Kingston County House Price Prediction Project

By _S.Karapia,B.Kinya,E.Omondi,P.Riungu,S.Kimutai and S.Gathai

Problem Statement

- In this project, we explored the different factors that impact the price of homes in Kingston County. We examined these features, constructed them and evaluated models to forecast the home prices.

Hypothesis Statement

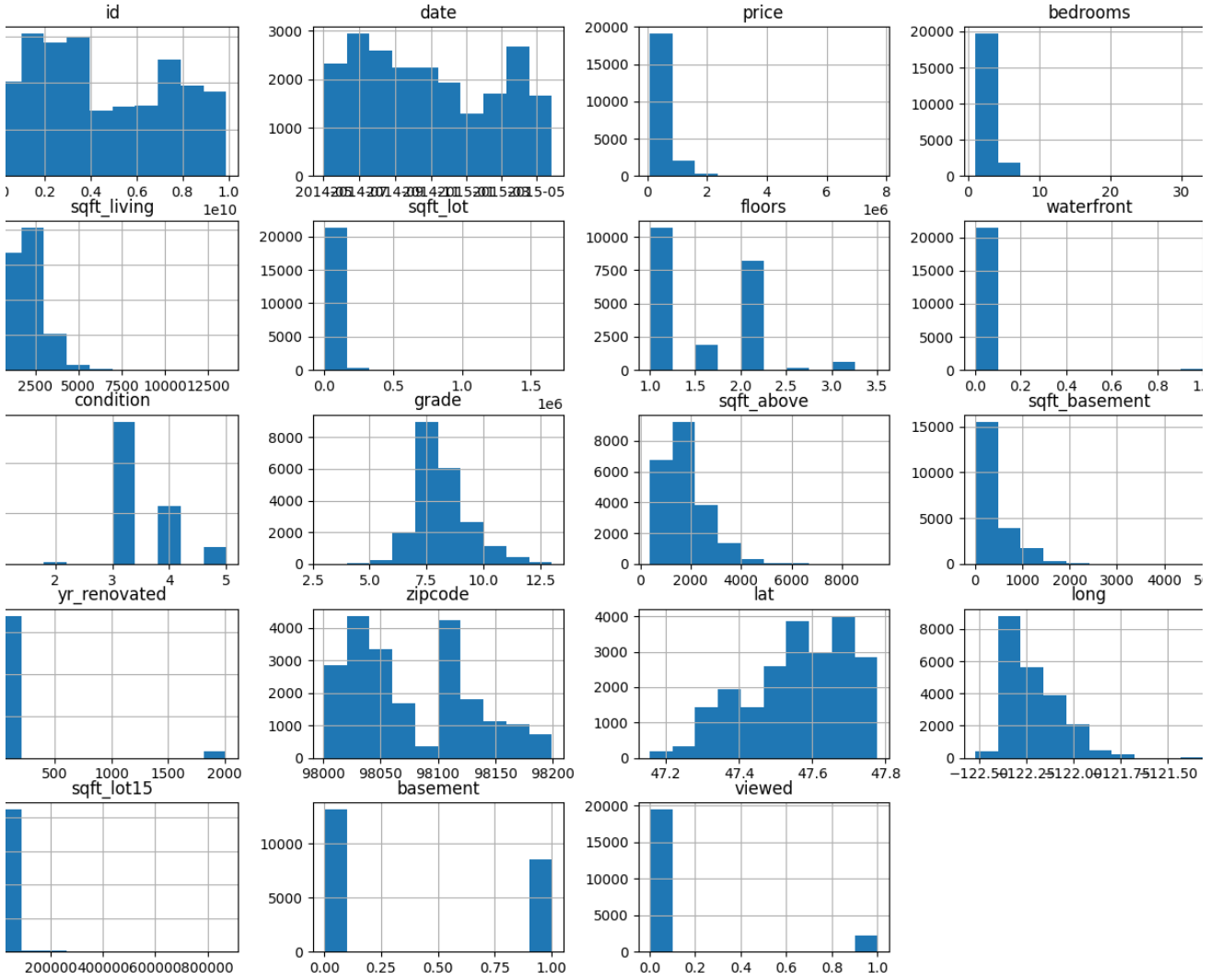
- Ha: There is a significant relationship between the independent variables (bathrooms, sqft_living, waterfront) and the price of houses in the dataset.
- Ho: There is no significant relationship between the independent variables (bathrooms, sqft_living, waterfront) and the price of houses in the dataset.
- The purpose of this project is to investigate and determine whether the selected independent variables have a meaningful impact on house prices. Through data analysis and regression modeling, we aim to either support the alternative hypothesis (Ha) by finding evidence of a significant relationship, or fail to reject the null hypothesis (Ho) if there is insufficient evidence to establish a significant relationship.

BUSINESS UNDERSTANDING

- This presentation aims to address the needs of stakeholders who require guidance on buying or selling houses and insights into the impact of house renovations on their estimated value. We will provide an in-depth analysis to identify houses with better renovation potential. By the end of this presentation, you will have a clear understanding of which houses offer promising opportunities for renovations and how these renovations can enhance the estimated value of the properties. This information will empower you to make informed decisions in the dynamic housing market.

DATA UNDERSTANDING

We analyzed housing data from Kingston County which constituted the following data



DSC PROJECT ON APPLICATION OF DATA ANALYSIS TOOLS AND PREDICTIVE MODELS ON HOME OWNERSHIP AND OR INVESTMENTS BY_S.Karapia,B.Kinya,E.Omondi,P.Riungu,S.Kimutai and S.Gitau

Key variables that were considered in our analysis

- # ***Column Names and Descriptions for King County 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
    * Includes Duwamish, Elliott Bay, Puget Sound, Lake Union, Ship Canal, Lake Washington, Lake Sammamish, other lake, and river/slough waterfronts
  * `**view**` - Quality of view from house
    * Includes views of Mt. Rainier, Olympics, Cascades, Territorial, Seattle Skyline, Puget Sound, Lake Washington, Lake Sammamish, small lake / river / creek, and other
  * `**condition**` - How good the overall condition of the house is. Related to maintenance of house.
    * See the [King County Assessor Website](https://info.kingcounty.gov/assessor/esales/Glossary.aspx?type=r) for further explanation of each condition code
  * `**grade**` - Overall grade of the house. Related to the construction and design of the house.
    * See the [King County Assessor
- Website](https://info.kingcounty.gov/assessor/esales/Glossary.aspx?type=r) for further explanation of each condition code
  * `**grade**` - Overall grade of the house. Related to the construction and design of the house.
    * See the [King County Assessor Website](https://info.kingcounty.gov/assessor/esales/Glossary.aspx?type=r) for further explanation of each building grade code
  * `**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

BUSINESS VALUE

- This project aims to enhance our understanding of buyer preferences and enhance our knowledge of the real estate market in Kings County. By developing a predictive model for house prices, we can refine our suggested sales price and make more informed decisions

METHODOLOGY

- The dataset utilized in this project comprises of approximately 21000 house sale prices in Kings County. The initial step involved cleaning the data to ensure it sustainability for modeling purposes. Subsequently, we identified the influential features that would yield valuable insights and constructed a model capable of accurately predicting house prices.

HOUSE FEATURES

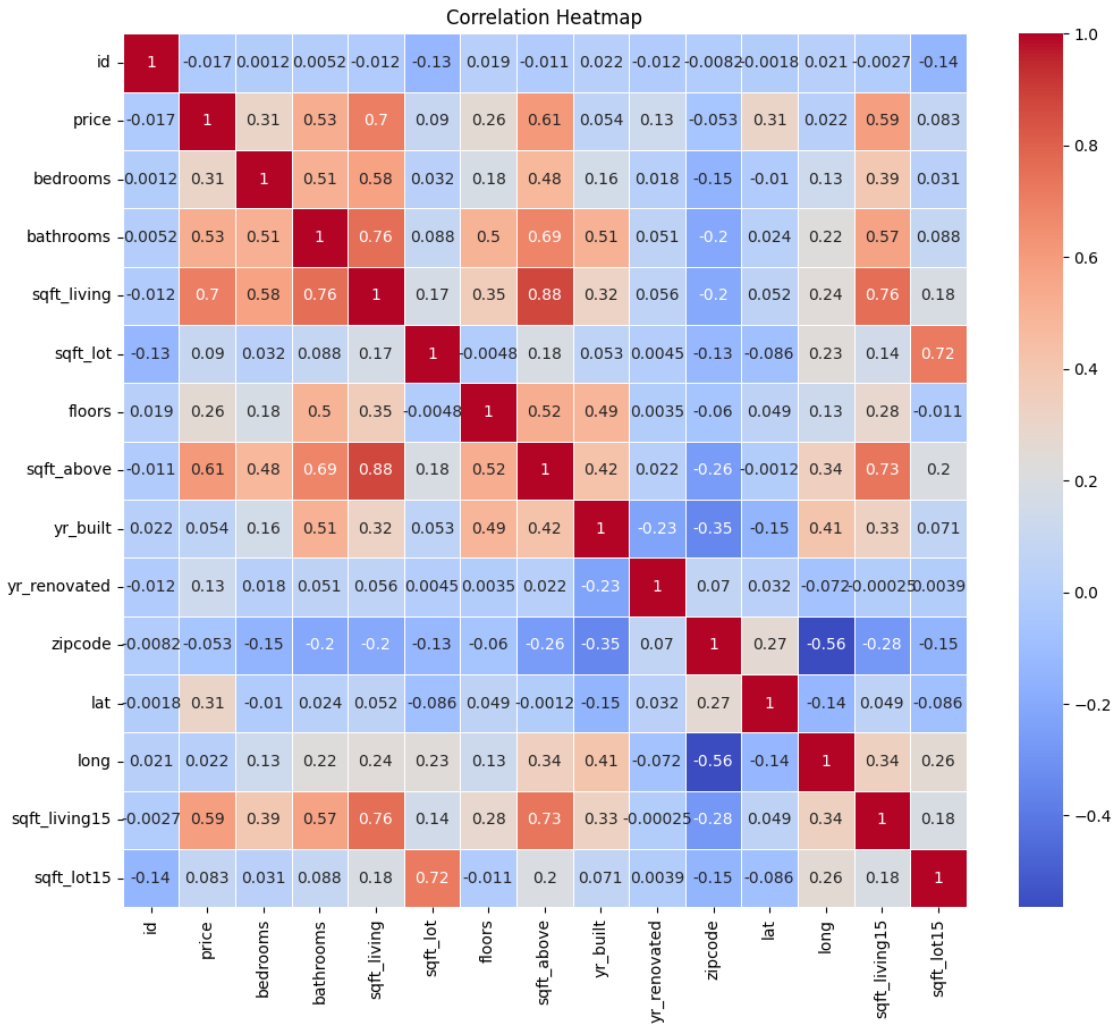
- We examined various house features to determine the factors that contribute to an increase in price and identify the most effective predictors of price.
- The house features we used in the final model include:

1. Bathrooms
2. Square foot living
3. Waterfront

RESULTS

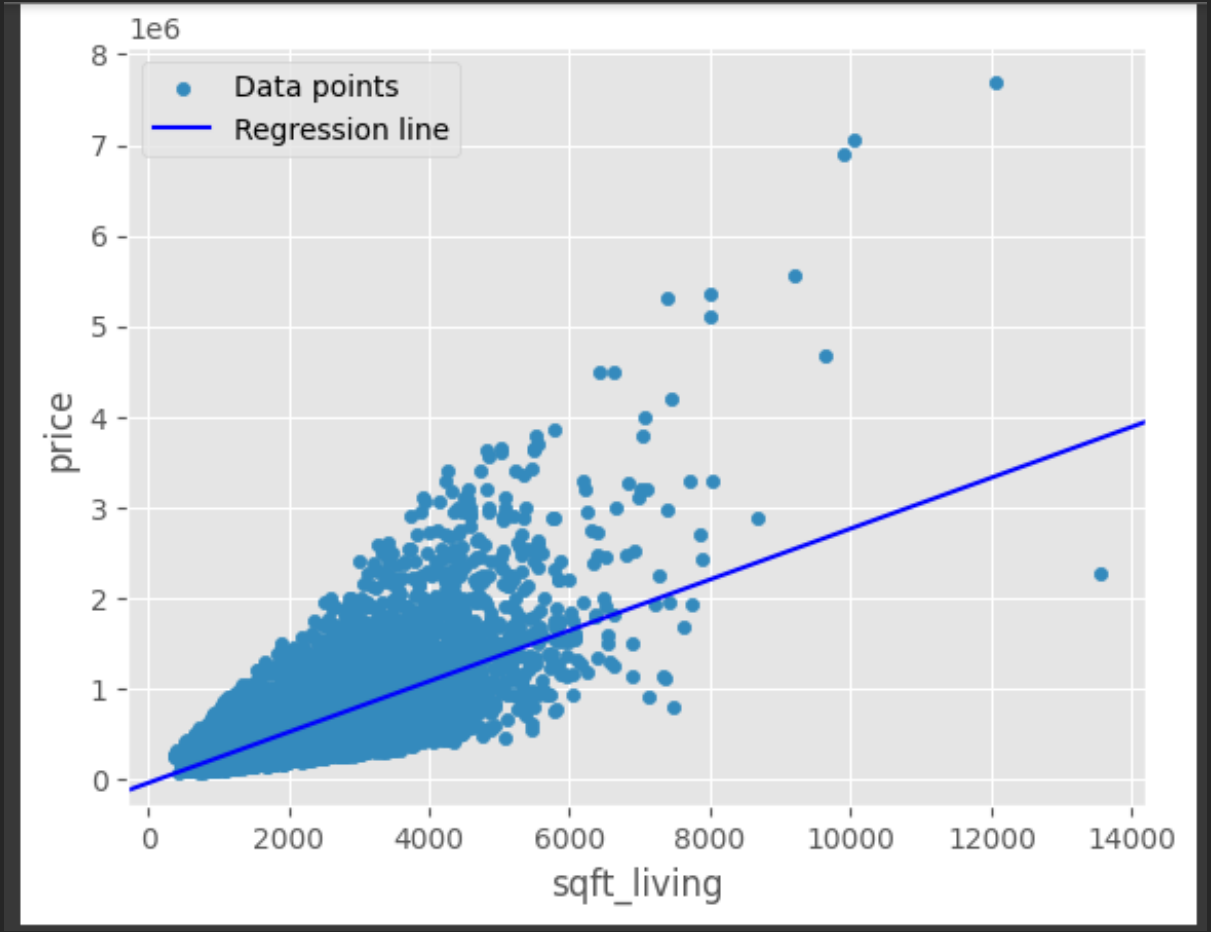
- These were the results of our final model:
- R-squared (uncentered): The R-squared value of 0.899 indicates that the model explains approximately 89.9% of the variance in the dependent variable (price). This suggests that the independent variables included in the model (bathrooms, sqft_living, waterfront) collectively have a strong association with the price.
- Adjusted R-squared (uncentered): The adjusted R-squared value is also 0.899, which means that the inclusion of the three independent variables in the model is not significantly impacting the overall goodness of fit. The adjusted R-squared value is useful for comparing models with different numbers of predictors.
- F-statistic: The F-statistic has a very large value of 6.378e+04, and the associated probability (Prob (F-statistic)) is 0.00. This indicates that the overall model is statistically significant, suggesting that at least one of the independent variables has a significant impact on the price.
- Coefficients: The coefficients for the independent variables indicate the magnitude and direction of their relationship with the dependent variable (price).
- Bathrooms: The coefficient for the "bathrooms" variable is 4.4689, indicating a positive relationship with the price. A one-unit increase in the number of bathrooms is associated with an increase in the price by approximately 4.4689 units.
- Sqft_living: The coefficient for the "sqft_living" variable is 0.0011, indicating a positive relationship with the price. A one-unit increase in the square footage of living area is associated with an increase in the price by approximately 0.0011 units.
- Waterfront: The coefficient for the "waterfront" variable is -1.3921, indicating a negative relationship with the price. A property with a waterfront location is associated with a decrease in the price by approximately 1.3921 units.
- All three variables have p-values close to zero, indicating that they are highly statistically significant in relation to the price.

Correlation of different house features



❖Heatmap Interpretation;

- From the previous slide, In a correlation heatmap, each variable is represented by a row and a column, and the cells show the correlation between them.
- The color of each cell represents the strength and direction of the correlation, with darker colors indicating stronger correlations.The correlation coefficient ranges from -1 to 1.
- A value of 1 indicates a perfect positive correlation,
- -1 indicates a perfect negative correlation, and 0 indicates no correlation.
- Correlation coefficients close to +1 (e.g., 0.8 or 0.9) indicate a strong positive correlation, where the variables tend to increase or decrease together.
- Correlation coefficients between 0 and +1 indicate a positive correlation, where an increase in one variable tends to be associated with an increase in the other variable, but the relationship may not be as strong.



Simple Regression

- From the correlation Heatmap, sqft living had the highest correlation to price

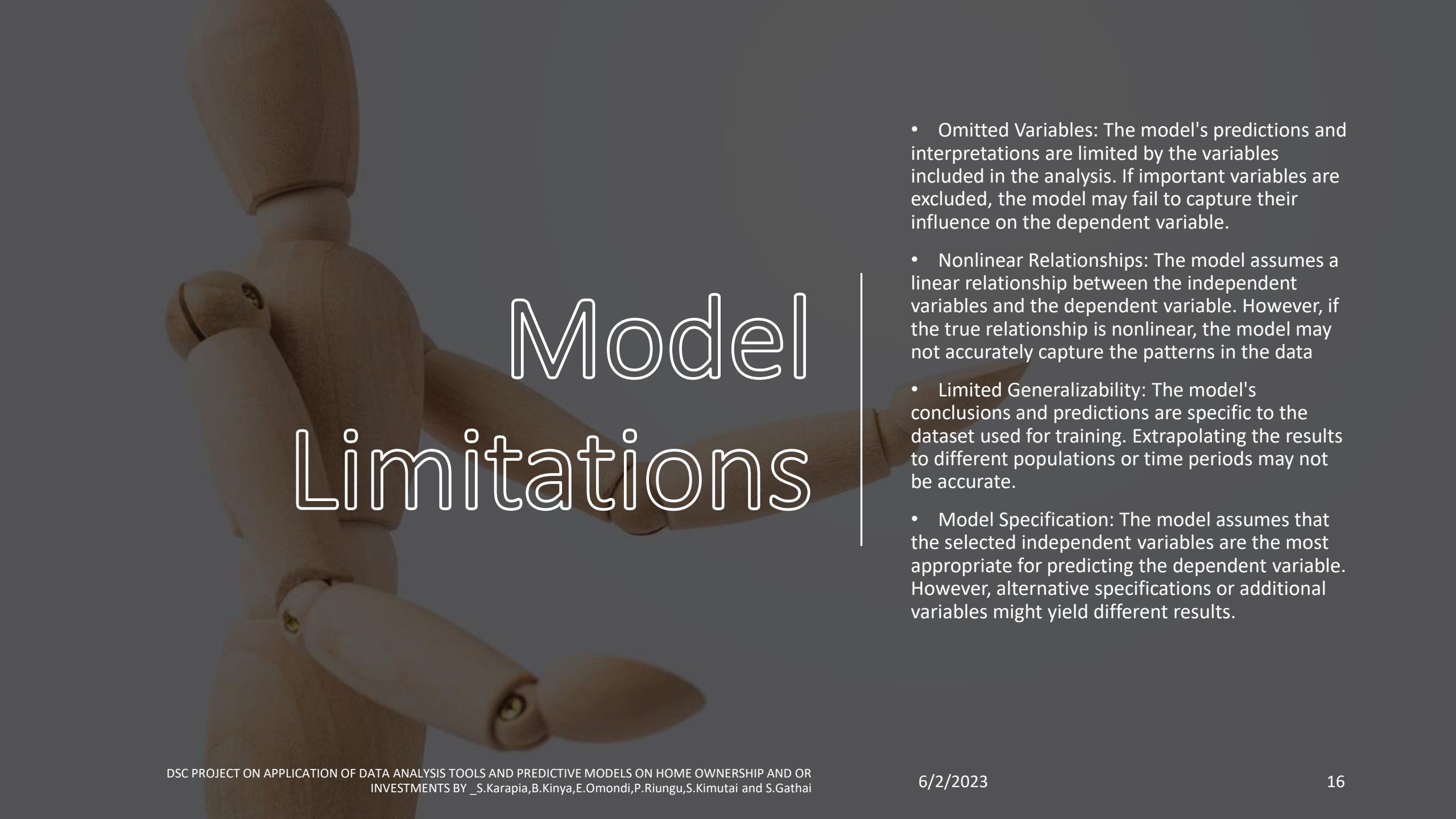
❖Scatter plots of house features vs price

- Here we displayed multiple scatter plots of house features with the price.

❖ *Predicting House Prices*

- We used two models to predict house prices and below is their comparison summary:
- The first OLS regression model has an R-squared value of 0.851, while the second model has an R-squared value of 0.899. This indicates that the second model explains a higher proportion of the variability in the price compared to the first model. In other words, the second model provides a better fit to the data.
- Examining the coefficients, we can observe the following differences between the two models:

1. Bathrooms: In the first model, an increase in the number of bathrooms is associated with a decrease in price (-$13,250), while in the second model, an increase in the number of bathrooms is associated with an increase in price ($4.4689). The coefficients have opposite signs, indicating a contrasting impact of bathrooms on price.
2. Sqft_living: Both models show a positive relationship between square footage of living space and price. However, the second model (coefficient: 0.0011) suggests a smaller impact compared to the first model (coefficient: 272.1194). The magnitude of the effect differs significantly between the two models.
3. Waterfront: Both models indicate that having a waterfront property increases the price. However, the second model (coefficient: -1.3921) suggests a larger negative impact compared to the first model (coefficient: 8.701e+05). The second model implies a more substantial reduction in price for waterfront properties.



Model Limitations

- Omitted Variables: The model's predictions and interpretations are limited by the variables included in the analysis. If important variables are excluded, the model may fail to capture their influence on the dependent variable.
- Nonlinear Relationships: The model assumes a linear relationship between the independent variables and the dependent variable. However, if the true relationship is nonlinear, the model may not accurately capture the patterns in the data
- Limited Generalizability: The model's conclusions and predictions are specific to the dataset used for training. Extrapolating the results to different populations or time periods may not be accurate.
- Model Specification: The model assumes that the selected independent variables are the most appropriate for predicting the dependent variable. However, alternative specifications or additional variables might yield different results.

CONCLUSION

- Key Findings:
- Our final model has a high R-squared value of 0.899, indicating that approximately 89.9% of the price variability can be explained by the included variables (bathrooms, sqft_living, waterfront).
- All three independent variables (bathrooms, sqft_living, waterfront) have significant impacts on the price, as demonstrated by their low p-values.
- The coefficients reveal that an increase in the number of bathrooms and square footage of living area leads to higher prices, while properties with a waterfront location tend to have lower prices.
- The F-statistic confirms the overall statistical significance of the model, indicating that at least one of the independent variables significantly affects the price.

RECOMMENDATIONS

1. Larger Living area(sqft) for their house renovations
2. Include more bathrooms in house renovation as they fetch a higher price
3. Houses with a waterfront seem to fetch a lower price depending on factors like the demand. Stakeholders should therefore consider excluding waterfronts in their house investments or renovations

Immanuel omondi

Gmail: immanuelomondi5@gmail.com

LinkedIn: http://linkedin.com/in/immanuel-omondi-388517279/

: