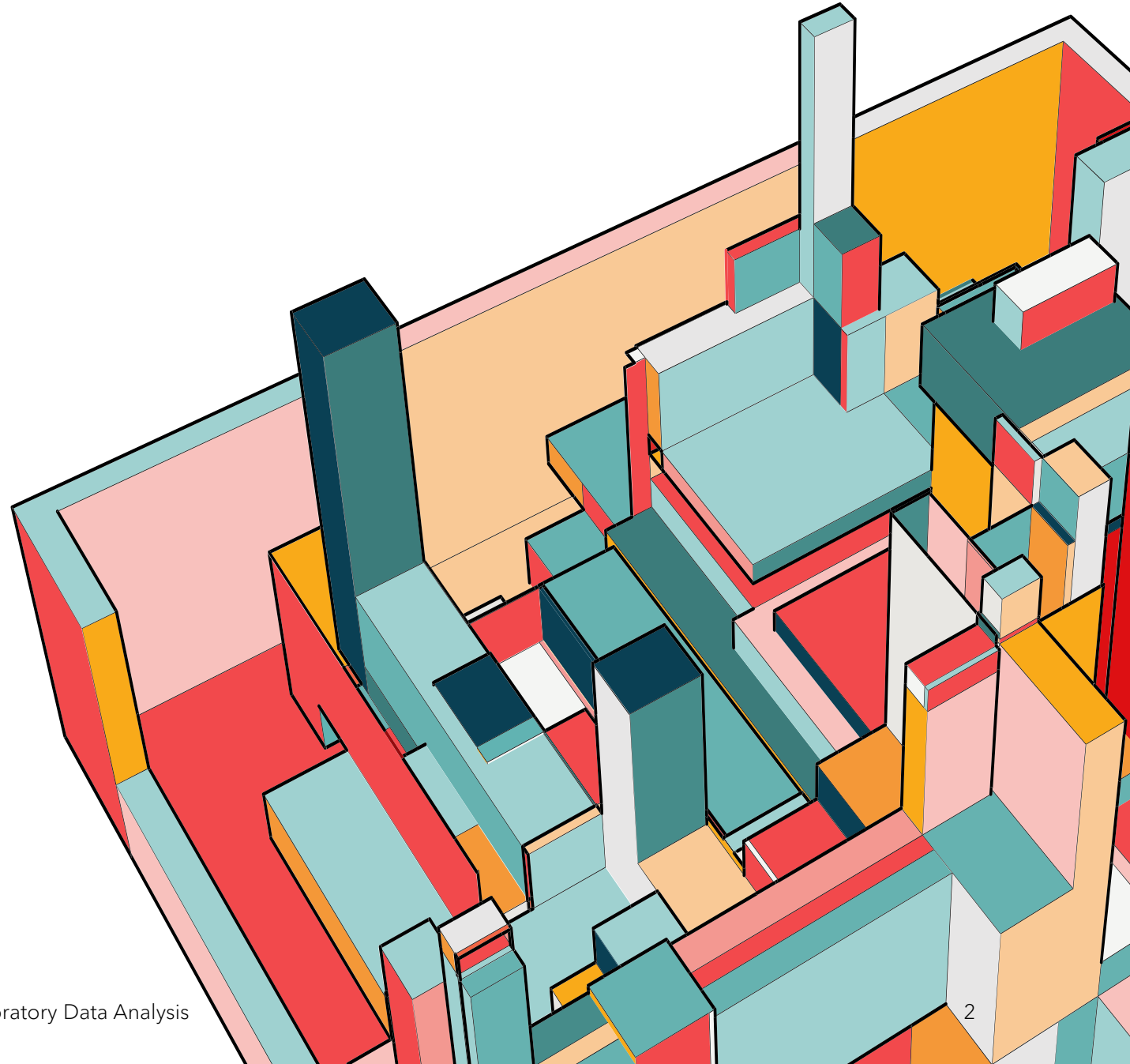
EXPLORATORY DATA ANALYSIS & DATA PRE-PROCESSING - HOUSING DATA

Padma Vasudevan

AGENDA

- Overview of the Dataset
- Exploratory Data Analysis
- Data Pre-Processing



EDA – HIGH-LEVEL DATA INSIGHTS

Dimension

5000 x 16
Numerical Features: 14
Categorical Features: 2

Data types and null values

df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 16 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   MLS               5000 non-null   int64
 1   sold_price        5000 non-null   float64
 2   zipcode           5000 non-null   int64
 3   longitude         5000 non-null   float64
 4   latitude          5000 non-null   float64
 5   lot_acres         4990 non-null   float64
 6   taxes             5000 non-null   float64
 7   year_built        5000 non-null   int64
 8   bedrooms          5000 non-null   int64
 9   bathrooms         4994 non-null   float64
 10  sqrt_ft           4944 non-null   float64
 11  garage            4993 non-null   float64
 12  kitchen_features  4967 non-null   object
 13  fireplaces        4975 non-null   float64
 14  floor_covering    4999 non-null   object
 15  HOA               4438 non-null   float64
dtypes: float64(10), int64(4), object(2)
```

Change data type from int64 to object

Highlighted cells have missing values. HOA with max missing values of 562

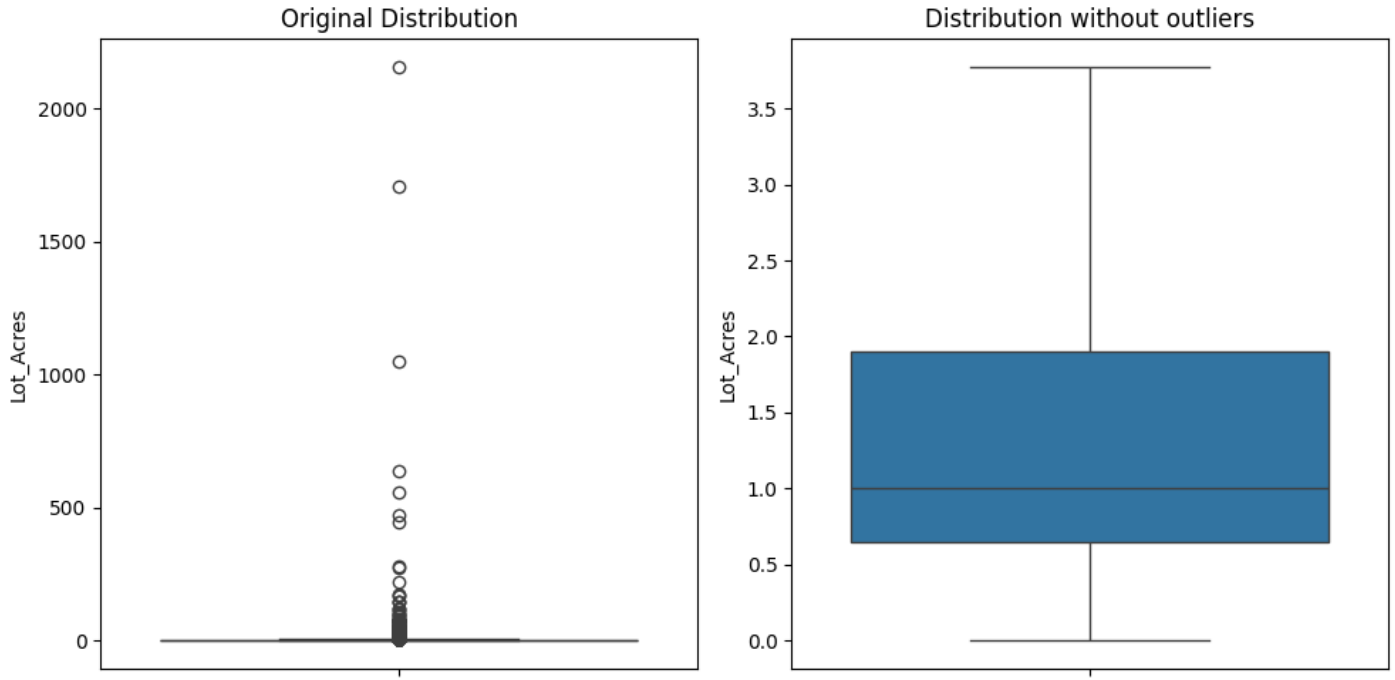
Data Dictionary

Column Name	Description
MLS	Unique Identifier
Sold_Price	Sold price of the listing, in USD
Zipcode	Postal Code
Longitude	Geographical coordinate specifying the property's east-west position, in decimal degrees.
Latitude	Geographical coordinate specifying the property's north-south position, in decimal degrees.
Lot_acres	Size of the property lot, in acres.
Taxes	Annual property taxes, in USD.
Year_Built	Year when the property was constructed.
Bedrooms	Number of bedrooms in the property.
Bathrooms	Number of bathrooms in the property (may include half-baths).
Sqft	Total square footage of the property's interior space.
Garage	Number of garage spaces available at the property.
Kitchen_Features	Description of list of features or upgrades in the kitchen.
Fireplaces	Number of fireplaces in the property.
Floor_Covering	Type of flooring materials used in the property, list.
HOA	House owner association fees, in USD

NUMERICAL FEATURES - DISTRIBUTION

Lot_Acres

- Lot_Acres is right skewed with a handful of values above 3.5 acres, extending up to 2154 acres.
- Around 90% of the data is roughly normally distributed with median at 0.99 and 90th percentile value at 3.77.
- Mean value is 4.66



NUMERICAL FEATURES - DISTRIBUTION

Sold_Price

99% of the properties have Sold_price between 169000 and 2M USD.

Max. value is 5.3M USD.

Year_Built

There were 5 records with 0 value.

Omitting 0 values, the distribution is left-skewed, with median value of 1997.

Max. value is 2019.

Bedrooms_New

Transformed variable with “>8” as value for bedrooms with value > 8.

Median & Mode = 4.

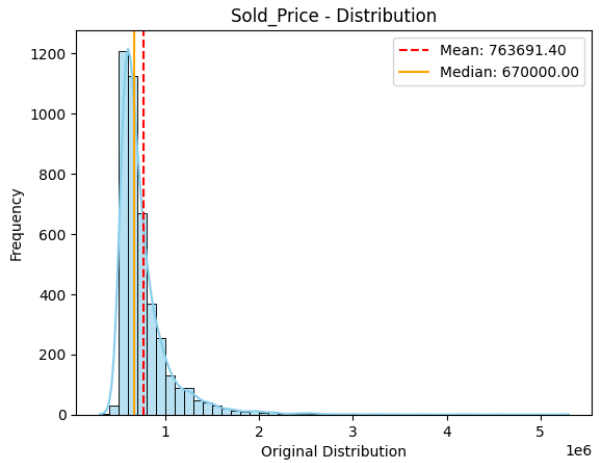
Max. value is 36.

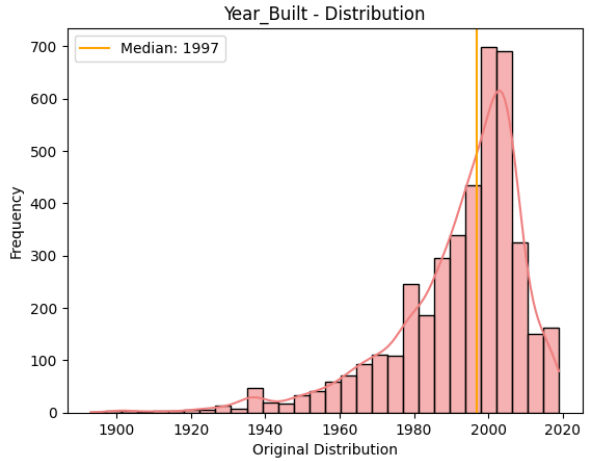
Bathrooms

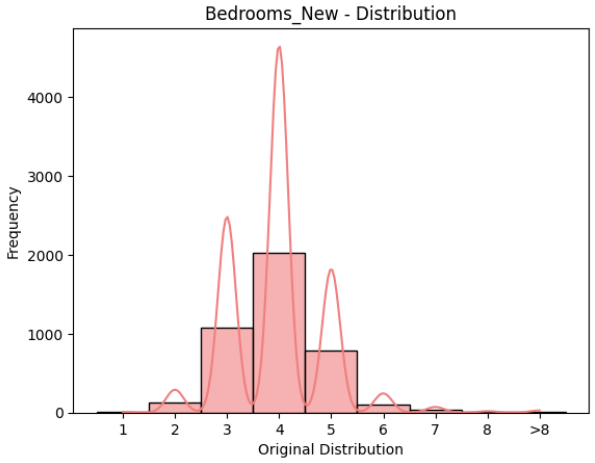
99% of the values are between 1 and 7.

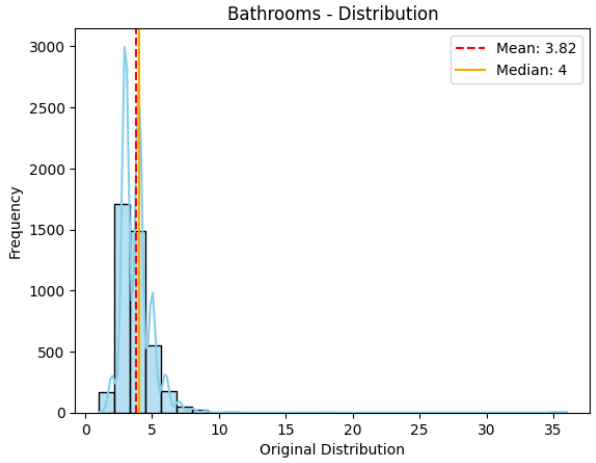
Median is 4.

Max. value is 36.

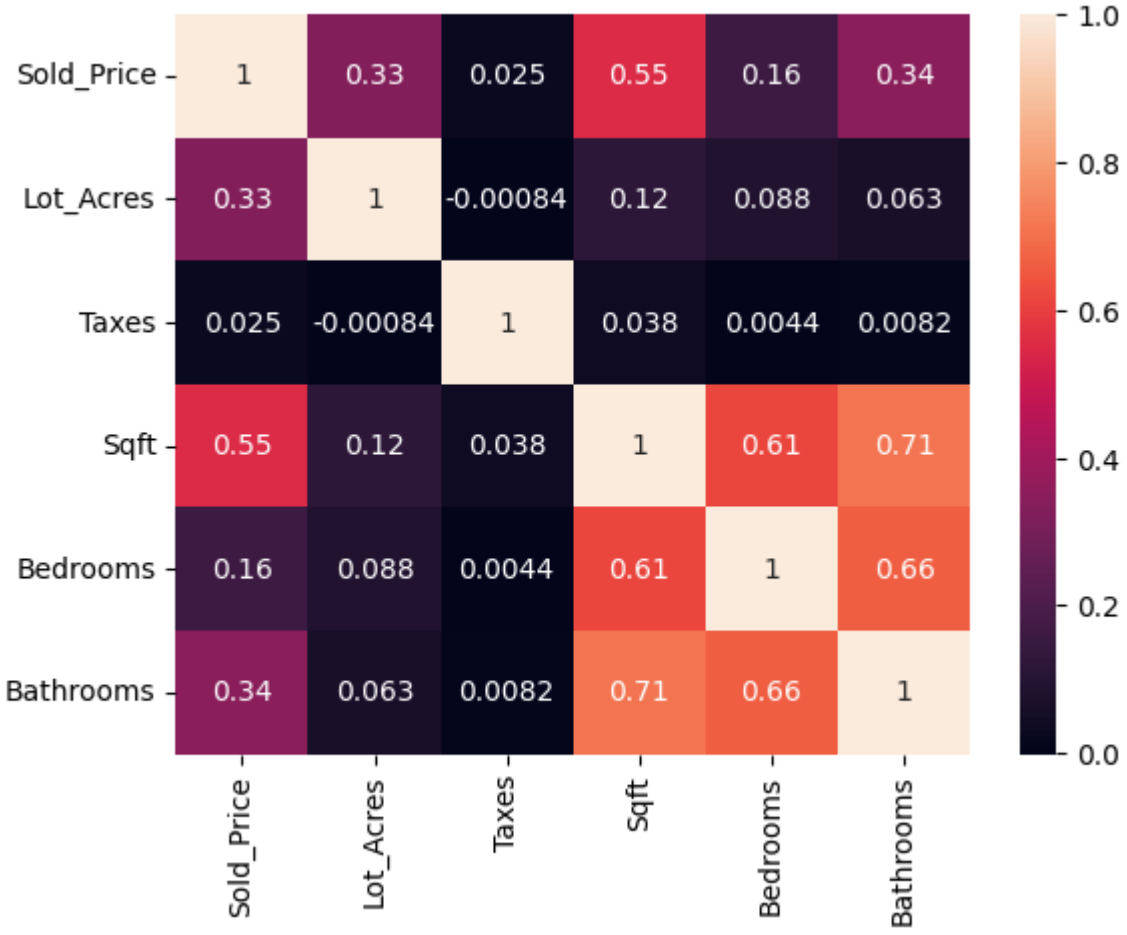








NUMERICAL FEATURES - CORRELATION

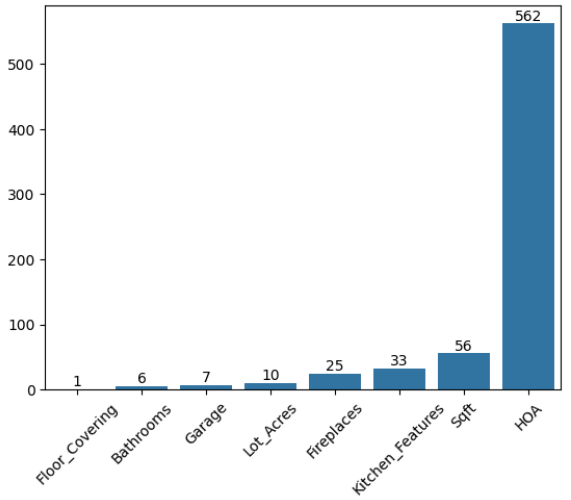


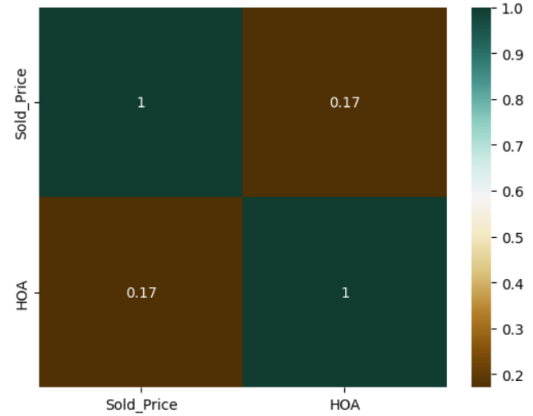
- Looking at Pairwise Correlation, we see that Sold_Price is highly correlated with Sqft.
- Sold_price is moderately correlated with Bathrooms and Lot_Acres.
- Taxes do not correlate with any other numerical feature.
- Sqft is highly correlated with Bathrooms and Bedrooms as can be expected.

DATA PREPROCESSING – HANDLING MISSING VALUE

HOA

- HOA has >10% missing values.
- To check if it is a significant attribute, its correlation with Sold_Price was measured, which is 0.17 (very mild correlation).
- To assess if zipcode-based imputation approach would be reasonable for replacing missing HOA, US zipcodes were reviewed in an online Zip repository. Each zipcode has 1000s of residential units, and are >100 sq.miles in area, as shown below.
- Hence, HOA has been removed from the dataframe due to large missing values.





Cities in ZIP code 85192

The list below includes the cities that the US Post Office accepts for ZIP code 85192. The preferred city may not be the city in which the ZIP is located. The city for 85192 is usually the name of the main post office. When mailing your package or letter, always include the preferred or acceptable cities. Using any city in the list of unacceptable cities may result in delays.

Primary/preferred city: Winkelman, AZ
Acceptable: Dudleyville

Stats and Demographics for the 85192 ZIP Code

ZIP code 85192 is located in southeast Arizona and covers a large land area compared to other ZIP codes in the United States. It also has a slightly less than average population density.

The people living in ZIP code 85192 are primarily white. The number of middle aged adults is extremely large while the number of seniors is extremely large. There are also a slightly higher than average number of single parents and a slightly less than average number of families. The percentage of children under 18 living in the 85192 ZIP code is slightly less than average compared to other areas of the country.

Population	2,120		**Land Area**	363.58	sq mi
Population Density	6	people per sq mi	**Water Area**	0.27	sq mi
Housing Units	997		**Occupied Housing Units**	804	
Median Home Value	$77,500		**Median Household Income**	$31,645	

DATA PREPROCESSING – HANDLING MISSING VALUE

Sqft

Transformed features:
Sold_Price_Bin
Bedrooms_New

Sold_Price_Bin	Bedrooms_New
Bin_4	>8
Bin_4	2
Bin_4	2
Bin_4	7
Bin_4	4

Calculated the mean Sqft value for each combination of Bedrooms_New and Sold_Price_Bin.

```
df.groupby(['Bedrooms_New', 'Sold_Price_Bin'])['Sqft'].mean()
```

3	Bin_3	3508.576923
	Bin_4	2974.956522
	Bin_1	2875.093240
	Bin_2	3029.123288
	Bin_3	3348.087097
	Bin_4	3852.866667
4	Bin_1	3220.283422
	Bin_2	3357.979661
	Bin_3	3655.614551
	Bin_4	4497.753022

- Sqft has >1% of missing values [56 records].
- To impute for these missing values, 2 new features were created:
  Sold_Price_Bin: Categorical feature for Sold_Price consisting of 4 equal Bins for the price range
  Bedrooms_New: Based on distribution, values >8 were grouped into a class called '>8'.
- For a missing value in Sqft, based on which Bedrooms_New value and price bin, the mean Sqft is imputed.

HANDLING DATA TYPES

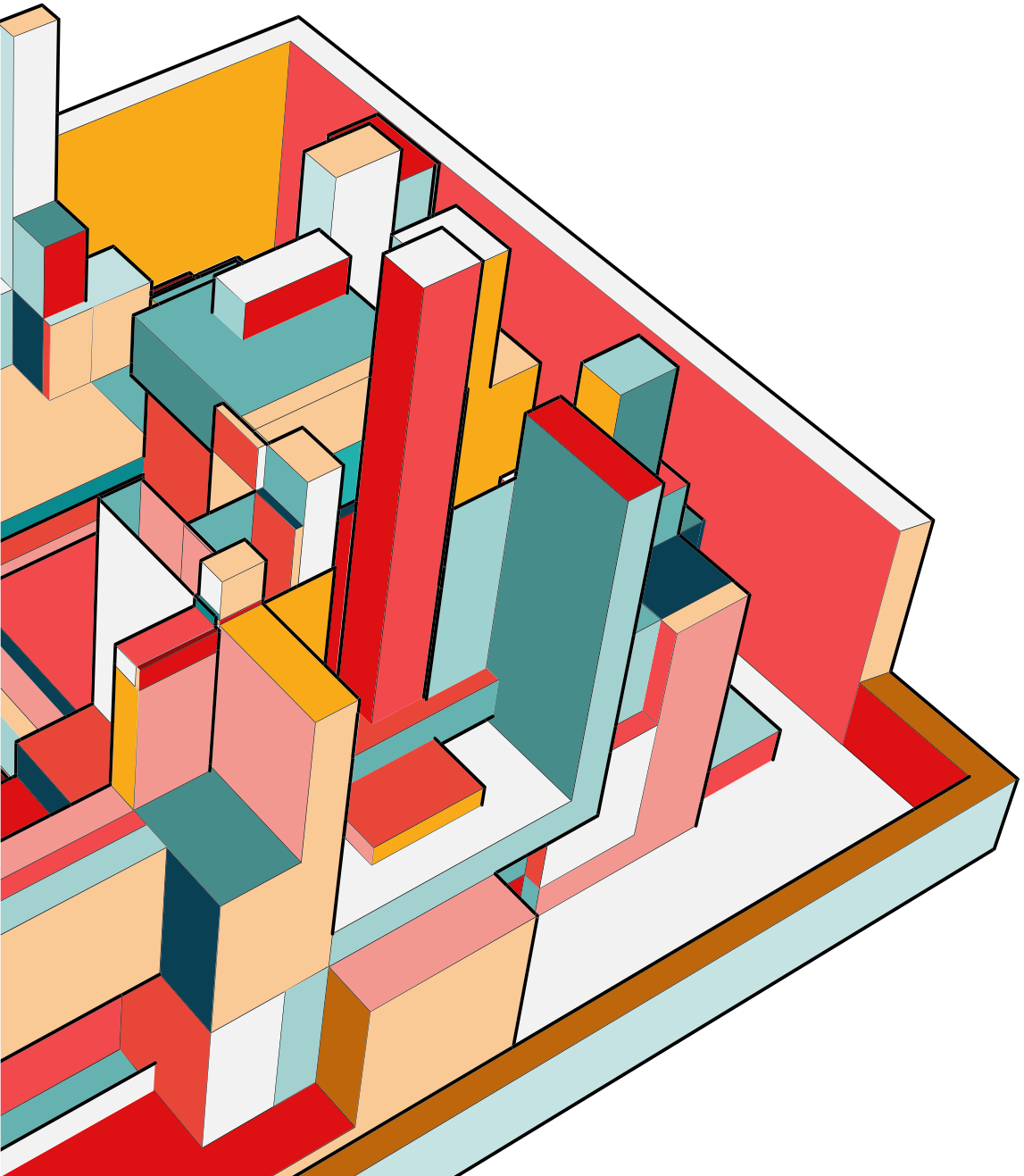
```
dtype_mapping = {
    'Year_Built': 'object',
    'Bedrooms': 'int64',
    'Bathrooms': 'int64',
    'Sold_Price': 'int64',
    'Garage':'int64'
}

df = df.astype(dtype_mapping)

print(df.dtypes)
```

```
MLS                int64
Sold_Price         int64
cell output actions int64
Longitude        float64
Latitude         float64
Lot_Acres        float64
Taxes            float64
Year_Built        object
Bedrooms           int64
Bathrooms          int64
Sqft             float64
Garage             int64
Kitchen_Features  object
Fireplaces       float64
Floor_Covering    object
Sold_Price_Bin  category
Bedrooms_New    category
dtype: object
```

ONE-HOT ENCODING FOR KITCHEN_FEATURES AND FLOOR_COVERING

Oven: mutiple	Microwave: stainless	Oven: GAS/STAINLESS	Countertops: TILE	Microwave: +Convection Built-in	Microwave: convection	Oven: Double w/Convection	Microwave: Counter Top	Microwave: Stainless built-In	Countertops: Chisled Edge Granite
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

Other: Parquet	Other: Travertine Accents	Other: Tile	Other: cork	Other: Travertine Tile	Other: NEW WOOD PLANK TILE	Other: Wood like	Other: Stained concrete	Other: 100% Porcelain Tile	Other: UPG Flooring
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0



NEXT STEPS

- To analyse 0 values in Lot_Acres and Year_Built and understand data quality issue.
- Devise suitable approach for imputation of the above 0 values.
- Some City and State values are showing as None. Need to assess reason, and update zipcode package.
- City-based analysis; groupby city and find count, avg(sold_price), etc.

THANK YOU