Exploratotry analysis and modeling of

KING COUNTY HOUSE PRICES

Presented by:
Eya Cherif, Alexandre Andrade and Marius Gören

Ironhack
DSML JAN2026

```
RangeIndex: 21613 entries, 0 to 21612
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   id             21613 non-null  int64
 1   date           21613 non-null  object
 2   price          21613 non-null  float64
 3   bedrooms       21613 non-null  int64
 4   bathrooms      21613 non-null  float64
 5   sqft_living    21613 non-null  int64
 6   sqft_lot       21613 non-null  int64
 7   floors         21613 non-null  float64
 8   waterfront     21613 non-null  int64
 9   view           21613 non-null  int64
 10  condition      21613 non-null  int64
 11  grade          21613 non-null  int64
 12  sqft_above     21613 non-null  int64
 13  sqft_basement  21613 non-null  int64
 14  yr_built       21613 non-null  int64
 15  yr_renovated   21613 non-null  int64
 16  zipcode        21613 non-null  int64
 17  lat            21613 non-null  float64
 18  long           21613 non-null  float64
 19  sqft_living15  21613 non-null  int64
 20  sqft_lot15     21613 non-null  int64
dtypes: float64(5), int64(15), object(1)
memory usage: 3.5+ MB
```

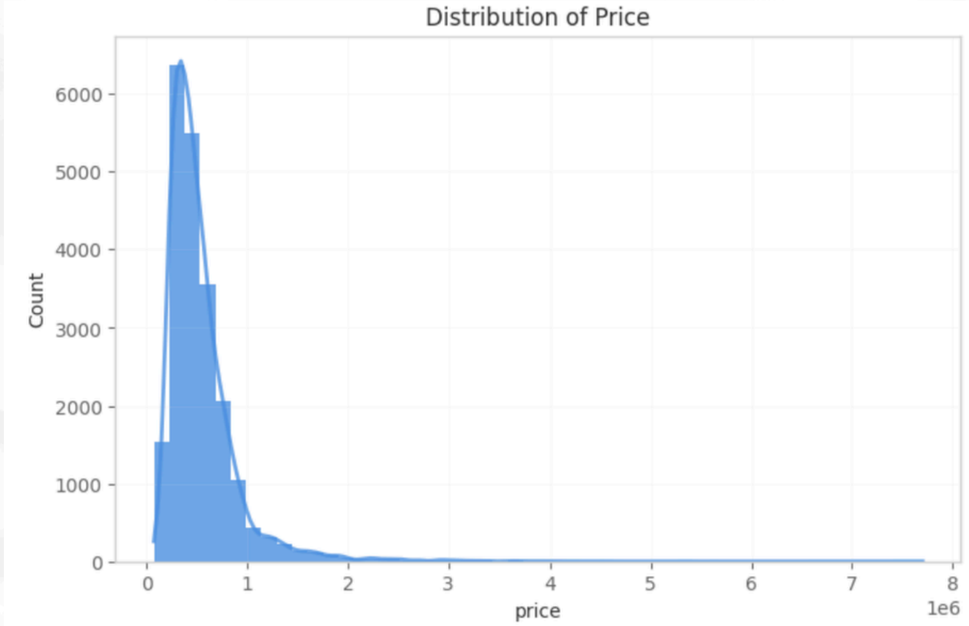
Dataset description

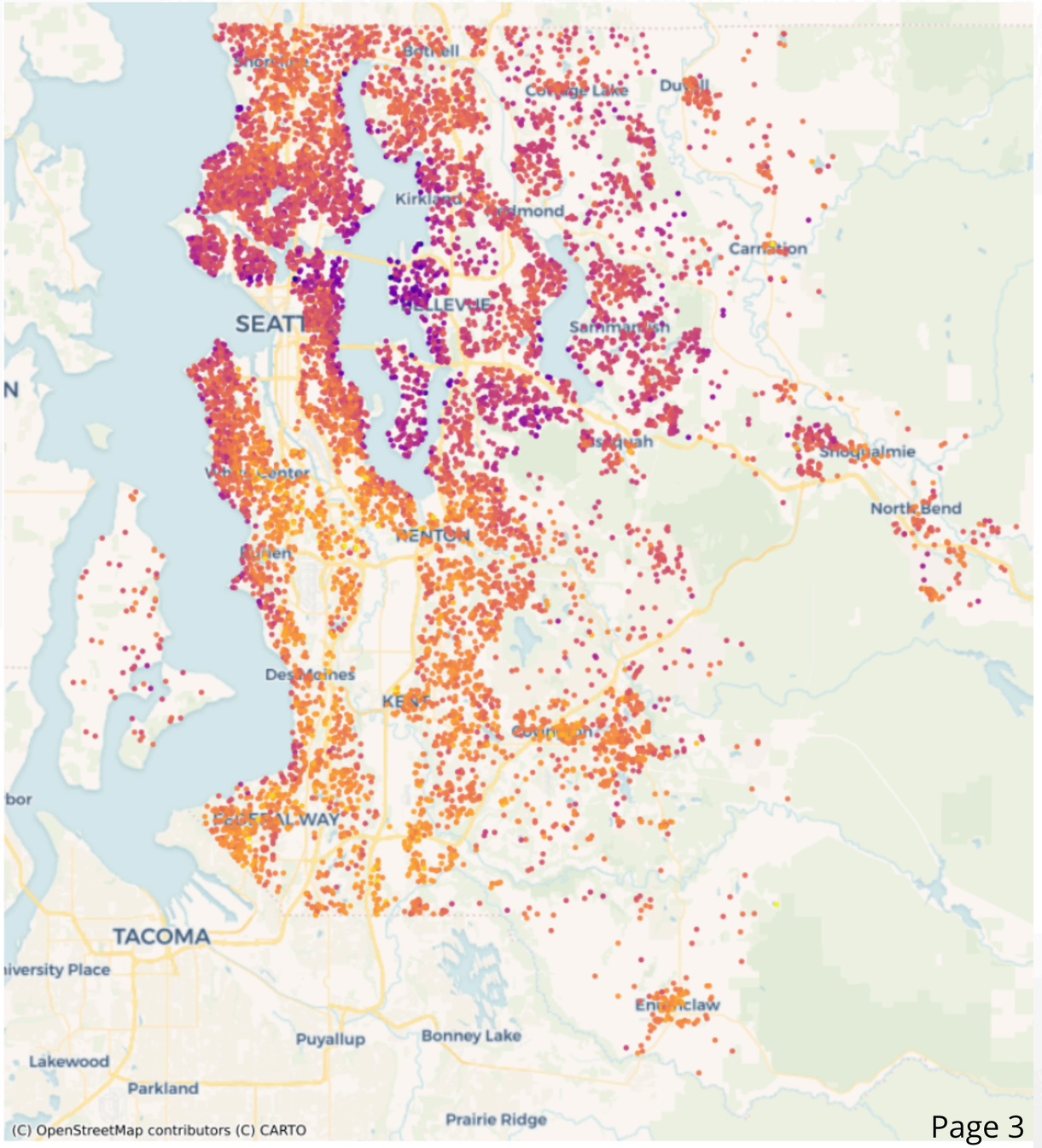
Target: Price

EDA

Data Quality

- No missing values
- Duplicated IDs (multiple sales) - needs handling
- Outliers in price, sqft features, and bedrooms
- Some suspicious rows





Baseline model

- Data cleaning:
  - date splitting into year_sold and month_sold

- We used the Linear model
- Performance:

LinearRegression

	r2_score	RMSE	MAE
train	0.701	200436.651	125635.378
test	0.703	201460.995	124748.824

Feature engineering

The cleaning the advances cleaning and processing:

- drop 0 bedrooms 0 bathrooms rows
- Processing of outliers insqft_living, sqft_lot, sqft_living15, sqft_lot15 with winsorization

LinearRegression

	r2_score	RMSE	MAE
train	0.748	159522.092	108867.509
test	0.753	158109.265	108572.783

New Features	Description
Renovated instead of yr_renovated	Binary column (renovated or not)
Loc_clusters instead of zipcode	Geographic clusters from K-Means on lat/long (captures neighborhood patterns) :4 clusters
Sold_occ instead of id (avoid duplicates)	counts number of times sold
Grade_clean instead of grade	7 grades instead of 13

Ensemble models: XgboostRegressor

Features:

sqft_living
waterfront
view
grade
sqft_living15
yr_built
sqft_above
lat
long
loc_clusters
renovated
condition
sqft_lot15
year_sold
sqft_lot
floors
bathrooms
sqft_basement
sold_occ
month_sold
bedrooms

BEST PARAMETERS:

• colsample_bytree: 0.8 • max_depth: 7 • subsample: 0.8
• learning_rate: 0.1 • n_estimators: 300

	r2_score	RMSE	MAE
train	0.983919	40264.038943	28547.181641
test	0.916751	91875.881492	58868.847656

Ensemble models: XgboostRegressor

Features ordered by importance:

	Feature	Ranking	Selected
2	sqft_living	1	True
5	waterfront	1	True
6	view	1	True
8	grade	1	True
13	sqft_living15	1	True
11	yr_built	1	True
9	sqft_above	1	True
17	lat	1	True
18	long	1	True
19	loc_clusters	1	True
12	renovated	2	False
7	condition	3	False
14	sqft_lot15	4	False
15	year_sold	5	False
3	sqft_lot	6	False
4	floors	7	False
1	bathrooms	8	False
10	sqft_basement	9	False
20	sold_occ	10	False
16	month_sold	11	False
0	bedrooms	12	False

Tested with top 10 features:

	r2_score	RMSE	MAE
train	0.914016	93105.516399	64248.941406
test	0.876386	111955.652935	75849.617188

top 10 features + 2 PLS linear combination of the rest

	r2_score	RMSE	MAE
train	0.979607	45343.120669	32114.087891
test	0.910115	95467.611764	61873.476562

Ensemble models: Catboost

Features:

grade_clean
loc_clusters
sold_occ
sqft_per_bedroom
years_since_renovation
renovated
house_age
month_sold
year_sold
sqft_lot15
sqft_living15
long
lat
zipcode
yr_renovated
yr_built
sqft_basement
sqft_above
grade
condition
view
waterfront
floors
sqft_lot
sqft_living
bathrooms
bedrooms

BEST PARAMETERS:

depth=8, learning_rate=0.05,
n_estimators=1000, loss_function='MAE'

Test	R2_score	MAE
String categorical	0.9112	58,780
Numerical categorical	0.9079	59,367

Categorical features

- waterfront
- view
- condition
- grade
- zipcode
- month_sold
- renovated
- sold_occ
- loc_clusters
- grade_clean

Ensemble models: Catboost

New Features:

New Features	Description
dist_seattle	How far the house is from downtown Seattle (closer = typically higher prices)
loc_clusters	Geographic clusters from K-Means on lat/long (captures neighborhood patterns) k= 50
log_price	Log-transformed price (reduces skewness, better for linear models)
total_living_ratio	Living space / lot size (indoor density indicator)
relative_size	House size / neighborhood avg size (relative to neighbors)
luxury_index	Grade + view + 2×waterfront (combined premium features)
quality_interaction	Grade × condition (overall quality score)
size_grade_interaction	Living space × grade (large + high-quality homes)

Categorical features

- waterfront
- view
- condition
- grade
- zipcode
- loc_clusters
- year_sold

Numerical features

- waterfront
- view
- condition
- grade
- zipcode
- loc_clusters
- year_sold
- bathrooms
- sqft_living
- sqft_lot
- sqft_above
- sqft_basement
- lat
- long
- sqft_living15
- sqft_lot15
- house_age
- dist_seattle
- total_living_ratio
- relative_size
- luxury_index
- quality_interaction
- size_grade_interaction

BEST PARAMETERS:

depth=10, learning_rate=0.015, l2_leaf_reg=10, n_estimators=5000, loss_function='RMSE'

	r2_score	RMSE	MAE
train	0.960151	62929.313704	41044.859203
test	0.916696	94482.024616	57474.073113

Ensemble models: GradientBoostingRegressor

base data:

	r2_score	RMSE	MAE
train	0.901612	115257.424719	72576.984177
test	0.877216	128163.254250	76518.730876

without feature enineering
(base data → only numerical
dropped “price” and “date”)

improving on data side:

Tried to improve the model by creating new columns like:

- location clusters
- renovated as boolean
- …

→ result didn’t change

Ensemble models: GradientBoostingRegressor

improving on model side:

1. using RandomizedSearchCV:
- looking for the best parameters
  → improved results

2. manual improvement:
- changing some parameter 1 by 1
  → improved results (less overfitting)

	r2_score	RMSE	MAE
train	0.971890	61561.621065	27030.593053
test	0.902597	114488.559859	61339.691364

```
(n_estimators= 600, min_samples_split=15,
 min_samples_leaf= 10, max_features= 'log2',
 max_depth= 10, loss= 'huber', learning_rate= 0.05)
```

	r2_score	RMSE	MAE
train	0.949469	82538.552972	44968.288611
test	0.905463	112791.656918	62105.108455

```
(n_estimators= 600, min_samples_split=15,
 min_samples_leaf= 10, max_features= 'sqrt',
 max_depth= 7, loss= 'huber', learning_rate= 0.05)
```

Thank You