## Data Wrangling: D.C. housing dataset

#### Cleaning steps:

-I combined the bathroom and half-bathroom columns into one column.

-I removed a column “Unnamed” which was equivalent to the index.

-I removed “State” and “City” because values were identical for each sale.

-I removed “Fulladdress” because we had “latitutde” and “longitude” columns, and missing data would be difficult to fill accurately.

-I removed “Nationalgrid” because we had “latitutde” and “longitutde” columns.

#### Outliers:

\*I removed outliers before NaNs because I used the data to make predictions to fill NaN values.

-Specific outliers:

-Stories: 250, 275, 826

-Year remodeled: 20

-Numerical data:

-Didn’t remove values outside of Q1-1.5IQR or Q3+1.5IQR for most columns because the data was reasonable.

-For GBA and LIVING\_GBA, I removed outliers according to technique above. It is likely that extreme values in other numerical categories were associated with higher GBA (gross building area).

-For LANDAREA and PRICE, the data was skewed right. I used fences at Q1-1.5IQR or Q3+2.5IQR to keep some of the higher values, keeping with the nature of the data.

-Categorical data:

-All categories and distributions seemed reasonable.

#### How did you deal with missing values, if any?

-Numerical data:

-Sorted data by location (x/y coordinates).

-Rolling mean with window of 500 so that no column had more than 1% missing data.

-Dropped remaining rows.

-Categorical data:

-Grouped variables by neighborhood.

-Replaced NaN values with the mode of that column in that neighborhood.

-Price:

-removed all observation with missing price, since the goal of the project is to predict price.