Dubai Real Estate Analysis & Modeling

1. Data collection

Sales market

Our dataset consists of approximately **1 million transactions** from the **Dubai property sales market**, covering the period from **1997 to 2022**.

We will analyze three main property types separately:

- **Unit**
- **Building**
- **Land**

Before conducting analytics and modeling, **data cleaning** is essential to ensure accuracy and reliability. Our data preparation process includes the following steps:

- **Categorical Attribute Cleaning**:
    - Standardizing values using a **dictionary-based renaming approach**.
    - Grouping **uncommon categories** under an "**OTHER**" label.
    - To analyze trends effectively, we compute the average target values within each category for the previous month, segmented separately for each real estate type.
- **Date Feature Processing**:
    - Converting date attributes into structured formats.
    - Generating new time-based features such as **year, quarter, and month**.
- **Numerical Feature Transformation**:
    - Refining numeric attributes and adding relevant flags.
    - Ensuring accurate calculations by **dividing total amounts by the number of units** to determine the **true rental price per property type**.

Rental market

Our dataset consists of approximately **5.5 million transactions** from the **Dubai rental market**, covering the period from **2010 to 2022**.

We will analyze four main property types separately:

- **Unit**
- **Virtual Unit**
- **Building** (includes **Villas**)
- **Land**

We will take the same steps to prepare data for the Rental market as for the Sales market.

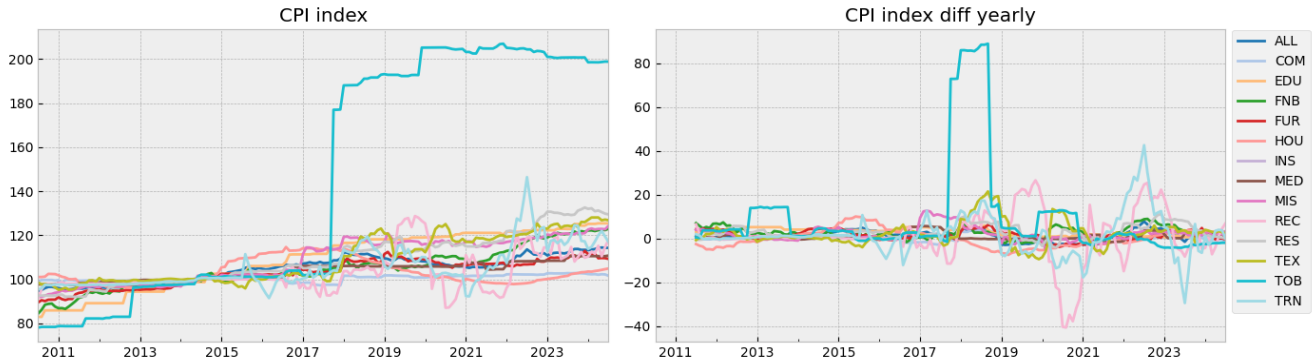
Additional Indicators for Real Estate Valuation

Below, we present additional indicators that can serve as valuable factors in assessing real estate value.

Consumer Price Index (CPI) Indicators

The figure below displays graphs of monthly CPI index indicators across different sectors.

- On the left, absolute values are shown, while on the right, the year-over-year changes are depicted.
- Notably, there was a sharp increase in the CPI index for **Tobacco in 2018** and **Transportation in 2022**.



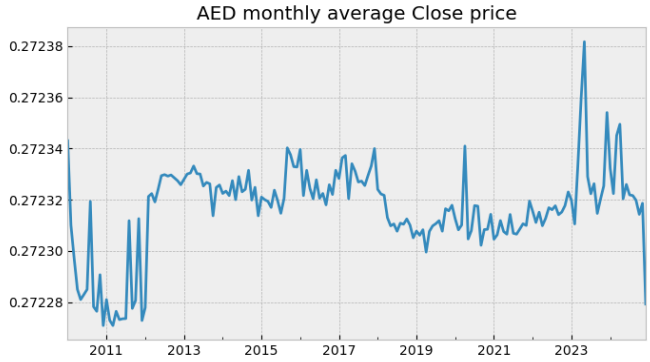
	CPI_DIV	CPI Division
0	TOB	Tobacco
1	MED	Medical Care
2	ALL	All Items
3	FUR	Furniture and Household Goods
4	EDU	Education
5	REC	Recreation and Culture
6	MIS	Miscellaneous Goods and Services
7	COM	Communications
8	FNB	Food and Beverages
9	TEX	Textiles, Clothing and Footwear
10	HOU	Housing, Water, Electricity, Gas
11	RES	Restaurants and Hotels
12	TRN	Transportation
13	INS	Insurance and Financial Services

When integrating this data with real estate figures, we will **apply a one-month data shift**. These attributes will be labeled with the prefix **CPI_**.

AED/USD Exchange Rate (Currency Strength)

The figure below illustrates the **average monthly value** of the AED/USD currency pair.

- The exchange rate remains stable at **0.27**.

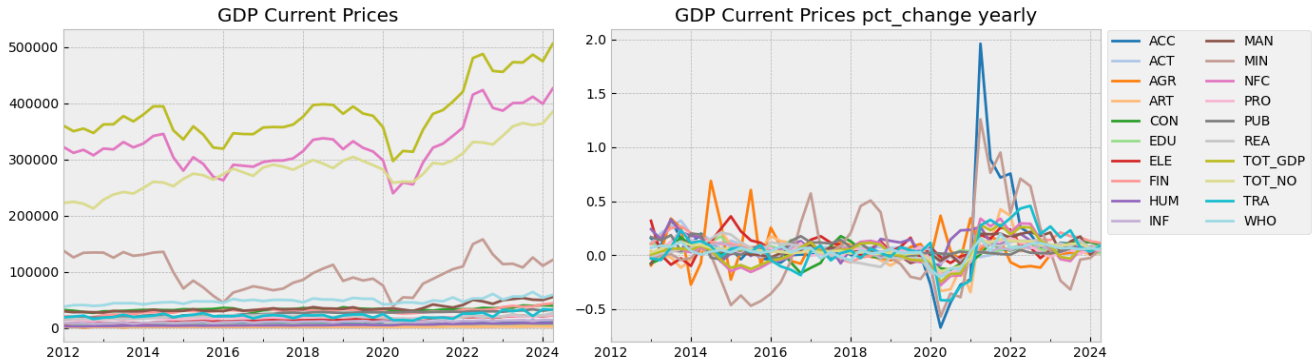


When incorporating this data into real estate analysis, we will **apply a one-month data shift**.

Gross Domestic Product (GDP) Indicators

The figure below presents the **quarterly GDP index** across different sectors.

- On the left, absolute values are displayed, while on the right, the **year-over-year percentage changes** are shown.
- During the **COVID-19 period**, we observe **significant fluctuations**, particularly in:
  - **Accommodation & Food Service Activities**
  - **Mining & Quarrying**
  - **Transportation & Storage**



	MEASURE	Measure
0	NFC	Non-Financial Corporations
1	FIN	Financial and insurance activities
2	WHO	Wholesale and Retail Trade
3	MIN	Mining and Quarrying
4	REA	Real Estate Activities
5	CON	Construction
6	PUB	Public Administration and Defence
7	HUM	Human Health and Social work Activities
8	PRO	Professional Activities
9	TOT_GDP	Gross Domestic Product
10	ART	Arts and Other Service Activities
11	MAN	Manufacturing
12	ELE	Electricity, gas, and Water Supply

13	AGR	Agriculture,Forestry and Fishing
14	EDU	Education
15	TOT_NO	Non-oil Gross Domestic Product
16	TRA	Transportation and Storage
17	INF	Information and Communication
18	ACT	Activities of Households as Employers
19	ACC	Accommodation and Food Service Activities

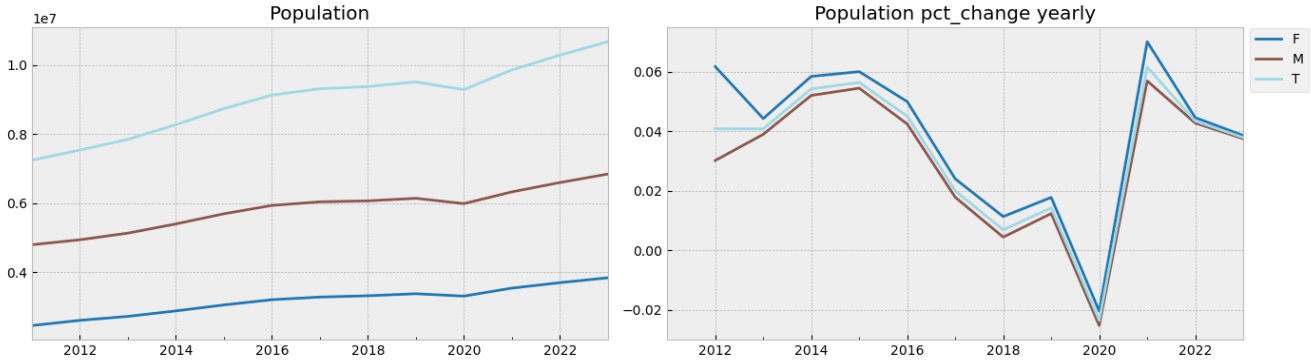
To align this data with real estate figures, we will:

- Convert quarterly GDP data to a **monthly format** using linear interpolation.
- Apply a **three-month data shift**.
- These attributes will be labeled with the prefix **GDP_**.

Population Indicators

The figure below illustrates the **annual population trends**.

- On the left, absolute values are presented, while on the right, the **year-over-year percentage changes** are shown.
- A **sharp decline** in population is observed during the **COVID-19 period**.
- The male population consistently accounts for approximately **65%** of the total.



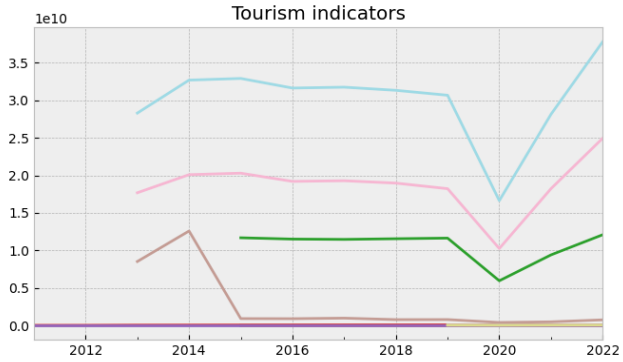
To integrate this data with real estate metrics, we will:

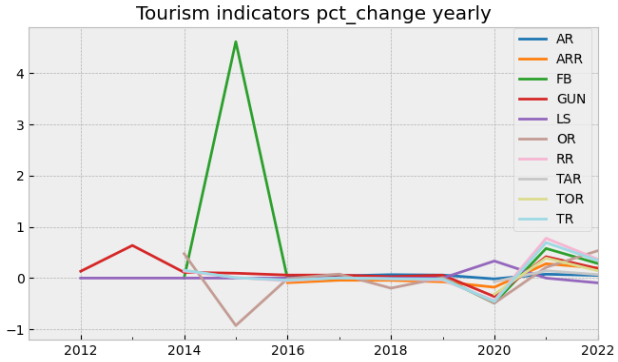
- Convert annual population data to a **monthly format** using linear interpolation.
- Apply a **12-month data shift**.
- These attributes will be labeled with the prefix **POP_**.

Tourism Indicators

The figure below showcases **annual tourism statistics**.

- On the left, absolute values are displayed, while on the right, the **year-over-year percentage changes** are shown.





	H_INDICATOR	Hotel Indicator	UNIT_MEASURE
0	RR	Room revenue	AED
1	TOR	Total Occupaied Rooms	NUMBER
2	GUN	Guest nights	NUMBER
3	LS	Length of Stay (Avg)	NUMBER
4	FB	Food and beverage revenue	AED
5	AR	Rooms (No.)	NUMBER
6	OR	Other revenue	AED
7	TR	Total revenue	AED
8	TAR	Total Available Rooms	NUMBER
9	ARR	Average room rate (ARR)	AED

For real estate analysis, we will **follow the same methodology as other annual indicators**.

Additional Annual Indicators from the World Bank

We have also incorporated a comprehensive dataset of **1,496 annual indicators** from the **World Bank**.

- After filtering out insufficiently populated indicators, **389 key indicators** were retained and added to the main dataset.
- When integrating these indicators with real estate data, we will **follow the same methodology as other annual indicators**.
- These attributes will be labeled with the prefix **WB_**.

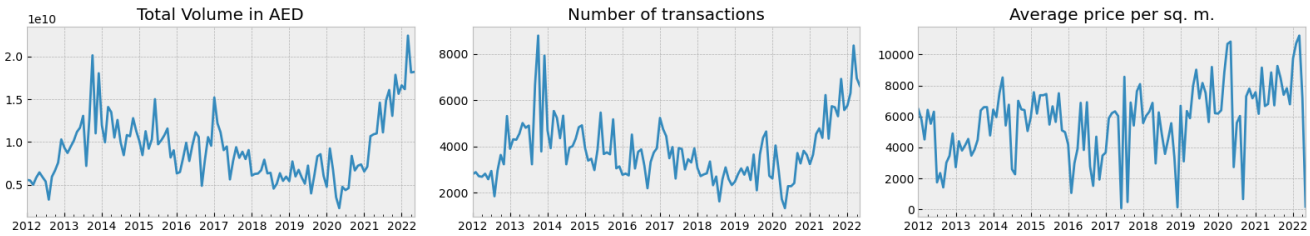
2. Exploratory Data Analysis

Sales market

Overall statistics

The figure below presents **monthly data** on transaction volumes, transaction counts, and the **average price per square meter**.

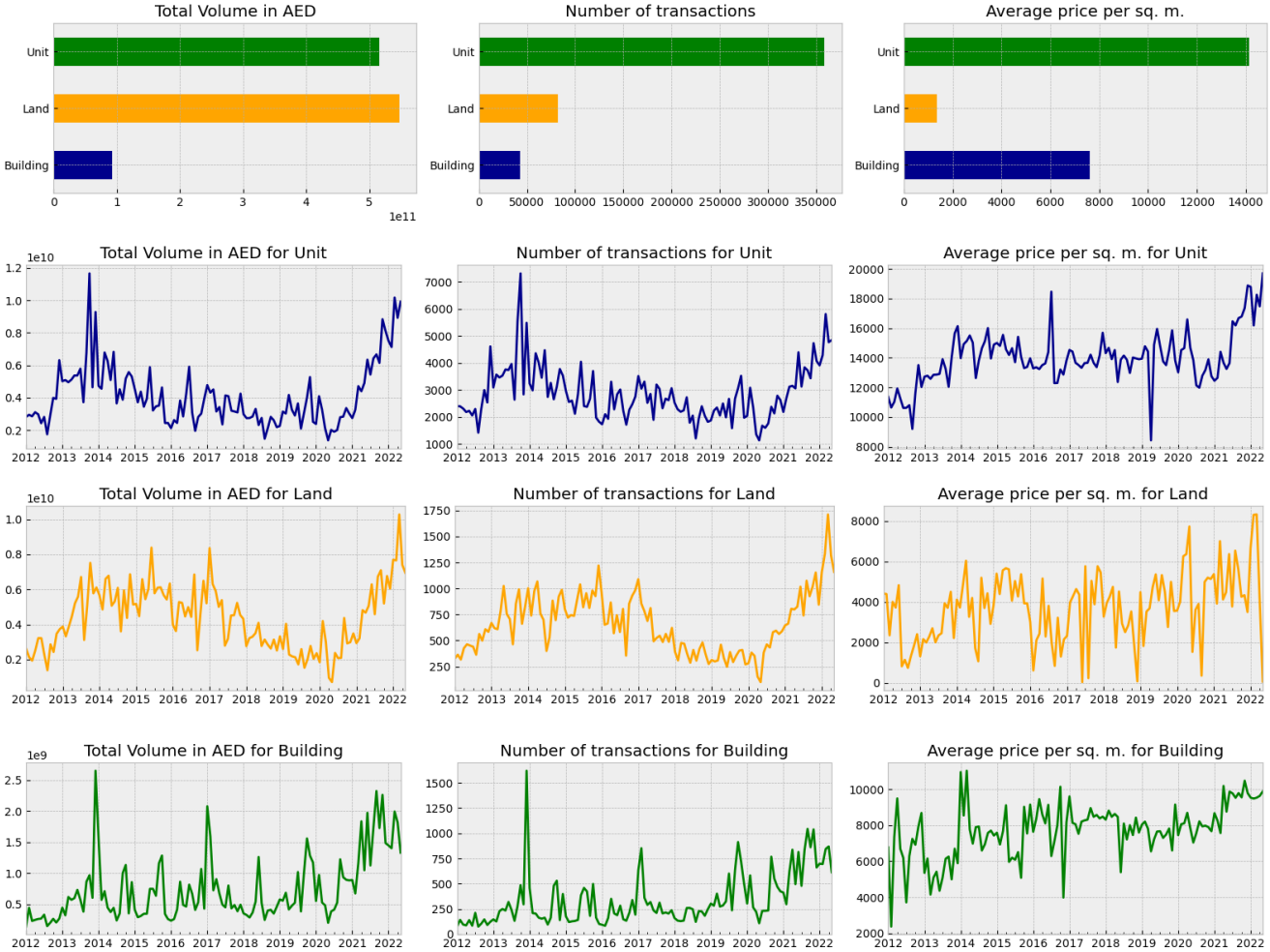
- We observe **significant growth** in both transaction counts and volumes since **2020**.
- Over the two-year period from **2020 to 2022**, these metrics have increased by approximately **four times**.



Variations by Property Type

However, the data varies significantly depending on the **type of real estate**. From the graph below, we can see that:

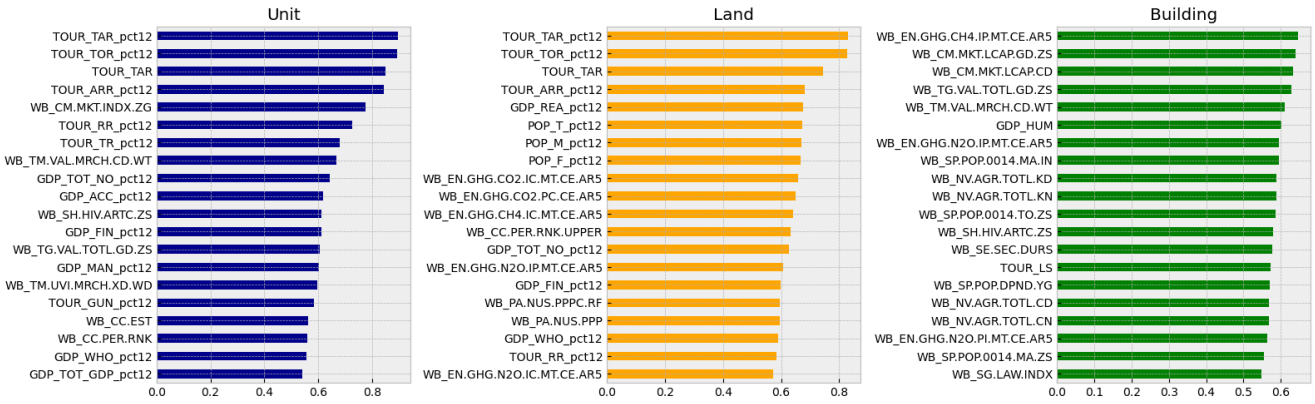
- **Units** represent the **largest market** in terms of transaction count, yet their total transaction volume is comparable to that of **Land**.
- The **lowest average price per square meter** is observed in **Land**, followed by **Buildings**, with **Units** having the highest price per square meter.



Feature Selection & Correlation Analysis

The dataset of additional attributes is **extensive**, so we will perform **feature selection** by calculating the **correlation coefficient** with the target variable.

The figure below highlights the **top 20 features** with the **highest correlation** for each **property type**.

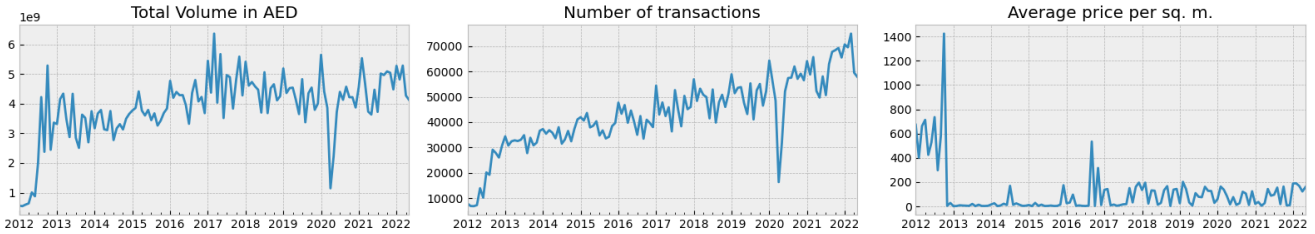


Rental Market

Overall Statistics

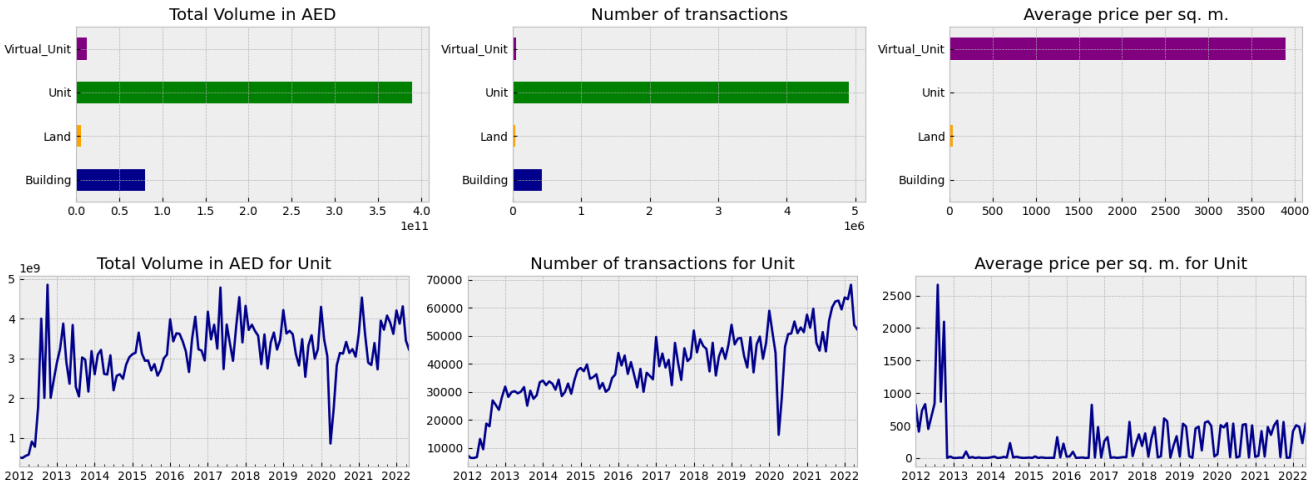
The figure below presents **monthly data** on transaction volumes, transaction counts, and the **average price per square meter**.

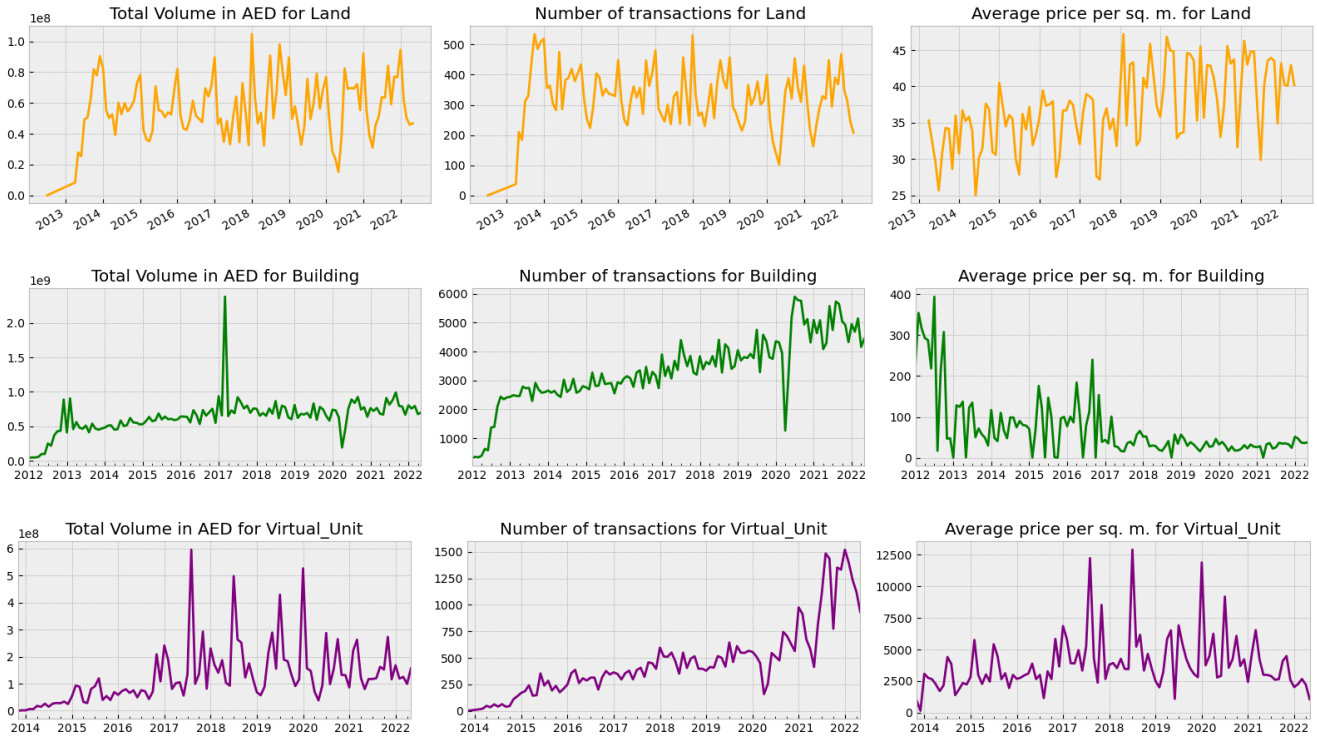
- Over the past **10 years**, we observe a **steady increase** in key metrics.
- A **sharp decline** occurred in **2020**, mirroring trends in the broader real estate transaction market.



Variations by Property Type

However, the data varies significantly depending on the **type of real estate**.



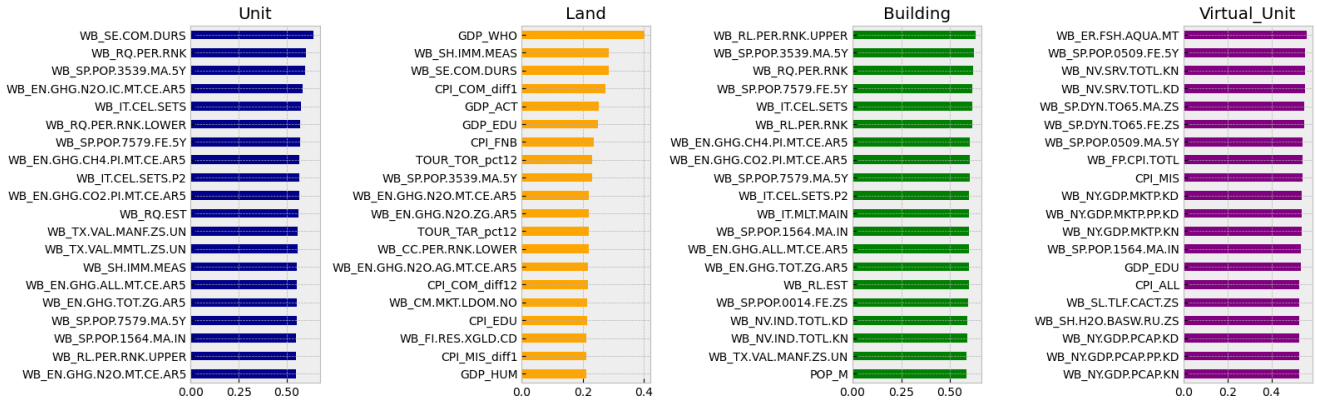


Due to **anomalous values**, the **average price per square meter** does not reflect meaningful trends.

Feature Selection & Correlation Analysis

Given the **large dataset** of additional attributes, we will perform **feature selection** by calculating the **correlation coefficient** with the target variable.

The figure below highlights the **top 20 features** with the **highest correlation** for each **property type**.



3. Modeling

Sales market

Target Analysis

Let's examine the **distribution of the target variable** that we will be forecasting.

Due to the presence of **extremely high values**, we decided to **clip** the target into **fixed**

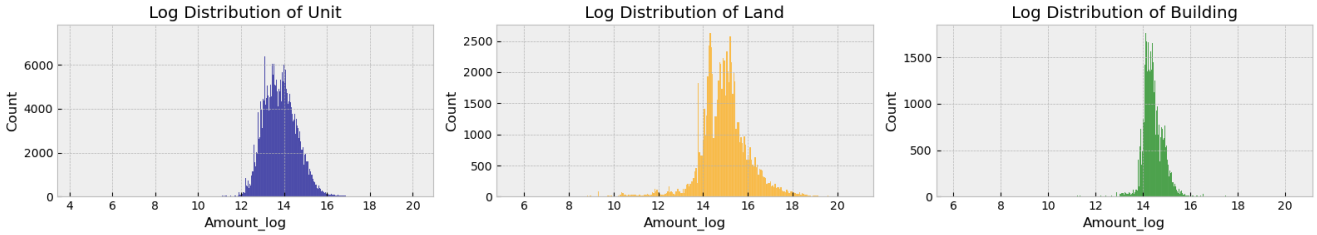
intervals.

The table below presents **summary statistics** for different **property types**.

Defining Target Clipping Boundaries

To determine appropriate clipping thresholds, we follow these steps:

- Log Transformation:
  - We apply a **logarithmic transformation** to the target variable.
- Statistical Analysis:
  - We analyze values within the range of **mean B± 3 standard deviations**.
  - Based on these calculations, we select **well-defined lower and upper bounds**.
- Impact on Outliers:
  - As seen from the **lower and upper percentiles**, only a **small number of transactions** fall into the outlier range.
  - These transactions will be clipped to the **minimum and maximum** values accordingly.



Property Type	Building	Land	Unit
count	42,627	82,344	358,487
mean	2,181,542	6,646,610	1,438,908
std	5,579,646	20,978,677	2,473,631
min	400	251	68
0.2%	319,137	10,825	99,508
50%	1,750,888	3,050,000	977,299
99.8%	20,000,000	189,030,695	17,800,025
max	750,000,000	1,125,000,000	575,000,000
mean-3std	451,691	100,292	96,423
mean + 3std	7,661,078	96,384,734	10,644,605
lower	100,000	100,000	100,000
upper	10,000,000	100,000,000	10,000,000
lower_cnt	39	1,028	718
upper_cnt	198	447	2,145

Modeling Approach

We will develop **separate models** for each **property type**, as our experiments indicate that

this approach yields **higher accuracy** compared to a **single model** for all types.

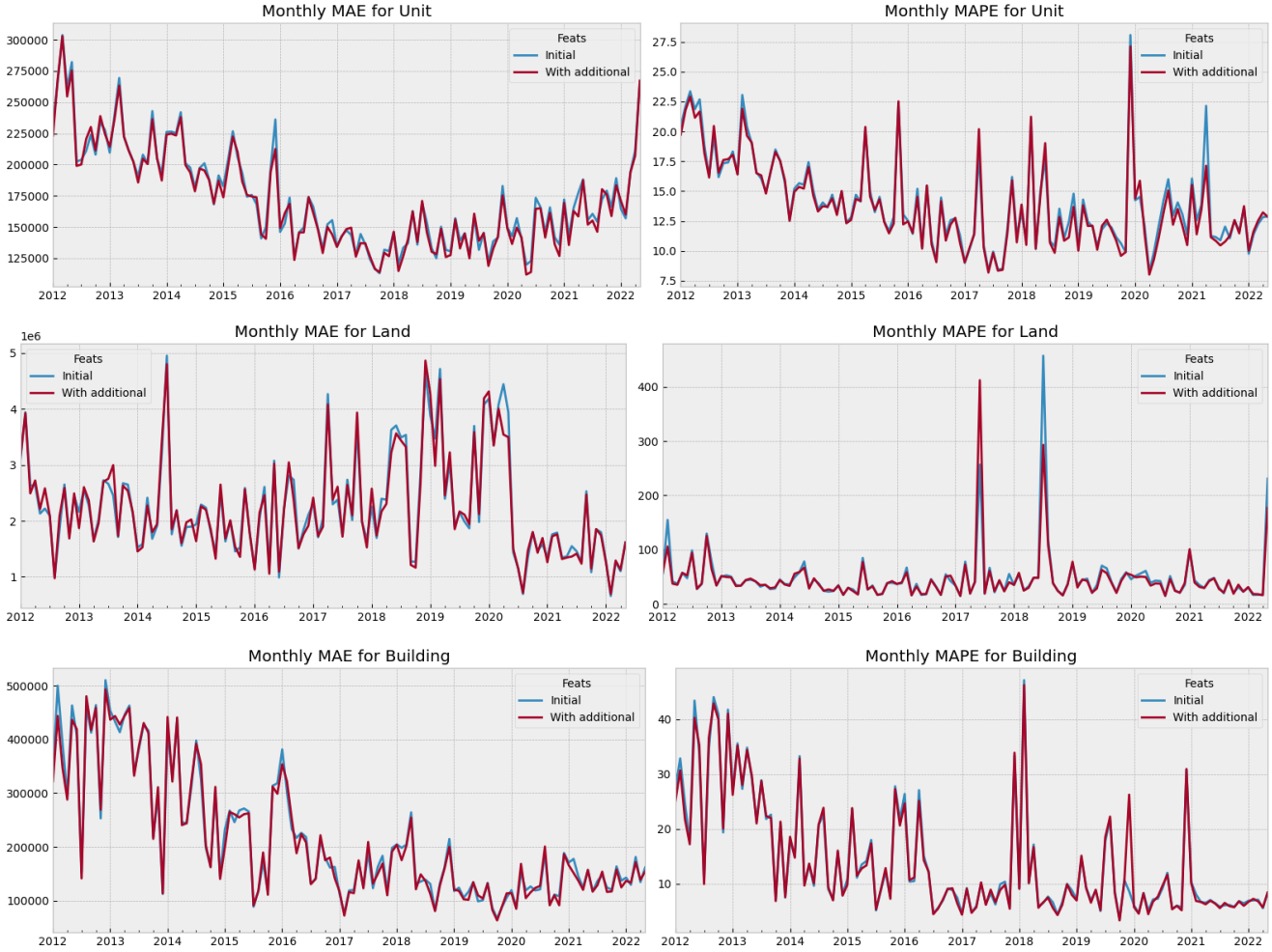
We will train the models using **Gradient Boosting (XGBoost)** and evaluate performance using the following metrics:

- **MAE (Mean Absolute Error)**
- **MAPE (Mean Absolute Percentage Error)**

Model Performance Insights

- **Overall, we achieved strong model performance across all property types.** except **Land**.
- The **best results** (in terms of MAPE and MAE) were observed for the **Building** property type.
- The **lowest performance** was recorded for the **Land** property type.
- Model accuracy has **significantly improved** in **recent years**, particularly for the **Building** category.

	MAE		MAPE	
Feats	**Initial**	**With additional**	**Initial**	**With additional**
Property Type				
Building	170,716.39	168,324.85	10.65	10.76
Land	2,041,145.59	2,034,784.49	45.04	43.40
Unit	178,319.93	175,984.26	14.08	13.85



Rental market

Target Analysis

Let's examine the **distribution of the target variable** that we will be forecasting.

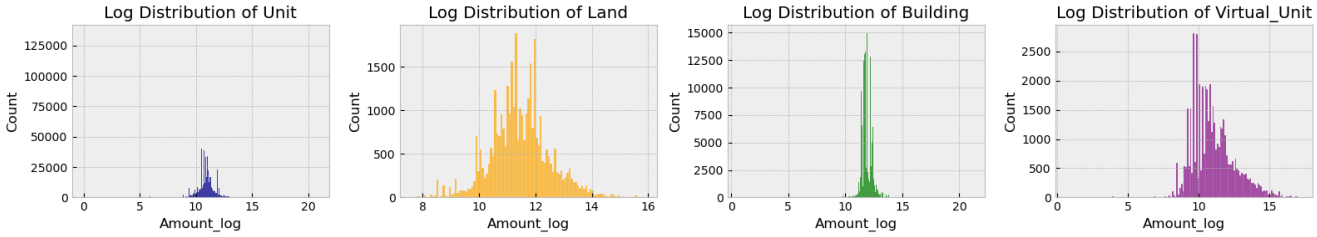
Due to the presence of **extremely high values**, we decided to **clip** the target into **fixed intervals**.

The table below presents **summary statistics** for different **property types**.

Defining Target Clipping Boundaries

To determine appropriate clipping thresholds, we follow these steps:

- Log Transformation:
  - We apply a **logarithmic transformation** to the target variable.
- Statistical Analysis:
  - We analyze values within the range of **mean B± 3 standard deviations**.
  - Based on these calculations, we select **well-defined lower and upper bounds**.
- Impact on Outliers:
  - As seen from the **lower and upper percentiles**, only a **small number of transactions** fall into the outlier range.
  - These transactions will be clipped to the **minimum and maximum** values accordingly.



Property Type	Building	Land	Unit	Virtual_Unit
count	421,395	36,097	4,904,728	47,306
mean	188,953	175,261	79,367	276,529
std	2,738,228	335,550	1,400,668	1,111,630
min	1	2,000	0	1
0.2%	13,200	4,000	113	2,400
50%	150,000	90,035	52,000	48,000
99.8%	3,007,434	3,109,842	1,295,072	12,576,374
max	1,610,378,071	8,000,000	1,116,000,000	30,356,872
mean-3std	26,045	3,955	4,257	692
mean+3std	828,682	2,248,094	640,390	5,448,659
lower	10,000	1,000	1,000	1,000

upper	1,000,000	2,000,000	1,000,000	1,000,000
lower_cnt	417	0	20,429	46
upper_cnt	2,265	204	14,218	2,545

Modeling Approach

We will develop **separate models** for each **property type**, as our experiments indicate that this approach yields **higher accuracy** compared to a **single model** for all types.

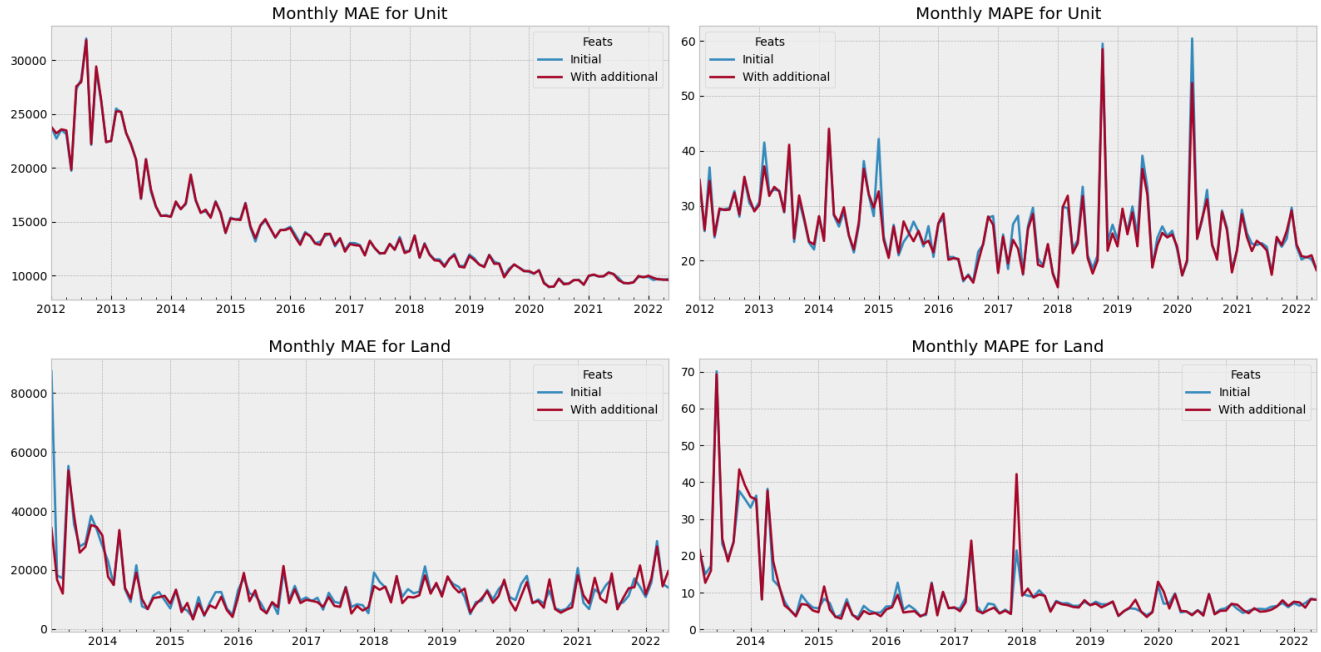
We will train the models using **Gradient Boosting (XGBoost)** and evaluate performance using the following metrics:

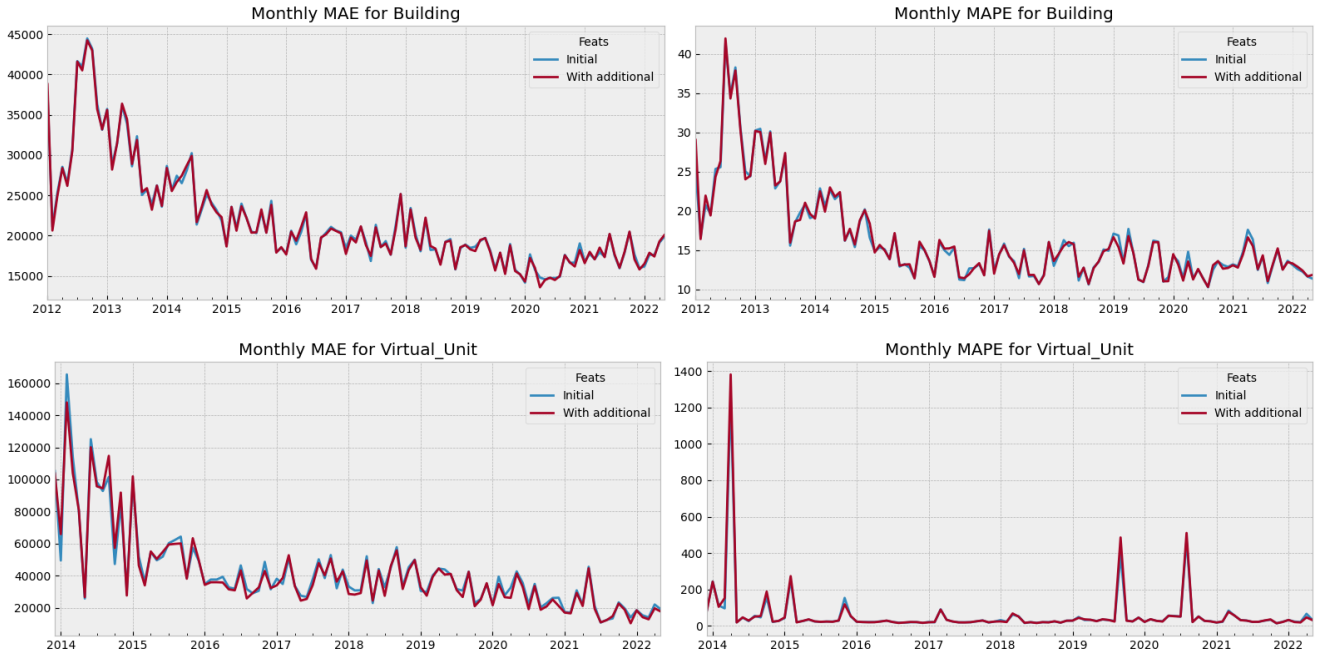
- **MAE (Mean Absolute Error)**
- **MAPE (Mean Absolute Percentage Error)**

Model Performance Insights

- **Overall, we achieved strong model performance across all property types.** except **Virtual_Unit**.
- The **best results** (in terms of MAPE and MAE) were observed for the **Land** property type.
- The **lowest performance** was recorded for the **Virtual_Unit** property type.
- Model accuracy has **significantly improved** in **recent years**.

	MAE		**MAPE**	
Feats	**Initial**	**With additional**	**Initial**	**With additional**
Property Type				
Building	20,441.50	20,395.86	15.25	15.23
Land	14,031.49	13,504.47	9.73	9.86
Unit	13,066.90	13,048.80	25.44	25.06
Virtual_Unit	30,411.23	29,239.37	41.64	42.65





4. Bonus and recommendations

Forecasting & Analysis

We will take **all transactions from the past six months** and generate **predictions** using our models.

To analyze performance, we will:

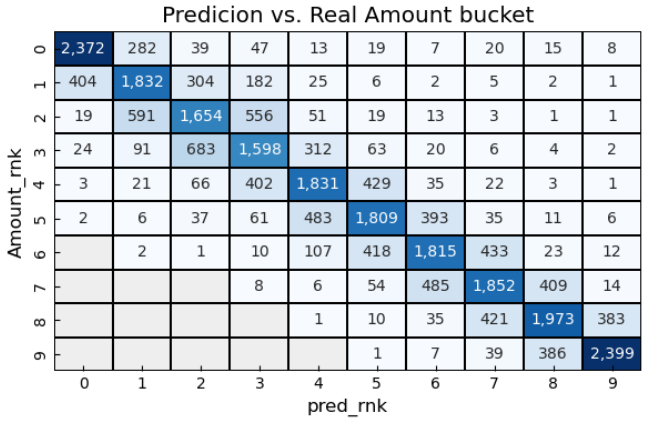
- **Divide both predictions and actual values into 10 evenly distributed buckets**.
- Construct a **matrix comparing predicted vs. actual values**.

Interpreting the Results

- If a **prediction is high**, but the **actual value is low**, this could indicate:
  - A **model error** in overestimating the propertyвЂ™s value.
  - A **potentially lucrative deal** that warrants further investigation.
- Therefore, a **detailed review** of such properties is necessary.

Methodology & Use Case

- This approach is useful for **sorting and identifying promising candidates**.
- However, it requires **additional analysis** to confirm opportunities.
- Below, we present the **top listings for each property type** based on our predictions.



Property Type	Property ID	Amount	dt_month	pred	Amount_rnk	pred_rnk
Building	859514334	100,000.00	2022-03-01	8,210,458.25	0	9
Land	1253380035	142,707.68	2022-05-01	7,679,044.50	0	8
Unit	276319953	500,000.00	2022-05-01	2,960,526.30	1	8

Output

We have analyzed the property rental and sales markets of Dubai for the period from 2012 to 2022. These markets are very similar in terms of features impacting the rental/transactional price of properties, but there are some important differences that we can highlight:

1. The rental market has been **constantly growing** since 2012 for all property types except **Land**, whereas the **sales market** started actively growing from the **second quarter of 2020** (coinciding with the start of COVID-19).
2. **Annual rental prices have been declining since 2017**, when they peaked, while **sales property prices have been increasing**.
3. **Models for predicting sales prices perform better** in terms of accuracy compared to models for predicting annual rental amounts.
4. **Additional attributes provided only a slight improvement** in model quality across both markets.
5. The developed **model can be used to identify potentially attractive real estate**, helping to **detect undervalued properties** and support **data-driven investment decisions**.