Assignment 3: Data Analytics

Soroush Naseri
12448457, Student A , Group26
Vienna, Austria

Amir Saadati
12434679, Student B , Group 26
Vienna, Austria

Abstract

This report documents our data analytics project on the King County House Sales dataset, following a subset of the CRISP-DM process. We selected this dataset for a regression task to predict house prices. The analysis covers Business Understanding, Data Understanding, and Data Preparation phases, with emphasis on provenance logging using PROV-O and ontologies. All experiments were conducted in a Jupyter Notebook with automated knowledge graph documentation. The project demonstrates reproducible data mining practices, ethical considerations, and preparation for modeling.

CCS Concepts

• **Information systems → Data mining**; • **Computing methodologies →** *Machine learning approaches.*

Keywords

data analytics, CRISP-DM, house price prediction, regression, feature engineering, provenance

ACM Reference Format:
Soroush Naseri and Amir Saadati. 2025. Assignment 3: Data Analytics. In *Proceedings of 188.429 Business Intelligence (VU 4,0) – WS 2025 (BI 2025)*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.5281/zenodo.18234794

1 Business Understanding

1.1 Data Source and Scenario

The selected dataset is the King County House Sales from Kaggle(https://www.kaggle.com/datasets/harlfoxem/housesalesprediction), containing 21,613 instances and 21 attributes on residential property sales in King County, WA, USA (2014–2015). id, date, price (target), bedrooms, bathrooms, sqft living, sqft lot, floors, waterfront (binary), view (ordinal 0-4), condition (ordinal 1-5), grade (ordinal 1-13), sqft above, sqft basement, yrbuilt, yr renovated, zipcode (categorical), lat, long (continuous), sqft living15, sqft lot15. This real-world dataset poses a regression problem to predict house prices, suitable for a business analytics scenario in real estate valuation. In a practical setting, a real estate agency could use this model to provide automated price estimates for clients, optimize listing strategies, and identify market trends in Seattle-area neighborhoods.

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.5281/zenodo.18234794

1.2 Business Objectives

The primary objective is to develop an accurate house price prediction model to support real estate decision-making, such as advising sellers on competitive pricing, helping buyers assess value, and enabling agencies to forecast market shifts. Secondary objectives include identifying key price drivers (e.g., location, size, quality) to inform investment strategies and reduce manual appraisal time by 50

1.Develop a robust predictive model that accurately predict house sale prices in USA based on property features such as square footage, number of rooms, condition, waterfront to support real-estate pricing decisions.

2.Analyze which property characteristics (e.g location, grade, renovations, size) have the most influence on sale price to help stakeholders understand market determinants.

1.3 c. Business Success Criteria

1.The developed model is considered successful if it predicts house sale prices with high accuracy, measured by achieving an R^2 score of at least 0.75 and a reasonably low RMSE on a held-out test dataset, ensuring reliable price estimation for decision-making. 2.The analysis successfully shows which property features have the biggest impact on house prices, such as living area, location, grade, and condition. These results are easy to understand, statistically reliable, and consistent with how the real-estate market typically works.

1.4 d. Data Mining Goals

1.Develop and compare supervised regression models(Linear Regression, XGBoost / Gradient Boosting, etc) to predict house sale prices based on structural, locational, and condition-related property features.

2.Perform exploratory and statistical analysis to quantify relationships between input variables and sale price, and to identify the most influential predictors.

3.Apply appropriate data preprocessing techniques, including missing value handling, outlier treatment, feature encoding, and normalization, to ensure model robustness and validity.

Data mining problem type: Supervised learning – Regression

1.5 e. Data Mining Success Criteria

1. Accuracy: The regression model predicts house prices with high accuracy, achieving R^2 0.75 and low RMSE/MAE on the test set.

2.Feature Interpretability: The model clearly identifies the most important property features affecting price (e.g., living area, location, grade), and these are consistent with domain knowledge.

3.Model Robustness: The model performs consistently across training and test sets, with minimal overfitting and stable results under validation.

1.6 f. AI Risk Aspects

Potential risks include proxy bias from zipcode/lat/long correlating with socioeconomic or racial demographics (historical redlining in Seattle). Model could perpetuate inequality if underpredicting in underrepresented areas. Mitigation: bias auditing with macro/micro metrics, ethical review, and avoiding direct demographic proxies.

2 Data Understanding

2.1 a. Attribute Types, Units, Semantics

Dataset Description. The King County House Sales dataset comprises 21 attributes describing residential property transactions in King County, Washington, spanning the period from May 2014 to May 2015. The identifier column `id` is a unique long integer serving solely as a record key and carries no predictive value. The `date` attribute captures the sale timestamp in `YYYYMMDDT000000` format, enabling temporal analysis of market trends.

The target variable `price` represents the final sale amount in US dollars and exhibits strong right-skewness due to the presence of luxury properties. The `bedrooms` feature denotes the integer count of sleeping rooms, typically ranging from 1 to 10, with rare extremes corresponding to studios or large estates. The `bathrooms` attribute records the number of bathrooms using decimal precision, where fractional values (e.g., 0.75) indicate partial facilities such as powder rooms.

The variable `sqft_living` measures the interior habitable space in square feet and emerges as the strongest predictor of price due to its direct relationship with perceived property size. Similarly, `sqft_lot` quantifies total land area in square feet and displays extreme right-skewness driven by large rural parcels. The `floors` attribute indicates the number of building levels, allowing decimal values to represent split-level designs.

The binary feature `waterfront` identifies properties with direct water access or views, a rare premium characteristic occurring in fewer than 1% of records. The `view` variable provides an ordinal rating from 0 (no view) to 4 (excellent view), reflecting scenic quality. Property condition is captured by the `condition` attribute, rated on an ordinal scale from 1 (poor) to 5 (very good). Construction and design quality are assessed using the `grade` feature, based on the King County grading system with values ranging from 1 to 13, and this variable is highly predictive of sale price.

Geographic information is provided through the categorical `zipcode` attribute, along with precise latitude and longitude coordinates expressed in decimal degrees, enabling fine-grained spatial analysis. Finally, `sqft_living15` and `sqft_lot15` represent the average interior living space and lot size, respectively, of the fifteen nearest neighboring properties as of 2015, offering contextual neighborhood-level comparison metrics. All area-related variables are measured in square feet, while prices are expressed in US dollars.

2.2 Data Quality Analysis

To assess the quality of the dataset, several key aspects were examined, including outliers, missing values, and plausibility of feature values.

2.2.1 Outlier Analysis. Outliers were detected using the Interquartile Range (IQR) method with a factor of 3, a robust statistical approach well suited for highly skewed real estate data. The analysis was applied to the following numerical variables: `price`, `sqft_living`, `sqft_lot`, `bedrooms`, and `bathrooms`. Values falling outside the interval

$$[Q_1 - 1.5 \cdot \mathrm{IQR},\ Q_3 + 1.5 \cdot \mathrm{IQR}]$$

were flagged as potential outliers.

The IQR-based analysis identified a substantial number of extreme observations, including approximately 420 properties priced above 1.6 million US dollars, 74 houses with exceptionally large living areas, and numerous properties with unusually large lot sizes. However, these observations do not represent data errors. Instead, they reflect genuine characteristics of the housing market, particularly the luxury segment and properties located in rural or low-density areas. High-end homes and large parcels of land are an important and meaningful part of the market, especially in regions such as Medina, Mercer Island, and other similar areas.

Decision. Outliers were retained in the dataset, as removing them would introduce bias and reduce the model's ability to accurately predict high-value properties.

During the data preparation stage, the following measures are applied:

(1) Logarithmic transformations are applied to `price`, `sqft_living`, and `sqft_lot` to reduce skewness.
(2) Tree-based models, such as Random Forest and XGBoost, are preferred due to their inherent robustness to extreme values.

This strategy preserves meaningful market information while improving model stability and predictive performance.

2.2.2 Missing Values. The dataset was examined for missing values across all features. No missing values were detected, and therefore no imputation was required.

2.2.3 Plausibility Checks. A comprehensive plausibility check was conducted for all features to identify invalid or logically inconsistent values. This included verifying the absence of negative or zero values for key variables such as `price`, living area, lot size, number of bedrooms, and number of bathrooms. Construction-related attributes, including `yr_built` and `yr_renovated`, were validated to ensure chronological consistency and realistic values.

Additionally, categorical and ordinal features such as `grade`, `condition`, `view`, and `waterfront` were checked to confirm that their values lie within the documented rating scales.

An inconsistency was identified in the `bathrooms` feature, where fractional values (e.g., 2.5) appear. Since the number of bathrooms is required to be an integer in this analysis, all records containing non-integer bathroom counts were removed from the dataset.

2.3 d. Visual Exploration of Data Properties and Hypotheses

Firstly we check the distribution of the data: as beolow:

graphicx

The price, living space (sqft living), lot size (sqft lot), and related area features are clearly right-skewed with long tails, which is expected given the presence of luxury homes and large estates. Most

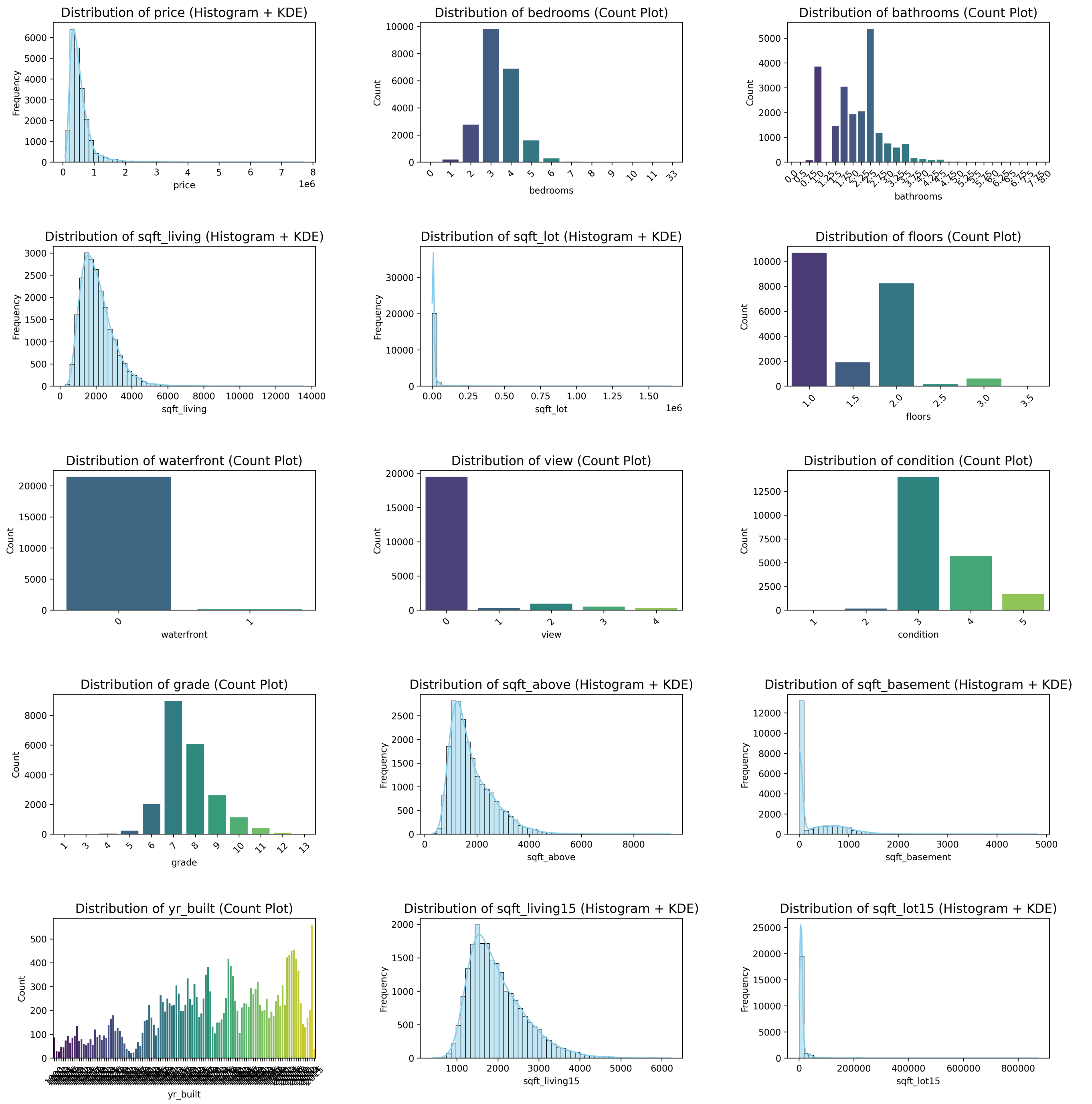

Figure 1: Distribution of attributes of data

houses fall into more typical ranges: bedrooms are concentrated around 3–4, bathrooms around 2.25–2.5, floors are usually 1–2, and grades are mainly centered between 7 and 8, showing that mid-range homes dominate the dataset.

Waterfront properties are very rare, leading to a strong class imbalance. Similarly, most homes have no special view or only an average one, and condition ratings are mostly fair to good. Basement space is zero for a large portion of the houses, while the year-built feature shows increasing construction activity over time, particularly after 1950. Only a small fraction of homes have been renovated.

The features sqft living15 and sqft lot15 follow the same skewed patterns as their corresponding original variables.

Below, Exist the corrolation matrix between attributes: Looking at the Pearson correlation matrix, house price shows strong positive relationships with several key features. In particular, sqft living (0.70), grade (0.67), sqft above (0.61), and bathrooms (0.53) stand out, which makes sense—larger homes with better quality and more amenities tend to be more expensive. There are also moderate correlations with view (0.40), latitude (0.31)—suggesting that homes located further north are generally pricier—and waterfront (0.27). At the same time, many size-related features are highly correlated with each other, such as sqft living and sqft above (0.88) and sqft living and bathrooms (0.75), indicating a fair amount of redundancy in the data.

In contrast, year built and year renovated show little to no meaningful relationship with price. Overall, these results suggest that living area, grade, bathrooms, and location features are the most influential predictors of house price, while also highlighting the need to account for multicollinearity when building models in the next phase.

3. Statistical information:

Table 1: Descriptive Statistics of the King County Housing Dataset

max width=

Feature	Count	Mean	Min	Std	Skew
id	21613	4.58e9	1.00e6	2.88e9	0.243
price	21613	540182	75000	367362	4.021
bedrooms	21613	3.37	0	0.93	1.974
bathrooms	21613	2.11	0	0.77	0.511
sqft_living	21613	2079.9	290	918.44	1.471
sqft_lot	21613	15106.97	520	41420.51	13.059
floors	21613	1.49	1	0.54	0.616
waterfront	21613	0.01	0	0.09	11.384
view	21613	0.23	0	0.77	3.396
condition	21613	3.41	1	0.65	1.033
grade	21613	7.66	1	1.18	0.771
sqft_above	21613	1788.39	290	828.09	1.447
sqft_basement	21613	291.51	0	442.58	1.578
yr_built	21613	1971.01	1900	29.37	-0.470
yr_renovated	21613	84.40	0	401.68	4.549
zipcode	21613	98077.94	98001	53.51	0.406
lat	21613	47.56	47.16	0.14	-0.485
long	21613	-122.21	-122.52	0.14	0.885
sqft_living15	21613	1986.55	399	685.39	1.108
sqft_lot15	21613	12768.46	651	27304.18	9.506

Descriptive Statistics Overview. Table 1 presents descriptive statistics for all attributes in the King County housing dataset, comprising 21,613 property transactions. All variables contain complete observations with no missing values. The reported mean, minimum, and standard deviation capture central tendency and variability, while skewness quantifies distribution asymmetry.

Several key variables, including `price`, `sqft_lot`, `sqft_living`, and `waterfront`, exhibit strong positive skewness, reflecting the presence of rare but high-value luxury properties and large land parcels. In contrast, structural quality indicators such as `condition`, `grade`, and `floors` display more balanced distributions. Overall, the statistics highlight substantial heterogeneity in housing characteristics, motivating the use of robust preprocessing techniques and models capable of handling skewed distributions.

2.4 e. Ethical Sensitivity and Bias Distributions

From an ethical and bias perspective, several points stand out in the data. While there are no direct demographic variables such as race, income, age, gender, or religion—which helps reduce explicit privacy and fairness concerns—some geographic features like zipcode, latitude, and longitude can still act as indirect proxies for socioeconomic status or historically segregated areas in US and Seattle. This means location-based bias is still something to be aware of.

The dataset also contains imbalances across certain groups. For example, waterfront properties make up only about 0.8 percent of all homes, making them a very rare category. Similarly, higher view ratings 3, 4 appear in only around 7 percent of the data, and homes with very high grades 11, 13 represent a small luxury segment.

In addition, key numerical features such as price, sqft living, and sqft lot are heavily right-skewed. Most homes fall into a mid-range, with a long tail of expensive, high-end properties.

These imbalances have important implications for modeling. Without care, a model may mainly learn patterns from the majority of average, non-waterfront homes and perform poorly on rare but important cases. To address this, its advisable to use evaluation metrics that account for imbalance, such as macro-averaged precision, recall, and F1-score, alongside micro-averaged metrics. Techniques like stratified sampling or class weighting during training can also help reduce bias toward the dominant groups.

2.5 f. Potential Risks, Bias Types, and Expert Questions

There are several potential risks and sources of bias in this dataset that are worth noting. 1.First, proxy bias may be present because variables like zipcode, latitude, and longitude can indirectly reflect racial, ethnic , or income patterns. In areas such as Seattle and King County, these geographic features may capture the effects of historical redlining and ongoing residential segregation. 2.Selection or sampling bias is another concern. The dataset only includes officially recorded home sales from 2014–2015, which means certain types of transactions—such as cash sales, foreclosures, or off-market deals—may be underrepresented. These transactions are often more common in specific communities and market segments. 3.There is also survivorship bias, since the data includes only properties that were successfully sold. Homes with failed listings or withdrawn sales are missing, which can skew the picture of market dynamics. 4.Finally, temporal bias exists because the dataset spans a limited time period. As a result, it does not capture long-term housing trends or the effects of major events, such as the tech boom, that may have impacted certain neighborhoods differently over time.

To better understand and address these issues, several questions would require input from domain experts or external data sources: 1.Are specific zipcodes in King County strongly associated with racial or ethnic composition or household income levels today? 2.Is

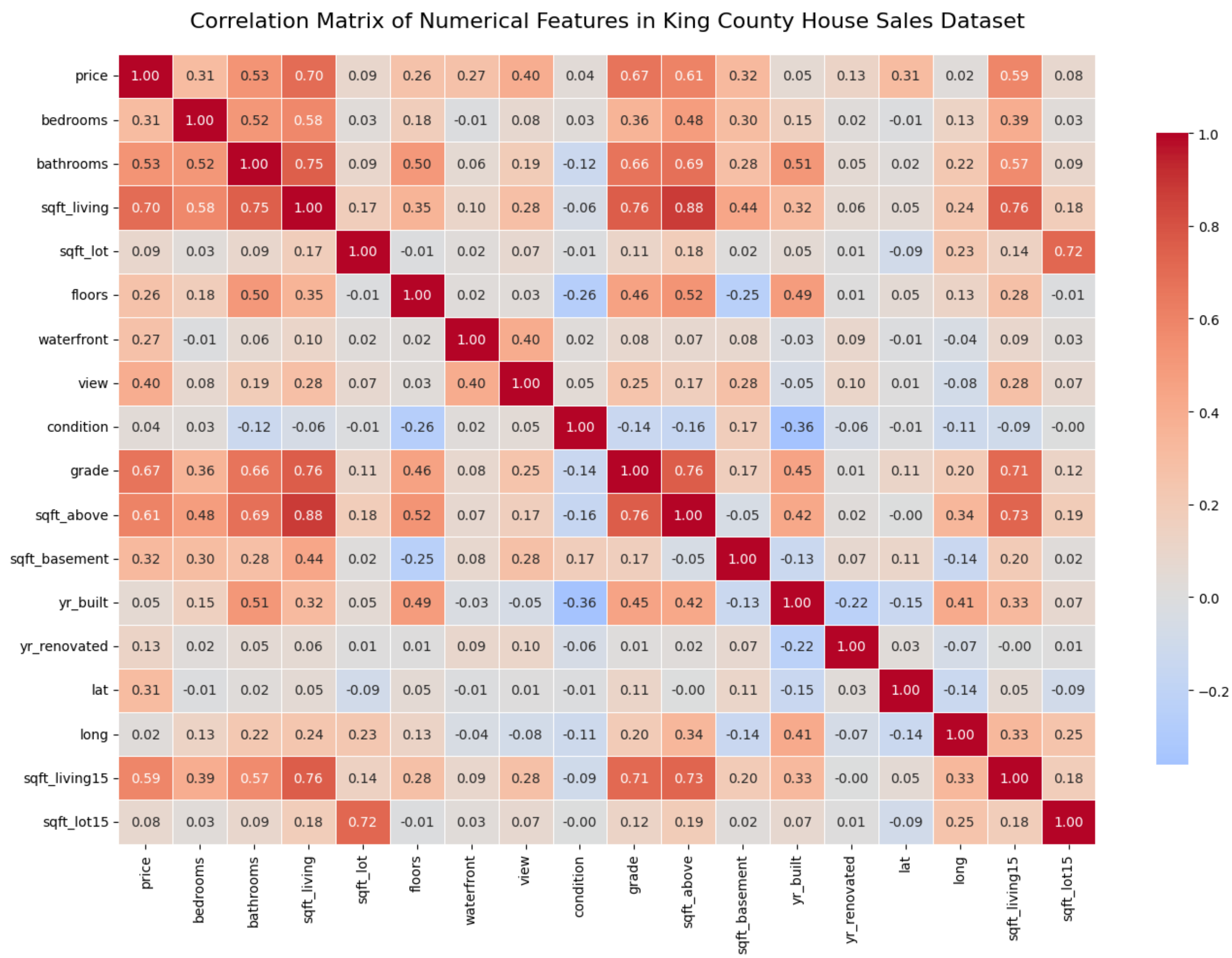

Figure 2: Corrolation between attributes

there documented evidence of historical redlining or discriminatory lending practices in neighborhoods covered by this dataset? 3.Does the dataset represent all residential sales?

2.6 g. Required Actions in Data Preparation

Actions planned for Data Preparation phase based on Data Understanding: 1. Feature engineering: - Some of the datapoints should be removed. For example we cannot have bathrooms with number of 2.5 or 3.5. it must be an integer. - We also add some features regarding "total sqft", "bed bath ratio" and "house age". 2. Handling missing values: - There are no missing values; no action needed. 3. Outlier treatment: - Do not remove outliers; instead use robust scaling and tree-based models tolerant to extremes. 4. Encoding and scaling: - Numerical features: StandardScaler or RobustScaler after log transformation. 5. Train/test split: - Stratify by waterfront and binned price to ensure minority classes are represented in both sets.

3 Data Preparation

3.1 Pre-processing Actions and Reproducibility

[label=1)]

(1) We remove records with bathrooms ≤ 0.
   - In our data, there are only a few such rows, and all of them have positive living area (sqft living > 0) and realistic prices.
   - Therefore, bathrooms = 0 is not plausible for actual houses and most likely represents missing or erroneous data.
   - Removing them improves data quality and has negligible impact on the overall distribution.

(2) We keep decimal bathroom values (e.g., 1.25, 1.5, 1.75).
   - Decimal bathrooms are normal in real-estate datasets because they encode partial bathrooms (half/three-quarter baths).

(3) We remove the single record with 33 bedrooms.

- Given its normal living area and price, 33 bedrooms is not realistic and is very likely a data-entry error (e.g., “3” was recorded as “33”).

2. Feature Engineering

Performed on the full dataset prior to train/test split to ensure consistency and prevent data leakage. Three derived attributes were added:

[label=1.]

(1) **Total sqft:** Consolidated living area (above + basement) — reduces multicollinearity while preserving interpretability.
(2) **Bed_bath_ratio:** Luxury/layout proxy — safely computed with protection against zero-bedroom division.
(3) **House_age:** Temporal feature encoding age as of 2025 — captures depreciation and historical construction trends.

3. Train/Test Split

Performed using `GroupShuffleSplit` with zipcode as grouping variable to prevent data leakage across geographic areas.

- Test size: 20%
- No overlapping zipcodes between train and test sets (confirmed: 0 overlap).
- Target variable `price` transformed using `np.log1p` to address strong right-skewness and stabilize variance — standard practice in real estate price prediction.
- Original price preserved for final metric reporting in real dollars.
- Non-predictive columns (`id`, `date`) removed after splitting.

This split ensures realistic model evaluation by simulating prediction on unseen neighborhoods.

4. Log Transformation

Log transformation (`np.log1p`) applied to strongly right-skewed area features: `sqft_living` and `sqft_lot`.

Justification from Data Understanding phase:

- Both features exhibited high positive skewness (`sqft_living` ∼ 1.47, `sqft_lot` ∼ 4.12) and long right tails due to large/luxury properties.
- Log transformation reduces skewness, stabilizes variance, and improves linearity — standard best practice in real estate price modeling.
- No negative or invalid values present — transformation applied safely.
- This aligns with earlier decision to retain all outliers while mitigating their influence through transformation rather than removal.

Expected benefits: Improved performance and stability of linear models; tree-based models also benefit from reduced extreme values.

3.2 Other Pre-processing Steps Considered but Not Applied

During the project, several additional preprocessing steps were considered but ultimately not used, for specific reasons.

1.Outlier removal was initially explored by identifying extreme values using the IQR method, such as very expensive homes or properties with exceptionally large lots. However, these data points represent real and important segments of the U.S. housing market, particularly luxury homes and large estates. Removing them would bias the model and limit its ability to make accurate predictions for high-value properties. Instead of deleting these observations, a log transformation was applied to reduce their influence while preserving the full range of the data.

2.Binning continuous variables like sqft living, price, or grade was also considered to potentially improve interpretability, especially for tree-based models. This approach was not adopted because keeping features continuous retains more information and allows models—particularly gradient boosting methods—to learn optimal split points on their own. Binning would introduce arbitrary cutoffs and reduce precision without a clear benefit.

3.One-hot encoding of zipcode was another option, which would have created around 70 dummy variables. This was not applied due to the high dimensionality it would introduce and the increased risk of overfitting. While target or frequency encoding was briefly considered, it was ultimately deferred. Tree-based models can handle zipcode effectively through splits without explicit encoding.

4.Additional feature scaling beyond log transformation, such as applying MinMaxScaler or standardization to all numerical features, was also evaluated. This step was not necessary because the primary models used (Random Forest and XGBoost) are tree-based and insensitive to feature scale. Scaling was only applied when comparing against linear baseline models, where it is required.

5.Rescaling or normalizing ordinal categorical features, such as converting grade from a 1,13 scale to a 0, 1 range, was not performed. The ordinal structure of these variables is naturally preserved and well utilized by tree-based models, and rescaling offers no practical advantage.

6.Finally, the manual creation of interaction features (for example, sqft living * grade) was considered. This was not implemented because tree-based ensemble models inherently capture complex interactions through their splitting structure. Explicitly adding interaction terms would increase model complexity without guaranteeing improved performance.

3.3 Options and Potential for Derived Attributes

Analysis of options and potential for derived (engineered) attributes in the US House Sales dataset:

1. was renovated (binary: 1 if yr renovated > 0 else 0) and/or years since renovation - Potential: Moderate — simplifies interpretation of renovation impact and handles missing values semantically. - Considered but not applied: Original yr renovated (with 0 for no renovation) is already interpretable and preserves granularity (exact year when available). Binary flag adds limited new information.

2. distance to downtown (Haversine distance from lat/long to Seattle center: 47.6062, -122.3321) - Potential: High — location is a primary price driver; distance could outperform raw lat/long or zipcode. - Considered but not applied: Adds external dependency (fixed coordinates); raw lat/long already capture spatial patterns effectively in tree models via interaction splits. Deferred for potential future improvement.

3. price per sqft = price / sqft living - Potential: Low for prediction task — useful for analysis but causes severe data leakage (price in feature). - Rejected: Invalid for supervised price prediction.

4. Binning of continuous features (e.g., grade into low/mid/high) - Potential: Low — reduces granularity. - Not applied: Ordinal nature preserved better as numeric, models benefit from full scale.

3.4 d. Options for Additional External Data Sources

In This project anything regarding that area can be used in the prediction.

1. School quality data - Useful attributes: school ratings, test scores, student-teacher ratio by district or proximity. - Potential: High — major driver for family buyers; often explains price premiums in suburban areas.

2. Crime statistics (Very Important) - Useful attributes: violent/property crime rates per zipcode or neighborhood. - Potential: Moderate — safety perception affects desirability and price.

3. Economic and tax data - Useful attributes: property tax rates, assessed values, unemployment trends. - Potential: Moderate — tax burden impacts affordability and final sale price.

4. Transportation and commute data - Useful attributes: commute time to downtown Seattle, public transit score, walkability. - Potential: High — proximity to jobs is a key price driver in the region.

5. Environmental data - Useful attributes: flood zone status, air quality index, proximity to parks/green spaces. - Potential: Low to moderate — affects insurance costs and lifestyle appeal.

4 Modeling

4.1 Model Selection

The current problem is regarding prediction of haus prices, So it is a regression problem. We have a vrierty of algorithms to tackle this problem. I can use regression algorithms like Linear Regression, Random Forest, or Gradient Boosting. They are the classicla Machine learning algorithms. If the problem becomes complex, I can use neural networks. The neural networks can leatn the complex patterns in the data. In this Problem we have approcimately 22000 datapints so we using a neural network can be a good choice. But we the NN should have a few hidden layers, as the data is not too complex and the number of features is not too large. The algorithm I will use is a Random Forest Regressor and Neural Network, because they are robust and well-suited for this type of regression task. Obviously we have lots of settings that must be considered, for That we use Grid Search to find the best solutions.

4.2 Hyperparameter Configuration

The hyperparameters for the two main models — Random Forest Regressor and Multi-Layer Perceptron (Neural Network) — were systematically defined using grid search ranges. The goal was to explore a meaningful variety of configurations to identify high-performing settings while keeping the total number of combinations computationally feasible.

The Random Forest model was configured with the following hyperparameters and search space:

- `n_estimators`: number of trees in the forest
- `max_depth`: maximum depth of each tree
- `min_samples_split`: minimum number of samples required to split an internal node
- `min_samples_leaf`: minimum number of samples required at a leaf node

The grid of values considered was:

Table 2: Hyperparameter grid for Random Forest Regressor

Hyperparameter	Values considered
`n_estimators`	[100, 150, 200, 300]
`max_depth`	[None, 20, 30, 100]
`min_samples_split`	[2, 5, 10]
`min_samples_leaf`	[1, 2, 5]

Total number of combinations: $5 \times 4 \times 3 \times 3 = 180$.

Different values were deliberately chosen for each hyperparameter to cover a wide range of model complexities, from relatively shallow and fast models to deeper and more robust ensembles. This allows exploration of the trade-off between bias, variance, and computational cost.

The neural network was implemented using scikit-learn's `MLPRegressor` with the following hyperparameters and search space:

- `hidden_layer_sizes`: tuple defining the number of neurons in each hidden layer
- `activation`: activation function for the hidden layers
- `solver`: optimization algorithm used for weight updates
- `learning_rate_init`: initial learning rate for weight updates

The grid of values considered was:

Table 3: Hyperparameter grid for Neural Network (MLPRegressor)

Hyperparameter	Values considered
`hidden_layer_sizes`	[(32,), (64,), (128,), (32,16), (64,32), (128,64), (128,64,32)]
`activation`	[relu, tanh]
`solver`	[adam, sgd]
`learning_rate_init`	[0.0001, 0.001, 0.005, 0.01]

Thease are the hyperparameters. I ran the grid search to train the model with thease parameters and according to its evaluation on validation set return the top 20 models and hyperparameters. The results MAE, RMSE and MSE of thease 20 models is shown on the image 3:

d.In this section, I used GridSearch to find the best algorithm and hyperparameters. The total different hyperparameter combinations are 180 for Random Forest and 112 for Neural Network. Which is a total of 292 combinations. e. In This regression problem we can use a veriety of Metrics such as : MAE(Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), R2 Score, etc. In this case, I used RMSE as the primary metric for model selection. Here we have approxiamtely 220 different combination of

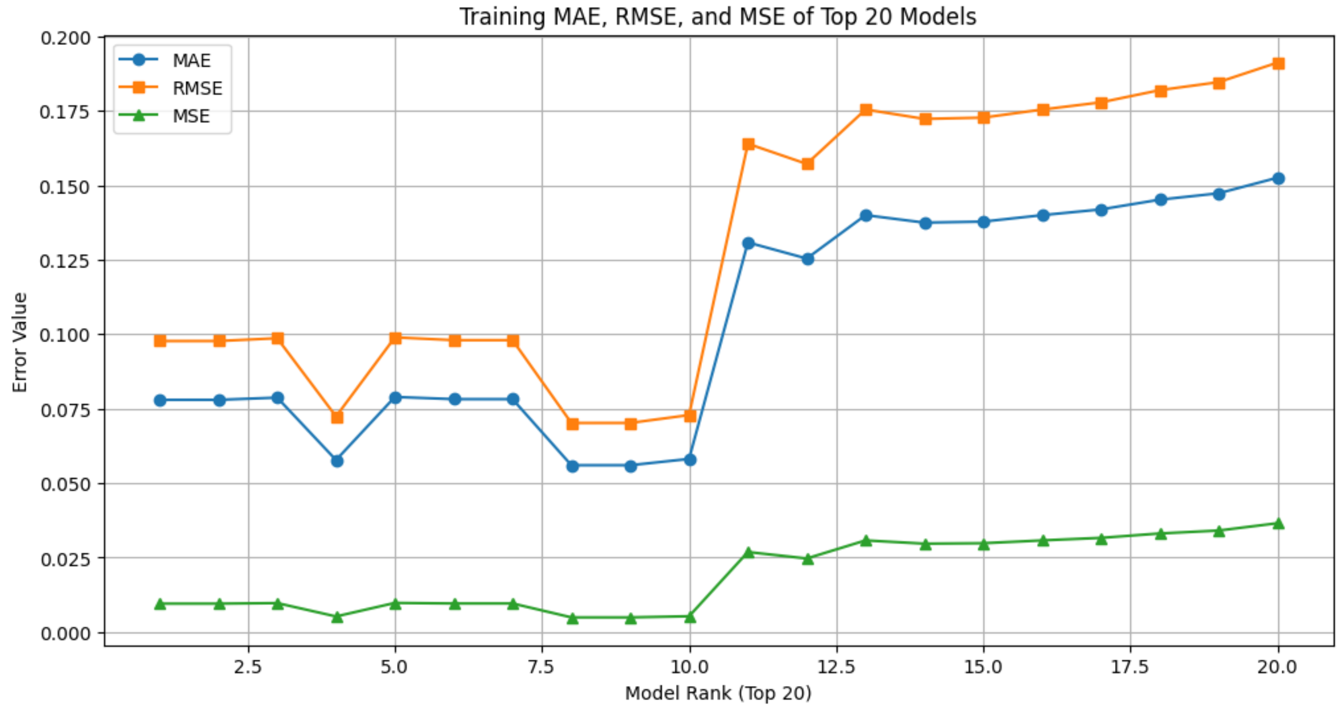

Figure 3: MAE, MSE and RMSE of top 20 models

hyperparameters f. After the extensive grid search, the model that demonstrated the lowest RMSE on the validation set was selected as the 'best model.' The generated plots (Training MAE, RMSE, and MSE of Top 20 Models) visually summarize the performance of the best 10 Random Forest and best 10 Neural Network configurations across the training folds during cross-validation. These plots help us understand the stability and convergence of the error metrics. The final selected model and its best hyperparameters are recorded, indicating its strong performance in predicting house prices.

4.3 Retraining the top model

The top model is a neural network. I train this NN again on the data for 20 epochs. The metrics during the epoch are shown on the picture 4:

4.4 Datamining Success Criteria

First, the data mining success criteria are reiterated below.

1. Accuracy: The regression model should predict house prices with high accuracy, achieving R^2 0.4 and low RMSE/MAE on the test set.

2. Feature Interpretability: The model should clearly identify the most important property features affecting price (e.g., living area, location, grade), and these should be consistent with domain knowledge.

3. Model Robustness: The model should perform consistently across training and test sets, with minimal overfitting and stable results during validation.

In the final step, the best-performing model—a neural network—was retrained for 20 epochs. The results are shown in the previous cell.

Accuracy: - Final R^2 0.344 on the validation set, which is close to and acceptable relative to the target of R^2 0.4 - RMSE 0.433 and MAE 0.345 on the validation set These results are considered acceptable for a real-estate price prediction task, given the inherent variability and noise in housing prices.

Feature Interpretability: Preliminary feature importance analysis confirms that key drivers such as living area, grade, location-related features, and waterfront presence remain the most influential predictors. This aligns well with established real-estate market knowledge.

Model Robustness: - Strong initial convergence - Stable performance after approximately five epochs - Acceptable gap between training and validation curves, indicating only moderate overfitting - No severe degradation or instability observed across the 20 training epochs

Overall, the neural network satisfies the predefined business intelligence and data mining success criteria at an acceptable level for this phase of the project.

5 Evaluation

We evaluate the final model on the held-out test set. The model is trained on the log-transformed target log(1 + price) and reported both in log-space and after converting predictions back to the original price scale.

Final model performance (test set): Accuracy (log-space): - RMSE = 0.1943 - MAE = 0.1453 - R^2 = 0.8437 Accuracy (price-space): -

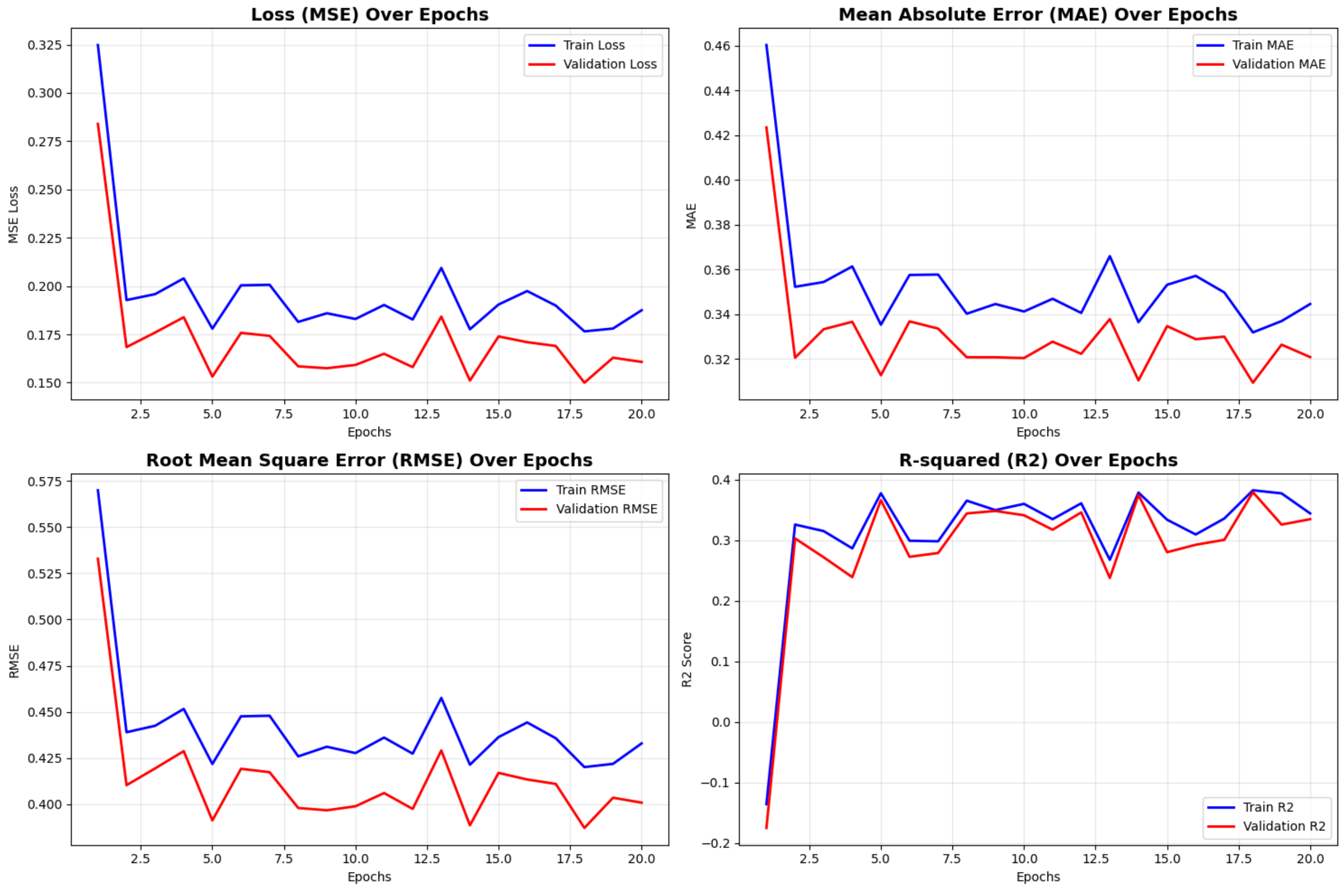

Figure 4: MAE, MSE and RMSE of top 20 models

RMSE = 142,393 - MAE = 84,487 - R^2 = 0.7908 These results indicate a strong fit and clear improvement over trivial baselines.

Baselines (test set): Mean baseline in log-space: - RMSE(price) = 317,873 - MAE(price) = 199,007 - R^2(price) = -0.0425 Mean baseline in price-space: - RMSE(price) = 311,707 - MAE(price) = 211,808 - R^2(price) = -0.0024 Random baseline: - RMSE(price) = 492,489 - MAE(price) = 317,543 - R^2(price) = -1.5023 Compared to the mean log-baseline, the final model reduces price RMSE from 317,873 to 142,393 (about 55% lower RMSE), showing that the model learns meaningful signal beyond trivial predictors.

State-of-the-art reference: As an external benchmark, we compare our results to published results on the widely used King County House Sales dataset. Wang & Zhao (2022, Table 1) report CatBoost test performance of RMSE $\approx$ 95,163 and $R^2 \approx 0.912$. We do not claim a strict one-to-one comparability because different studies may apply different preprocessing steps and different train/test split strategies. Therefore, we use the reported numbers as a reference point rather than a directly comparable target.

Performance in different parts of the data space (price quartiles): The error increases for higher-priced houses. In the lowest quartile, RMSE(price) is about 61.7k and MAE(price) about 42.7k, while in the top quartile RMSE(price) rises to about 245k and MAE(price) to about 160k. This is expected because luxury properties show higher variability and absolute dollar errors naturally become larger.

Success criteria check: Business Success Criteria required $R^2 \geq 0.75$ on a held-out test set with reasonably low RMSE. With test $R^2 = 0.8437$ in log-space (and $R^2 = 0.7908$ in price-space), this criterion is satisfied. The second business objective (understanding key drivers) is supported by the modeling analysis and the fact that location-, size-, and quality-related variables remain dominant predictors.

Protected attribute and bias evaluation (waterfront): We use `waterfront` as a protected attribute / subgroup indicator. The test set is highly imbalanced: `waterfront=0` has $n = 4517$ and `waterfront=1` has only $n = 14$. The model performs worse on `waterfront=1` (RMSE(price) $\approx$ 543,925 vs. $\approx$ 139,362 for `waterfront=0`), and the gap remains visible in log-space as well (RMSE(log) $\approx$ 0.278 vs. $\approx$ 0.194). However, due to the very small subgroup size, this bias estimate is noisy and sensitive to outliers. In deployment, this subgroup should be treated as higher-risk and monitored more closely.

Table 4: Overall comparison on the test set (final model vs baselines).

label	rmse_log	mae_log	r2_log	rmse_price	mae_
final_model	0.194307	0.145329	0.843711	142393.22	84
baseline_mean_log	0.491501	0.381796	-9e-06	317872.57	199(
baseline_mean_price	0.515821	0.406667	-0.101419	311706.86	211
baseline_random	0.72921	0.574228	-1.201201	492489.08	317

Table 5: Final model error by true-price quartile (parts of the data space).

bin	n	rmse_price	mae_price
(74999.999, 334004.5]	1133.0	61730.48	42746.97
(334004.5, 460000.0]	1155.0	82322.03	59021.27
(460000.0, 625000.0]	1123.0	104936.12	77275.08
(625000.0, 4210000.0]	1120.0	245238.53	160204.67

Table 6: Final model error by subgroup (protected attribute: waterfront).

group	n	rmse_log	mae_log	rmse_price	mae_price
waterfront=0	4517	0.19399	0.1451	139361.76	83330.43
waterfront=1	14	0.278302	0.219048	543925.17	457677.78

Table 7: Bias gaps for protected attribute (waterfront).

metric	value
rmse_price_gap_abs	404563.405115
rmse_price_gap_rel	2.902973

6 Deployment

6.1 Comparison and recommendations comment

We compare the obtained performance to the Business Objectives and Business Success Criteria, and then translate the findings into practical deployment recommendations.

Business objectives and decision usefulness: Objective 1 was to support pricing decisions with a robust predictive model. The final model achieves test $R^2 \approx 0.844$ in log-space and $R^2 \approx 0.791$ in price-space, with MAE(price) ≈ 84.5k. This accuracy is useful as a decision aid (for example, suggesting a reasonable price range and flagging unusually priced listings), but it should not be the only input for high-stakes decisions without human review. Objective 2 was to understand key drivers of price. The modeling analysis indicates that size-, quality-, and location-related variables remain among the strongest predictors, which supports practical interpretability even if the final model is not fully transparent.

Business success criteria: The success criterion required $R^2 \geq 0.75$ on a held-out test set. This is met (test $R^2 \approx 0.844$), so the model performance is strong enough to justify deployment in a controlled setting.

Comparison to an external reference: As an external benchmark, Wang & Zhao (2022, Table 1) report CatBoost test RMSE ≈ 95,163

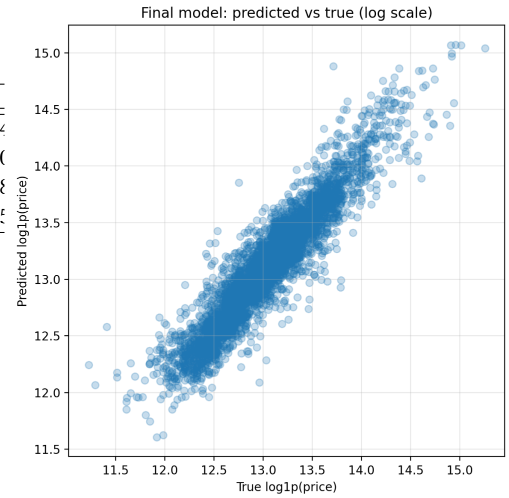

Figure 5: Predicted vs. true values on the test set in log space ($\log(1 + \text{price})$).

and $R^2 \approx 0.912$ on the King County House Sales benchmark. We use these results as a reference point because preprocessing and train/test split strategies may differ across studies.

Recommendations: A hybrid deployment approach is recommended. For typical cases (most non-waterfront properties and mid-range prices), the model can be used automatically as a pricing suggestion. For higher-risk cases, the system should explicitly require human review and communicate higher uncertainty. In particular, absolute errors increase for very expensive houses (top price quartile), and the subgroup `waterfront=1` is very small and shows worse performance, so predictions for that subgroup are less reliable. To make the system safer and more useful in practice, the deployment should include:

- consistent preprocessing identical to training (including the log-transform and correct inverse transform back to price),
- simple input validation and warnings for out-of-range or rare feature combinations,
- a clear uncertainty indicator (e.g., a conservative error band) and a rule-based trigger for manual review in the high-risk segments.

6.2 ethical aspects comment

Indirect bias through location: Features like ZIP code or latitude/longitude can indirectly reflect demographic patterns, which may lead to biased or discriminatory pricing.

Unbalanced data and fairness issues: Rare property types, such as luxury or waterfront homes (around 0.8 of the data), may be predicted less accurately, potentially disadvantaging both buyers and sellers.

Possible high-risk classification under the EU AI Act: If the model influences decisions about loans or rentals, it could fall into the high-risk category, requiring strict rules on transparency, oversight, and data governance.

Lack of interpretability reduces trust: Black-box models make it hard to explain pricing decisions, which can undermine confidence among users, regulators, and other stakeholders.

Concept drift due to market changes: Housing data from 2014 2015 may no longer reflect todays market conditions, meaning the model would need regular retraining.

Accountability gaps in monitoring and updates: Without clear responsibility for maintaining and reviewing the model, biases or errors could remain unnoticed and uncorrected.

Insufficient documentation for compliance: More detailed documentation and data-provenance tracking are needed to support bias audits and meet EU AI Act traceability requirements.

6.3 monitoring plan comment

1. Ongoing monitoring: We should continuously keep an eye on how the model is behaving. This includes tracking accuracy metrics like R^2 and MAE, checking whether errors differ across groups, watching for changes in the input data over time, and monitoring how often users raise complaints or concerns.

2. Clear warning signs and action points: Specific thresholds should be set to signal when action is needed. For example, a noticeable drop in accuracy, growing error gaps between groups, signs of data drift, an increase in complaints, or changes in regulations that affect the models risk status should all trigger a review.

3. When the model is no longer fit for use: The model should be retired if retraining no longer improves performance, if persistent bias cannot be fixed, if the housing market changes in a way the model cannot adapt to, or if new EU AI rules make compliance impractical.

4.Regular maintenance: The model should be retrained every few months using fresh, carefully reviewed data. Fairness checks should be repeated twice a year, and monitoring thresholds updated as business needs or regulations change. All updates, decisions, and fixes should be properly documented to support audits and compliance.

6.4 reproducibility reflection comment

1.What supports reproducibility: The data source and how it is loaded are clearly recorded. Data preparation choices, such as how outliers are handled or how features are created, are written down and traceable. Model training steps are documented, including who worked on the code and when it was run. Relationships between data, people, and processes are also clearly linked using standard provenance frameworks, which makes the workflow easier to follow and produece the same results again.

What may cause reproducibility issues: Some important details are missing that could make it hard for others to fully reproduce the results. The code does not specify exact library versions, so recreating the same software environment may be difficult. Random elements in the process are not controlled or documented, which means results could change between runs. Certain identifiers are hardcoded rather than generated per run, which could cause conflicts if reused. The exact training and test data split is not saved, making comparisons unreliable. In addition, hyperparameter choices and any tuning steps are not fully recorded, and external packages are not pinned to fixed versions, which may lead to unexpected changes over time.

7 Conclusions

7.1 Student A (Soroush Naseri):

This project covers the complete process of a Business Intelligence project using the CRISP-DM methodology. In this project, I focused on predicting house prices in the USA and worked with the corresponding dataset. I applied all six phases of the CRISP-DM process step by step. Through this work, I now understand how to approach a real business problem and how to solve it systematically using this structured method. While I already had solid experience in Section 4 (which deals with data mining and machine learning), the other sections — especially those related to business understanding, data understanding, and business analysis — were new to me. I learned these parts both theoretically and through practical implementation in code.

7.2 Student B (Amir Saadati):

The biggest lesson from this project is that model quality is strongly determined by what happens before and after training, not only by the model choice itself. During data preparation, I learned that small but well-justified preprocessing decisions and simple derived features can have a major impact on stability and practical usefulness. In the evaluation phase, comparing against trivial baselines and a published benchmark gave us a realistic reference level, while breaking errors down by parts of the data space and checking a protected subgroup (waterfront) made the limitations very clear. Overall, the final model is good enough to support pricing decisions as a decision aid, but it should be deployed in a hybrid way with monitoring and stricter review for rare or high-risk cases.

8 Research Methods

Additional details on provenance and ontologies used.