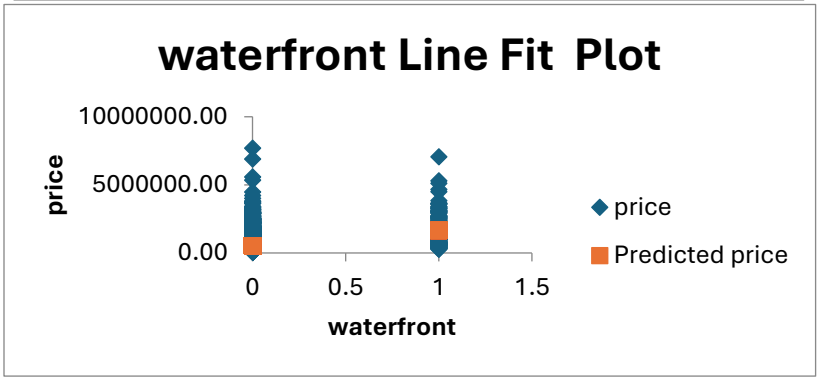
a. Average price of houses with waterfront are **higher** than those without a waterfront.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	531563.5998	2416.194414	220.000343	0	526827.6805	536299.5191	526827.6805	536299.5191
waterfront	1130312.425	27822.46472	40.6258912	0	1075778.342	1184846.508	1075778.342	1184846.508

Regression Results:

- **Intercept**: $531,563.60
- **Waterfront Coefficient**: $1,130,312.43
- **Standard Error**: $27,822.46
- **t-Stat**: 40.63
- **P-value**: 0.00
- **95% Confidence Interval**: $1,075,778.34 to $1,184,846.51





Interpretation:

- The coefficient for the waterfront variable is $1,130,312.43, which is highly significant (P-value = 0.00). This indicates that, on average, houses with waterfronts are priced $1,130,312.43 higher than those without waterfronts.
- The t-statistic of 40.63 is very high, further confirming the significance of the waterfront variable.
- The 95% confidence interval does not include zero, reinforcing the conclusion that waterfront properties are significantly more expensive.

Conclusion:

- The hypothesis that the average price of houses with waterfront is higher than those without a waterfront is strongly supported by the regression results.

b. Older House have Lower price.

Creating the Age Variable:

- The age of each house was calculated by subtracting the year built from the years 2014 and 2015.
- The mean age of the houses is 43 years.

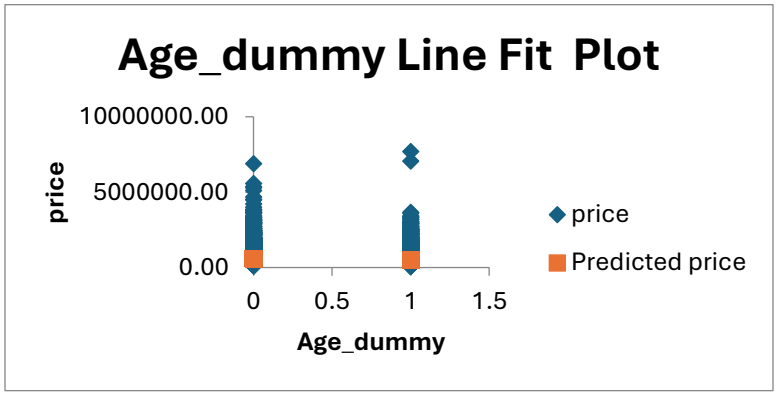
Dummy Variable for Age:

- A dummy variable was created where houses older than 43 years were coded as 1, and those 43 years or younger were coded as 0.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	571124.4609	3426.326101	166.687129	0.00	564408.609	577840.3128	564408.609	577840.3128
Age_dummy	-65615.56941	4981.923346	-13.17073	0.00	-75380.50665	-55850.63218	-75380.50665	-55850.63218

Regression Results:

- **Intercept**: $571,124.46
- **Age Dummy Coefficient**: -$65,615.57
- **Standard Error**: $4,981.92
- **t-Stat**: -13.17
- **P-value**: 0.00
- **95% Confidence Interval**: -$75,380.51 to -$55,850.63



Interpretation:

- The coefficient for the age dummy variable is -$65,615.57, which is highly significant (P-value = 0.00). This indicates that, on average, houses older than 43 years are priced $65,615.57 lower than those 43 years or younger.
- The negative sign of the coefficient supports the hypothesis that older houses have lower prices.
- The t-statistic of -13.17 is very high in absolute value, confirming the significance of the age dummy variable.
- The 95% confidence interval does not include zero, reinforcing the conclusion that older houses are significantly less expensive.

Conclusion:

- The hypothesis that older houses have lower prices is strongly supported by the regression results.