1.0 – Data Cleaning

Starting Point:

A. 1.0 of data set.csv

- BROKERTITLE column had too much information, “Brokered by Douglas Elliman - 111 Fifth Ave”.
- TYPE column is too detailed “Condo for sale
- Price , Bed, Property square feet is input as general text rather than numerical value
- ADDRESS, STATE, MAIN_ADDRESS is repetitive.
- LOCALITY, SUBLOCALITY, STREET_NAME, LONG_NAME, FORMATTED_ADDRESS can be reduced to just one column.
- Critical information such as Payment amounts and Occupancy.

Process:

1.0 Full Data Cleaning & Wrangling.ipynb

1. Load Data, inspect current data frame information
2. Remove Duplicates
3. Explore Data, Further exploration for Bathroom
4. Convert Bathroom from integer to string
5. Replace Strange Bathroom # to a standardized integer using .str
6. Reformatting the column inputs for “BROKERTITLE” using .str.split
7. Renaming Brokerages to only display Brokerage name, no additional information
8. Reformatting “Type” – Property Type
9. Drop Un-Necessary columns – Excess information
10. Reorder columns
11. Identifying Borough by extracting Borough information out of “STATE”
12. Removing STATE column, extracting zip code out
13. Renaming columns
14. Identify & standardize Data Types
15. Inputting Neighborhoods to match Zip codes
16. Fixing Zip codes with Nan – no neighborhood matches
17. Renaming columns
18. Merging column from one Data frame to another DataFrame
19. Export Data

Results:

B. Clean_Data.csv

- Only the columns that are necessary remain, no repetitive information.
- Duplicates removed
- Data is clearly named and Data types matches
- Critical information can be added such as Down Payment, Mortgage Rate, Years, Loan Amount, Loan Payment per month, Property Tax Per Month, Homeowner Insurance, Utility, Total Payments, Recommended Occupancy, Maximum Occupancy