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, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

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.

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

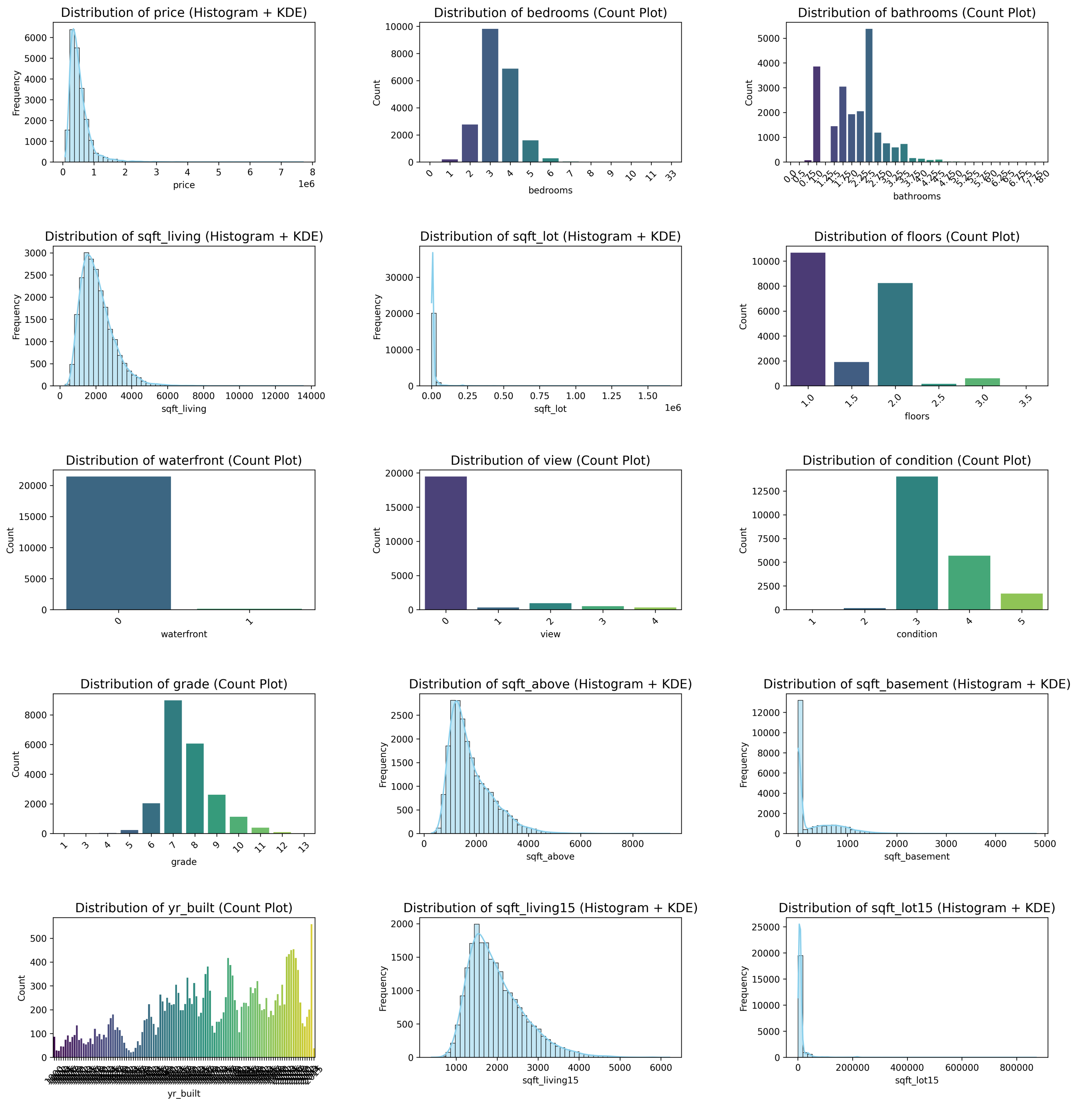

Figure 1: Distribution of attributes of data

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

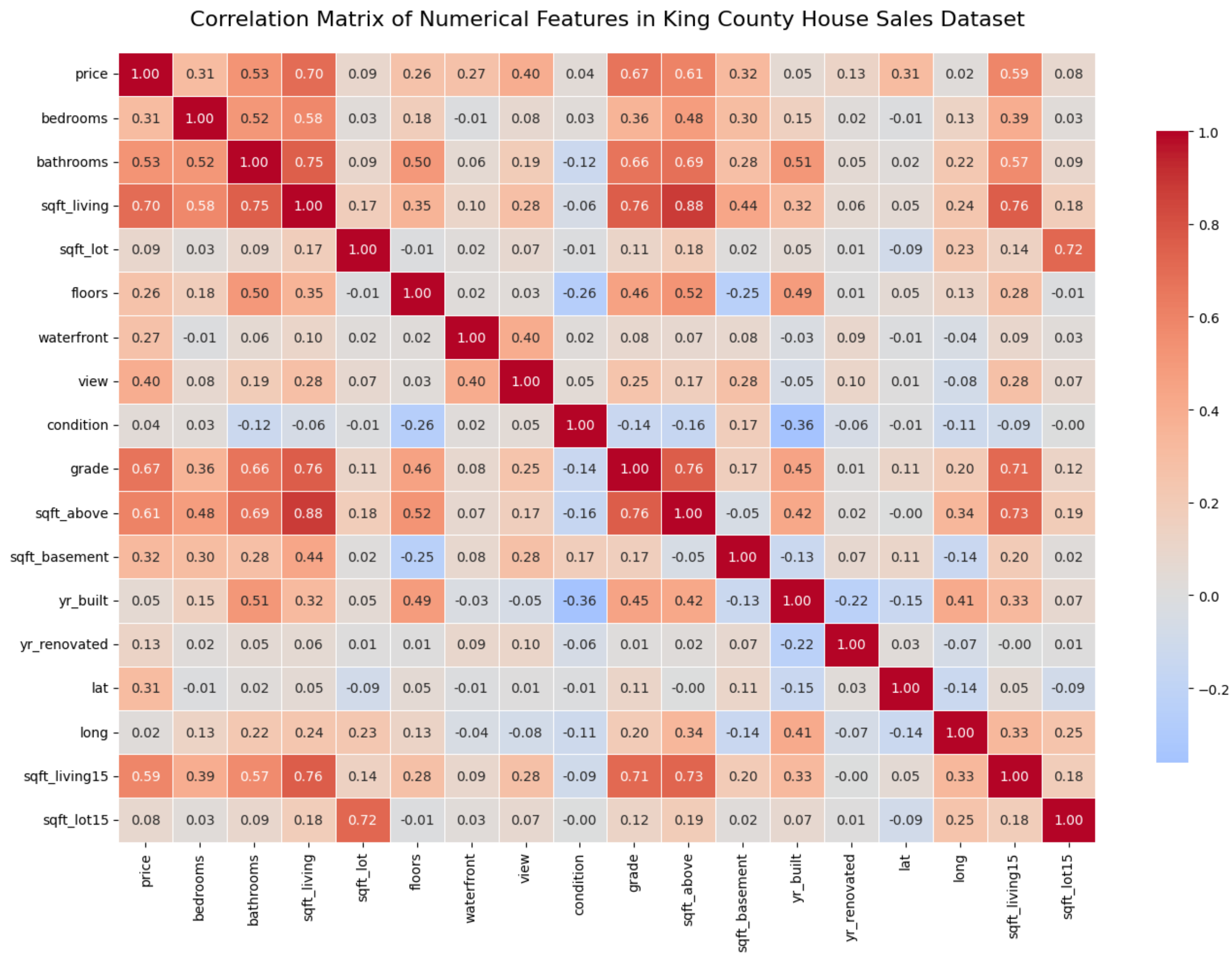

Figure 2: Corrolation between attributes

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 Conclusions

References

A Research Methods

Additional details on provenance and ontologies used.