INTRODUCTION

Welcome to the presentation on predicting house prices using data analysis! Today, we'll discuss how we used data to gain insights into the housing market and predict house prices.

A Presentation by:

Sammy Warah

PROJECT OVERVIEW

This project focuses on predicting house prices based on various factors such as the number of bedrooms, bathrooms, square footage of living space, and more. Our goal is to provide valuable insights to real estate agencies, property investors, and individuals looking to buy or sell houses.

Understanding the Market

OUR DATASET PROVIDES INSIGHTS INTO MARKET TRENDS AND PROPERTY CHARACTERISTICS, HELPING STAKEHOLDERS NAVIGATE THE HOUSING MARKET EFFECTIVELY.

Before diving into the analysis, it's crucial to understand our stakeholders' needs. Real estate agencies, investors, and homebuyers rely on accurate information to make informed decisions.

THIS HEATMAP SHOWS HOW VARIOUS FACTORS RELATE TO THE PRICE

BUILDING THE MODEL

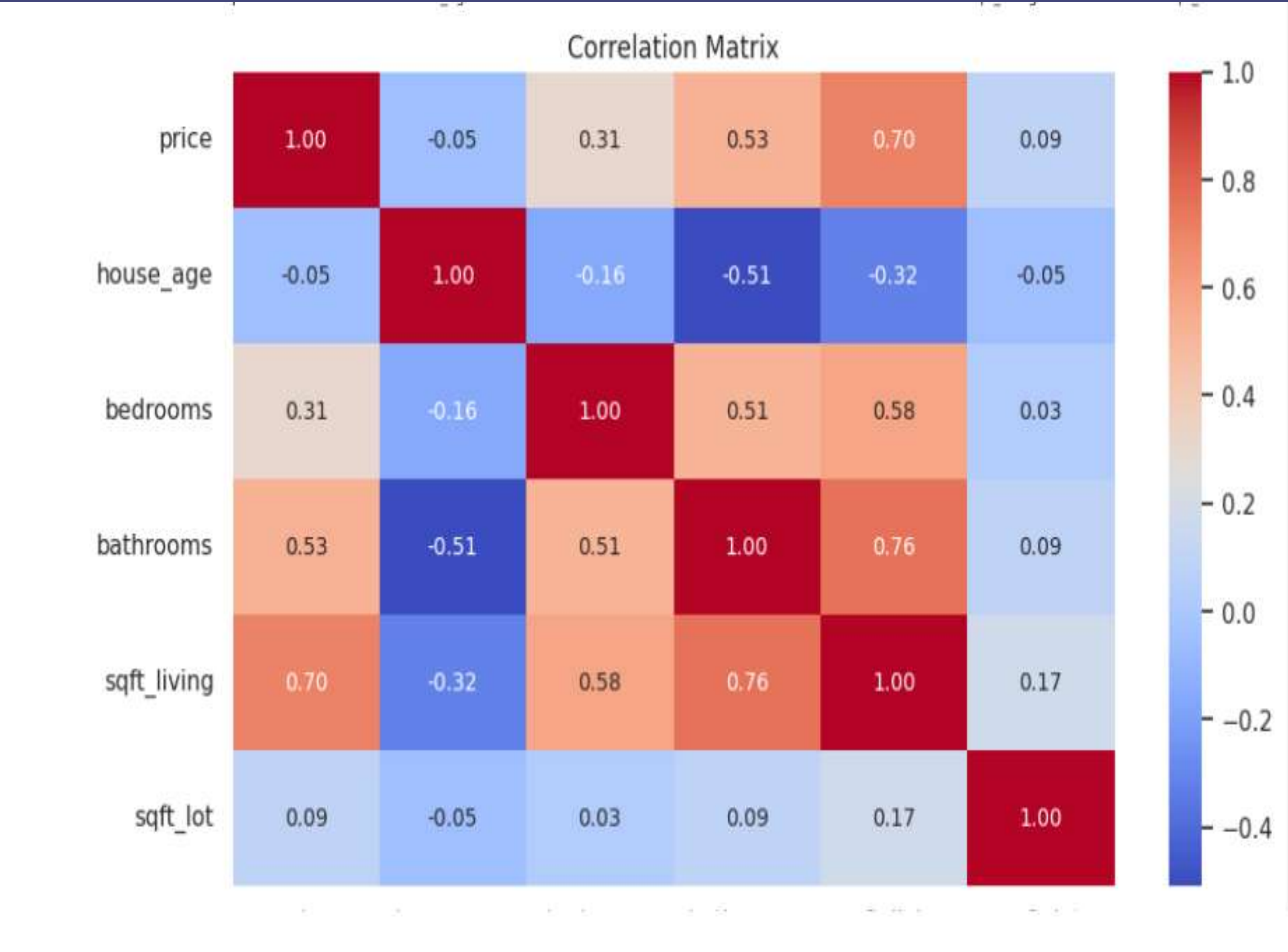
We utilized multiple linear regression to predict house prices.

This model considers various features such as the number of bedrooms, bathrooms, and square footage of living space to make predictions.

By analyzing historical data, we can identify patterns and trends that influence house prices.

visualizing the relationship between the 'house_age' feature and the target variable ['price'].

This trend suggests that newer houses tend to have a higher value in the market, which diminishes over time as the property ages.

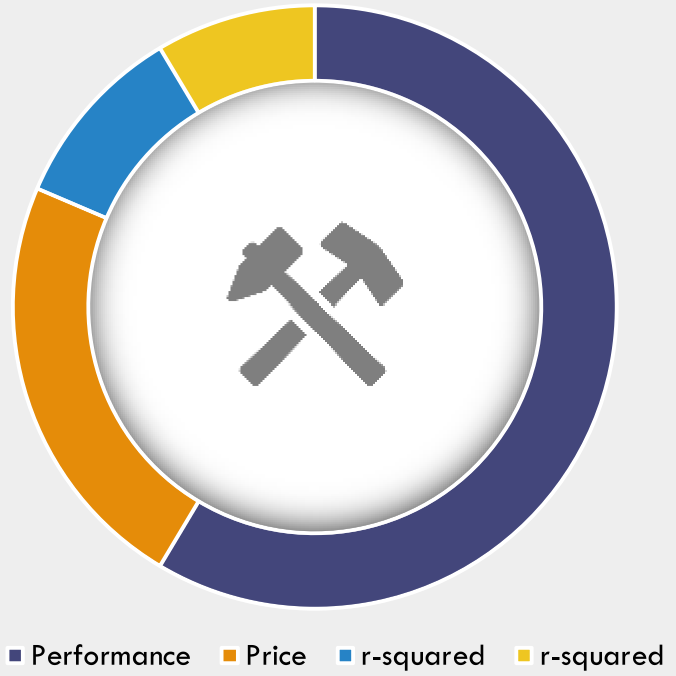


The pair plot shows the distributions and relationships between 'house_age' and other numerical features. We can observe some patterns:

- 'house_age' tends to decrease as the number of bedrooms and bathrooms increases, indicating that newer houses tend to have more bedrooms and bathrooms.
- There's a slight negative correlation between 'house_age' and 'sqft_living', suggesting that newer houses may have smaller living areas.
- 'house_age' doesn't seem to have a strong correlation with 'sqft_lot'.

ANALYZING PREDICTIONS

Our model achieved a strong performance, accurately predicting house prices with a high degree of confidence. We evaluated the model's performance using metrics such as R-squared, which measures how well the model fits the data. The results provide valuable insights into the factors driving house prices and help stakeholders make informed decisions.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.651
Model:                            OLS   Adj. R-squared:                  0.650
Method:                 Least Squares   F-statistic:                     2678.
Date:                Sat, 06 Apr 2024   Prob (F-statistic):               0.00
Time:                        03:51:39   Log-Likelihood:            -2.3690e+05
No. Observations:               17277   AIC:                         4.738e+05
Df Residuals:                   17264   BIC:                         4.739e+05
Df Model:                          12
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const       -6.042e+07   1.15e+07     -5.233      0.000   -8.31e+07   -3.78e+07
bedrooms    -4.686e+04   2377.188    -19.711      0.000   -5.15e+04   -4.22e+04
bathrooms    5.368e+04   3870.654     13.869      0.000    4.61e+04    6.13e+04
sqft_living   179.0715      3.725     48.072      0.000     171.770     186.373
sqft_lot       -0.2968      0.042     -7.007      0.000      -0.380      -0.214
floors       2.235e+04   3848.653      5.808      0.000    1.48e+04    2.99e+04
waterfront   7.385e+05   2.04e+04     36.233      0.000    6.99e+05    7.78e+05
condition    2.125e+04   2779.490      7.645      0.000    1.58e+04    2.67e+04
grade        1.315e+05   2408.922     54.594      0.000    1.27e+05    1.36e+05
yr_built    -3896.3245     74.222    -52.495      0.000   -4041.807   -3750.841
year         3.349e+04   5729.704      5.845      0.000    2.23e+04    4.47e+04
month         780.3010    860.123      0.907      0.364    -905.628    2466.230
day          -229.5988    193.839     -1.184      0.236    -609.542     150.344
==============================================================================
Omnibus:                    12781.051   Durbin-Watson:                   2.001
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           818312.708
Skew:                           2.970   Prob(JB):                         0.00
Kurtosis:                      36.188   Cond. No.                     2.99e+08
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.99e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.647
Model:                            OLS   Adj. R-squared:                  0.647
Method:                 Least Squares   F-statistic:                     3297.
Date:                Sat, 06 Apr 2024   Prob (F-statistic):               0.00
Time:                        04:42:09   Log-Likelihood:            -2.9615e+05
No. Observations:               21597   AIC:                         5.923e+05
Df Residuals:                   21584   BIC:                         5.924e+05
Df Model:                          12
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const       -5.749e+07   1.03e+07     -5.559      0.000   -7.78e+07   -3.72e+07
bedrooms    -4.279e+04   2047.112    -20.902      0.000   -4.68e+04   -3.88e+04
bathrooms    5.108e+04   3449.231     14.808      0.000    4.43e+04    5.78e+04
sqft_living   177.3511      3.298     53.783      0.000     170.888     183.814
sqft_lot       -0.2460      0.037     -6.701      0.000      -0.318      -0.174
floors       2.157e+04   3454.195      6.243      0.000    1.48e+04    2.83e+04
waterfront   7.555e+05   1.83e+04     41.216      0.000     7.2e+05    7.91e+05
condition    2.005e+04   2487.543      8.060      0.000    1.52e+04    2.49e+04
grade        1.303e+05   2152.852     60.521      0.000    1.26e+05    1.35e+05
yr_built     8097.8681   1710.907      4.733      0.000    4744.363    1.15e+04
year         2.005e+04   3421.215      5.860      0.000    1.33e+04    2.68e+04
month         790.9691    770.605      1.026      0.305    -719.473    2301.412
day          -406.8222    173.070     -2.351      0.019    -746.051     -67.593
house_age    1.195e+04   1710.957      6.985      0.000    8597.013    1.53e+04
==============================================================================
Omnibus:                    15874.124   Durbin-Watson:                   1.977
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1021679.453
Skew:                           2.940   Prob(JB):                         0.00
Kurtosis:                      36.178   Cond. No.                     7.25e+15
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 7.98e-19. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

The training R-squared score is 0.6505295038401273, and the testing R-squared score is 0.6314455116733808. These scores indicate that the model explains around 65.1% of the variance in the training data and around 63.1% of the variance in the testing data, which suggests that the model's performance is consistent across both sets.

INSIGHTS AND RECOMMENDATIONS

Based on our analysis, we have several recommendations for stakeholders in the housing market. These recommendations include pricing strategies, investment opportunities, and market trends to capitalize on. By leveraging data-driven insights, stakeholders can optimize their decision-making processes and achieve success in the real estate industry.

FUTURE OPPORTUNITIES

Moving forward, there are several opportunities to further enhance our analysis. This includes refining the model with additional data sources, exploring advanced machine learning techniques, and conducting more in-depth market research. By continually improving our methods, we can provide even more valuable insights to stakeholders in the housing market.

CONCLUSION

Thank you for joining me today! I hope this presentation has provided valuable insights into predicting house prices using data analysis. If you have any questions or would like to learn more, please don't hesitate to reach out. I look forward to continuing our collaboration and driving success in the real estate industry.