Phase 2 Project

Group 13:

Sharon Kaliku
Ezra Kipchirchir
Paul Kamau
Heri Kimotho
Kipkosgei Kiptui

King County, Washington

The Business Case

- A real estate company’s need to help its clients understand how prices of houses vary
- The clients: 1. Homeowners
  2. Potential house buyers

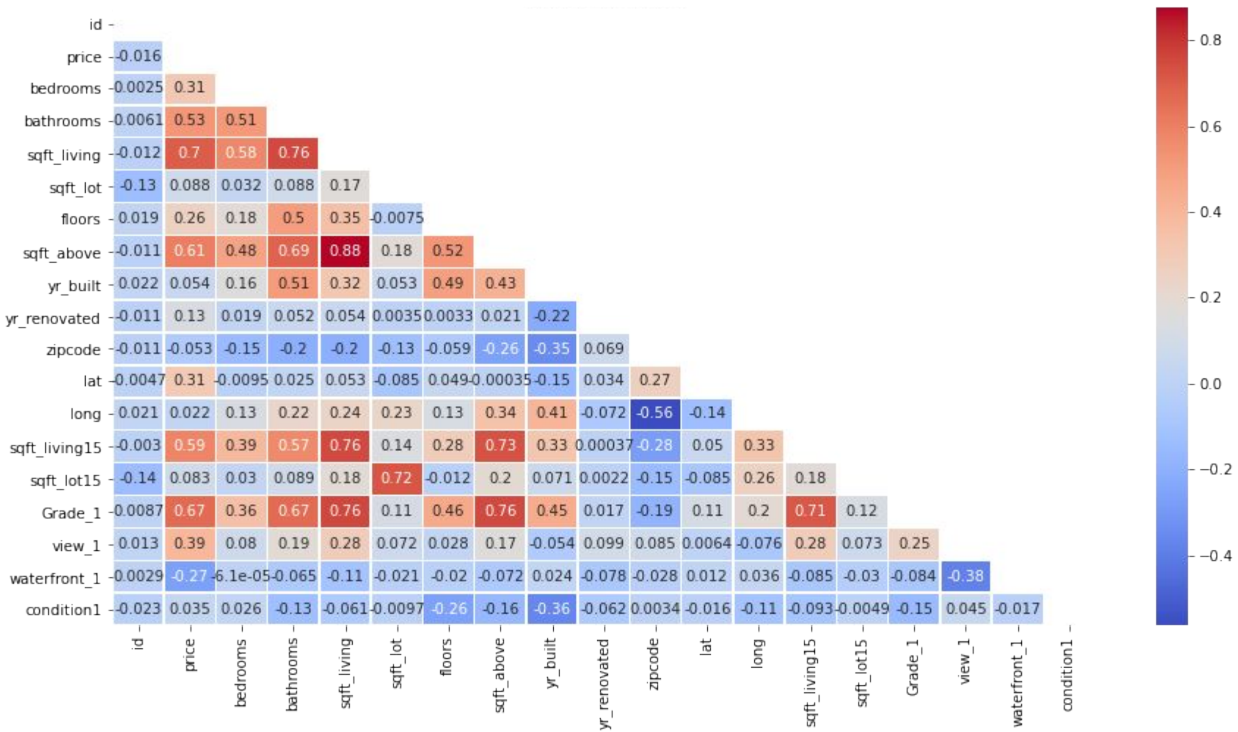
Objectives

- Analyze relationship between location and price of houses
- Analyze seasonal trends of price
- Predict prices of houses depending on features

Data Understanding

- Sources of data for analysis:
  1. KC house data - Various features against price
  2. Delivery locations - Zip-codes and their corresponding cities

Data Visualization



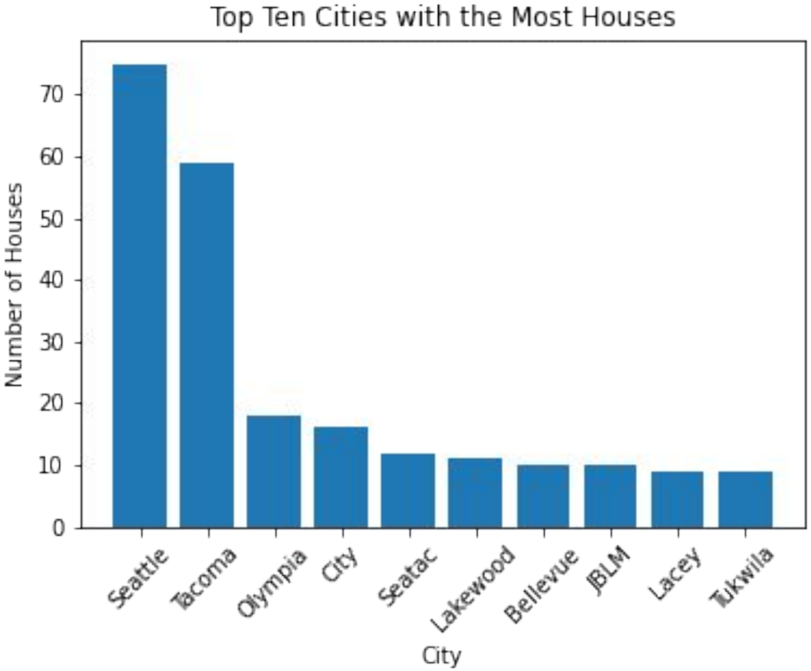
Correlation Matrix

	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	

Data Analysis

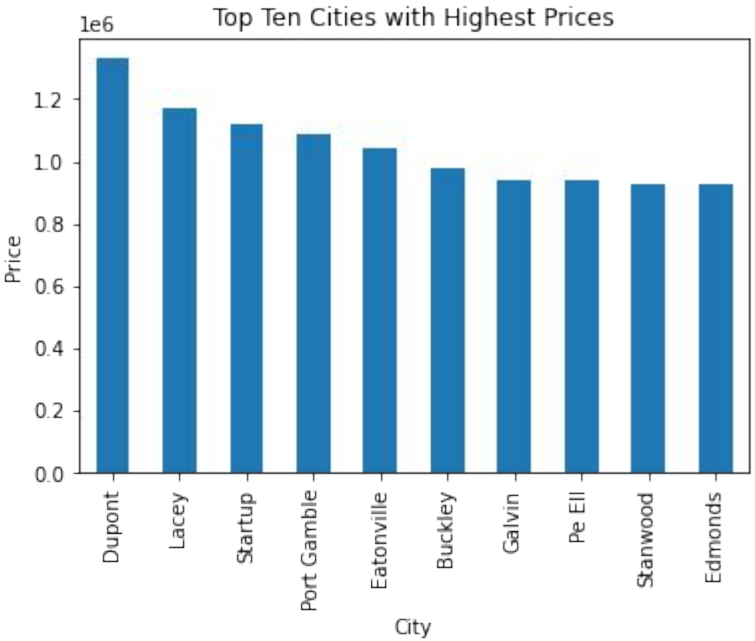
- Visualized top cities with the most houses.
- Showed the top ten cities with the highest prices.
- Analyzed seasonal trends in prices
  - ● Created seasonal variables
  - ● Visualized trends and differences
- ● Performed ANOVA test for seasonality

PRICE VS LOCATION



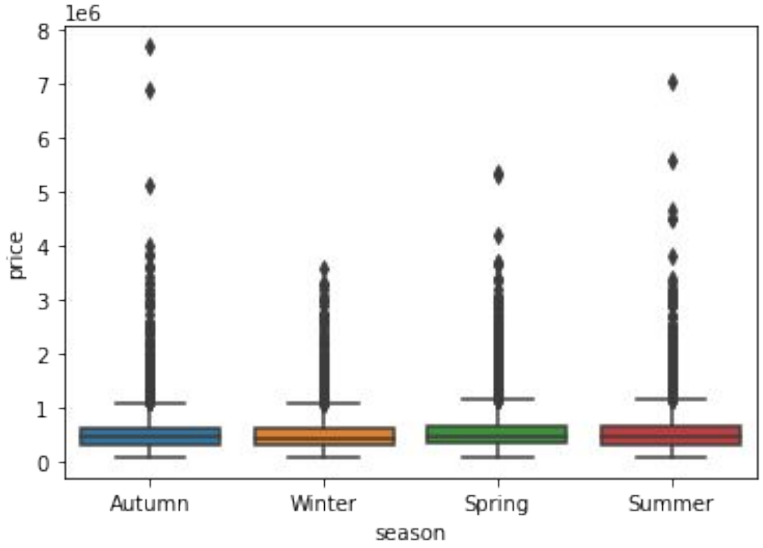
- Seattle has the most houses
- Tacoma is a close second

PRICE VS LOCATION


Top Ten Cities with Highest Prices

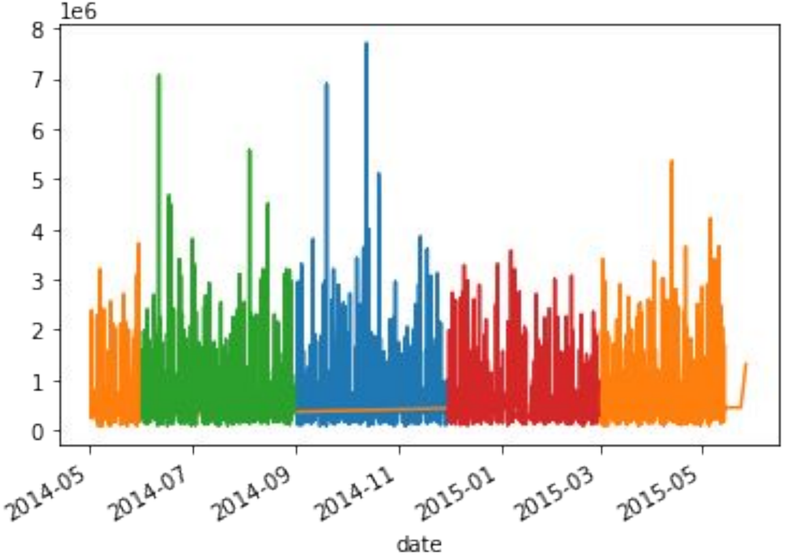
- Dupont has houses with the highest prices
- Lacey is a close second

Seasonal Trends vs. Prices



- Spring has the highest mean price
- Winter has the lowest mean price

Seasonal trends vs. Prices



- Spring and Autumn have the highest sales prices.

Linear Regression -Baseline Model

- Detailing the selection of the target and features for the baseline model.
- Explaining the process of splitting the data into training and test sets.
- Providing the model summary and interpreting the findings:
  - F-statistic and R-squared.
  - Interpretation of coefficients.
- Displaying the evaluation metrics:
  - Mean Absolute Error (MAE).
  - Mean Squared Error (MSE).
  - R-squared values for training and test sets.

Model Evaluation

- Metrics for model evaluation
- Compared R-squared values for training and test models
- Visualized residuals for normality

Log Transformations

- Discussed the need for target transformation
- Performed a log transformation on the target
- Visualized the transformed target
- Created a new model with the log-transformed target
- Model summary and evaluation

2nd Model (Multiple Linear Regression

- Feature selection
- Standardization of data
- Model creation and summary
- Model evaluation and metrics

Polynomial Transformation

- Introduction to polynomial transformation
- Transformation of features
- Model creation and summary for polynomial regression
- Analysis of coefficients and p-values

CONCLUSION RECOMMENDATIONS

1. The agency should be on the lookout for features such as square footage of the living area, square footage above,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 and Eatonville.

CONCLUSION RECOMMENDATIONS

4. The agency should be aware that the Spring season generally demands higher prices for the houses.

5. The agency should be aware that the Summer season generally demands lower prices for the houses.

NEXT STEPS

1. The agency should look for more data in regards to other house features.

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 other