Steps

1. Omitted some columns.
2. Imagesrc – url not required.
3. Detailurl – url not required.
4. Price column is not required as we have unformatted price column.
5. Should decide on address or address rd.
6. Converting unformatted price column from categorical to numerical by selecting the red arrow on near columns and changing the data type.
7. Row 7106 is an outlier with value 474900000 so we are excluding at since it is the only highest outlier.
8. Marking for the outliers.
9. Diamond is for having address state as DC since it is having only one row with the value.
10. Triangle pointing upwards is for “Pending”.
11. Imputed missing values for unformatted price as it has 8 null values replaced it with 373643.708737864.
12. Address st has 1 missing value we can omit it row 3056.
13. Addresscity has only 1 missing value so we can omit the row 6064.
14. AddressZipcode has 2 missing values row 6064 and row 3974.
15. Beds has 2696 with values.
16. VariableDataType has 1001 values with NULL.
17. We also have VariableDataText with other values.
18. badgeInfo has 6458 NULL values. So excluding as it is missing more than 50 percent of data
19. Pgapt has 5 NULL values rows 929, 2468, 4096, 6539 and 7958
20. Pgapt has 2 similar values called coming soon and comingsoon so merging and also merging for For Sale (Broker) & For Sale (Agent) and For Sale by owner and For Sale .
21. Pgapt 5 NULL values and 3 Unknown Listed By are excluded with ‘X’ symbol.
22. Zestimate has 3229 NULL values.
23. Brokername excluding.
24. hasAdditionalAttributions has 5 NULL values.
25. Variabledatatext can be excluded as it has values like 11 days on Zillow with this we cant do analysis with price
26. Statustext has too many values some of them cant be used so excluding.
27. Address is excluded.
28. Addressstreet is excluded.
29. Addresscity is excluded.
30. Performing imputation on beds since it has 2,696 null values out of 10,000 got value 3.11872233981143 making it into 3.
31. Performing imputation on baths since it has 2,933 null values out of 10,000 got value 2.4279872331937 making it into 2.5 since it is closer.
32. Performing imputation on area since it has 3,054 null values out of 10,000.
33. Badgeinfo has 6458 NULL values so excluding.
34. Statustext we can exclude since it is very complicated.
35. addressZipcode can be excluded since we are using addressstate.