Section 3: Data Cleaning Report

This section documents the data cleaning process applied to the King County housing dataset before any modeling or visualization took place. It covers four key areas: identifying and removing data entry errors, explaining how those observations were handled, identifying problematic variables as recorded in the dataset, and describing how those variables were transformed.

1. Identifying Observations with Clear Data Entry Errors

We started by inspecting the dataset using summary statistics tables. This helped us understand the range and frequency of values for each variable. This initial review helped us spot patterns and inconsistencies, leading to the identification of three types of data errors.

A. Bedrooms

- 13 homes were listed as having 0 bedrooms, and one home was listed with 33 bedrooms.
- A home with 0 bedrooms is not considered livable by typical residential standards, and 33 bedrooms is extremely unrealistic and likely a typo (perhaps a misplaced digit, such as 3 instead of 33).

B. Bathrooms

- Several homes in the dataset were recorded as having 0 bathrooms.
- Homes should have at least one bathroom. A bathroom count of 0 indicates either a data entry error or an incomplete information in the original listings.

C. Future Construction or Renovation Dates

- We checked for homes where the year built or year renovated was later than 2016.
- Any construction or renovated after 2016 could not have been sold in this window and would therefore be inconsistent with the sale date.

We also checked other fields, including square footage and price, and found no invalid values such as zeros or negatives in those variables.

2. How these Errors Handled

Once these errors were identified, we applied the following cleaning actions:

- Removed any rows where the number of bedrooms were equal to 0 or greater than 15.
- We also removed listings with 0 bathrooms, as these are likely errors or incomplete entries.
- Finally, we confirmed that no homes had construction or renovation dates beyond 2016.

These cleaning steps helped us eliminate invalid or inconsistent records, ensuring the dataset was reliable.

3. Identifying Variables That Have Issues in Their Recorded Form

In addition to errors at the individual row level, we also found several variables that were problematic in how they were recorded or defined, making them difficult to use in their original format:

A. Date

- The date was stored as a long character string (e.g., "20141013T000000"), which isn't easy to interpret or work with.

B. Zipcode

- Zip codes were stored as numeric values, which can misleadingly suggest a continuous or ordered relationship. However, zip codes are categorical identifiers for geographic areas and should be treated as such.

C. Year Renovated

- Uses 0 to indicate "never renovated," but it's stored as a numeric year, which is confusing.

4. How We Transformed These Variables

To make these variables more suitable for analysis and modeling, we applied the following transformations:

A. Date: Converted to Year Sold and Month Sold

We extracted the year and month from the original date string to create two new variables; Year Sold and Month Sold will be useful for analyzing sales trends over time and identifying any seasonal patterns.

B. Zipcode: Convert to Factor

We converted zipcode from numeric to categorical, so that models will interpret each zip code as a separate geographic area rather than as a number. There were 70 unique zip codes in the dataset. Converting this variable into a factor allowed us to treat them as separate categories.

C. Year Renovated: Convert to Binary Renovated

We created a new binary variable that indicates whether a home had ever been renovated. This simplified the original data and made it easier to use in analysis and modeling.