A blue and white logo

Description automatically generated

Business Analytics and Stats Fall 2024

**King County Housing Sales**

**Group**

Aakash Janardhan Jadhav([aj68160n@pace.edu](mailto:aj68160n@pace.edu))

Varun Shah([vs68126n@pace.edu](mailto:vs68126n@pace.edu))

[Ganai Izaan Hamid](javascript://)([ig79632n@pace.edu](mailto: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?**

A screenshot of a computer

Description automatically generated

**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**

A screenshot of a computer screen

Description automatically generated

A yellow line with black text

Description automatically generated

* **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**

A screenshot of a computer screen

Description automatically generated

A yellow line with black text

Description automatically generated

* **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:**

A screenshot of a computer

Description automatically generated

A yellow line on a white background

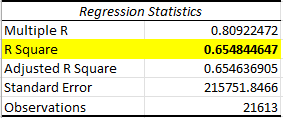
Description automatically generated

* **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**

A screenshot of a computer

Description automatically generated



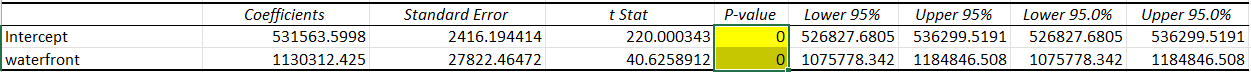
* **R Square**: 0.6548
* **Significant Variables**: bedrooms, bathrooms, sqft\_living, floors, waterfront, view, condition, grade, yr\_renovated, Year\_dummy, sqft\_living15
* **Removed Variable**: sqft\_lot

**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.



**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.

A yellow rectangular object with black text

Description automatically generated

**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.