Exploratory Data Analysis and Data Cleaning Presentation

- Data Analyst Team

21 October 2024

Hello!

Warm greetings as I share my findings from the specified House Dataset that was assigned for analysis and will be forwarded to the modeling team.

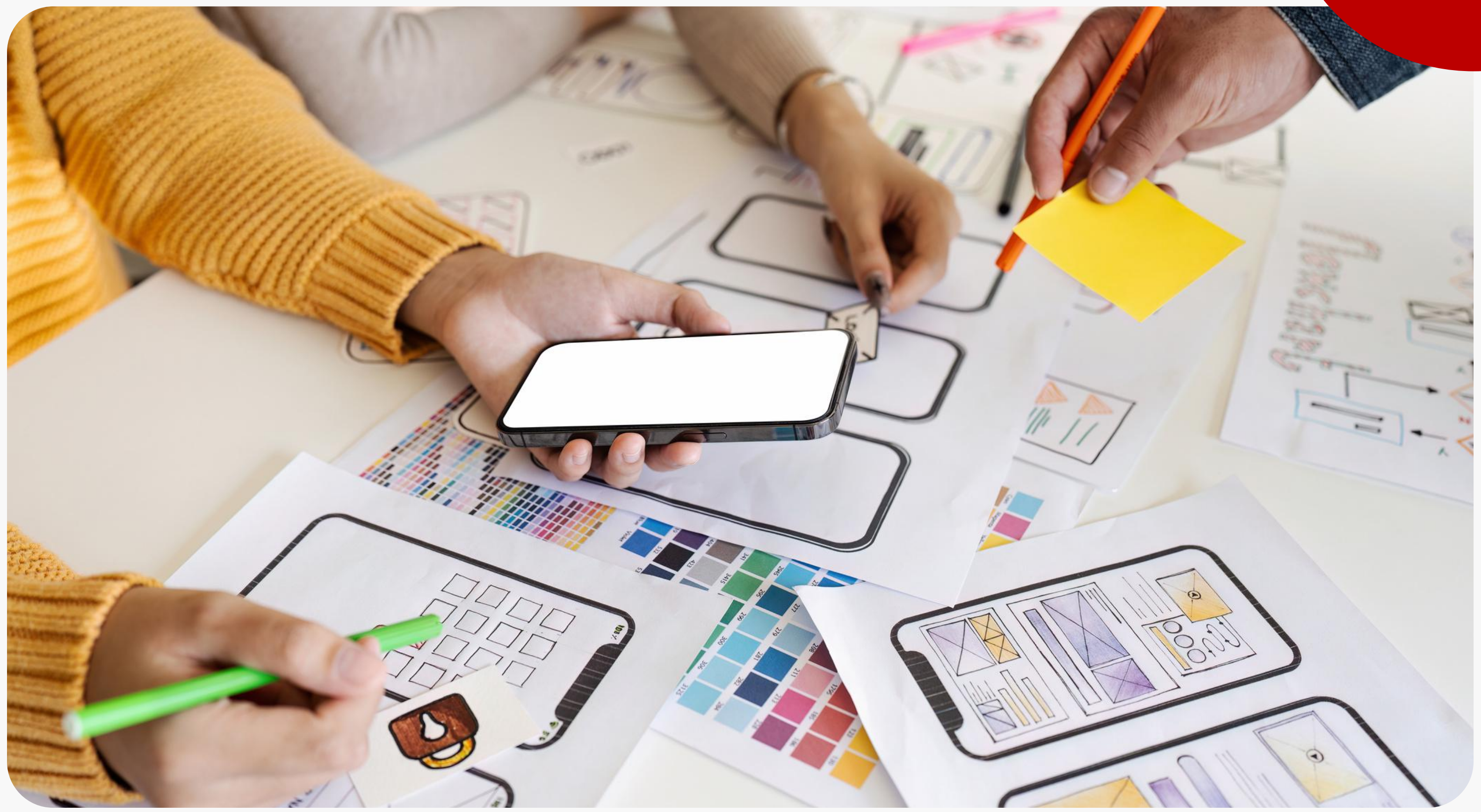
Dataset Overview

This dataset contains information about houses. It includes various features of the properties sold, such as price, location, size, and amenities. The data is intended for use in exploratory data analysis, data cleaning, and potentially for building predictive models for house prices.

Original dimensions: 5000 rows and 16 columns

Data types:

- **Numeric:** MLS, sold_price, zipcode, longitude, latitude, lot_acres, taxes, year_built, bedrooms, bathrooms, sqrt_ft, garage, fireplaces, HOA
- **Categorical:** kitchen_features, floor_covering



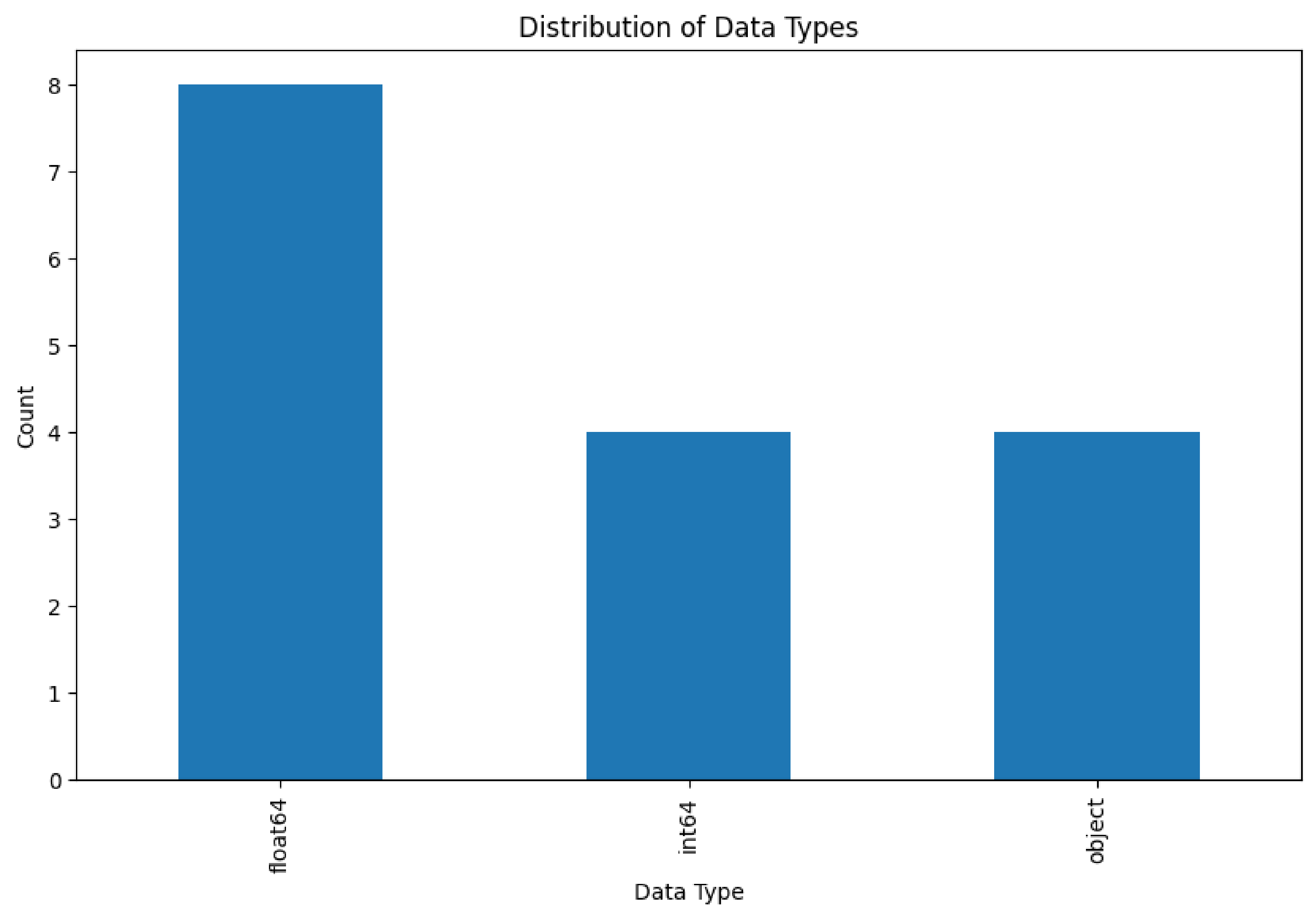
Column Description

- MLS: Multiple Listing Service number (unique identifier for the property)
- sold_price: The price at which the property was sold (in USD)
- zipcode: The ZIP code of the property's location
- longitude: Longitude coordinate of the property
- latitude: Latitude coordinate of the property
- lot_acres: Size of the lot in acres
- taxes: Annual property taxes (in USD)
- year_built: Year the house was built
- bedrooms: Number of bedrooms
- bathrooms: Number of bathrooms
- sqrt_ft: Square footage of the house
- garage: Number of garage spaces (0 if no garage)
- kitchen_features: List of features in the kitchen
- fireplaces: Number of fireplaces
- floor_covering: Types of floor coverings in the house
- HOA: Homeowners Association fees (0 if no HOA)

Statistical Analysis

Performed initial data exploration, including examining the first few rows and basic statistics of the dataset

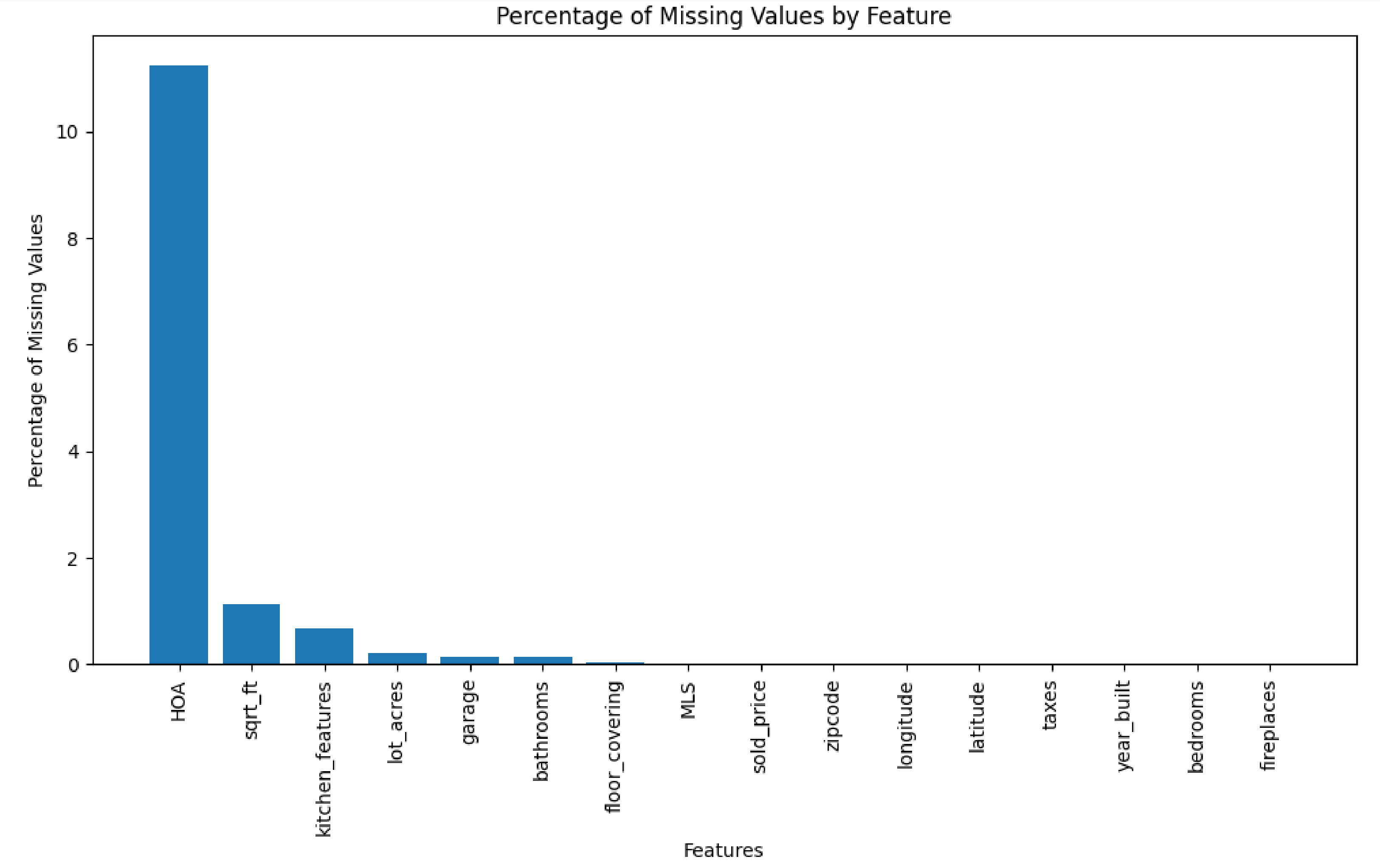
- Original dimensions: 5000 rows and 16 columns
- Numeric: MLS, sold_price, zipcode, longitude, latitude, lot_acres, taxes, year_built, bedrooms, bathrooms, sqrt_ft, garage, fireplaces, HOA
- Categorical: kitchen_features, floor_covering



Missing Value Analysis

Identifying missing values is essential for data integrity. We need to understand the extent of missing data to decide on appropriate imputation strategies.

- HOA: 11.24% (562 entries)
- sqrt_ft: 1.12% (56 entries)
- kitchen_features: 0.66% (33 entries)
- lot_acres: 0.20% (10 entries)
- garage: 0.14% (7 entries)
- bathrooms: 0.12% (6 entries)
- floor_covering: 0.02% (1 entry)
- All other features: 0% (no missing values)



EDA - Insights and Actions

OBSERVATIONS	VALUES	SUMMARY
HOA (Homeowners Association)	11.24% (562 entries) – Large percentage of Missing values	• Some properties not being part of HOA • Incomplete data • Information not available
sqrt_ft (Square Footage)	1.12% (56 entries) – Not a large percentage	• Important feature in real estate analysis
Kitchen_features	0.66% (33 entries) – Incomplete Listings	• Properties without notable kitchen features
Low-level missing data	< 0.2%	• Data entry errors • Truly missing information
Complete data	MLS,year_built,taxes,sold_price,bedrooms,etc..,	• Which is excellent for the analysis

OUTCOMES	Analysis	SUMMARY
HOA (Homeowners Association)	• Filling up with 0	• Using fillna method
sqrt_ft (Square Footage)	• Mean imputation	• Group by [bedrooms + bathrooms] from that calculating the mean values • It might help in the KNN or Regression Model
Low percentage	Median or Mode	• Mean : average of the dataframe • Median : middle value of the highest/lowest • Mode : most repeat value in the dataframe
Missing values	• MCAR , MAR , MNAR	• Missing completely at random • Missing at random • Missing not at random

Outliers Detections

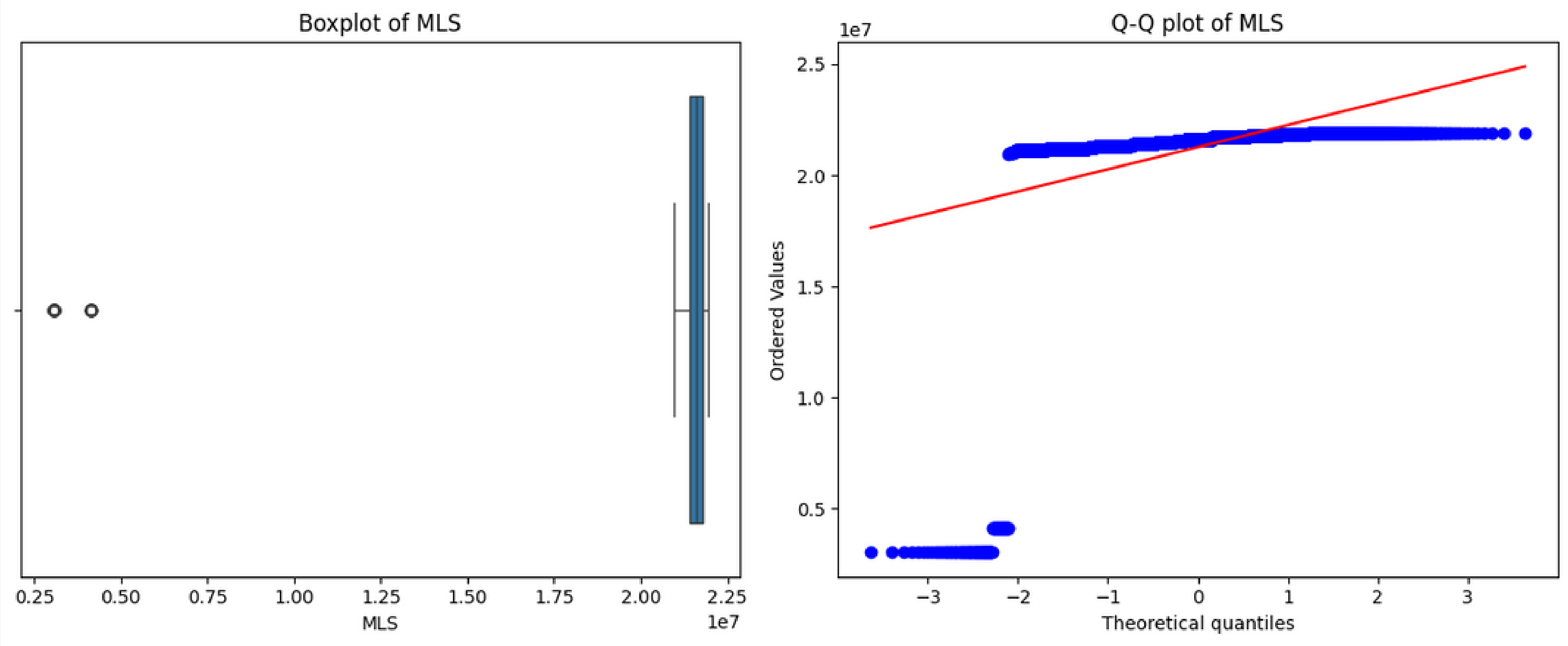
Implemented outlier detection using box plots and Q-Q plots for numeric features

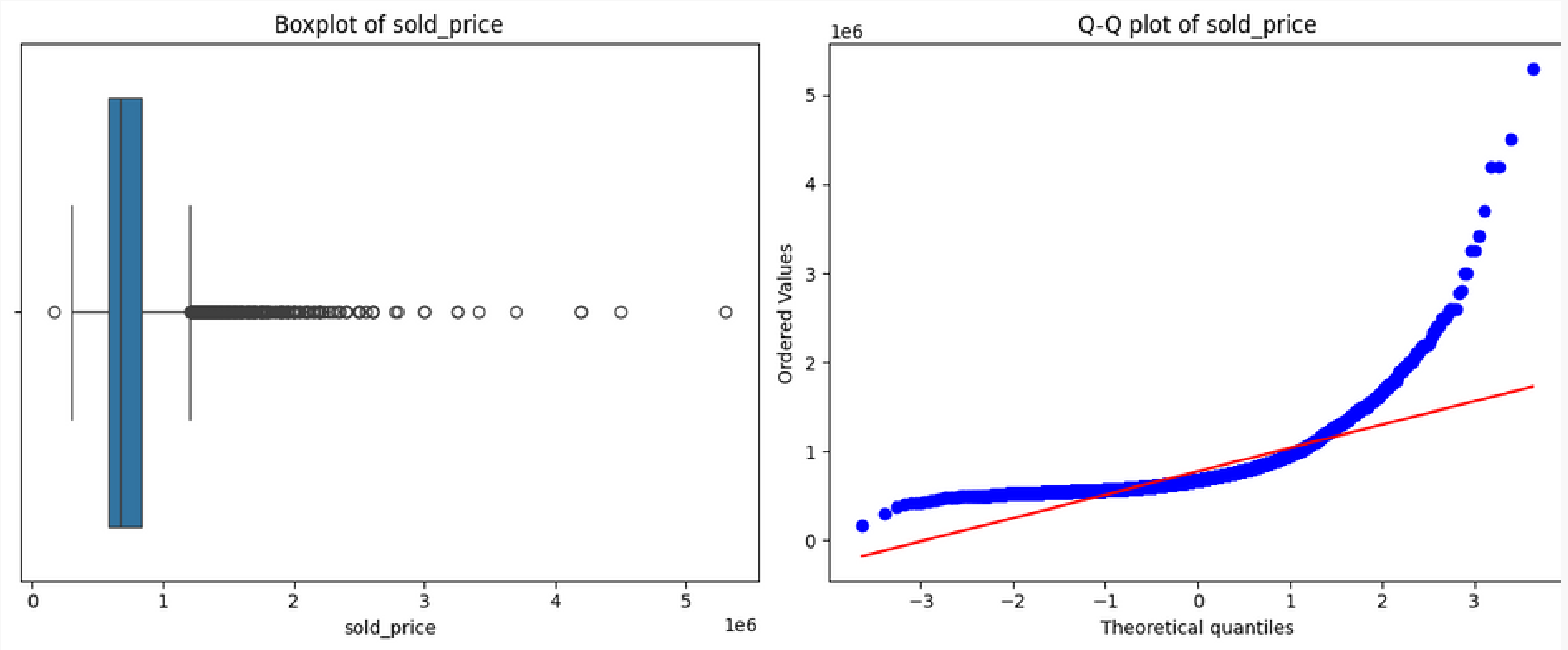
Summary Statistics:

	MLS	sold_price	zipcode	longitude	latitude
count	5.000000e+03	5.000000e+03	5000.000000	5000.000000	5000.000000
mean	2.127070e+07	7.746262e+05	85723.025600	-110.912107	32.308512
std	2.398508e+06	3.185556e+05	38.061712	0.120629	0.178028
min	3.042851e+06	1.690000e+05	85118.000000	-112.520168	31.356362
25%	2.140718e+07	5.850000e+05	85718.000000	-110.979260	32.277484
50%	2.161469e+07	6.750000e+05	85737.000000	-110.923420	32.318517
75%	2.180480e+07	8.350000e+05	85749.000000	-110.859078	32.394334
max	2.192856e+07	5.300000e+06	86323.000000	-109.454637	34.927884

	lot_acres	taxes	year_built	bedrooms	bathrooms
count	5000.000000	5.000000e+03	5000.00000	5000.000000	5000.000000
mean	4.653974	9.402828e+03	1992.32800	3.933800	3.830100
std	51.633769	1.729385e+05	65.48614	1.245362	1.386243
min	0.000000	0.000000e+00	0.00000	1.000000	1.000000
25%	0.580000	4.803605e+03	1987.00000	3.000000	3.000000
50%	0.990000	6.223760e+03	1999.00000	4.000000	4.000000
75%	1.750000	8.082830e+03	2006.00000	4.000000	4.000000
max	2154.000000	1.221508e+07	2019.00000	36.000000	36.000000

	sqrt_ft	garage	fireplaces	HOA
count	5000.000000	5000.000000	5000.000000	5000.000000
mean	3719.206533	2.816400	1.885800	73.367618
std	1154.166087	1.192131	1.133762	90.725940
min	1100.000000	0.000000	0.000000	0.000000
25%	3050.000000	2.000000	1.000000	0.000000
50%	3506.000000	3.000000	2.000000	44.000000
75%	4125.000000	3.000000	3.000000	122.000000
max	22408.000000	30.000000	9.000000	925.000000





Corr - Insights and Actions

Significant Outliers:

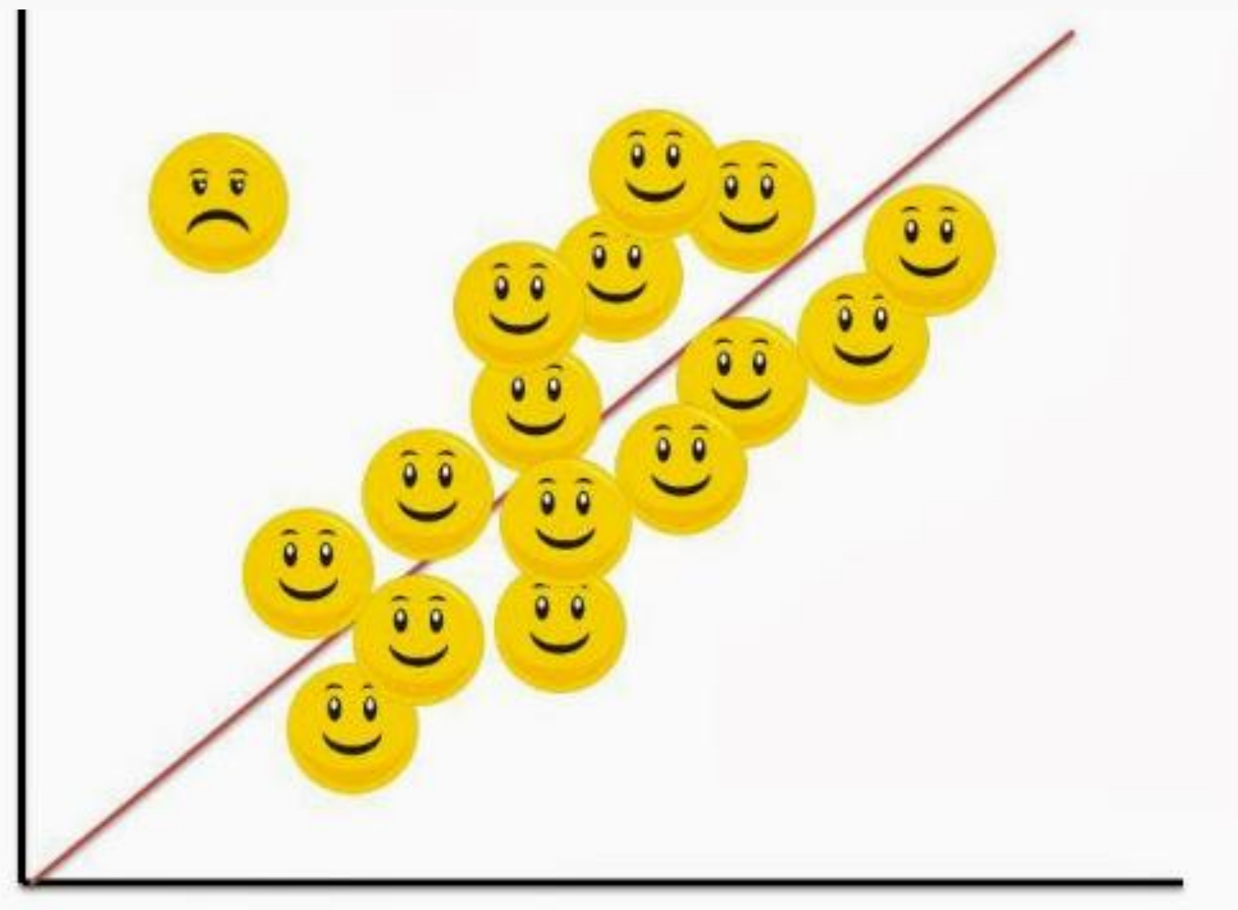
- lot_acres:Max: 2154 acres, Median: 0.99 acres
- taxes: Range: 0 to 12,215,080 (Upper outliers)
- sold_price: Max: 5,300,000, Median: 675,000
- sqrt_ft: Max: 22,408, Median: ~3,719
- HOA: Max: 925 (Upper outliers)

Patterns

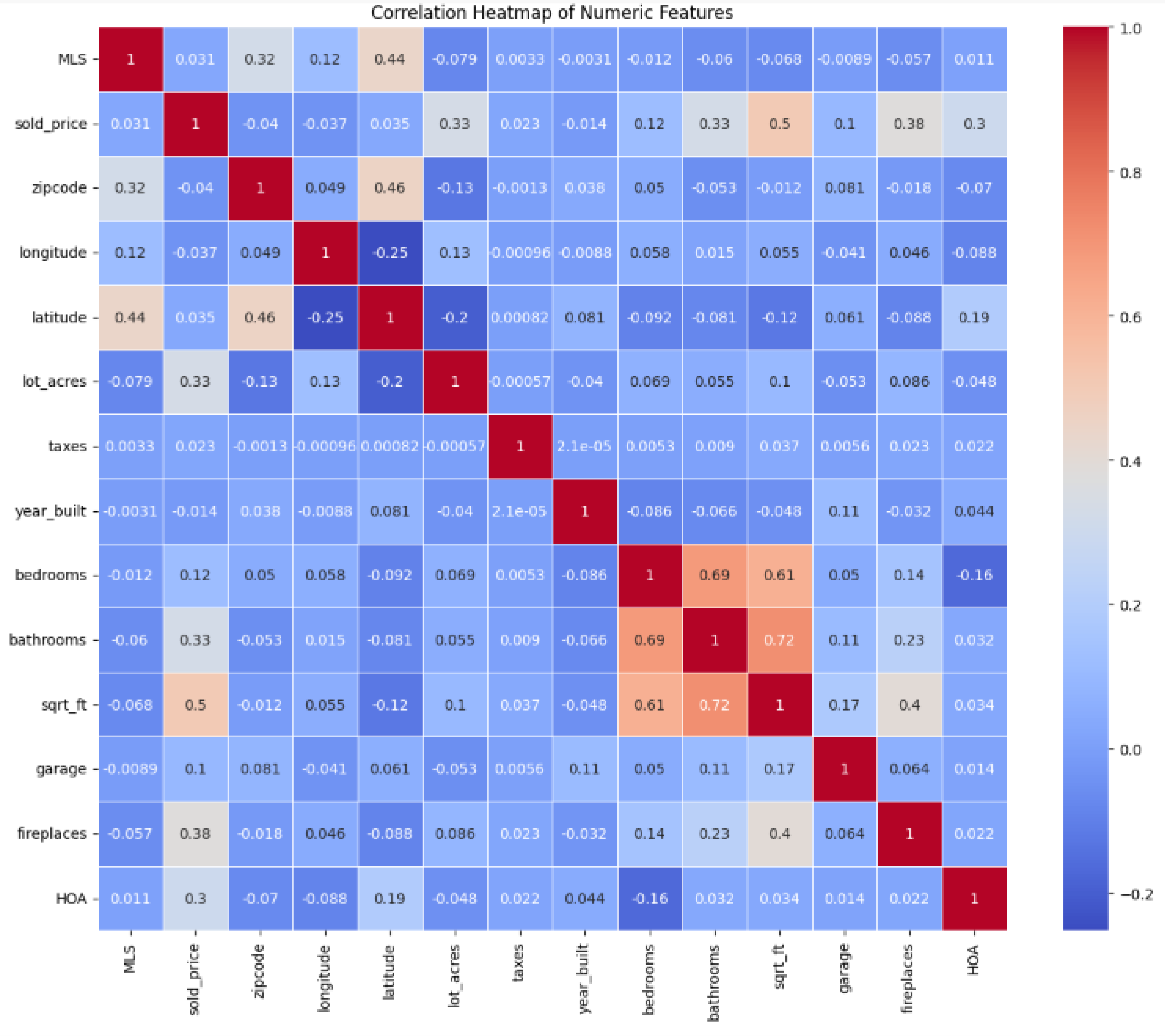
- Right-skewed disributions, most outliers are upper-end
- Extreme outliers in lot_acres and taxes suggest unique properties
- Notable outliers in solid_price,sqrt_ft and HOA

Strategies for Handling Outliers

- Investigate extreme outliers for data accuracy.
- Use robust statistical methods for modeling.
- Consider separate models for luxury vs. standard properties.
- Assess the impact of removing outliers on dataset integrity.



Correlation Analysis



Correlation Table:

	MLS	sold_price	zipcode	longitude	latitude	lot_acres	taxes	year_built	bedrooms	bathrooms	sqrt_ft	garage	fireplaces	HOA
MLS	1.00	0.03	0.32	0.12	0.44	-0.08	0.00	-0.00	-0.01	-0.06	-0.07	-0.01	-0.06	0.01
sold_price	0.03	1.00	-0.04	-0.04	0.04	0.33	0.02	-0.01	0.12	0.33	0.50	0.10	0.38	0.30
zipcode	0.32	-0.04	1.00	0.05	0.46	-0.13	-0.00	0.04	0.05	-0.05	-0.01	0.08	-0.02	-0.07
longitude	0.12	-0.04	0.05	1.00	-0.25	0.13	-0.00	-0.01	0.06	0.01	0.05	-0.04	0.05	-0.09
latitude	0.44	0.04	0.46	-0.25	1.00	-0.20	0.00	0.08	-0.09	-0.08	-0.12	0.06	-0.09	0.19
lot_acres	-0.08	0.33	-0.13	0.13	-0.20	1.00	-0.00	-0.04	0.07	0.06	0.10	-0.05	0.09	-0.05
taxes	0.00	0.02	-0.00	-0.00	0.00	-0.00	1.00	0.00	0.01	0.01	0.04	0.01	0.02	0.02
year_built	-0.00	-0.01	0.04	-0.01	0.08	-0.04	0.00	1.00	-0.09	-0.07	-0.05	0.11	-0.03	0.04
bedrooms	-0.01	0.12	0.05	0.06	-0.09	0.07	0.01	-0.09	1.00	0.69	0.61	0.05	0.14	-0.16
bathrooms	-0.06	0.33	-0.05	0.01	-0.08	0.06	0.01	-0.07	0.69	1.00	0.72	0.11	0.23	0.03
sqrt_ft	-0.07	0.50	-0.01	0.05	-0.12	0.10	0.04	-0.05	0.61	0.72	1.00	0.17	0.40	0.03
garage	-0.01	0.10	0.08	-0.04	0.06	-0.05	0.01	0.11	0.05	0.11	0.17	1.00	0.06	0.01
fireplaces	-0.06	0.38	-0.02	0.05	-0.09	0.09	0.02	-0.03	0.14	0.23	0.40	0.06	1.00	0.02
HOA	0.01	0.30	-0.07	-0.09	0.19	-0.05	0.02	0.04	-0.16	0.03	0.03	0.01	0.02	1.00

Correlation Analysis

1. **Strong Positive Correlations**

- Sold Price & Taxes: Higher prices = higher property taxes.
- Square Footage & Bedrooms: Larger homes typically have more bedrooms.
- Bathrooms & Bedrooms: More bedrooms often mean more bathrooms.

2. Strong Negative Correlation

- Latitude & Longitude: Expected due to geographical factors.

3. Unexpected Correlations

- Year Built: Weak correlation with most features, including sold price; age may not strongly predict price.
- Lot Acres: Weak correlation with sold price; larger lots don't necessarily equate to higher prices.

4. Feature Selection & Engineering Strategies

- Combine correlated features (e.g., create a 'rooms' feature from bedrooms and bathrooms).
- Engineer new features from 'year_built' (e.g., 'age', 'decade_built').
- Create 'price_to_tax_ratio' from the strong sold price-tax correlation.
- Investigate non-linear relationships or interactions for 'lot_acres' and sold price.

These observations will guide our feature selection and engineering process in the next steps of our analysis.

References

- Intro_to_Python.ipynb
- The_NumPy_Stack.ipynb
- EDA_ipynb
- python.org https://www.python.org/
- numpy.org https://numpy.org/
- pandas.pydata.org https://pandas.pydata.org/
- matplotlib.org https://matplotlib.org/
- seaborn.pydata.org https://seaborn.pydata.org/

Thank You

- Data Analyst Team