IRON KEGGLE MINI PROJECT

A real prediction on real estate

Predicting the future, one home at a time.

RAÍSSA FRANCO
P. KATHERINE TORIAN

Overview

☆	Introduction
★	Objectives and Goals
☆	Data Cleaning
★	Analysis
☆	Modeling and Prediction
★	Evaluation
☆	Final Conclusions
★	Future Outlook

01 The rationale is to predict house prices accurately and identify the key drivers of value, particularly for homes above $650K.

02 The project scope includes data exploration, feature analysis, and predictive modeling using advanced regression techniques.

03 Expected outcomes are insights into price determinants and robust models for predicting property values across different market segments.

Introduction

This project analyzes King County housing sales (2014–2015) with the objective of predicting property prices and identifying the most influential factors. To account for differences between market segments, the dataset was divided into properties below $650K and properties above $650K, enabling targeted modeling and deeper insights into value drivers.

Objectives and Goals

Goal # 1

Predict house prices accurately in King County.

Goal # 2

Identify key factors influencing property values.

Goal # 3

Provide insights for homes valued at $650K and above.

Data Cleaning

Steps taken for an optimized use of the data

Dropped irrelevant columns

Removed columns that aren't useful for prediction like id and date.

Handled numeric and target separation

Extracted numeric features (_get_numeric_data()) and separated the target variable price.

Feature engineering

- Created an age column: current_year - yr_built.
- Created a renovated binary column: 1 if yr_renovated > 0, else 0.
- Removed redundant columns (yr_built, yr_renovated)

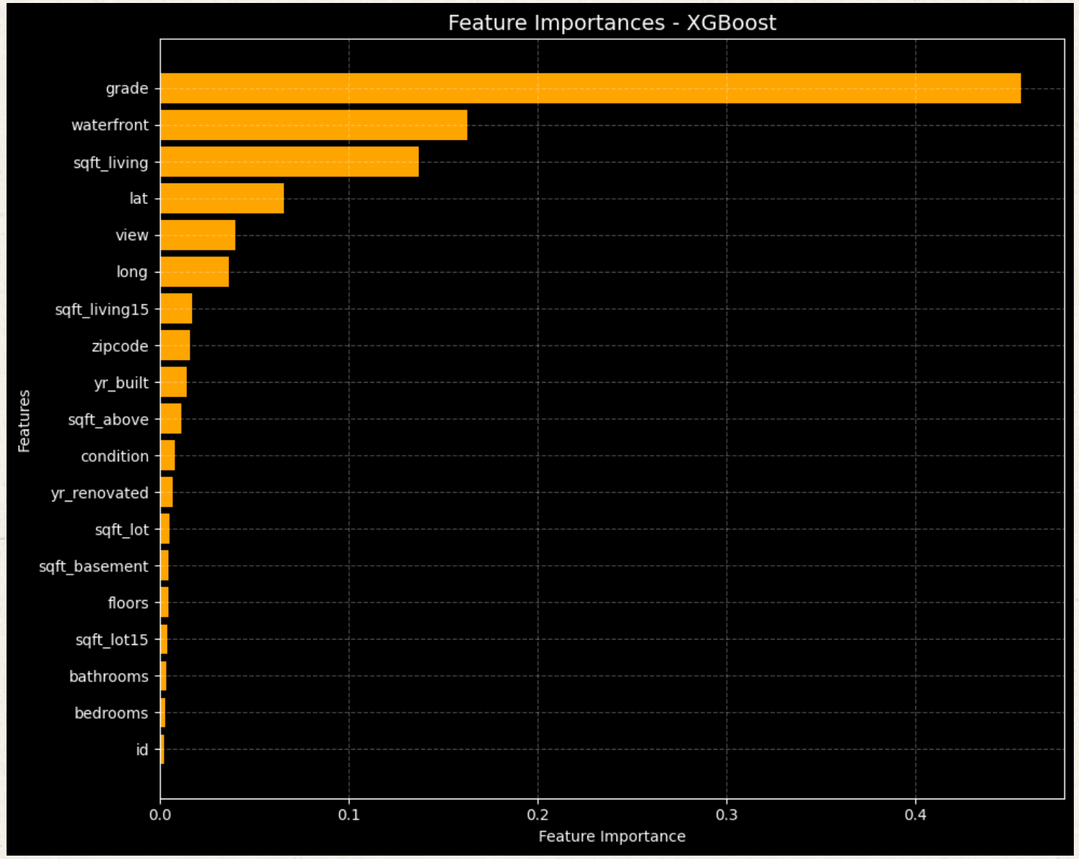
Converted columns

Changed the column types to be categorical so we can handle the data without receiving any errors with our models.

Analysis

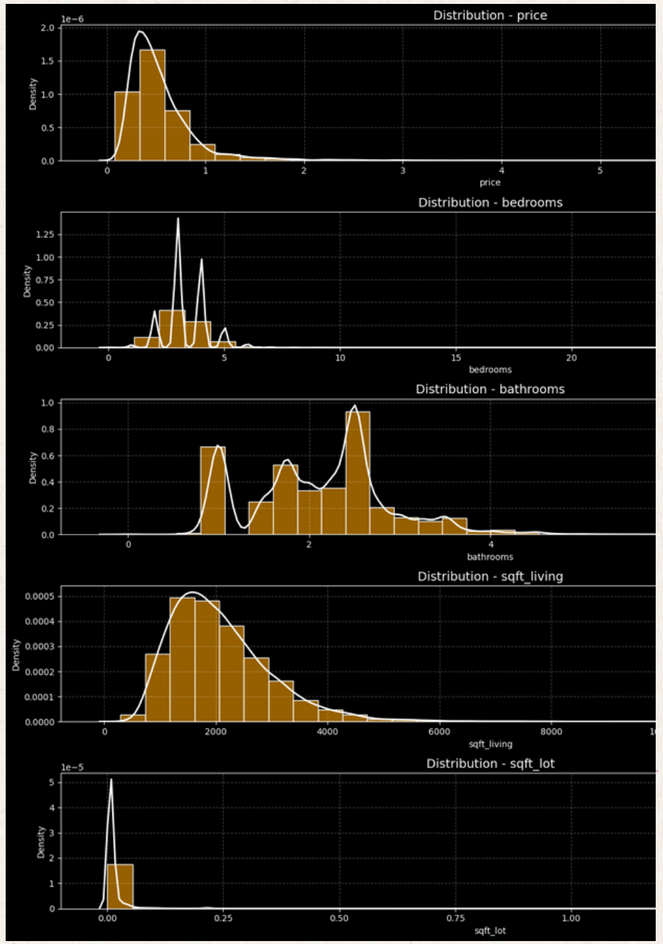
We first conducted various univaried analysis so we could get information on which columns we should consider to proceed with our prediction models.

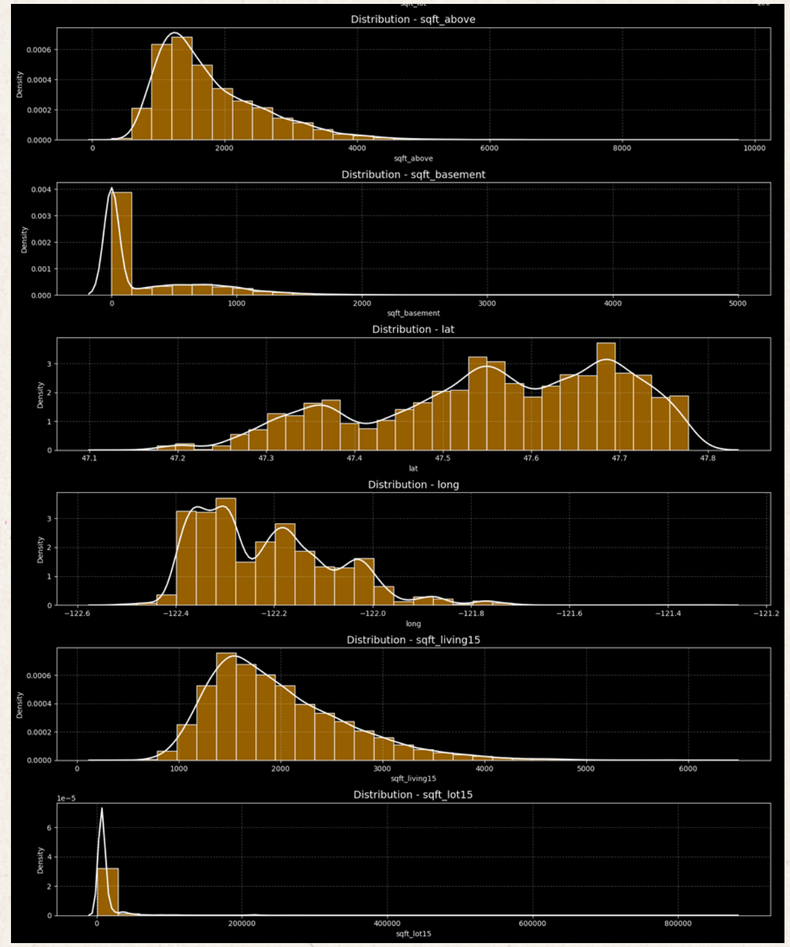
A house's grade shouldn't be the most important feature when valuing a house, so we concluded that our features have to be enhanced and improved so we can work with them.

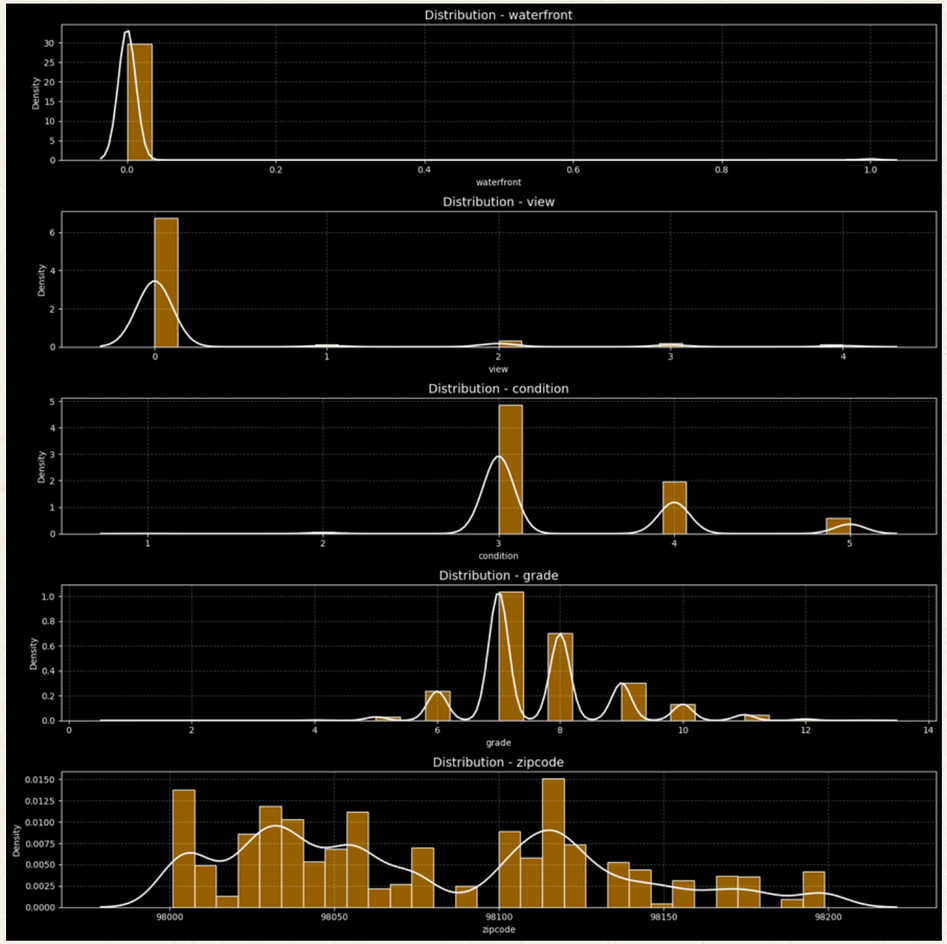


Analysis

After seeing the distribution, we can clearly see that some of the alleged most important features lose importance and are lower and less relevant in these plots.

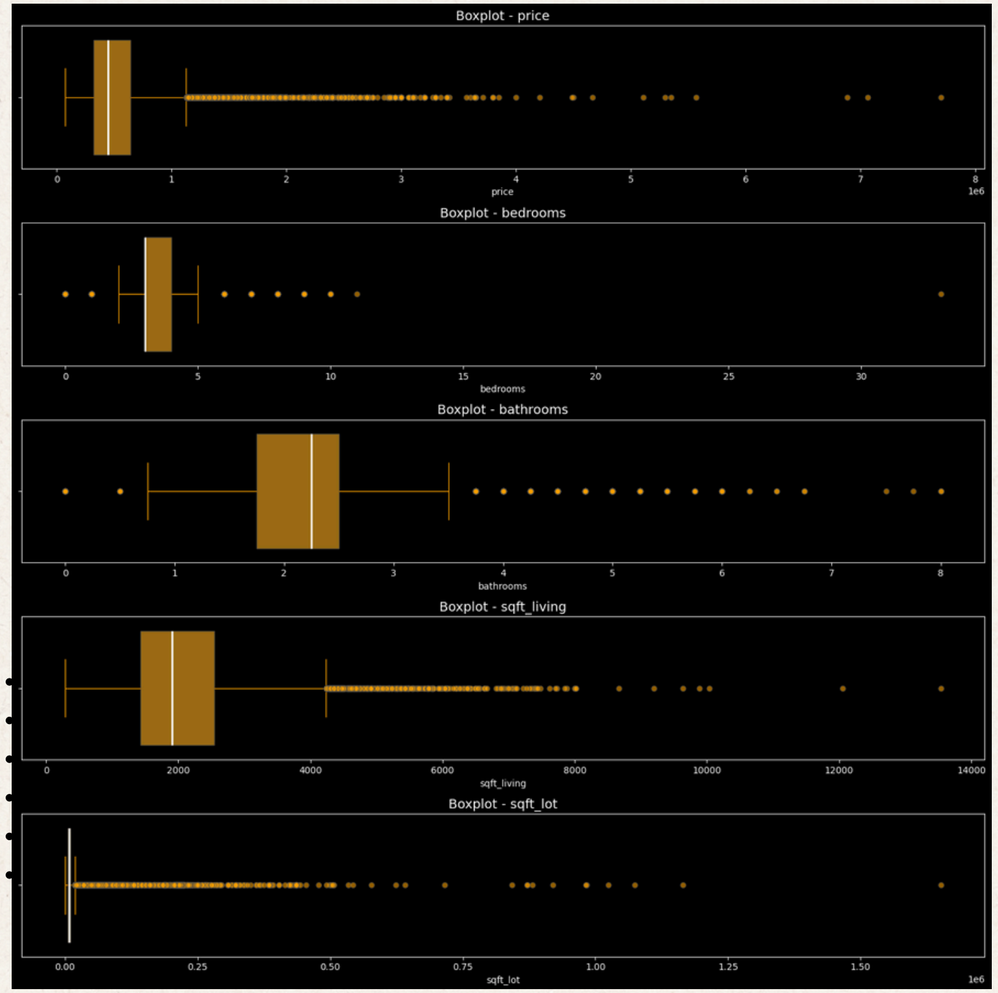


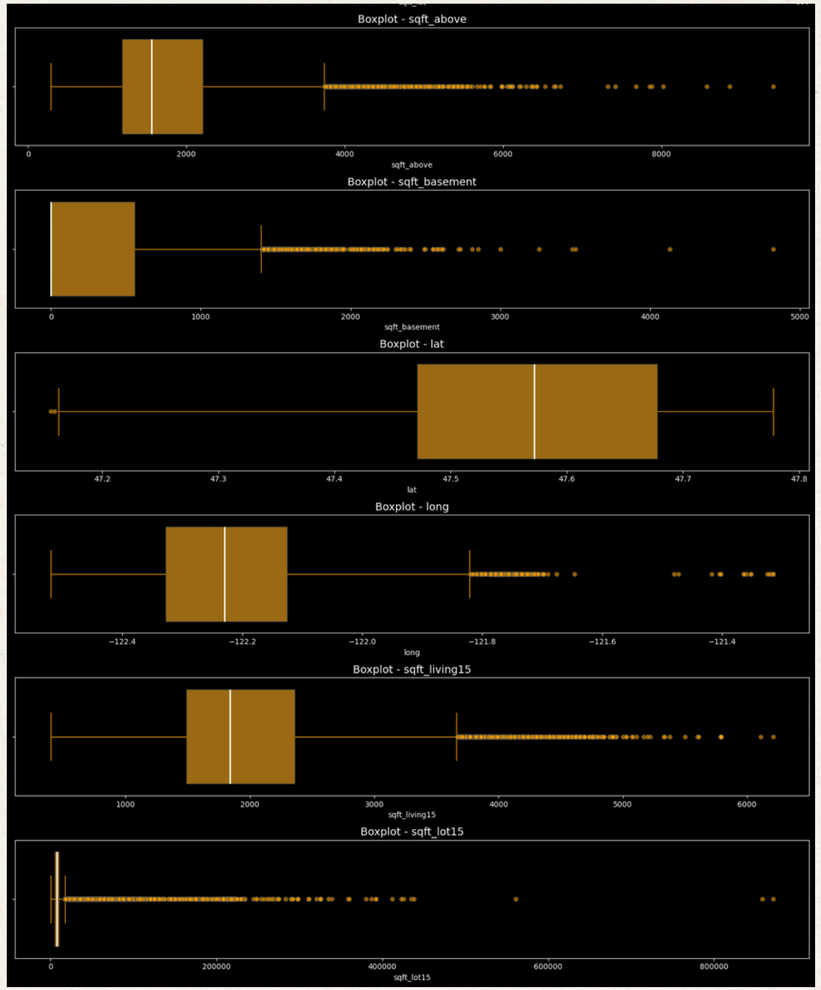




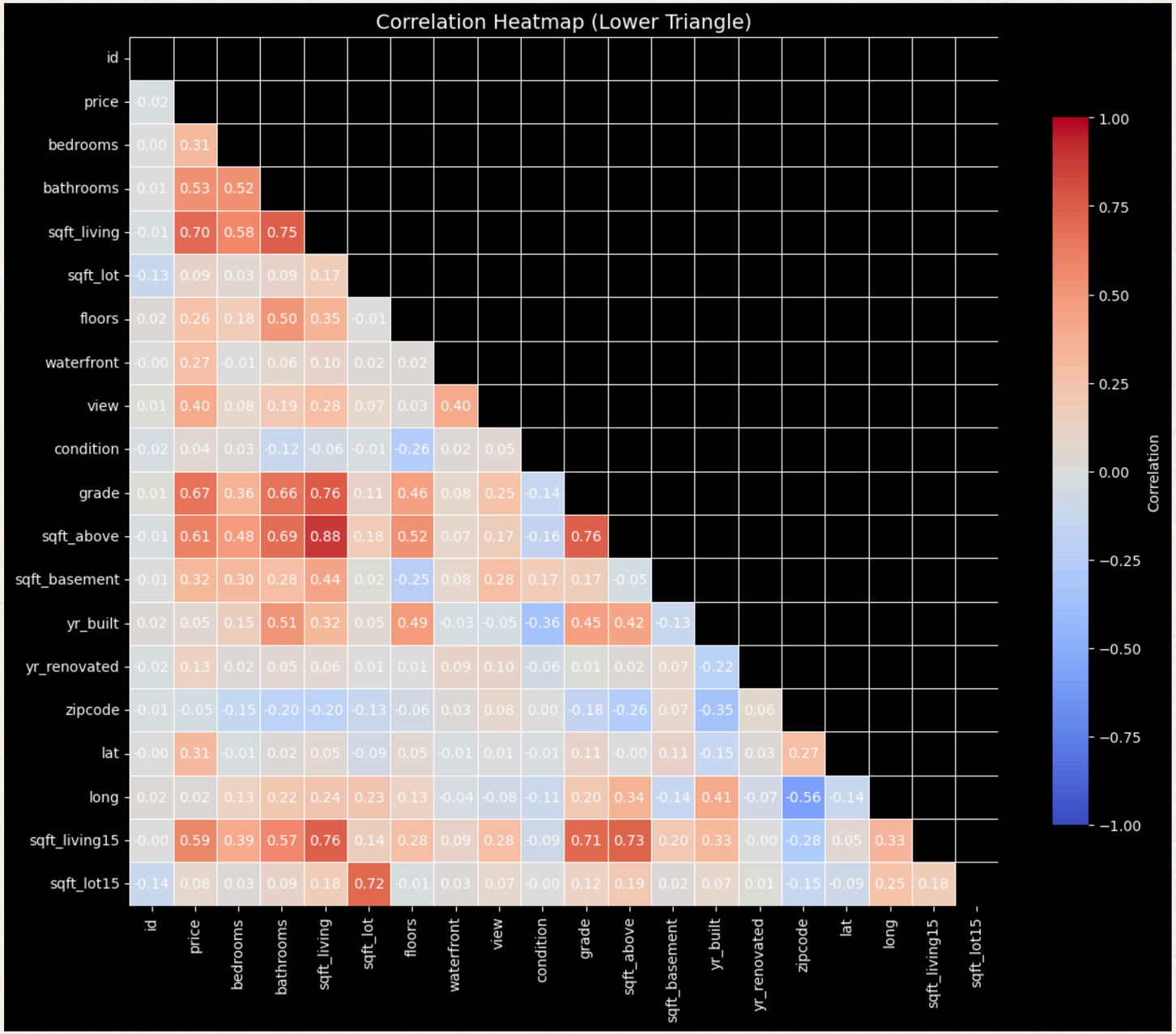
Analysis

As we observed some columns with outliers, we determined we'd have to treat them after to receive better results.





Analysis


Correlation Heatmap (Lower Triangle)
id
price -0.02
bedrooms -0.00 0.31
bathrooms -0.01 0.53 0.52
sqft_living -0.01 0.70 0.58 0.75
sqft_lot -0.13 0.09 0.03 0.09 0.17
floors 0.02 0.26 0.18 0.50 0.35 -0.01
waterfront -0.00 0.27 -0.01 0.06 0.10 0.02 0.02
view -0.01 0.40 0.08 0.19 0.28 0.07 0.03 0.40
condition -0.02 0.04 0.03 -0.12 -0.06 -0.01 -0.26 0.02 0.05
grade 0.01 0.67 0.36 0.66 0.76 0.11 0.46 0.08 0.25 -0.14
sqft_above -0.01 0.61 0.48 0.69 0.88 0.18 0.52 0.07 0.17 -0.16 0.76
sqft_basement -0.01 0.32 0.30 0.28 0.44 0.02 -0.25 0.08 0.28 0.17 0.17 -0.05
yr_built -0.02 0.05 0.15 0.51 0.32 0.05 0.49 -0.03 -0.05 -0.36 0.45 0.42 -0.13
yr_renovated -0.02 0.13 0.02 0.05 0.06 0.01 0.01 0.09 0.10 -0.06 0.01 0.02 0.07 -0.22
zipcode -0.01 -0.05 -0.15 -0.20 -0.20 -0.13 -0.06 0.03 0.08 0.00 -0.18 -0.26 0.07 -0.35 0.06
lat -0.00 0.31 -0.01 0.02 0.05 -0.09 0.05 -0.01 0.01 -0.01 0.11 -0.00 0.11 -0.15 0.03 0.27
long 0.02 0.02 0.13 0.22 0.24 0.23 0.13 -0.04 -0.08 -0.11 0.20 0.34 -0.14 0.41 -0.07 -0.56 -0.14
sqft_living15 -0.00 0.59 0.39 0.57 0.76 0.14 0.28 0.09 0.28 -0.09 0.71 0.73 0.20 0.33 -0.00 -0.28 0.05 0.33
sqft_lot15 -0.14 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 0.01 -0.15 -0.09 0.25 0.18
id price bedrooms bathrooms sqft_living sqft_lot floors waterfront view condition grade sqft_above sqft_basement yr_built yr_renovated zipcode lat long sqft_living15 sqft_lot15
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Correlation

Modeling and Prediction

Linear Regression

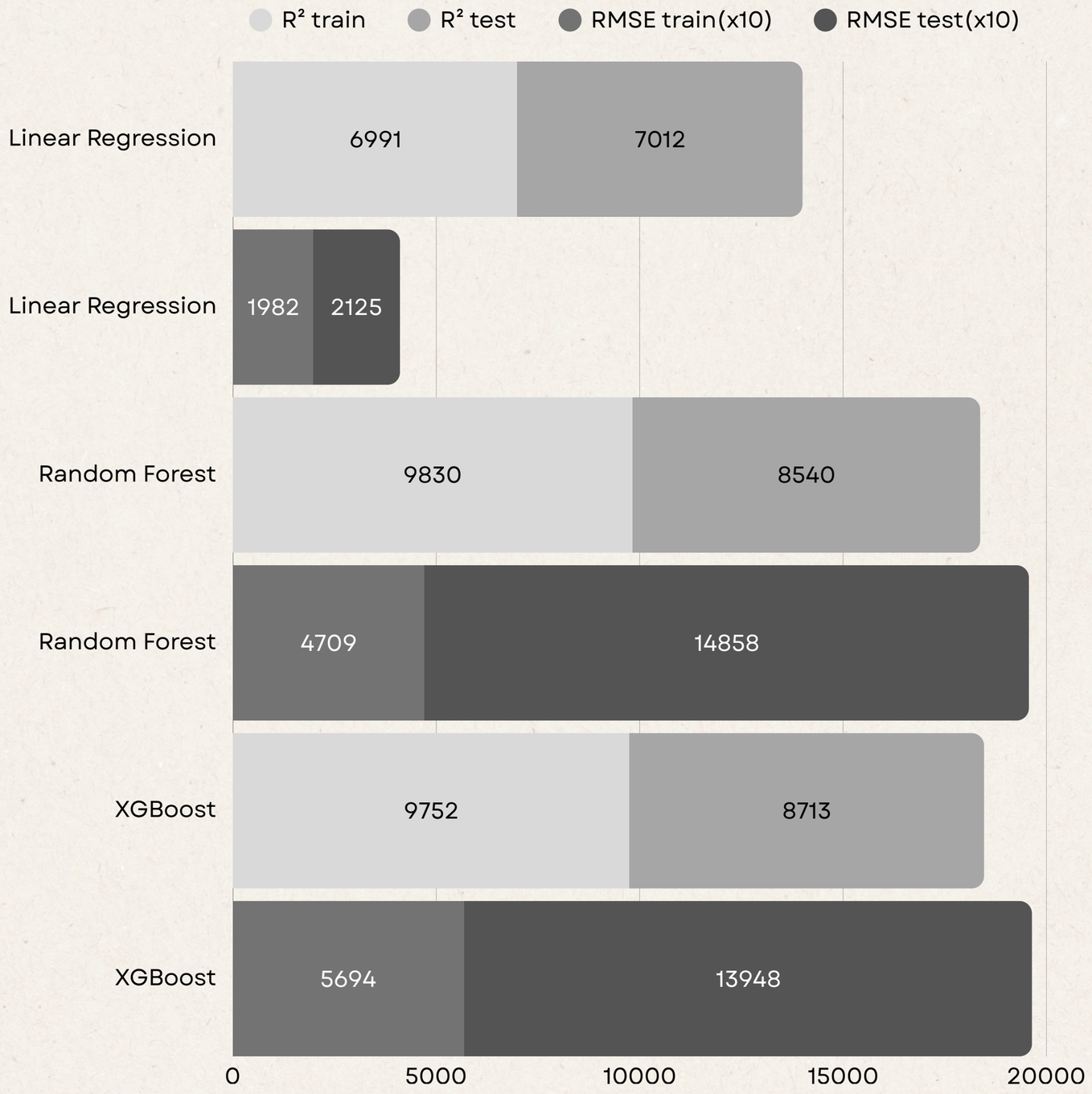
- Model explains ~70% of the variance in the target.
- Model generalizes well.

Random Forest

- More powerful than Linear Regression: better test R^2 (0.854 vs 0.701) and lower RMSE (~149k vs ~212k).
- Slight overfitting is present, but it's within an acceptable range.

XGBoost Regressor

- Best predictive performance among the three models:
- Highest test R^2 → Most variance on unseen data.
- Lowest test RMSE → Smallest average prediction error.
- Slight overfitting is present but less severe than Random Forest.



Modeling and Prediction

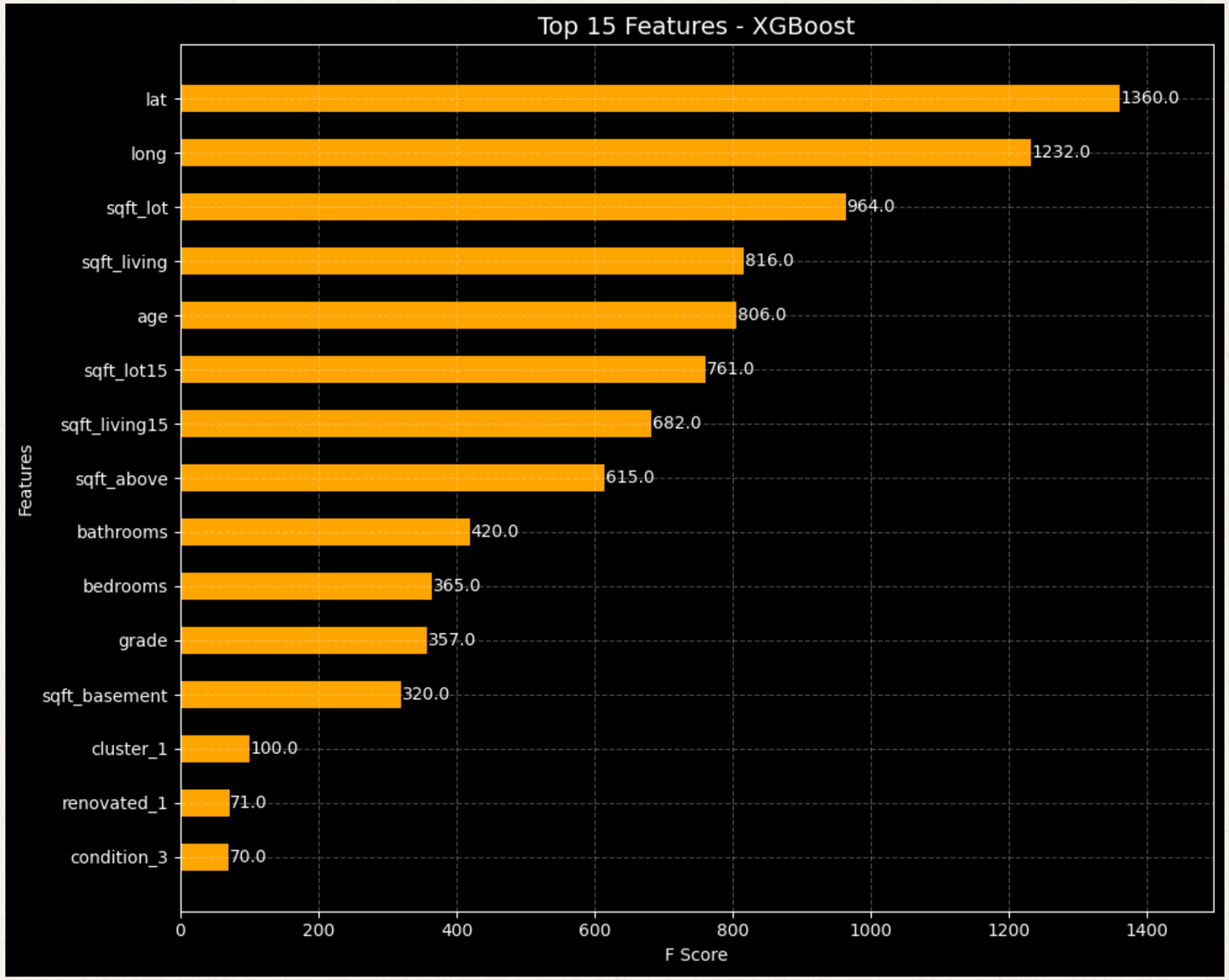
XGBoost Regressor

- Feature engineering
- Hyperparameters

```
colsample_bytree': 0.7, 'learning_rate': 0.1,
'max_depth': 6, 'n_estimators': 200, 'subsample': 1.0}
```

Evaluation

The most important features, after carefully deciding and cleaning the data, include the location (latitude and longitude) and the size of the house (sqtf_lot and sqtf_living). The age of the house is also an influencing factor.
It'd be best to use these features in our models.



Houses over $650k

As our XGBoost Regressor model was the most accurate before, we're using that one to predict the values of the houses that have prices over $650k.

The challenge? Avoiding overfitting, as we have very little data.

We created two subsets, so we can also analyze and predict the prices of houses under $650k.

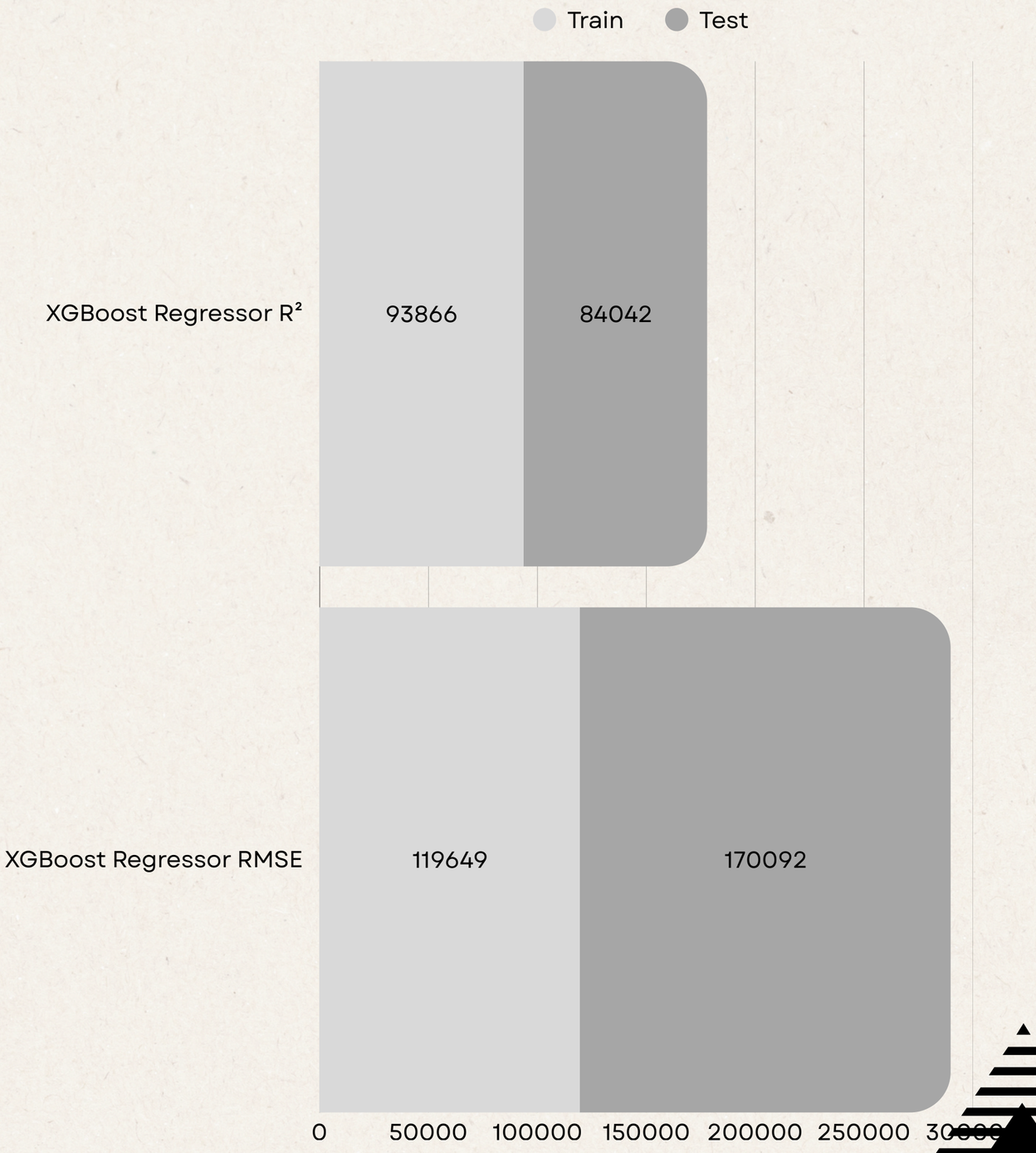
Focusing on the houses over $650k...

R^2:

- The model fits the training data very well.
- Shows some overfitting, but the model still explains a large portion of the variance in unseen data.

RMSE:

- The test shows some overfitting, but not severe.



Final Conclusion

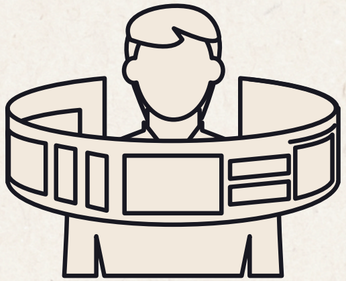
General project conclusions

- Location matters: Latitude and longitude strongly influence prices.
- Property features count: Size, age, bathrooms, bedrooms all contribute significantly.
- Feature engineering helped: Adding derived variables like age improved performance and balanced feature importance.
- Model is robust: XGBoost achieves high test R^2 and low RMSE, generalizing well to new data.
- Our model can guide pricing and investment decisions effectively.

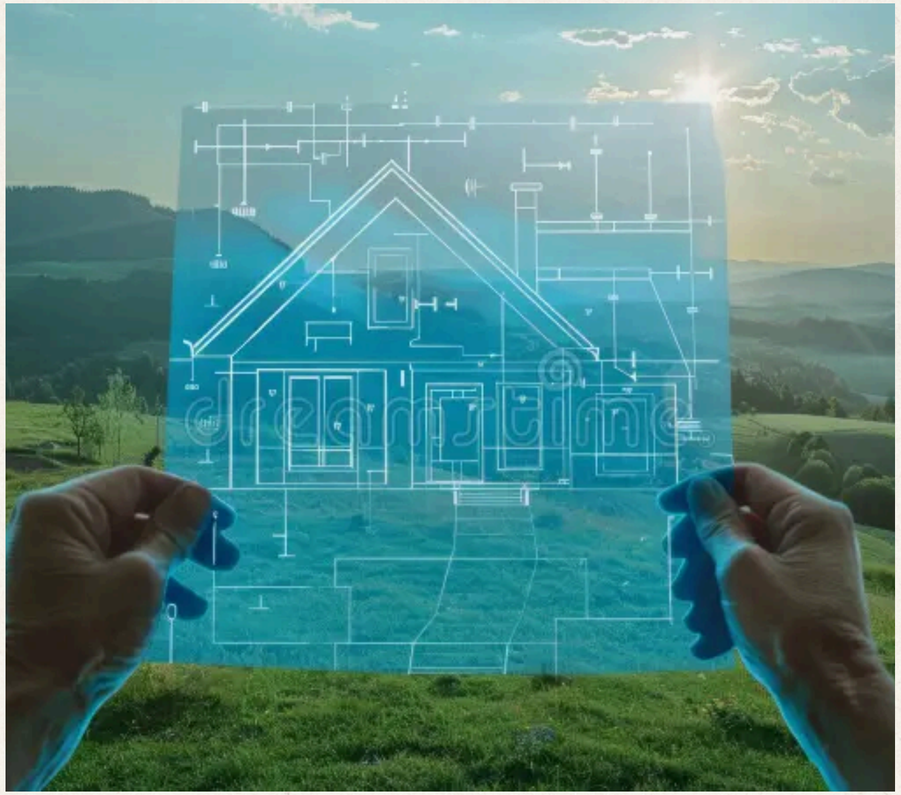
Houses over $650k

- The model performs well in terms of explaining variance (R^2 0.84 on test).
- It overfits a little, but that's common with high-priced predictions because the data range is wider and more variable.
- Absolute prediction errors are large (~170k), but relative to the high-price range, this might be reasonable.

Future Outlook



- Housing prices are expected to rise steadily, with well-located areas seeing the strongest growth.
- Larger homes with modern amenities and high-quality features will continue to be in higher demand.
- Neighborhood and spatial factors will remain critical in determining home values.
- Overall market trends are likely to stay stable and predictable, reflecting multiple contributing factors.
- Premium properties in desirable locations are predicted to appreciate faster than average.



Thank you

Any questions?