

# King County Real Estate Market Analysis: May 2014 - May 2015

By: Jithin Kumar and Natalia Dominguez



# Overview

-  Project Overview 01
-  Dataset Description 02
-  Key Insights 03
-  Process 04
-  Models Tested 05
-  Metrics & Evaluation 06
-  Best Model 07
-  Improvements Made 08
-  Key Factors 09
-  Conclusions and Future Steps 10





# Project Overview

This project analyzes house sale prices in King County, including Seattle, over one year (May 2014 – May 2015).

The main goal is to predict house prices and identify the most influential factors affecting them. The project involves data exploration and modeling, simulating real-world real estate analysis while applying Python skills and analytical techniques to extract actionable insights from the housing market.



# Dataset Description

- House sales in King County (Seattle), May 2014 – May 2015
- 21 features per property: bedrooms, bathrooms, sqft, floors, waterfront, view, condition, grade, year built/renovated, location
- Target variable: price (sale price of the house)



# Key Insights

Before moving into modeling, we explored the dataset to detect inconsistencies and patterns. During this exploration, we found some relevant aspects that shaped how we approached the analysis.

01

Converted the date column into proper datetime format for accurate analysis.

02

Detected duplicate house IDs, indicating multiple sales of the same property.

03

Decided to follow two approaches:

1. Using the full dataset.
2. Keeping only the most recent transaction per house, since it typically had the highest sale price.

# Process

- 01 **Data Exploration & Cleaning:**  
overview, handle nulls/duplicates
- 02 **EDA:** visualize patterns
- 03 **Prepare Data:**  
split features/target, encode categoricals
- 04 **Modeling & Tuning:**  
train models and improve performance
- 05 **Feature Importance:** identify key predictors

# Models Tested

Linear Regression

ADAboost

XGBoost

Random Forest

Gradient Boost

Cat Boost

# Metrics & Evaluation

## Linear Regression

All Data

R2 Linear Regression Train: 0.69910  
R2 Linear Regression Test: 0.70119

Drop data

R2 Linear Regression Train: 0.69740  
R2 Linear Regression Test: 0.70614

## Random Forest

All Data

Mean Squared Error: 2228496940.85  
R2 Random Forest Train: 0.98294

Mean Squared Error: 21807588291.06  
R2 Random Forest Test: 0.85575

Drop Data

Mean Squared Error: 2391235579.01  
R2 Random Forest Train: 0.98208

Mean Squared Error: 15992427088.73  
R2 Random Forest Test: 0.88639

## ADABOOST

R2 ADARegressor Train: 0.70886  
R2 ADARegressor Test: 0.65693

R2 ADARegressor Train: 0.70671  
R2 ADARegressor Test: 0.68257

## Gradient Boost

R2 Gradient Train: 0.98294  
R2 Gradient Test: 0.85575

R2 Gradient Train: 0.69062  
R2 Gradient Test: 0.69180

## XGBoost

R2 XGboost Train: 0.97646  
R2 XGboost Test: 0.85513

R2 XGboost Train: 0.97673  
R2 XGboost Test: 0.89680

## Cat Boost

Mean Squared Train Error: 68784.50  
R2 Cat Train: 0.96379

Mean Squared Test Error: 125663.72  
R2 Cat Test: 0.89554

Mean Squared Train Error: 68536.94  
R2 Cat Train: 0.96480

Mean Squared Test Error: 111021.33  
R2 Cat Test: 0.91244

# Best Models

CatBoost

XGBoost

Random Forest

# Improvements made

Before finalizing the model, we added new features and refined the data to help the algorithm capture more patterns and improve predictions.

01

Added month of sale as a new feature to keep both house sale prices for duplicate entries.

02

Removed irrelevant columns and split data into train/test sets (80/20).

03

Trained a CatBoost regressor to improve predictions on house prices.

## Results

Cat Boost

Mean Squared Train Error: 66879.13

R2 CatBoost Train: 0.96576

---

Mean Squared Train Error: 124742.05

R2 CatBoost Test: 0.89707

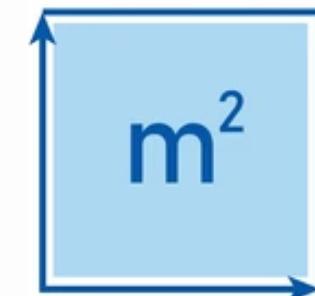
# Feature Importance



**Lat**

Latitude coordinate.

28.2737



**Sqft\_living**

Square footage of the interior living space.

17.9468



**Grade**

The overall grade given to the house, based on the King County grading system.

13.488

# Conclusions and Future Steps

01

## Conclusions:

- Data analysis was performed on King County housing dataset (21k+ records).
- After cleaning, key features affecting house prices were identified.
- Latitude, living area (sqft\_living), and grade are the strongest predictors of price.
- Features like bedrooms and floors showed little influence on pricing.
- The CatBoost model provided reliable insights into price prediction.
- Results can help homeowners, buyers, and real estate agents make informed decisions.

02

## Future Steps:

- Perform hyperparameter tuning on the model to optimize performance.
- Experiment with different train-test splits to evaluate model stability.
- Address overfitting (currently ~10%) by reducing the training set size, which may help improve test accuracy.

# THANK YOU!

