PREDICTIVE ANALYSIS IN PROPERTY VALUATION

A King County Case Study

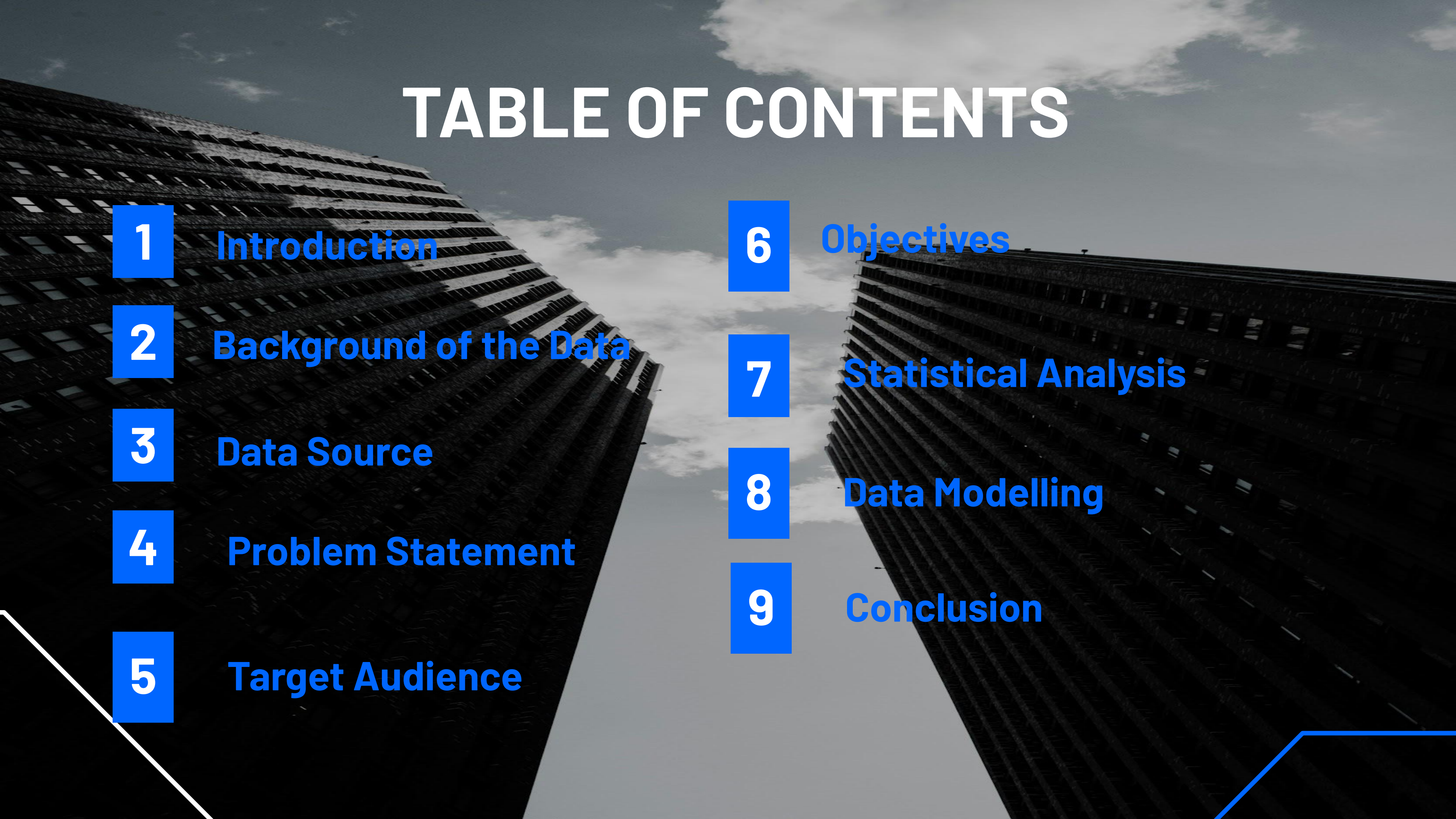


TABLE OF CONTENTS

1 Introduction

2 Background of the Data

3 Data Source

4 Problem Statement

5 Target Audience

6 Objectives

7 Statistical Analysis

8 Data Modelling

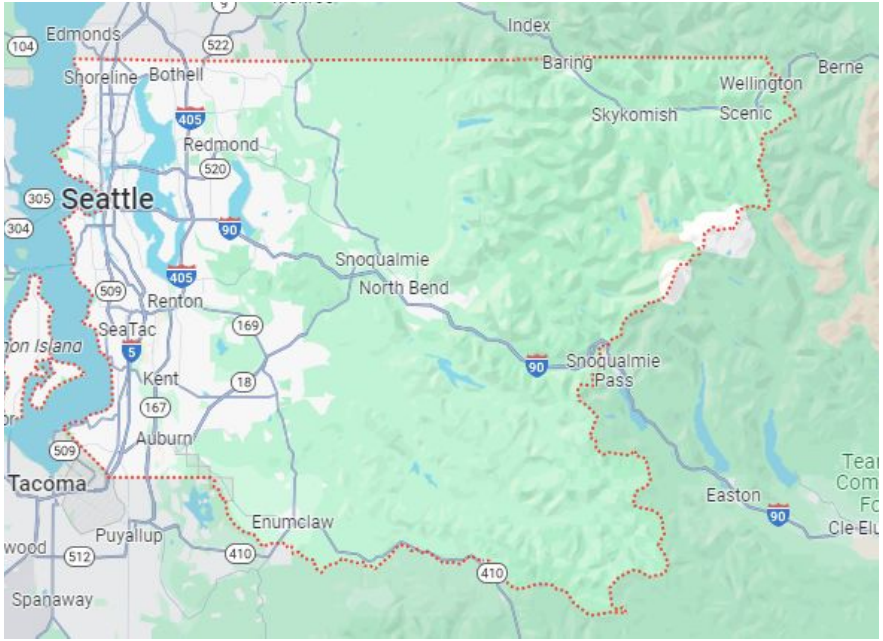
9 Conclusion

INTRODUCTION

- Real estate agencies guide homeowners in crucial decisions like buying, selling, or renting properties.
- The project empowers agencies with a regression-based model predicting property value increases based on features like bedrooms, year built, floors, living space, condition, and location.
- The model assists in pricing strategies, market analysis, and property inspections, maximizing return on investment for clients.

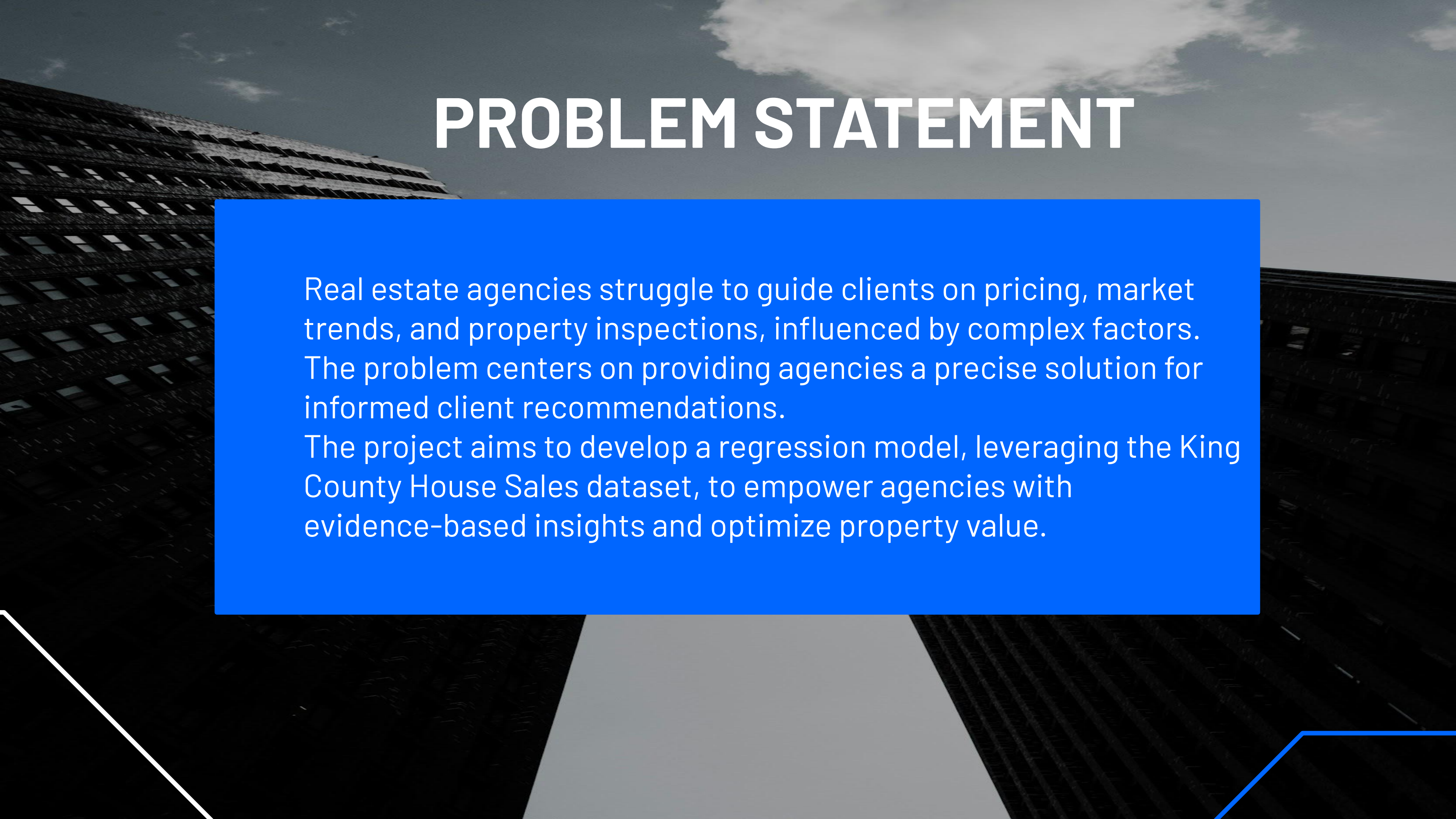
BACKGROUND OF THE DATA

- The housing market in King County, Washington, has a rich history, marked by significant growth and changes.
- Situated in the northwestern U.S., King County includes the major economic and cultural hub of Seattle.
- King County's real estate market is recognized for its competitiveness, with diverse neighborhoods providing a range of housing options from historic homes to modern developments.



DATA SOURCE

- King County
- Located in the US state of Washington
- 21,597 listings from May 2014 to May 2015



PROBLEM STATEMENT

Real estate agencies struggle to guide clients on pricing, market trends, and property inspections, influenced by complex factors.
The problem centers on providing agencies a precise solution for informed client recommendations.
The project aims to develop a regression model, leveraging the King County House Sales dataset, to empower agencies with evidence-based insights and optimize property value.

TARGET AUDIENCE

- Real estate professionals seeking insights on pricing, market trends, and property inspections.
- Data scientists interested in developing regression models with the King County House Sales dataset.
- Real estate industry professionals aiming to enhance advisory capabilities.
- Stakeholders in the real estate landscape looking to optimize property value.

OBJECTIVES

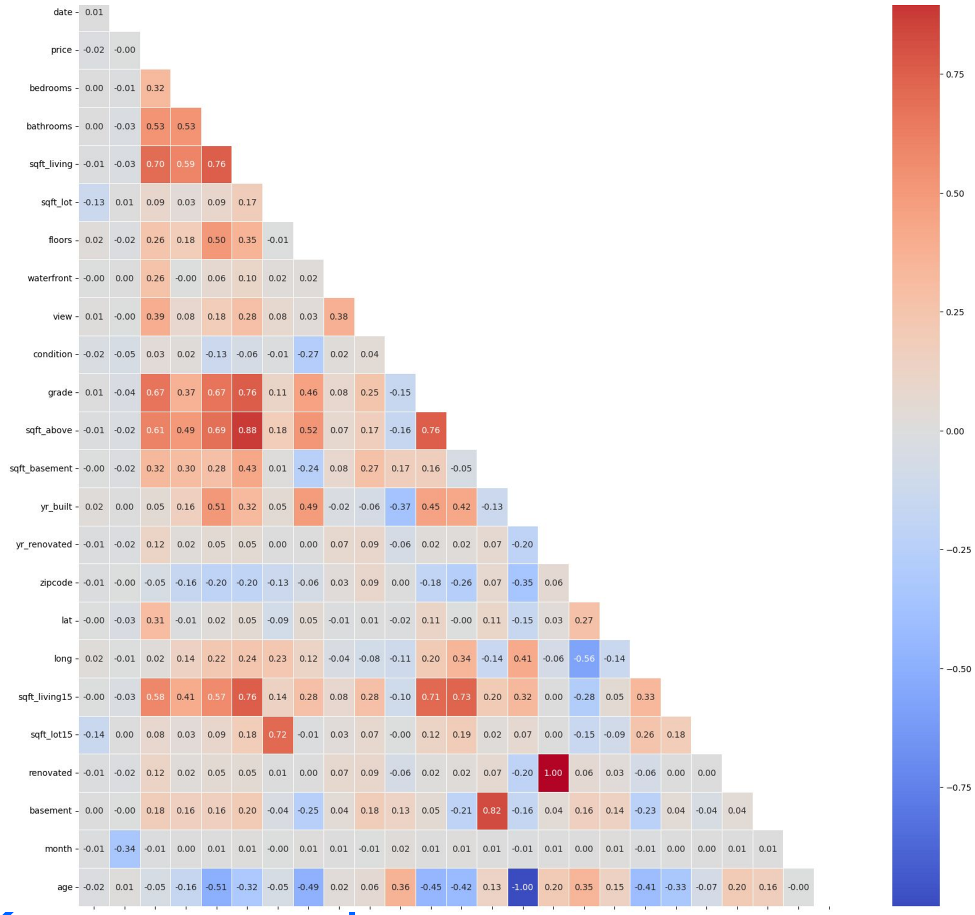
- To develop a predictive regression model that assists real estate agencies in advising clients on house prices
- To identify key factors influencing house prices in King County, California, to provide valuable insights for precise pricing strategies.
- To analyze model performance using metrics such as mean squared error, R-squared values, and residual analysis to gauge the model's effectiveness.
- To provide actionable recommendations to the Real Estate Agency for improving profitability and market presence, leveraging insights from the model.

Statistical Analysis

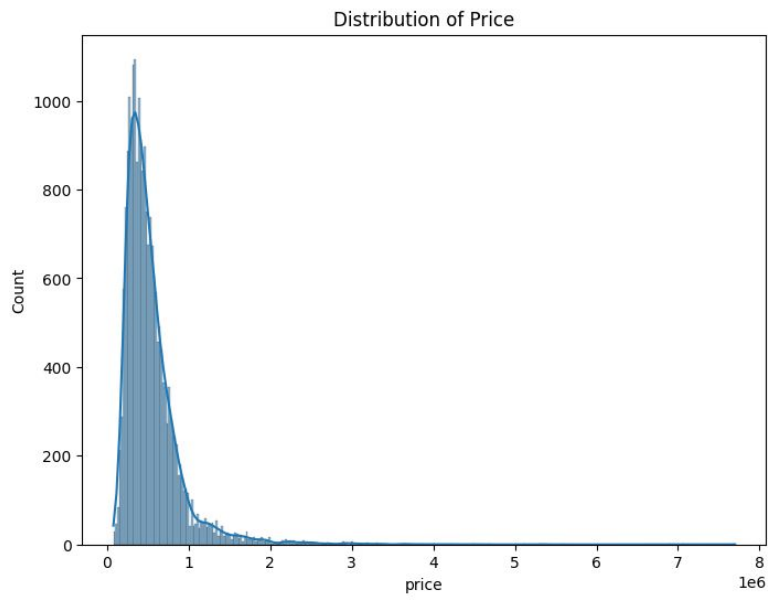
Triangular Correlation Matrix

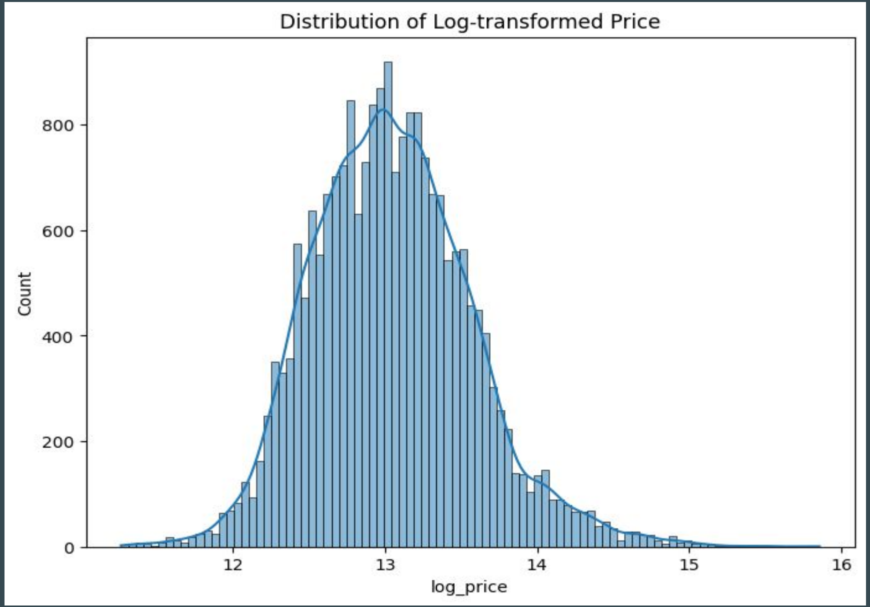
Analysis of correlation between the variables

For instance, Year Built' and 'Age' show a perfect negative correlation (-1.00), indicating as the property's construction year increases, its age decreases linearly.



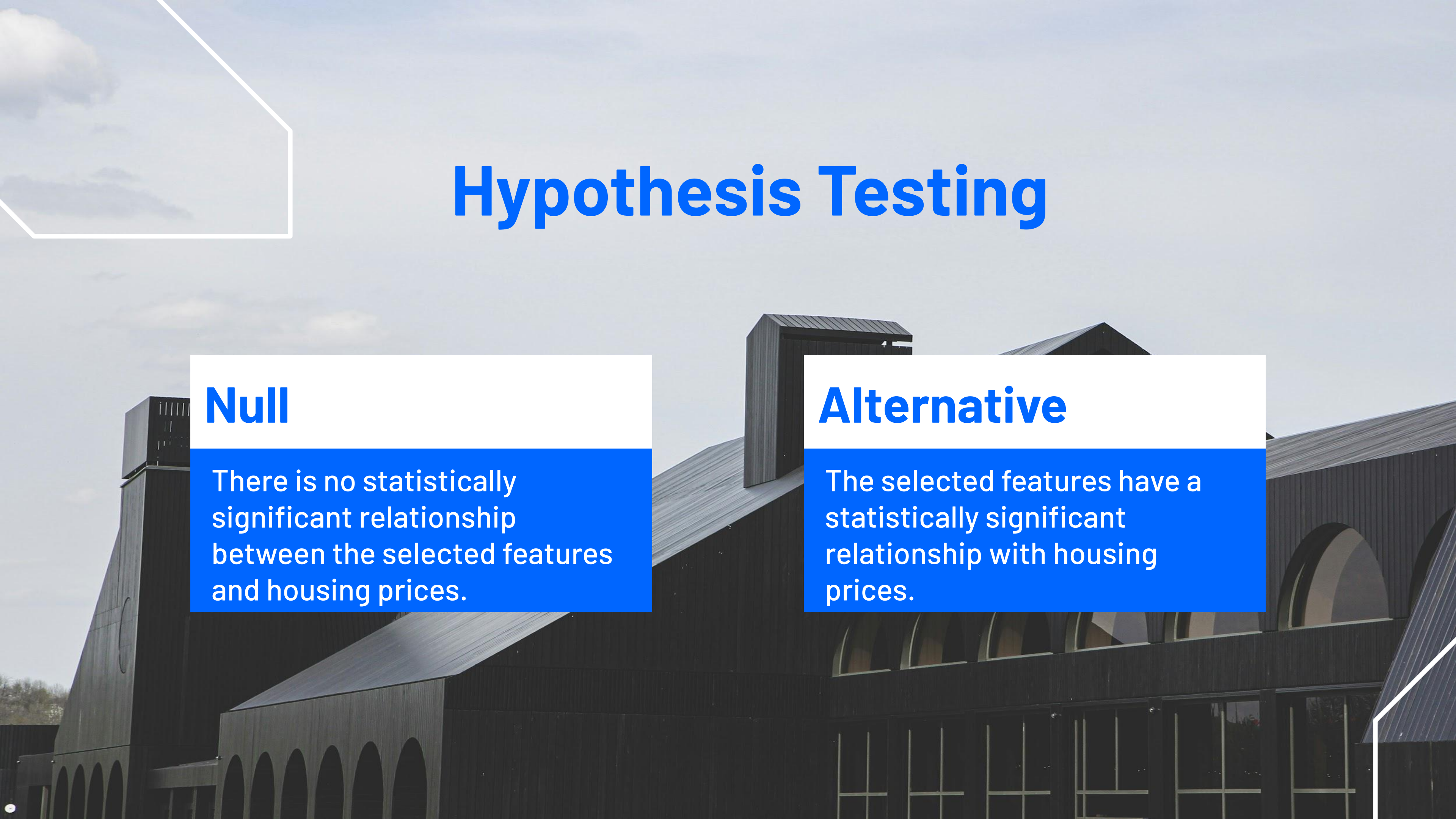
date	0.01																						
price	-0.02	-0.00																					
bedrooms	0.00	-0.01	0.32																				
bathrooms	0.00	-0.03	0.53	0.53																			
sqft_living	-0.01	-0.03	0.70	0.59	0.76																		
sqft_lot	-0.13	0.01	0.09	0.03	0.09	0.17																	
floors	0.02	-0.02	0.26	0.18	0.50	0.35	-0.01																
waterfront	-0.00	0.00	0.26	-0.00	0.06	0.10	0.02	0.02															
view	0.01	-0.00	0.39	0.08	0.18	0.28	0.08	0.03	0.38														
condition	-0.02	-0.05	0.03	0.02	-0.13	-0.06	-0.01	-0.27	0.02	0.04													
grade	0.01	-0.04	0.67	0.37	0.67	0.76	0.11	0.46	0.08	0.25	-0.15												
sqft_above	-0.01	-0.02	0.61	0.49	0.69	0.88	0.18	0.52	0.07	0.17	-0.16	0.76											
sqft_basement	-0.00	-0.02	0.32	0.30	0.28	0.43	0.01	-0.24	0.08	0.27	0.17	0.16	-0.05										
yr_built	0.02	0.00	0.05	0.16	0.51	0.32	0.05	0.49	-0.02	-0.06	-0.37	0.45	0.42	-0.13									
yr_renovated	-0.01	-0.02	0.12	0.02	0.05	0.05	0.00	0.00	0.07	0.09	-0.06	0.02	0.02	0.07	-0.20								
zipcode	-0.01	-0.00	-0.05	-0.16	-0.20	-0.20	-0.13	-0.06	0.03	0.09	0.00	-0.18	-0.26	0.07	-0.35	0.06							
lat	-0.00	-0.03	0.31	-0.01	0.02	0.05	-0.09	0.05	-0.01	0.01	-0.02	0.11	-0.00	0.11	-0.15	0.03	0.27						
long	0.02	-0.01	0.02	0.14	0.22	0.24	0.23	0.12	-0.04	-0.08	-0.11	0.20	0.34	-0.14	0.41	-0.06	-0.56	-0.14					
sqft_living15	-0.00	-0.03	0.58	0.41	0.57	0.76	0.14	0.28	0.08	0.28	-0.10	0.71	0.73	0.20	0.32	0.00	-0.28	0.05	0.33				
sqft_lot15	-0.14	0.00	0.08	0.03	0.09	0.18	0.72	-0.01	0.03	0.07	-0.00	0.12	0.19	0.02	0.07	0.00	-0.15	-0.09	0.26	0.18			
renovated	-0.01	-0.02	0.12	0.02	0.05	0.05	0.01	0.00	0.07	0.09	-0.06	0.02	0.02	0.07	-0.20	1.00	0.06	0.03	-0.06	0.00	0.00		
basement	0.00	-0.00	0.18	0.16	0.16	0.20	-0.04	-0.25	0.04	0.18	0.13	0.05	-0.21	0.82	-0.16	0.04	0.16	0.14	-0.23	0.04	-0.04	0.04	
month	-0.01	-0.34	-0.01	0.00	0.01	0.01	-0.00	0.01	0.01	-0.01	0.02	0.01	0.01	0.01	-0.01	0.01	0.00	0.01	-0.01	0.00	0.00	0.01	0.01
age	-0.02	0.01	-0.05	-0.16	-0.51	-0.32	-0.05	-0.49	0.02	0.06	0.36	-0.45	-0.42	0.13	-1.00	0.20	0.35	0.15	-0.41	-0.33	-0.07	0.20	0.16	-0.00





The log transformation successfully normalized the positively skewed price distribution, addressing extreme values and enhancing the data's suitability for more reliable statistical analysis.

Hypothesis Testing



Null

There is no statistically significant relationship between the selected features and housing prices.

Alternative

The selected features have a statistically significant relationship with housing prices.

FINDINGS

From hypothesis testing

ANOVA TEST

-The null hypothesis that there is no statistically significant relationship between the selected features and house price was rejected.

-Most of the features were found to be statistically significant apart from a few which include longitude,square footage of the lot a nd square footage of the land lots of the nearest 15 neighbors.

DATA MODELLING

Models

Baseline Model

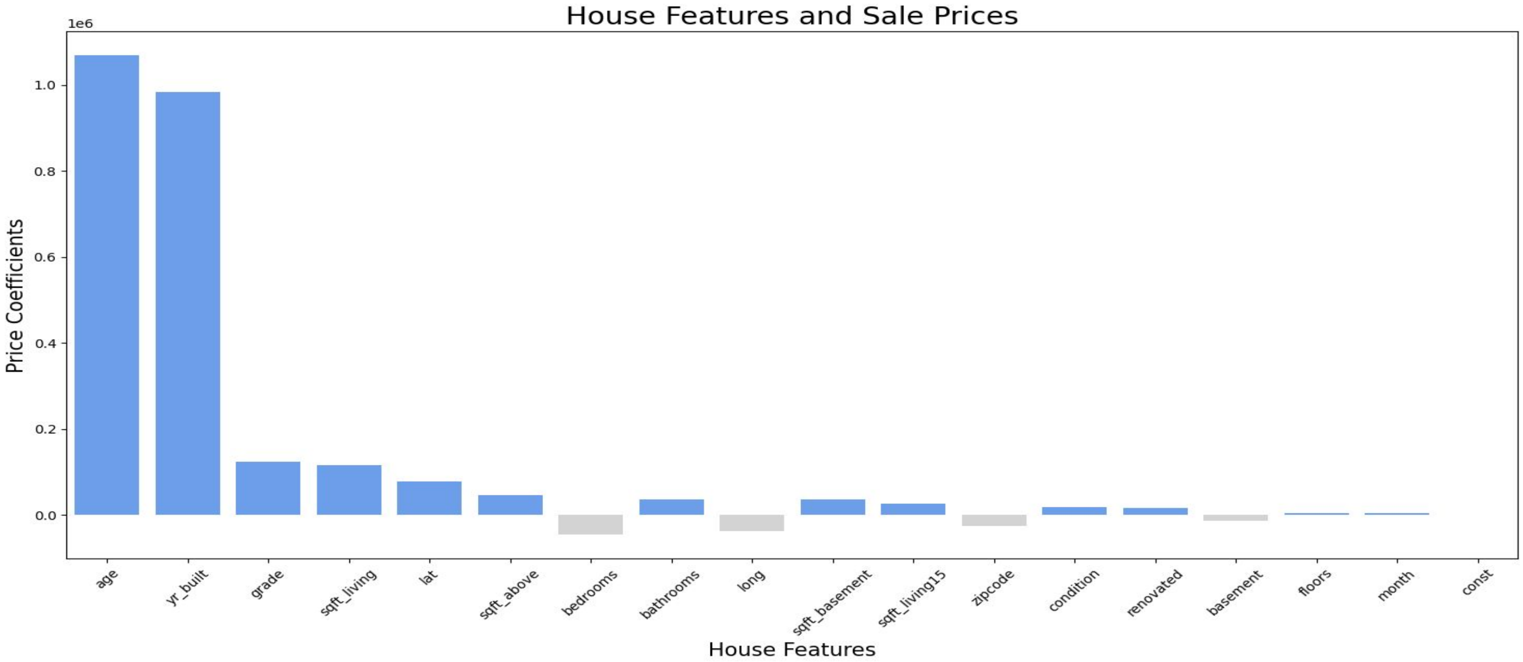
A simple linear regression that considers a single predictor variable to estimate house price.Moderate level of prediction accuracy

Polynomial regression

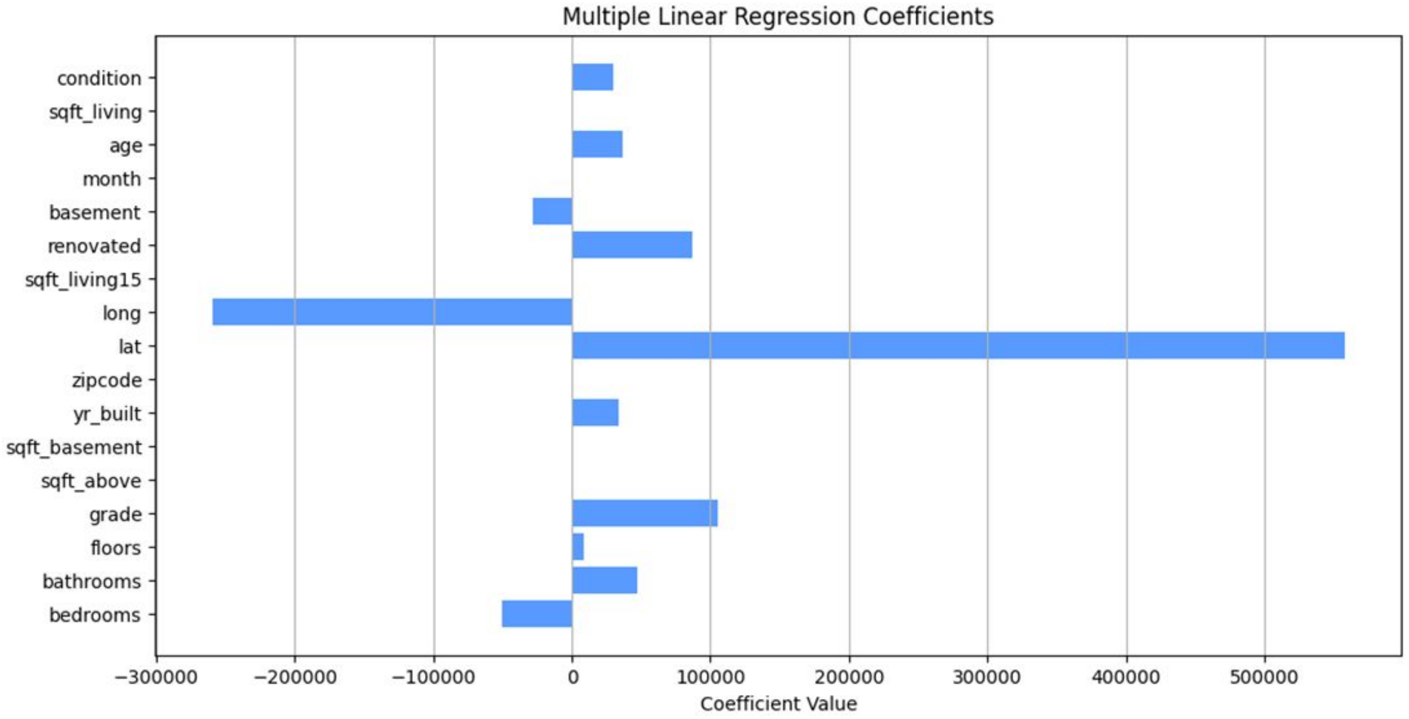
The model has a good explanatory power.Involved transforming features into higher order polynomial terms to model nonlinear relationships.

Log Transformation

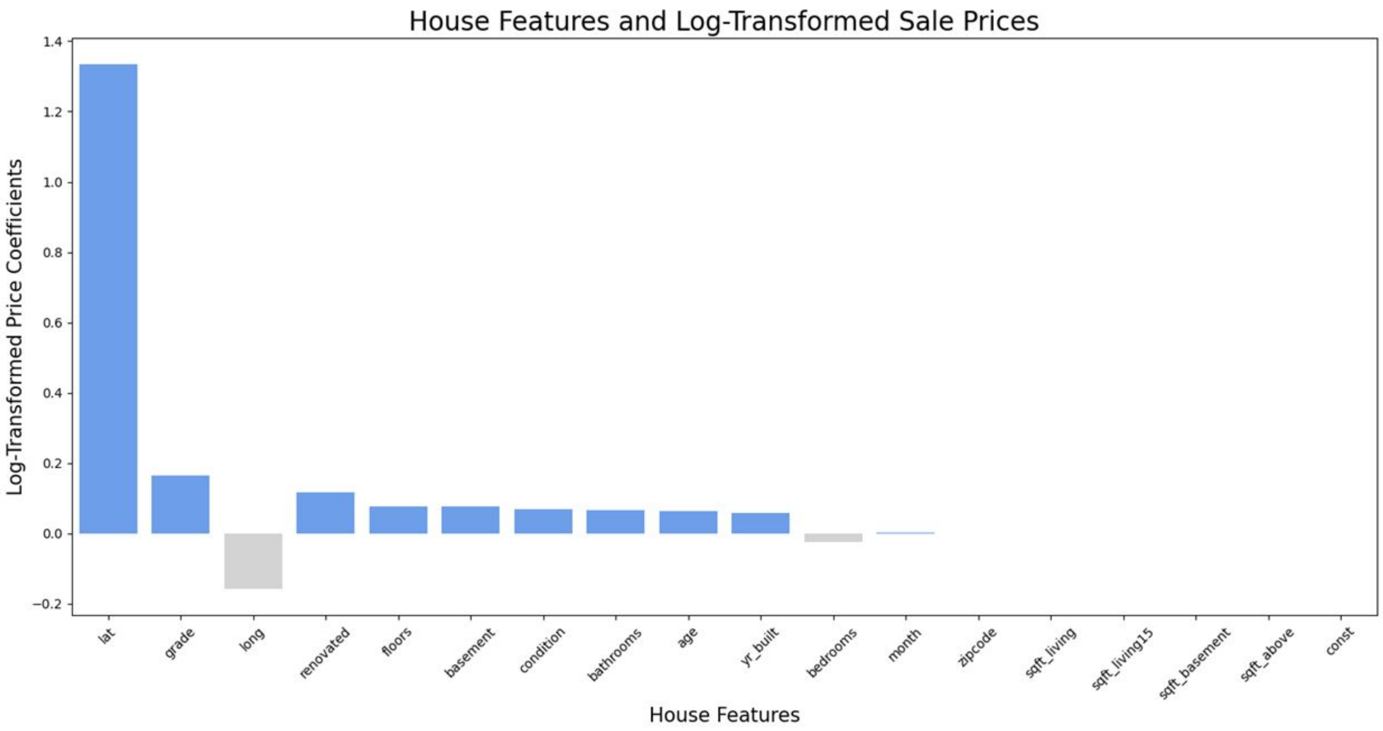
This model aimed to linearize relationships and stabilize the variance.It has improved model performance.



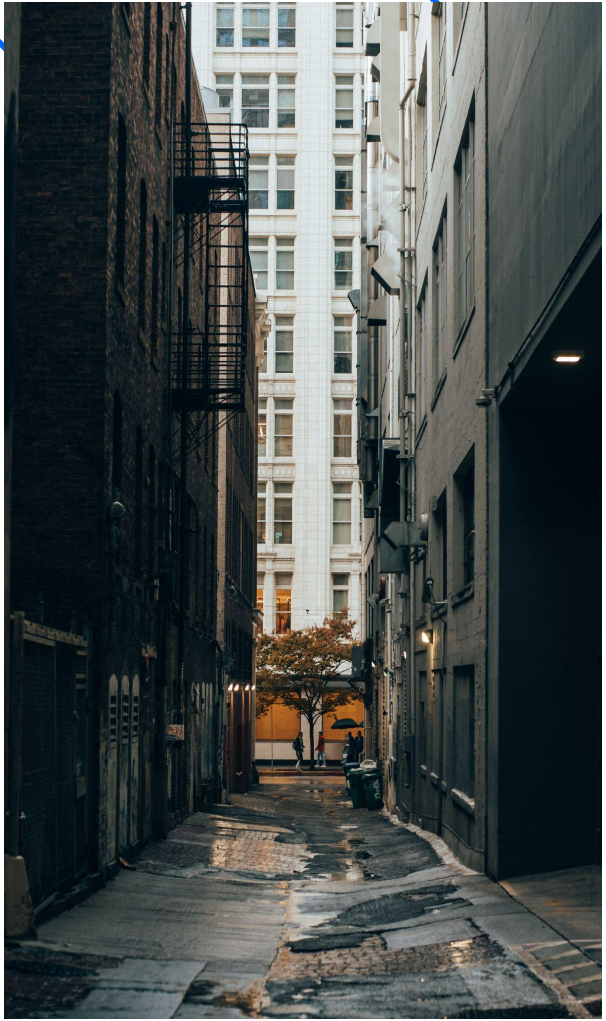
The R-squared is approximately 0.668, suggesting that the model explains around 66.8% of the variance in home prices. Positive coefficients (e.g., sqft_living, bathrooms, grade) suggest an increase in these features corresponds to an increase in home price.



Bar graph illustrating the impact of features on the dependent variable in a multiple linear regression model, where positive and negative bar heights denote the direction and magnitude of influence, providing a concise overview of variable significance.



The R-squared is approximately 0.761, suggesting that the model explains around 76.1% of the variance in home prices. Positive coefficients (e.g., renovated, floors, grade) suggest an increase in these features corresponds to an increase in home price.



Findings

- The log transformation of the variable has improved the models performance as indicated By the low RMSE and higher R squared .
- It has helped capture the underlying patterns in the data more effectively.

MODEL COMPARISON

	MSE	R-Squared	F-stat
Baseline	54.1 b	0.668	4333
Polynomial	40.2 b	0.663	1985
Log Trans	35.1 b	0.761	4637

Model 3 is the preferred model because:
From the evaluation metrics, the models have close performance in terms of MAE and RMSE. However, Model 3, which includes log transformations has the highest R-squared value .

KEY TAKEAWAYS

- As the size of the living space increases, the estimated price of the house also increases. This indicates that larger houses are generally priced higher.
- Older houses fetch higher prices. This could be due to factors such as historical significance or architectural value associated with older houses.
- The more the bedrooms the more expensive the house, and he more space/land a house occupies,the more expensive it is
- As the size of the living space increases, the estimated price of the house also increases. This indicates that larger houses are generally priced higher.

CONCLUSION

RECOMMENDATIONS

Feature Enhancement

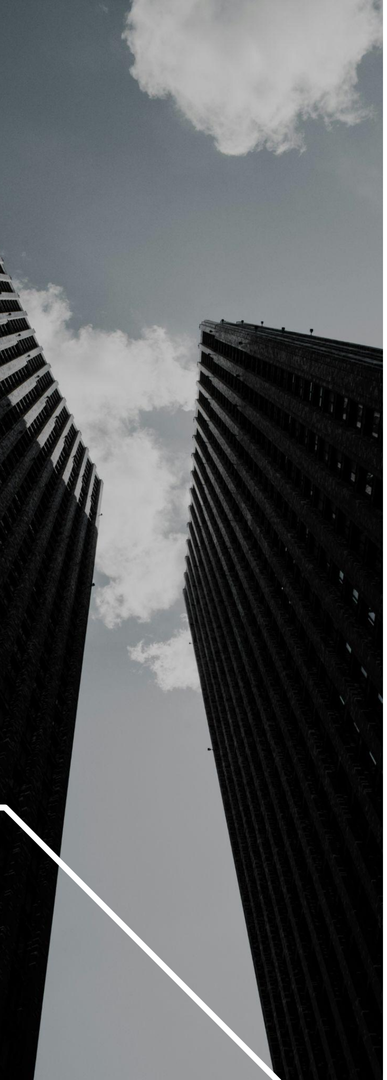
Upgrade features like living space, property grade, or bathrooms to positively impact house prices

Data Collection

Collect additional data on location-specific factors, amenities, property age, and neighborhood characteristics for a more accurate regression model.

Market Segmentation

Analyze correlations between independent variables and house prices to identify market segments, such as luxury properties associated with higher grades



LIMITATIONS

- **Non-Linearity**: Multiple linear regression struggles with non-linear relationships, leading to inaccurate predictions.
- **Causation vs. Correlation**: Correlation doesn't imply causation; the model identifies associations but can't establish cause-and-effect links.
- **Independence Assumption:** The model assumes independent observations; a violation can result in biased standard errors.
- **Linearity Assumption:** The model assumes a linear relationship; if not met, predictions may be biased, and coefficients inaccurate.

NEXT STEPS

- **Alternative Models**: Explore complex models like random forests or neural networks to capture non-linear relationships effectively.
- **Time Trends and Seasonality:** If relevant, use time series models or include time-related features to better capture temporal patterns in the data.
- **Regular Monitoring:** Implement periodic model updates to incorporate new data, ensuring ongoing relevance and effectiveness.

THANK YOU!

The Outliers