PHASE 2 PROJECT

OVERVIEW

Business Domain

As home sales are returning to pre-pandemic levels across the United States, the average cost of a home in Western Washington is still much higher than the national average.
In the United States, the average is about $515,000, but in King County, it's about $840,000. According to the Census Bureau, home sales across the country are up around 6% over this year.
Buyers are being very specific right now because the housing market inventory is low and interest rates are high.

Business Case

The local agents usually face such questions from their clients as: whether home certain features such as number of bathrooms, number of bedrooms, living area, area above and year built have an influence on the price of houses.

OBJECTIVES

REAL ESTATE AGENCY

- To identify the relationship between different features and price using data visulaization tools such as scatter plots.
- To identify the features that have direct influence on the prices of houses by checking for those with with high correlation to the price.
- To identify the locations with the highest sales prices.
- To observe seasonal trends in the data in regards to house prices.
- To predict prices of houses depending on the features using linear regression analysis, to determine the significance and impact of each independent variable on the dependent variable, i.e., the house price.

PROBLEM STATEMENT

A real estate agency wants to analyze the factors that influence the prices of houses in order to provide accurate pricing estimates to their clients. The agency aims to understand the relationship between various features of a house, such as the number of rooms, living area, basement area, overall quality, and other relevant factors, and how they affect the sale price.

DATA UNDERSTANDING

Understanding the columns in our data set

- **id** - Unique identifier for a house
- **date** - Date house was sold
- **price** - Sale price (prediction target)
- **bedrooms** - Number of bedrooms
- **bathrooms** - Number of bathrooms
- **sqft_living** - Square footage of living space in the home
- **sqft_lot** - Square footage of the lot
- **floors** - Number of floors (levels) in house
- **waterfront** - Whether the house is on a waterfront
- **view** - Quality of view from house
- **condition** - How good the overall condition of the house is. Related to maintenance of house.
- **grade** - Overall grade of the house. Related to the construction and design of the house.
- **sqft_above** - Square footage of house apart from basement
- **sqft_basement** - Square footage of the basement
- **yr_built** - Year when house was built
- **yr_renovated** - Year when house was renovated
- **zipcode** - ZIP Code used by the United States Postal Service
- **lat** - Latitude coordinate
- **long** - Longitude coordinate
- **sqft_living15** - The square footage of interior housing living space for the nearest 15 neighbors
- **sqft_lot15** - The square footage of the land lots of the nearest 15 neighbors

Further understanding of columns in the data set

Data columns : 21

Number of data entries : 21597

Number of data types : 3

Column	**Non- null count**	**Data type**
id	21597	Integer
Date	21597	Object
Price	21597	Float
Bedrooms	21597	Integer
Bathrooms	21597	Float
Sqft_living	21597	Integer
Sqft_lot	21597	Integer
Floors	21597	Float
Waterfront	19221	Object
View	21534	Object
Condition	21597	Object
Grade	21597	Object
Sqft_above	21597	Object
Sqft_basement	21597	Object
Yr_built	21597	Integer
Yr renovated	21597	Float
Zip code	21597	Integer
Latitude	21597	Float
Longitude	21597	Float
Sqft_living15	21597	Integer

Sqft_lot15	21597	Integer

Data cleaning

Null Values

Looking at the information above we can see only three columns have missing values, that is; "waterfront", "view" and "yr_renovated".
Every house has its own unique features and not all are the same. Some houses contain certain features while others lack them.
Since this is real world data, we can account for missing values in "waterfront" and "view" columns by saying not all houses are built the same and those lacking the two features have caused our data on the two columns to be inconsitent with the rest of the other columns.
The "yr_renovated" column can also be accounted for by saying not all houses undergo renovation. Houses built earlier might need renovation but recent houses do not require renovation hence the missing values in the column

Becase the null values were not too many, we found it best to replace them with the mode value instead of dropping them entirely.

Duplicated Data

There were no duplicated values in our data set.

DATA ANALYSIS

Upon analyzing the King County house sales data, we will help local real estate agency to answer the questions from their clients, i.e. homeowners. We will help figure out what are the key features that determine the sales price of houses.

Data Visualization

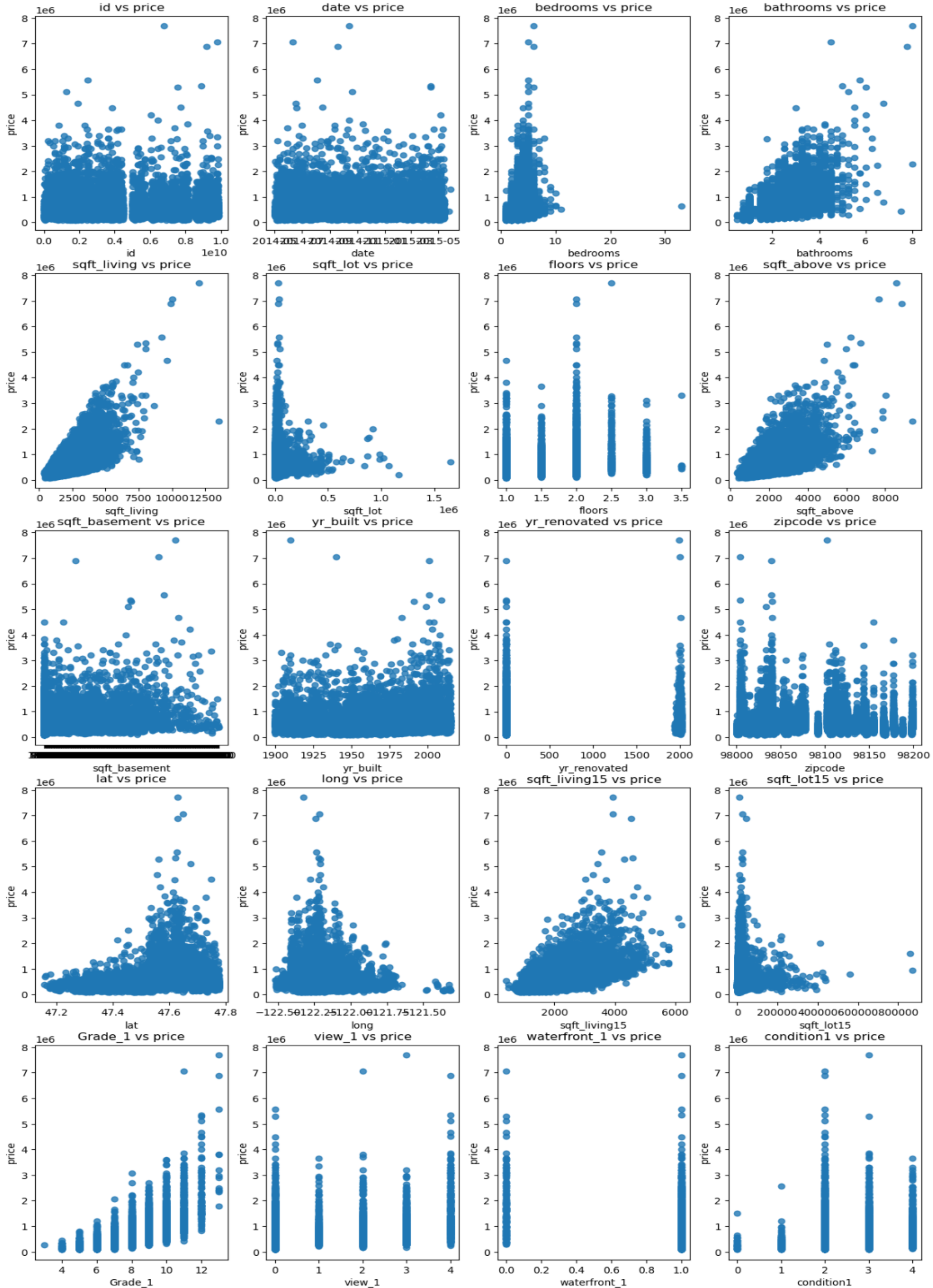
1. **Scatter Plot**

A scatter plot is a graph that displays the relationship between two variables. Each point on the graph represents a pair of values, one for each variable.
Scatter plots are commonly used to identify trends or patterns in data and to determine the strength and direction of the relationship between the two variables.
The horizontal axis is usually used to represent the independent variable, while the vertical axis represents the dependent variable.

The resulting plot can show if there is a correlation between the variables and if it is positive or negative.

We created a function to plot out scatter plots that compared each feature in the data set with price so as to see clearly the relationship between each
The function gave us the following results:

Scatter plot of Independent variables vs price



id vs price

date vs price

bedrooms vs price

bathrooms vs price

sqft_living vs price

sqft_lot vs price

floors vs price

sqft_above vs price

sqft_basement vs price

yr_built vs price

yr_renovated vs price

zipcode vs price

lat vs price

long vs price

sqft_living15 vs price

sqft_lot15 vs price

Grade_1 vs price

view_1 vs price

waterfront_1 vs price

condition1 vs price

Conclusions drawn from the scatter plots

1. Id vs price
   - No correlation
2. Date vs price
   - No correlation
3. Bedroms vs price
   - Very steep linear correlation
4. Bathrooms vs price
   - Positive correlation
5. Sqft living vs price
   - Positive correlation
6. Sqft lot vs price
   - *** correlation
7. Floors vs price
   - No correlation
8. Waterfront vs price
   - No correlation
9. Views vs price
   - No correlation
10. Condition vs price
    - No correlation
11. Grade vs price
    - No correlation
12. Sqft above vs price
    - Positive correlation
13. Sqft basement vs price
    - No correlation
14. Yr built vs price
    - No correlation
15. Yr renovated vs price
    - No correlation
16. Zipcode vs price
    - No correlation
17. Latitude vs price
    - No correlation
18. Longitude vs price
    - No correlation
19. Sqft living15 vs price

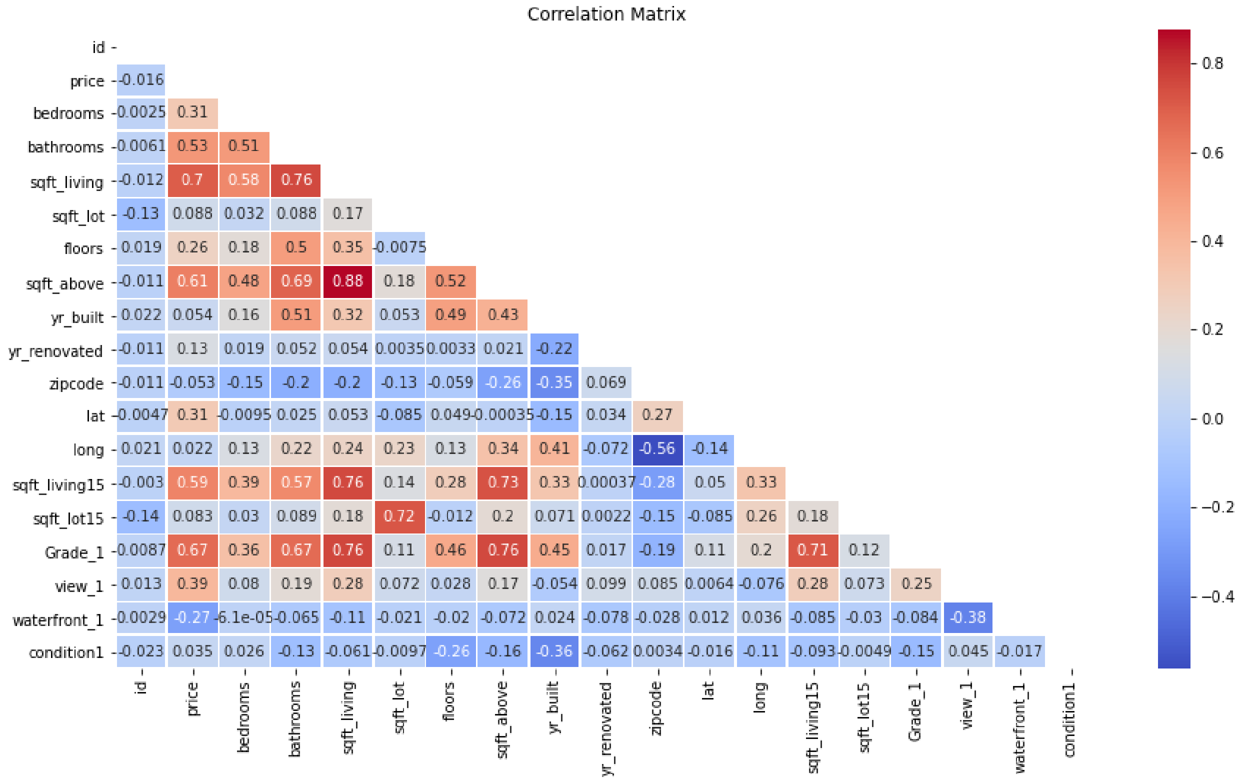
- Positive correlation

20. Sqft lot15 vs price
    - No correlation

2. **Heat Map**

Heat maps are graphical representations of data where values in a matrix are represented by colors. They are used to visualize data patterns and relationships. Heat maps are particularly useful when dealing with large datasets or when trying to identify patterns or trends in data.
In a heat map, each cell in the matrix is assigned a color based on its value. Typically, a color gradient is used, with low values represented by lighter colors and high values represented by darker colors. This color coding allows for easy identification of patterns and variations in the data.

We created a heat map to visualize further the correlation between the features
The results were:



	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	sqft_above	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15	Grade_1	view_1	waterfront_1	condition1
id																			
price	-0.016																		
bedrooms	-0.0025	0.31																	
bathrooms	-0.0061	0.53	0.51																
sqft_living	-0.012	0.7	0.58	0.76															
sqft_lot	-0.13	0.088	0.032	0.088	0.17														
floors	0.019	0.26	0.18	0.5	0.35	-0.0075													
sqft_above	-0.011	0.61	0.48	0.69	0.88	0.18	0.52												
yr_built	0.022	0.054	0.16	0.51	0.32	0.053	0.49	0.43											
yr_renovated	-0.011	0.13	0.019	0.052	0.054	0.0035	0.0033	0.021	-0.22										
zipcode	-0.011	-0.053	-0.15	-0.2	-0.2	-0.13	-0.059	-0.26	-0.35	0.069									
lat	-0.0047	0.31	-0.0095	0.025	0.053	-0.085	0.049	-0.00035	-0.15	0.034	0.27								
long	0.021	0.022	0.13	0.22	0.24	0.23	0.13	0.34	0.41	-0.072	-0.56	-0.14							
sqft_living15	-0.003	0.59	0.39	0.57	0.76	0.14	0.28	0.73	0.33	0.00037	-0.28	0.05	0.33						
sqft_lot15	-0.14	0.083	0.03	0.089	0.18	0.72	-0.012	0.2	0.071	0.0022	-0.15	-0.085	0.26	0.18					
Grade_1	-0.0087	0.67	0.36	0.67	0.76	0.11	0.46	0.76	0.45	0.017	-0.19	0.11	0.2	0.71	0.12				
view_1	0.013	0.39	0.08	0.19	0.28	0.072	0.028	0.17	-0.054	0.099	0.085	0.0064	-0.076	0.28	0.073	0.25			
waterfront_1	-0.0029	-0.27	-6.1e-05	-0.065	-0.11	-0.021	-0.02	-0.072	0.024	-0.078	-0.028	0.012	0.036	-0.085	-0.03	-0.084	-0.38		
condition1	-0.023	0.035	0.026	-0.13	-0.061	-0.0097	-0.26	-0.16	-0.36	-0.062	0.0034	-0.016	-0.11	-0.093	-0.0049	-0.15	0.045	-0.017	

Conclusions drawn from the heat map

1. Sqft living had the highest positive correlation to price with 70%
2. Sqft above had a positive correlation of 61%

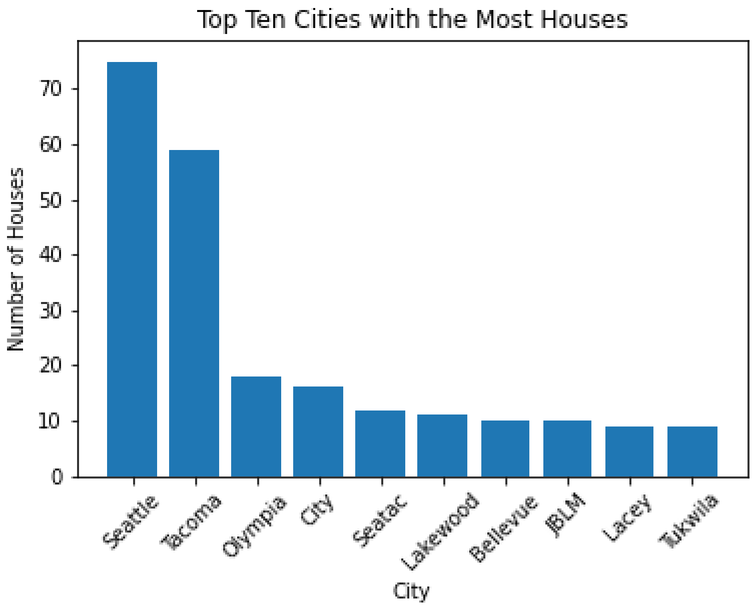
3. Bathrooms had a positive correlation of 53%
4. Bedrooms has a positive correlations of 31%
5. Year built has a negative correlation to price of 5%

PRICE VS LOCATION

To identify the relation between price and locations of the houses, we analyzed data form an outside source and created bar graphs.

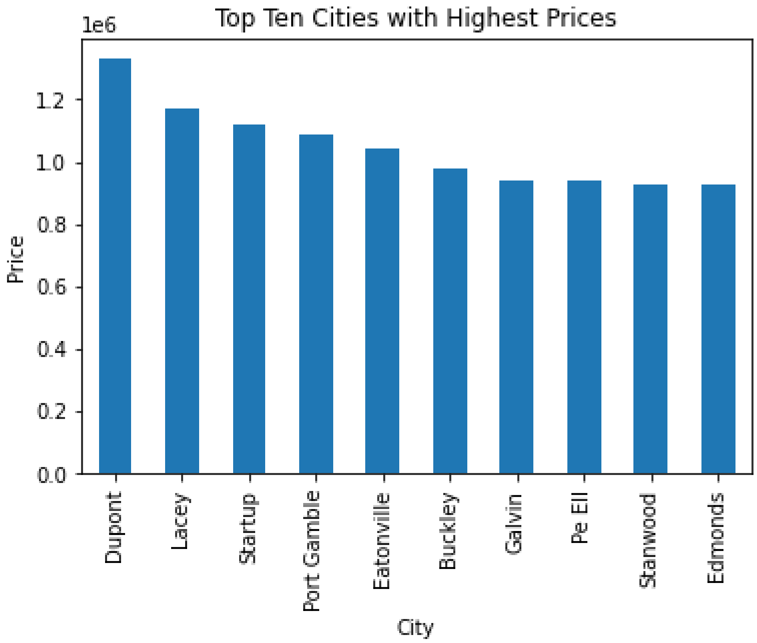
Top 10 cities with the highest number of houses

1. Seattle
2. Tacoma
3. Olympia
4. City
5. Seatac
6. Lakewood
7. Bellevue
8. JBLM
9. Lacey
10. Tukwila

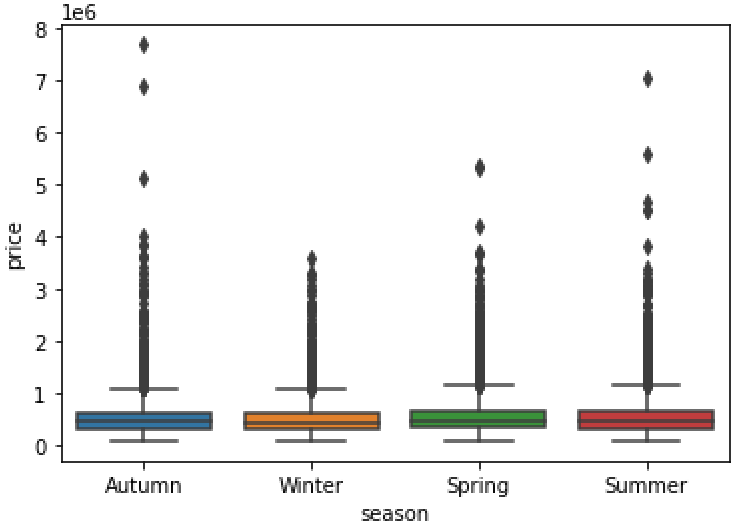


Top 10 cities with the highest sales prices

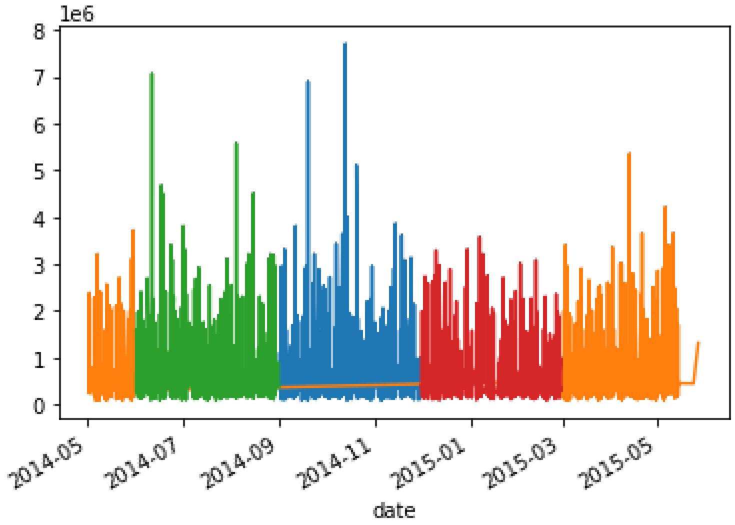
1. Dupont
2. Lacey
3. Startup
4. Port Gamble
5. Eatonville
6. Bucklery
7. Galvin
8. Pe Ell
9. Stanwood
10. Edmonds



SEASONAL TRENDS VS PRICE



From this we can generally see that Spring has the highest mean price whereas whereas Winter has the lowest.



From this we can see that generally Spring and Autumn have the highest sales prices.

LINEAR REGRESSION

By using linear regression analysis, the agency seeks to determine the significance and impact of each independent variable on the dependent variable, i.e., the house price. The goal is to develop a predictive model that can accurately estimate house prices based on the given features, enabling the agency to offer informed pricing recommendations to their clients.

Since we now have a better understanding of the correlation between our target("price") and our features("independent variables"), we proceed to building regression models to further understand the magnitude our features have on price and predict whether this model can give us accurate house prices when fitted with the said features. We will explore a few features from our data set which we have investigated and come to a conclusion that they have a significance on our target.Steps involved here are as follows;

1. Feature Selection
2. Model Selection
3. Model training
4. Model evaluation
5. Model interpretation
6. Model validation and testing
7. **feature engineering

MODEL 1

1. **Feature Selection**
   As per our correlation analysis from the above heat map, we concluded that “Sqft living” had the highest correlation to price.
   X variable (independent variable) - “Sqft living”
   Y variable (dependent variable) - “Price”

2. **Model Selection**
   Therefore, we used simple linear regression compared this two features to create our baseline model

3. **Model Training**
   In the train_test_split function we used, we split 80% of the data into training set and 20% of the data into test set.

4. **Model Evaluation**
   R- squared value = 0.497
   P- value = 0.00

Constant = -5.245e+04
Slope = 285.1471

5. **Model interpretation**
   The model is statistically significant overall, with an F-statistic p-value well below 0.05.
   The model explains about 50% of the variance in price.
   The model const and sqft_living coefficients are both statistically significant, with t-statistic p-values well below 0.05.
   If a house had 0 sqft of living , we would expect the price to be about -43,990.
   For each increase of 1 sq ft in the living , we see an associated increase in price of about 280.

6. **Model validation and testing**
   We now have our baseline model which we created using the train sets obtained from splitting our data.
   To analyze our model further we will predict our target("price") using the trained model then compare metrics to determine if our model is efficient or we need to adjust it.
   We also check if our model is under fitted or over fitted. To perform comparison we will import another module from sci-kit learn

   After running our prediction model and comparing the r-squared values from the test model and the prediction model we can note a difference.
   Our training model has a R-squared value of almost 49% while our test model has a value of 46%

MODEL 2

1. **Feature Selection**
   To better our regression model we decided to use more features to be able to make better predictions.
   From our correlation analysis results we chose to use:
   X variables (independent variables) = “Bedroom”, “ Bathroom”, “Sqft living”, “Sqft above”, “Yr built”
   Y variable (dependent variables) = “ Price”

2. **Model Selection**
   Therefore, we used multiple linear regression to compare these features to create our model

3. **Model training**
   In the train_test_split function we used, we split 60% of the data into a training set and 40% of the data into a test set.

4. **Model evaluation**
   R-squared value = 0.545
   P- value = a) Bedroom = 0.00
   b) Bathroom = 0.00
   c) Sqft living = 0.00
   d) Sqft above = 0.00
   e) Yr built = 0.00

   Constant = 6.27e+06
   Slope = a) Bedroom = -6.964e+04
   b) Bathroom = 7.593e+04
   c) Sqft living = 279.5577
   d) Sqft above = 38.0934
   e) Yr built = -3199.2205

5. **Model interpretation**
   The model includes five independent variables, namely bedrooms, bathrooms, square footage of living area, square footage of the above-ground area, and year built.
   The dependent variable is the price of the house.
   The R-squared value of 0.545 indicates that 54.5% of the variance in house prices can be explained by the independent variables.
   The F-statistic of 3100 suggests that the model is statistically significant.
   The p-value of 0.00 indicates that the model is a good fit for the data.

6. **Model validation and testing**
   To analyze our model further we will predict our target("price") using the trained model then compare metrics to determine if our model is efficient or we need to adjust it.

   After running our prediction model and comparing the r-squared values from the test model and the prediction model we can note a difference.
   Our training model has a R-squared value of almost 54%% while our test model has a value of 56%
   Our conclusion is that the prediction test was a success

3. POLYNOMIAL TRANSFORMATION OF FEATURES

We then used a polynomial transformation to see if our model will improve or not.
One would use a polynomial transformation in regression models to capture non-linear relationships between variables.
While a linear regression assumes a linear relationship, sometimes the relationship between the independent and dependent variables is better represented by a polynomial equation.
By transforming the independent variable into higher-degree polynomials (e.g., quadratic, cubic), we can capture more complex patterns and improve the model's fit to the data.
This can help us better understand the relationship and make more accurate predictions.

1. **Model Evaluation**
   R- squared = 0.719
   Adjusted R-squared = 0.719
   F-statistic = 815.8
   P- value = x1 = 0.00
   x2 = 0.00
   x3 = 0.017
   x4 = 0.00
   x5 = 0.00
   x6 = 0.00
   x7 = 0.021
   x8 = 0.00
   x9 = 0.003
   x10 = 0.00
   x11 = 0.025
   x12 = 0.00
   x13 = 0.764
   x14 = 0.00
   x15 = 0.00
   x16 = 0.00
   x17 = 0.067
   x18 = 0.00
   x19 = 0.00

x20 = 0.00
Constant = 8.921e+07

2. **Model interpretation**
This indicates that the model explains about 72% of the variance in house prices.
The F-statistic shows the overall model is statistically significant.
The coefficients provide insight into how each feature impacts predicted price.
Living square footage had the highest positive coefficient, indicating it strongly increases predicted price.
Number of bathrooms also had a large positive coefficient.
Interestingly, being on the waterfront had a negative coefficient, suggesting it decreases predicted price after controlling for other features.
Higher grades and view ratings increased predicted prices.

CONCLUSION

RECOMMENDATIONS

1. The agency should be on the lookout for features such as square footage of the living area, square footage above, number of bedrooms, number of bathrooms and the year the house was built when advising and valuing house for homeowners because they have strong correlations to price.

2. The agency should be on the lookout for houses in the areas: Seattle, Tacoma, Olympia, City, Seatac, Lakewood, Bellevue, JBLM, Lacey, Tukwila because they have the highest number of houses

3. When advising homeowners, the agency should be aware that the areas: Dupont, Lacey, Startup, Port Gamble, Eatonville, Bucklery, Galvin, Pe Ell, Stanwood, Edmonds

4. The agency should be aware that the Spring season generally demands higher prices for the houses. They should therefore advice home sellers to enlist their homes during this season to take advantage of the seasonal demands.

5. The agency should be aware that the Summer season generally demands lower prices for the houses. They should therefor advice potential home buyers to make their purchases during this season

NEXT STEPS

1. The agency should look for more data in regards to other features contained in a house so as to make more accurate predictions
2. The agency should conduct surveys to look find specific factors that cause this seasonal variation so as to understand the market better
3. The agency should conduct research to find location specific data, such as social amenities, neighborhoods and political stability to understand why certain areas command higher prices as compared to others