

# Comprehensive Report on Variable Selection and Data Acquisition for House Price Prediction

---

This document provides a comprehensive overview of the key variables selected for a house pricing model, along with insights into their acquisition and the rationale behind their inclusion. By understanding these variables, real estate professionals, data scientists, and stakeholders can build more robust predictive models that accurately estimate a property's market value.

---

## 1. Introduction

The real estate market is inherently complex, with house prices influenced by a multitude of factors ranging from physical property characteristics (e.g., area, number of floors) to local market conditions (e.g., neighborhood averages, renovation status). In this report, we outline the essential variables chosen to capture these dynamics for a house pricing predictive model.

Our goal is to ensure that:

1. **Each variable is well-defined and consistently measured** across properties.
  2. **Data acquisition procedures** align with practical, real-world sources (e.g., county records, MLS data, or proprietary databases).
  3. **Model accuracy and interpretability** are balanced by focusing on the most influential and readily available features.
- 

## 2. Data Collection Strategy

Data for this model can be sourced from local real estate transaction records, property listing databases, tax assessors' offices, or online APIs. The variables described here are typically available through such channels or via field inspections. Where direct measurements (e.g., square footage) are lacking, derived variables or proxy information from official records can serve as reliable substitutes.

---

## 3. Detailed Variable Descriptions

### 3.1 Core Property Characteristics

1. **bedrooms**
  - **Description:** Number of bedrooms in the house, covering both standard and additional bedroom spaces.
  - **Type & Range:** INT, ranging from 0 to 36. The upper range accounts for large estates or unusual properties.
  - **Data Acquisition:** Typically sourced from official listings, tax records, or home inspection reports.
2. **bathrooms**
  - **Description:** Number of bathrooms, including both full and partial bathrooms.

- **Type & Range:** `FLOAT`, from `0.0` to `8.0`. Partial baths (e.g., half-baths) justify the float format.
- **Data Acquisition:** MLS listings and property tax assessments commonly document this.

### 3. `sqft_living`

- **Description:** The total living area of the house in square feet, computed as the sum of `sqft_above` and `sqft_basement`.
- **Type & Range:** `INT`, from `290` to `13540`.
- **Formula:** `sqft_living = sqft_above + sqft_basement`.
- **Data Acquisition:** Official records often provide `sqft_above` and `sqft_basement`, enabling a derived calculation for `sqft_living`.

### 4. `sqft_lot`

- **Description:** The lot size in square feet, representing the total area of the land.
- **Type & Range:** `INT`, spanning `[520, 1651359]`.
- **Data Acquisition:** Parcel or tax records generally document the lot area in square feet.

### 5. `sqft_above`

- **Description:** Square footage of the house, excluding basement areas.
- **Type & Range:** `INT`, from `290` to `9410`.
- **Data Acquisition:** Builder or county documentation typically lists above-grade square footage separately.

### 6. `sqft_basement`

- **Description:** Basement area in square feet.
- **Type & Range:** `INT`, from `0` to `4820`.
- **Data Acquisition:** County records or property inspections. Some properties may have no basement (`0`).

### 7. `floors`

- **Description:** Number of floors (stories) in the property.
- **Type & Range:** `FLOAT`, `[1.0, 3.5]`.
- **Data Acquisition:** Documented in MLS listings or property descriptions, where partial floors (e.g., split-level designs) justify a float range.

## 3.2 Additional Property Features

### 1. `waterfront`

- **Description:** Binary indicator (1 = yes, 0 = no) denoting if the property has a waterfront view.
- **Type & Range:** `INT`, `[0, 1]`.
- **Data Acquisition:** Generally clear from listing details or geographical data (e.g., proximity to lakes, rivers, or coastal areas).

### 2. `view`

- **Description:** A discrete rating (0 to 4) reflecting the quality or number of views the house offers.
- **Type & Range:** `INT`, `[0, 4]`.

- **Data Acquisition:** Subjective but often recorded in property listings. Field inspections or third-party data may refine this measure.

### 3. **condition**

- **Description:** An ordinal rating (1 to 5) of the property's overall condition (from poor to excellent).
- **Type & Range:** `INT`, [1, 5].
- **Data Acquisition:** Typically assigned by assessors or property evaluators. Condition assessment can be somewhat subjective.

### 4. **grade**

- **Description:** An overall grade (1 to 13) based on construction quality and design.
- **Type & Range:** `INT`, [1, 13].
- **Data Acquisition:** Similar to **condition**, gleaned from official assessments, appraisals, or MLS remarks.

## 3.3 Neighborhood and Market Context

### 1. **avg\_nbhd\_sqft\_living15**

- **Description:** Average living area within the neighborhood over the last 15 years.
- **Type & Range:** `INT`, [399, 6210].
- **Data Acquisition:** Aggregated from historical property sales or municipal records. Useful for contextual comparisons.

### 2. **avg\_nbhd\_sqft\_lot15**

- **Description:** Average lot size in square feet for properties in the neighborhood over the past 15 years.
- **Type & Range:** `INT`, [651, 871200].
- **Data Acquisition:** Collected from local property databases or real estate analytics platforms.

## 3.4 Temporal and Renovation Factors

### 1. **date\_year**

- **Description:** Year in which the house was sold.
- **Type & Range:** `INT`, [2014, 2015].
- **Data Acquisition:** Pulled from transaction records, reflecting a targeted period for model training (e.g., a 2-year snapshot).

### 2. **date\_month**

- **Description:** Month of the house sale (1 for January to 12 for December).
- **Type & Range:** `INT`, [1, 12].
- **Data Acquisition:** Also from official sales records. Potentially relevant for seasonal pricing patterns.

### 3. **built\_since**

- **Description:** Number of years since original construction.
- **Type & Range:** INT, [0, 115].
- **Data Acquisition:** Calculated using building year data from tax records or historical permit data.

#### 4. has\_been\_renovated

- **Description:** Indicates if the property has undergone any significant renovation (1 = yes, 0 = no).
- **Type & Range:** INT, [0, 1].
- **Data Acquisition:** Often available in MLS details or building permit records.

#### 5. renovated\_since

- **Description:** Number of years since the last renovation; 0 if no renovation occurred.
- **Type & Range:** INT, [0, 115].
- **Data Acquisition:** Derived from permit or listing data. Helps account for the impact of recent upgrades.

### 3.5 Target Variable

#### 1. zipcode

- **Description:** Zip code in which the house is located.
- **Type & Range:** INT, [98001, 98199].
- **Data Acquisition:** Standard location data from listing or county records.

#### 2. price

- **Description:** Final sale price of the house, expressed in thousands of dollars (k\$).
- **Type & Range:** FLOAT, [75.0, 770].
- **Significance:** This is the **target variable** our model aims to predict. Understanding how the above features correlate with or affect price is central to model accuracy.
- **Data Acquisition:** Drawn from official sales records or transaction logs.

---

## 4. Rationale for Variable Selection

#### 1. Physical Attributes (bedrooms, bathrooms, sqft\_living):

Capturing a home's capacity and size is critical for pricing. Buyers attach strong significance to living space, number of rooms, and property layout.

#### 2. Quality Indicators (condition, grade):

Even similarly sized properties can vary widely in value due to differences in construction quality, design elegance, and maintenance level.

#### 3. Neighborhood Context (avg\_nbhd\_sqft\_living15, avg\_nbhd\_sqft\_lot15):

Local market norms heavily influence what a buyer is willing to pay. By comparing a property's size to its neighborhood average, the model can detect relative advantage or disadvantage.

#### 4. Temporal Factors (date\_year, date\_month):

House prices can be cyclical or influenced by economic conditions. Knowing the sale year and month accommodates temporal trends and seasonality.

### 5. Renovation and Age (built\_since, renovated\_since):

A recently renovated home often commands a higher price, whereas an older, unrenovated home may sell at a discount. These variables help capture the property's effective age and upgrade status.

### 6. Location and Specialty Attributes (zipcode, waterfront, view):

Proximity to water or scenic views can add substantial premiums to a house's value. Zip code data situates the property in a localized market environment, tying into socioeconomic and infrastructure aspects.

---

## 5. Data Acquisition and Integration

### 1. Primary Data Sources:

- **County Records:** Typically provide official figures for lot size, square footage, and transaction dates.
- **MLS Databases:** Offer details on home listings (bedrooms, bathrooms, photos for property condition, etc.).
- **Tax Assessor Files:** Great for recorded property valuations, renovations, building year, and historical sales.
- **Specialty APIs/Analytics Platforms:** Some third-party vendors compile data on neighborhood averages, local indexes, and historical trends.

### 2. Field Inspections:

For critical attributes (like condition or recent renovations), on-site evaluations or professional appraisal reports can validate or supplement official records.

### 3. Data Cleaning and Validation:

- Check consistency of derived attributes (e.g., `sqft_living` should match `sqft_above + sqft_basement`).
  - Ensure that no variable falls outside its expected range (e.g., negative or excessively large bedroom counts).
  - Use standardized criteria for condition and grade to maintain uniform interpretation.
- 

## 6. Conclusion and Next Steps

This formal report has presented a curated set of variables essential to a **House Pricing Predictive Model**. Each variable is accompanied by its definition, data source, and relevance to overall model performance. By prioritizing these features, analysts can develop models that are both accurate and interpretable.

**Next Steps** include:

- **Data Collection:** Acquire the latest records from county databases, MLS listings, and on-site inspections.
- **Feature Engineering:** Explore interactions among variables (e.g., the effect of a large lot size in combination with a basement).
- **Model Prototyping and Validation:** Use regression or machine learning approaches (random forests, gradient boosting) to test how well these variables explain and predict house prices.

By systematically incorporating these variables and adhering to rigorous data collection standards, stakeholders can confidently move forward in constructing robust pricing models that guide investment decisions, inform market analyses, and enhance real estate services.

---

### **References & Acknowledgments**

- MLS and County Websites for property details.
- Real estate analytics tools for aggregated neighborhood metrics.
- Appraisal standards for condition and grade classifications.