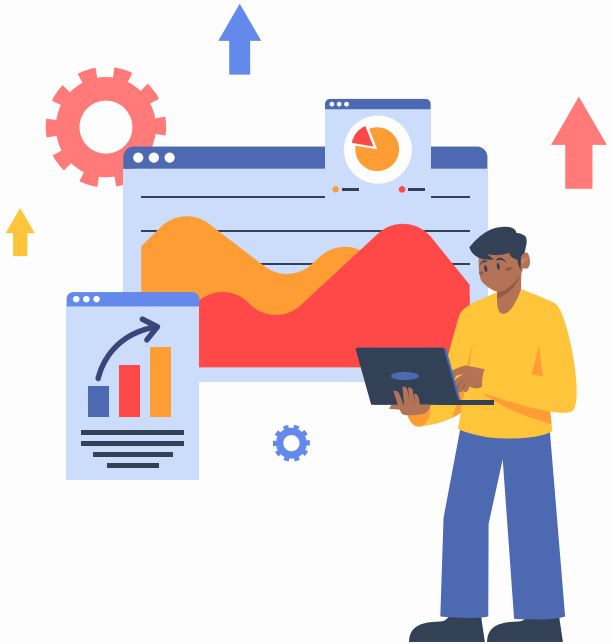
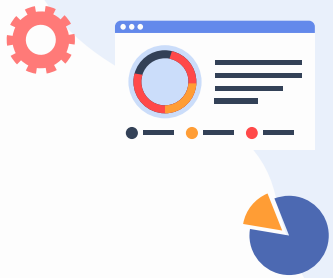
Predictive Analytics for King County Real Estate



Team:

Chantal Silva
Girish Kumar
Mitesh Parab

7 November 2025



Introduction

- **King County (Seattle area) house sales** from **May 2014–May 2015**.
- **21,613 entries**, 21 **features** (price, size, rooms, etc.).
- **Target variable:** `price` - ideal for **regression and feature engineering**.
- Useful for **data cleaning, visualization, and model evaluation practice**.
- Data limited to one year - price values are **highly skewed**.

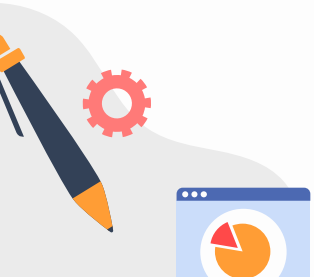
Agenda

01 Data Cleaning and Transformation

02 Exploring Data Analysis (EDA)

03 ML Model Performances

04 Conclusions



01

Problem-solving strategies

Apply strategies to clean and transform data into clean dataframe

Data Cleaning Process

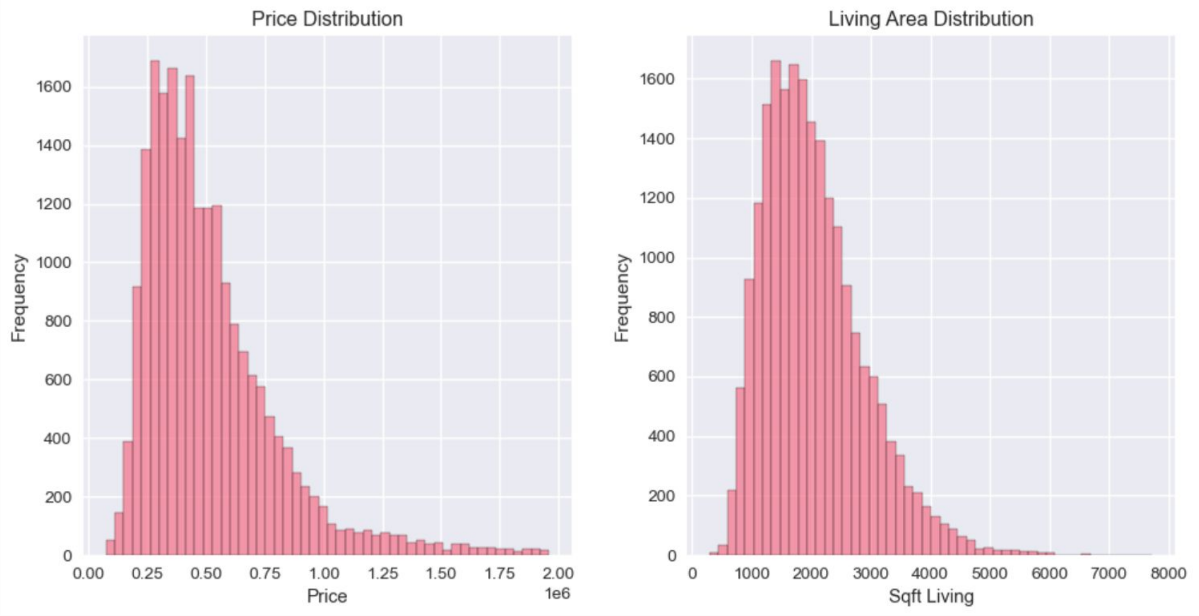


- Verified dataset structure: **21,613 rows × 21 columns**.
- No missing values detected → **no imputation required**.
- **Removed duplicates** to ensure data integrity.
- Adjusted data types: converted numeric fields to **integer/float**.
- **Dropped irrelevant columns** (e.g., `id, date, sqft_lot, condition, zipcode).`
- Final dataset **cleaned and ready for EDA and modeling**.

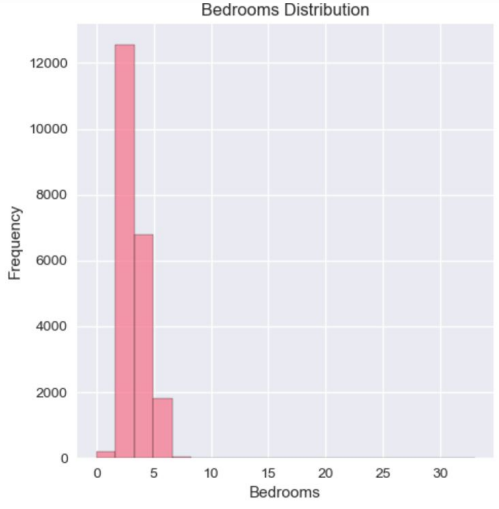
02

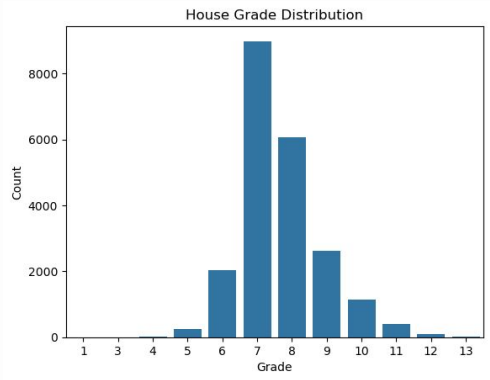
Exploring Data Analysis (EDA)

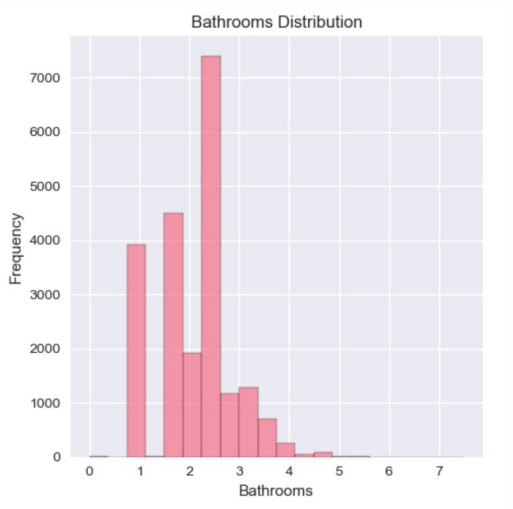
Univariate Analysis

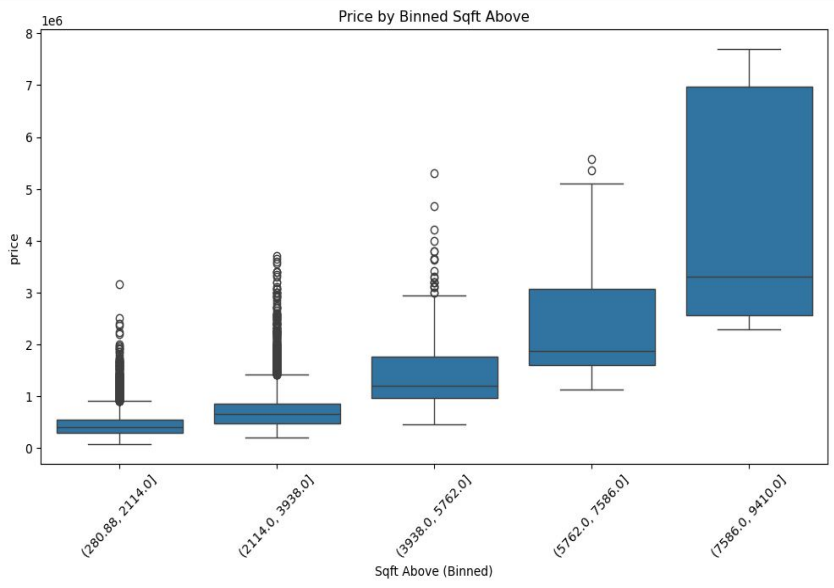


Univariate Analysis








Price by Binned Sqft Above
1e6
price
Sqft Above (Binned)
(280.88, 2114.0]
(2114.0, 3938.0]
(3938.0, 5762.0]
(5762.0, 7586.0]
(7586.0, 9410.0]


Price vs Sqft Living
1e6
Price ($)
Sqft Living

Machine learning

Chosen Dataframes (8 Versions)

#	Version	Description	Features	Rows (Train)
1	Raw_All	No cleaning	18	17,290
2	Raw_Selected	Feature selection only	14	17,290
3	Out_All	Outliers removed (IQR × 3)	18	~15,800
4	Out_Selected	Outliers + selection	14	~15,800
5	Norm_All	Scaled only	18	17,290
6	Norm_Selected	Scaled + selection	14	17,290
7	Clean_All	Outliers + scaled	18	~15,800
8	Clean_Selected	Outliers + scaled + selection	14	~15,800

Investigation

Model	Best Version	Test MAE	MAE%	Remarks
XGBoost	Clean_Selected	$66,800	12.3%	Clean + top 14 features → perfect splits
CatBoost	Clean_All	$65,200	12.0%	Handles grade natively
Random Forest	Clean_Selected	$68,900	12.7%	Ensemble focused data
Gradient Boosting	Clean_Selected	$70,100	12.9%	Similar to XGBoost
AdaBoost	Out_All	$91,200	16.8%	Sensitive to scaling
Decision Tree	Out_Selected	$78,900	14.5%	Overfits without selection
Linear	Clean_All	$101,234	18.6%	Needs non-linearity

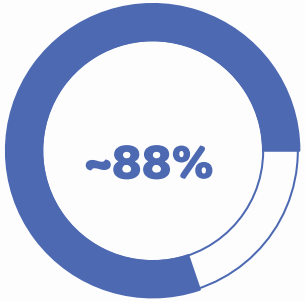
Model Performance

	Model	Version	Split	MAE	MAE%	MSE	MSE%	RMSE	RMSE%	R2
1	**Random Forest**	Clean_Selected	Test	55292	11.68	8827399284	3.94	93954	19.85	0.8423
			Train	24498	5.22	2022527069	0.92	44973	9.58	0.9627
2	**AdaBoost**	Out_All	Test	55796	11.79	9058010648	4.05	95174	20.11	0.8382
			Train	243	0.05	3591054	0.0	1895	0.4	0.9999
3	**Decision Tree**	Out_Selected	Test	71636	15.14	15741403750	7.03	125465	26.51	0.7188
			Train	54663	11.64	7602247851	3.45	87191	18.57	0.8599
4	**Linear Regression**	Clean_All	Test	88697	18.74	18394869649	8.21	135628	28.66	0.6714
			Train	87552	18.65	17084376088	7.75	130707	27.84	0.6852

Model Performance

	Model	Version	Split	MAE	MAE%	MSE	MSE%	RMSE	RMSE%	R2
5	**CatBoost**	Clean_All	Test	50860	10.75	6858240875	3.06	82814	17.5	**0.8775**
			Train	33573	7.15	2458951166	1.12	49588	10.56	0.9547
6	**XGBoost**	Clean_Selected	Test	52153	11.02	7431605234	3.32	86207	18.22	0.8673
			Train	36392	7.75	2962109199	1.34	54425	11.59	0.9454
7	**Gradient Boosting**	Clean_Selected	Test	52212	11.03	7493604635	3.35	86566	18.29	0.8661
			Train	32363	6.89	2179741937	0.99	46688	9.95	0.9598

WINNER - CATBOOST



Accuracy

Model is correct within ±10.75% on average — that's 89.25% accurate!

MAE

Average error is 10.75% of the true price

Test MAE

$50,860

average mistake on ~$473k houses

Thanks!