Phase 2 Project

Beatrice Kariuki - Lead
Rodgers Odhiambo
Mohamed Ali
Ahmed HAJI
Peter Muhia
Emmanuellar Karisa

Predicting House Prices in King County.

The main goal of this project is to create a model that can predict and estimate house prices in King County, Washington, USA, using various property features. The objective is to provide useful information through a comparative market analysis, which can benefit real estate analysts, investors, data scientists, machine learning practitioners, financial institutions, potential home buyers, and prospective tenants, giving them the ability to make informed decisions. The dataset used for this project is the King County House dataset.

Challenges in the Real Estate Industry

1 Runaway Inflation and Interest Rates

The interplay of factors such as runaway inflation, high interest rates, and bank stress is limiting consumer purchasing power, negatively affecting investment earnings in the area.

2 Intense Competition

The ever-increasing competition in the industry poses challenges in standing out and capturing the attention of potential leads.

3 Precise Property Pricing

Setting the right price from the beginning is crucial for a property to sell quickly and profitably.

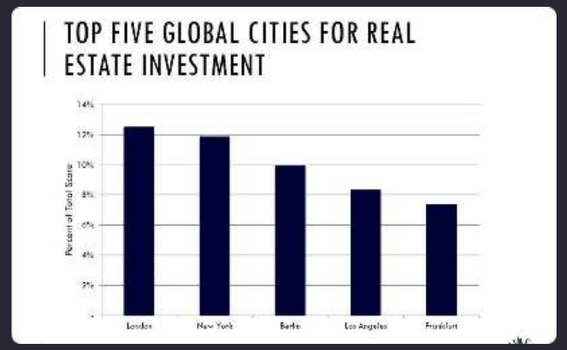
4 Demographic Shifts

Changes in population characteristics significantly influence demand, preferred locations, and overall market trends.

Stakeholders in the Real Estate Industry

Real Estate Analysts

They provide valuable insights and analysis to support decision-making processes.



Real Estate Investors

Individuals or institutions who invest in real estate to generate income.

Real Estate Agents

They facilitate property transactions between buyers and sellers.

Data Understanding

Dataset

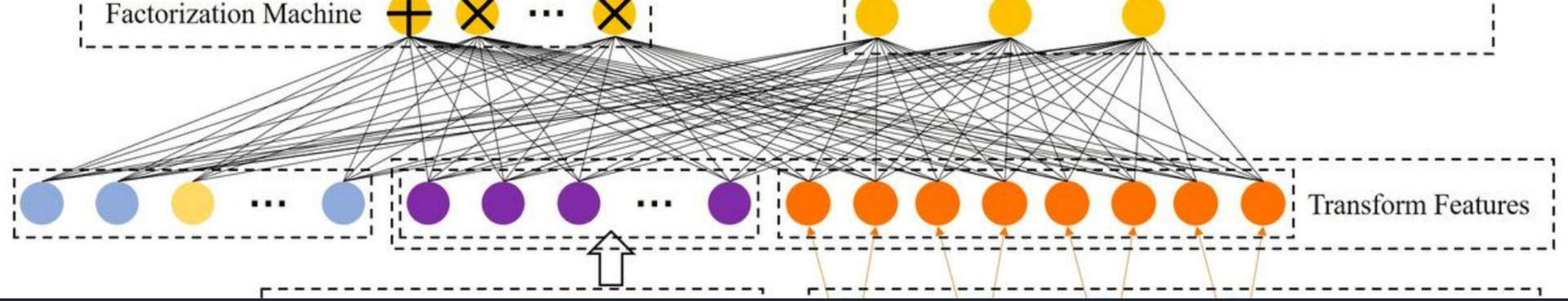
King County House Sales dataset with 20 columns and over 21500 records.

Selected Columns

Price, bedrooms, bathrooms, sqft living, condition, yr built, sell yr, house age.

Statistical Summary

Mean house price: USD 540297. No missing values. Standard deviation: USD 367368.



Feature Engineering

1

Date Conversion

Converting date to datetime format and extracting the year for sell year.

2

Categorical Columns

Encoding categorical columns for the linear regression model.

Univariate Analysis

1. **Histogram**

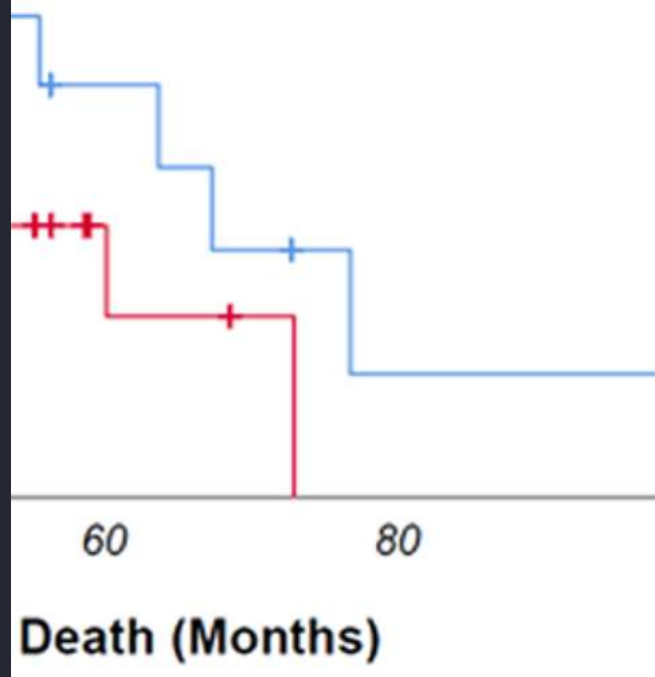
   Shows positively skewed distribution of house prices.

2. **Summary Statistics**

   Provides details on distribution of price and variability around the mean.

3. **Box Plot**

   Confirms the positively skewed distribution of house prices.



Bivariate Analysis

1

Scatter Plots

Show positive relationship between independent variables and house price.

2

Correlation Analysis

Examines the relationship between the target variable and other numeric features.

Multivariate Analysis

Heatmap

Visualizes the correlation matrix of features and their relationship to house price.

Modeling

1

Base Model

The simple linear regression model explains 49.3% of the price variation. The coefficient estimates show the relationship between the independent variable and the dependent variable.

2

Visualization

Plotting the model fit to assess the linearity of the relationship between the house price and the square footage of living space.

Analyzing the Models

Simple Linear Regression

The relationship between the independent variable (sqft_living) and the dependent variable (price) is statistically significant with an R-squared of 0.493.

Multiple Linear Regression

Utilizing multiple features to predict house prices resulted in a statistically significant model with an R-squared of 0.465.

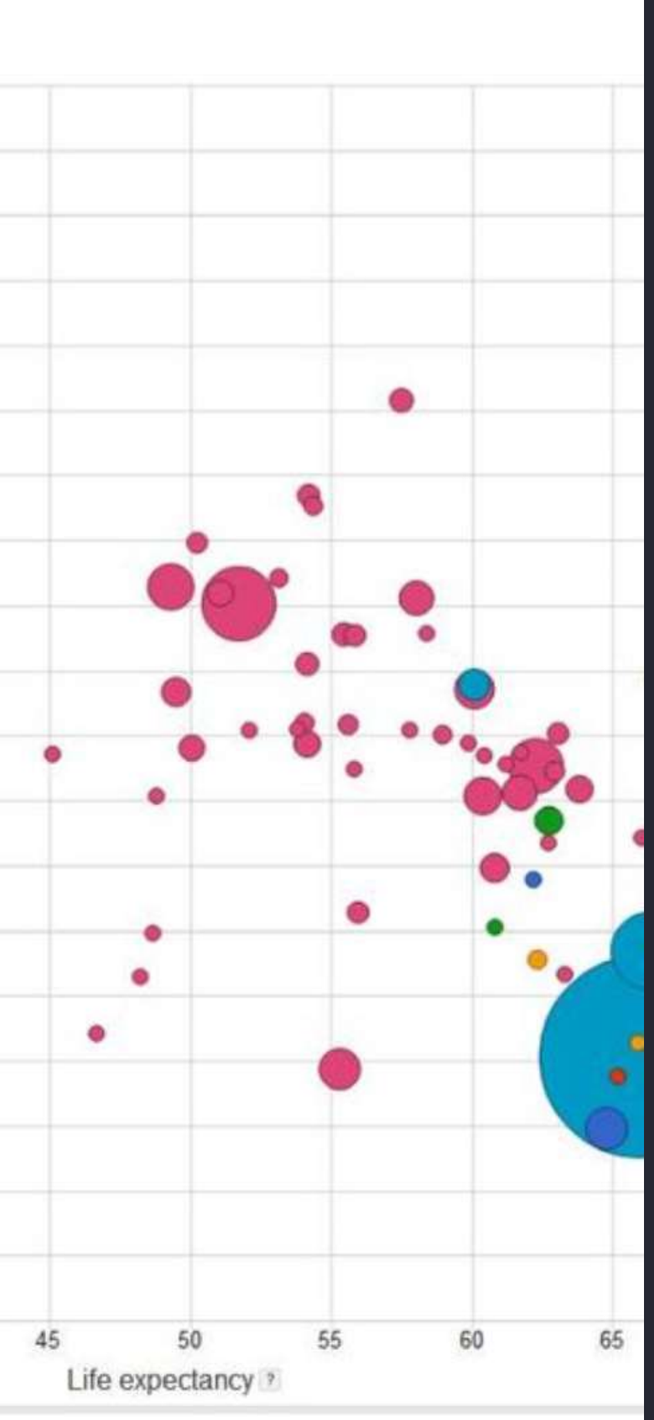
Modeling
Multiple Linear Regression

1 Model Results

Model 2 with multiple features has the highest R-squared value of 46.5%, indicating that it explains a larger proportion of the variance in the dependent variable compared to the other model.

2 Model Evaluation

The model's root mean squared error (RMSE) of approximately USD 253518.9 highlights the model's average prediction deviation from the actual prices.



Visualizing the Regression Line

1

Simple Linear Regression

The plot helps assess the linearity of the relationship between house price and living space.

2

Multiple Linear Regression

The visualization depicts the correlation between various predictors and housing prices.

Model Evaluation

Model 1

MAE: 0, RMSE: 0.4

Model 2

MAE: 0, RMSE: 0.3

Final Model Evaluation

The final model shows a Prob (F-statistic) of 0.000, indicating its statistical significance with an R-squared of 0.465.

Benefits of Informed Decisions

1

Data-Driven Strategy

Accessible comparative market analysis supports informed decisions and increases the likelihood of successful transactions.

2

Efficient Investment

Having accurate pricing information helps investors make efficient and profitable real estate investment decisions.

3

Enhanced Market Knowledge

Accessing data on property features and trends provides comprehensive market knowledge for all stakeholders.

Next Steps and Recommendations

1 Strategic Pricing

Implement a precise pricing strategy based on data and market knowledge to ensure competitive advantage.

2 Continuous Analysis

Regularly analyze market trends and data to adapt strategies and stay ahead in a dynamic real estate landscape.

3 Client-Centric Approach

Focus on providing personalized services tailored to the needs and preferences of potential buyers and tenants.

Conclusions

Runaway Inflation

High interest rates, bank stress, and tightening liquidity are impacting investment earnings in the area.

Competition

Standing out and capturing the attention of potential leads can be a challenging endeavor due to increasing competition.

Pricing Properties

Setting the right price for a property is crucial for a quick and profitable sale.

Demographics

Change in demographics can significantly impact the real estate industry, influencing demand and market trends.

THANK YOU