Section 3: Data Cleaning Summary Report

In this section of the project, I performed a thorough data cleaning process on the King County housing dataset. The purpose of this cleaning step was to identify and fix data entry errors, transform problematic variables, and prepare the dataset for future modeling. Instead of removing potentially incorrect records without context, I cross-verified each suspicious entry using the King County Parcel Viewer and corrected them based on available external evidence.

To begin, I loaded the dataset using 'read.csv()' and confirmed that it contained 21 variables across 21,613 observations. A quick inspection showed that there were no missing values in the dataset, which simplified the cleaning process. The next step involved identifying specific data entry errors, especially homes with unusual values such as 0 or 33 bedrooms, or 0 bathrooms. I found 17 such cases. After reviewing each one using the Parcel Viewer, I corrected 13 rows with reasonable bedroom and bathroom values, verified and kept one open-concept studio as-is, and removed 3 rows that couldn't be verified or were clearly invalid.

Beyond room counts, I also checked other important numeric variables. There were no records with 0 living area or lot size, no homes with a price of zero or less, and no homes that were built or renovated in the future (i.e., after 2015), so no cleaning was needed for those features.

Next, I focused on transforming variables that were either difficult to interpret or could lead to poor model performance. I extracted the year and month from the sale date column to create two new variables: 'year_sold' and 'month_sold'. This transformation made time-based analysis much easier and avoided complications from working with raw date strings. Another key transformation involved reducing the number of categories in the zipcode variable. Since there were more than 70 unique zipcodes, I grouped them based on their first three digits into broader regional clusters: "980" and "981." Group 980 primarily includes suburban and outer neighborhoods, while 981 covers the urban Seattle core. This grouping simplified the variable and helped reduce the risk of overfitting in future models.

I also created a new binary variable called renovated, based on the 'yr_renovated' column. If the year renovated was greater than zero, the value was set to 1; otherwise, it was set to 0. This made it easier to account for the impact of renovations on price or quality without using unnecessary numerical detail.

Lastly, I created a geographic feature called 'distance_to_downtown', which measures the Euclidean distance from each home to Pike Place Market in downtown Seattle (a well-known land mark). Since raw latitude and longitude values are hard to interpret directly, this new variable gave a clearer sense of location. When I compared this distance to the condition rating of homes using a boxplot, I observed that homes in poorer condition (rated 1–3) were generally closer to downtown, while those in better condition (rated 4–5) tended to be located farther out, possibly in more suburban or residential areas. This supports the idea that distance from the city center may influence neighborhood quality or perceived safety.

To address potential multicollinearity, I examined the relationships between four square footage variables: 'sqft_living', 'sqft_above', 'sqft_basement', and 'sqft_living15'. The correlation analysis showed a strong linear relationship between 'sqft_living' and 'sqft_above' ($r = 0.88$), and a similarly high correlation with 'sqft_living15' ($r = 0.76$). The correlation between 'sqft_living' and 'sqft_basement' was moderate ($r = 0.43$), while 'sqft_above' and 'sqft_basement' showed little to no correlation ($r = –0.05$).

Given that 'sqft_living' captures the total finished living area, including both above-ground and basement spaces, retaining all three variables would introduce redundancy and likely lead to multicollinearity in the regression model. To reduce this risk and simplify the model structure, I removed both 'sqft_above' and 'sqft_basement', keeping only 'sqft_living' as the most

comprehensive and interpretable measure of living space. The variable ‘sqft_living15’ was kept for now and will be evaluated for it’s effect on model performance later.

In conclusion, by verifying and correcting suspicious values and thoughtfully transforming key variables, I created a dataset that is better suited for both regression and classification modeling.