King County House Sales Analysis

Analysis, findings and recommendations

Presented by: Emily Owiti
Email: emily.owiti@student.moringaschool.com

Problem statement

The company wants to optimise the sale prices of the properties based on factors that are strongly associated with driving house prices

The company wants to:

- Identify the variables affecting house prices
- Create a linear model that quantitatively relates house prices with variables
- Know the accuracy of the model, i.e. how well these variables can predict house prices

Context - Real Estate Industry Overview

Key factors influencing price

- **Supply-Demand Dynamics:** Limited inventory boosts prices, while oversupply can lead to declines.
- **Economic Conditions:** Job growth, income levels, and consumer confidence impact housing demand and prices.
- **Interest Rates:** Lower rates stimulate demand and support higher prices, while higher rates may dampen demand.
- **Demographic Changes:** Population growth, migration, and generational preferences influence regional demand.
- **Location-Specific Factors:** Proximity to amenities, schools, transportation, and economic hubs contribute to price differentials.

Project Objectives

1. Build a model that accurately predicts house prices in King County
2. Provide top recommendations given model output

Data sources

King County house sales dataset

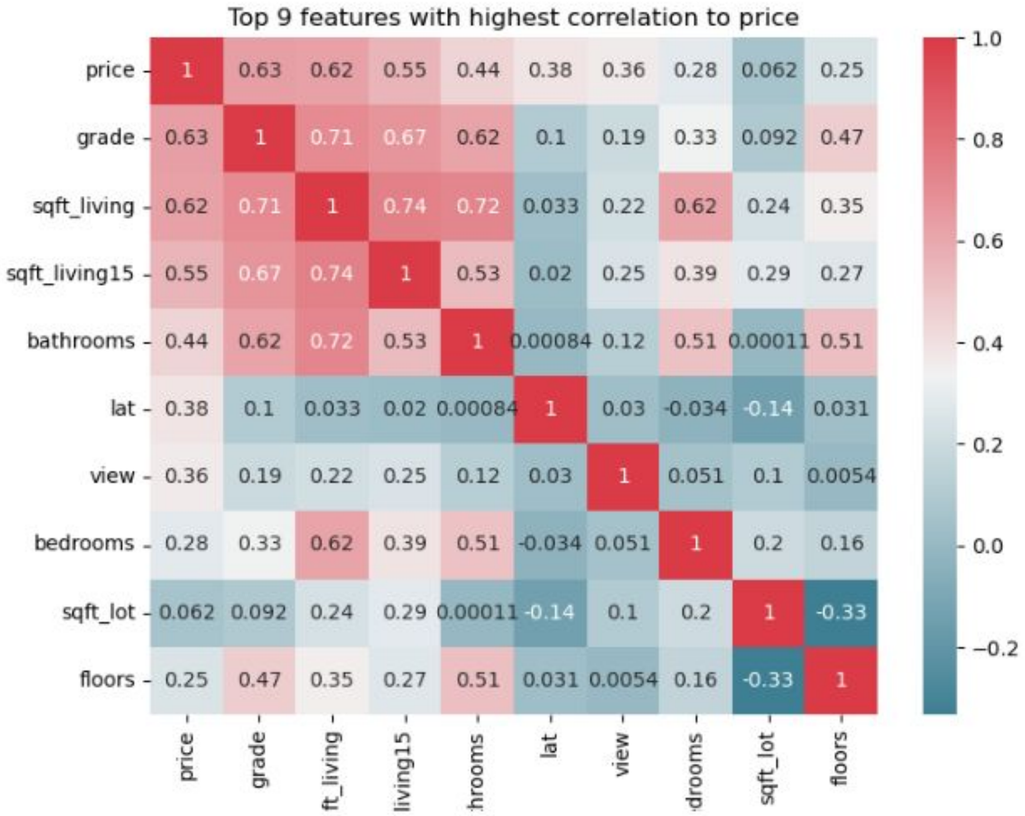
Data analysis approach

1. Loading the data to pandas and analyzing the dataframes
2. Cleaning the data by checking & handling:
   - Duplicates
   - Missing data
   - Anomalies
   - Invalid data
   - Other additional data cleaning procedures as needed
3. Performing exploratory analysis
4. Modelling
5. Drawing conclusions and making recommendations

Observations & Conclusions

See next pages

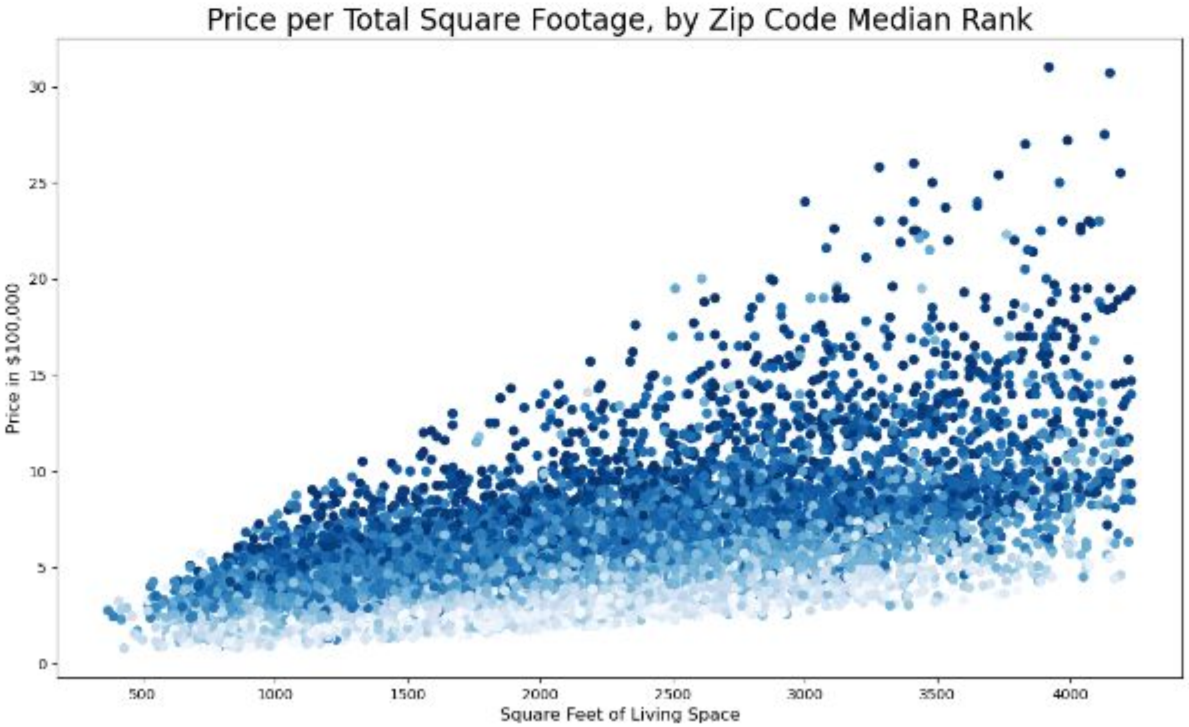
1. Top features with strongest correlation to price



- The raw data had 19 independent features

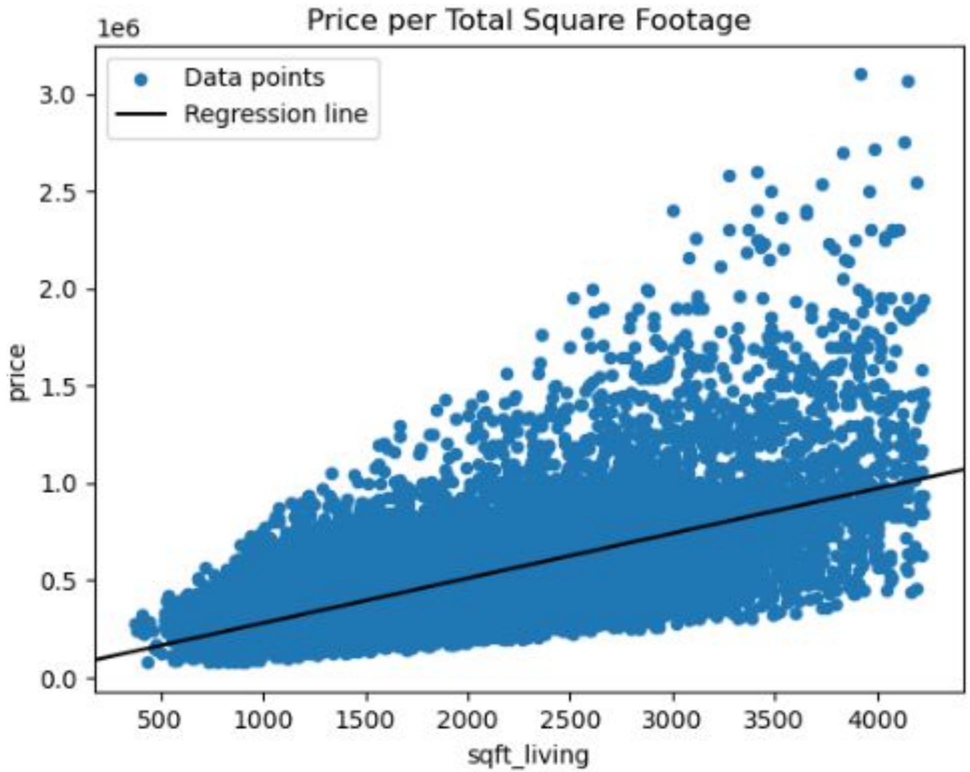
- The heatmap shows the top 9 features and the degree of correlation to price - hence most of these were used in modelling.

2. Relationship between price, sqft_living and zip code



- While the raw zip code feature showed limited correlation to price, when grouped by median price and ranked, they show a strong correlation with price.
- The zip code feature was utilized in modelling due to limited time, its important to note that its a key feature that should be further engineered and utilized in future.

3. Model 1 - Measuring the influence of Sqft_living on price



- The model explains about 38% of the variance in price
- This was an indicator that while sqft_living is highly correlated with price, we need more features to be able to explain price variance at a higher percentage.

3. Model 3 - Measuring the influence of Sqft_living, bathrooms, sqft_above, sqft_living15, & bedrooms on price

- With these five features this model was to explain ~42% of the houses price variance. This is a marginal improvement in predicting price from model 1’s 38%.
- Through this model we established that living spaces were not sufficient to predict price reliably. Other features such as grade, view should be added to the new model

4. Model 4(final) - Measuring the influence of top 9 features on price

- The final model included all the top 9 features. This model was able to predict explain variance by ~62%.

Regression Results

As indicated above, our final model was able to explain houses price variance by 62.39%. The best fit line from our final model (i.e. the equation for predicting price) is as below:

$$y = 498764.3 + (103300.58 \times sqft\_living) - (11052.18 \times bathrooms) + (22215.05 \times sqft\_living15) - (10988.05 \times bedrooms) + (89082.83 \times lat) - (11809.75 \times floors) - (16728.09 \times sqft\_lot) + (56165.34 \times view) + (83155.89 \times grade)$$

y = price
498764.3 -> y intercept
Other numbers - coefficients for each predictor feature

Interpreting the equation:

- Sqft_living coefficient: The size of the living area has a positive impact on the predicted house price. A larger living area, relative to the other features, contributes more to increasing the predicted price.
- Bathrooms & bedrooms coefficients: With each additional bathroom having a negative effect on the price. This suggests that more bathrooms may lead to a slightly lower predicted price.
- The lat (latitude) coefficient: indicates that the latitude of the property location plays a significant role in determining the predicted price. A higher latitude, relative to the other variables, has a positive impact on the price.
- Grade coefficient: A higher grade, relative to the other variables, has a positive impact on the predicted price.

Conclusions and Recommendations

Following observations in previous section, we recommended that:

1. The final model can explain 62.39% of the variance in house prices.

2. The top features that influence house prices in order of priority include: grade, sqft_living, sqft_living15, bathrooms, lat, view, bedrooms, sqft_lot and floors.

3. To improve model performance, we might recommend utilizing more features e.g., zip code as indicated earlier. We also recommend exploring the use of polynomial features.

Next steps

1. Hand over detailed repository of dataset, analysis and other documentation on the findings

Thank you!

Any questions?

Presented by: Emily Owiti
Email: emily.owiti@student.moringaschool.com