Luca Ordronneau

Dubai Real Estate

Ocean Data Challenge

Discord :
lucanew#3793

GitHub :
lucaordronneau

Version of
May 23, 2023

Introduction

The Dubai real estate market, a driver of economic growth, offers unique opportunities and challenges. This report focuses on employing advanced data analytics and machine learning to optimize pricing strategies and transactions, enabling efficient forecasting and evaluation of property prices in Dubai.

Our analysis is based on two comprehensive datasets, one detailing rental contracts and the other containing transaction records. Each dataset provides a range of attributes, such as property size, type, usage, geographical proximity to landmarks, and transactional details. This rich set of features not only enables a nuanced understanding of the market but also provides a robust basis for predictive modelling

The first part of our study examines the historical trends and correlations in the market, highlighting key insights and contributing factors affecting property prices. Special emphasis is placed on understanding the correlation between property size and sale/rental price, as well as identifying and ranking the factors influencing these prices over time.

In the second part, we design and implement machine learning models to predict rental prices based on a property's characteristics and forecast future trends. These models aim to assist stakeholders, such as buyers, sellers, and real estate agents, in making informed decisions.

These efforts aim to create a data-driven, efficient, and fair real estate market, directly contributing to Dubai's sustainable economic development.

0.1 Data Analysis

0.1.1 Data Preparation

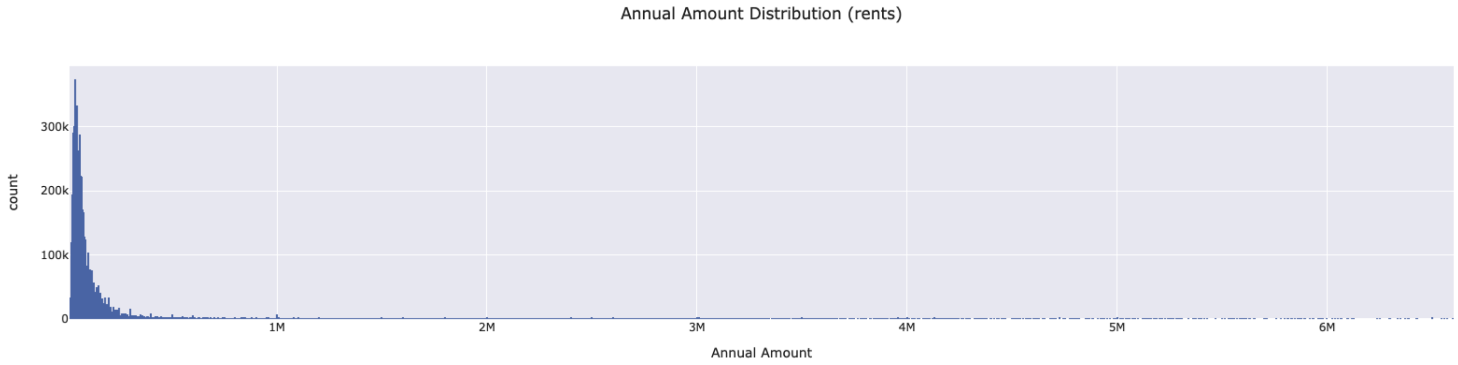
In the preparatory stage of the data analysis, meticulous data cleaning and preprocessing was carried out to ensure that the datasets were ready for detailed inspection and exploration. This process included rectifying the errors encountered in various data types such as datetime, float, and integer values, which may have been caused by inconsistencies during data collection. Furthermore, aberrant variables (ex. negative Contract Amount) and outliers were identified and effectively handled to avoid any skewed interpretations in the analysis. These outliers were detected by adopting a quantile-based approach, where values falling below the 0.01 quantile and above the 0.99 quantile were considered anomalous and consequently excluded. This strategy led to a more homogenized and streamlined dataset, which was critical for obtaining reliable and robust insights in the subsequent stages of the study.

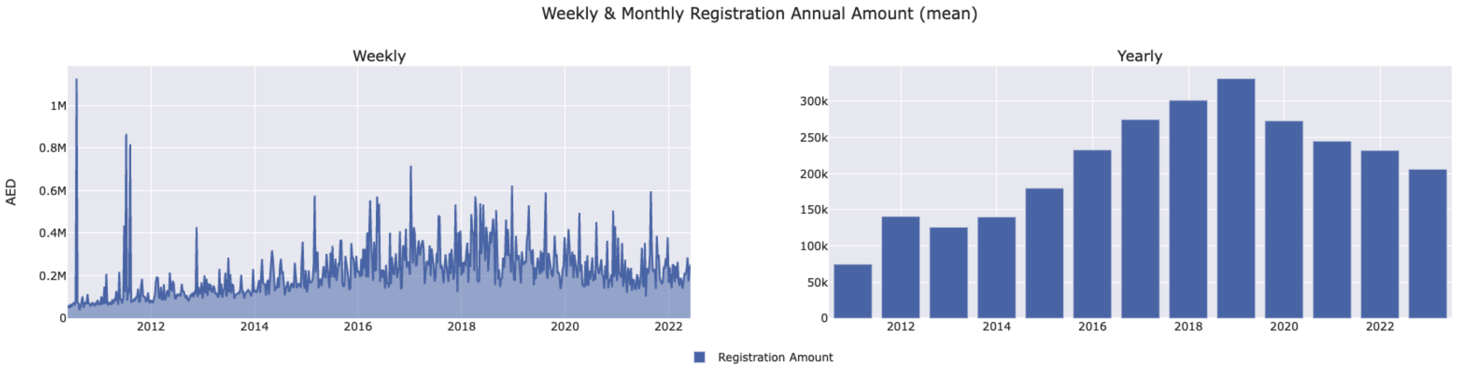
0.1.2 Global Real Estate Market Analysis

Rents

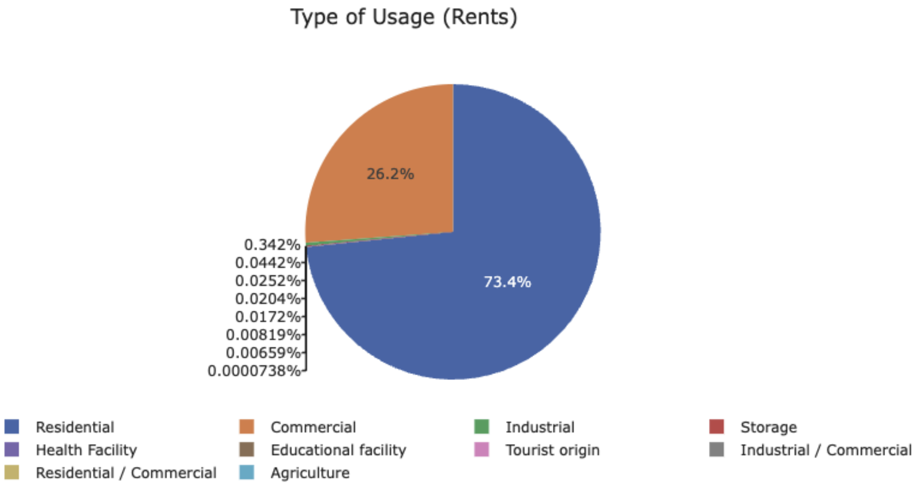
In our data-driven analysis of the real estate market in Dubai, we dissected the information into several components to reveal underlying trends, relations, and insights.

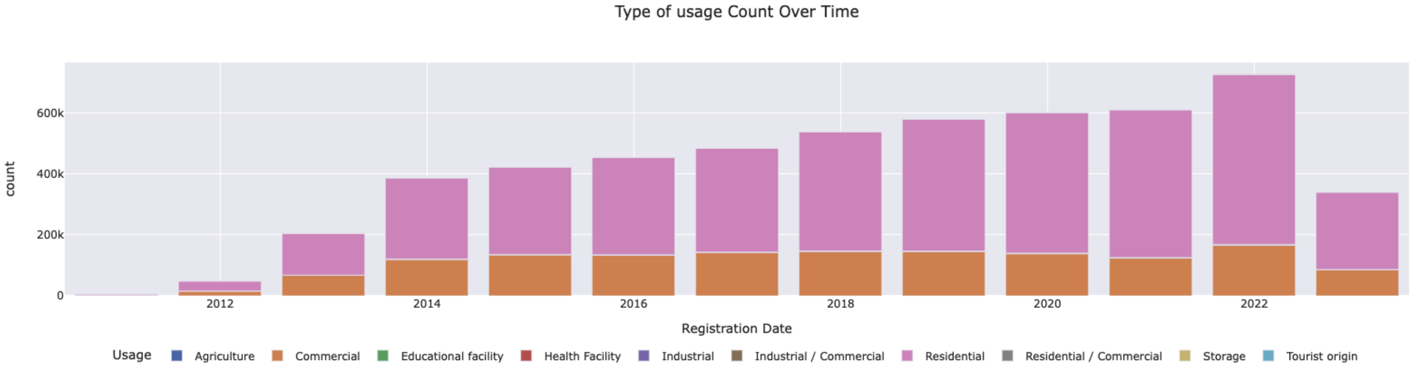
Our focus was primarily on the rental market, and the initial exploration revealed that the distribution of annual rents gravitates around the 200k to 300k AED range. Over the years, particularly from 2010 to 2019, there was a steady surge in rent prices, rising from a baseline of 100k to a peak of 400k AED. However, post-2019, rents exhibited a downward trend, returning to the 200k AED ballpark.

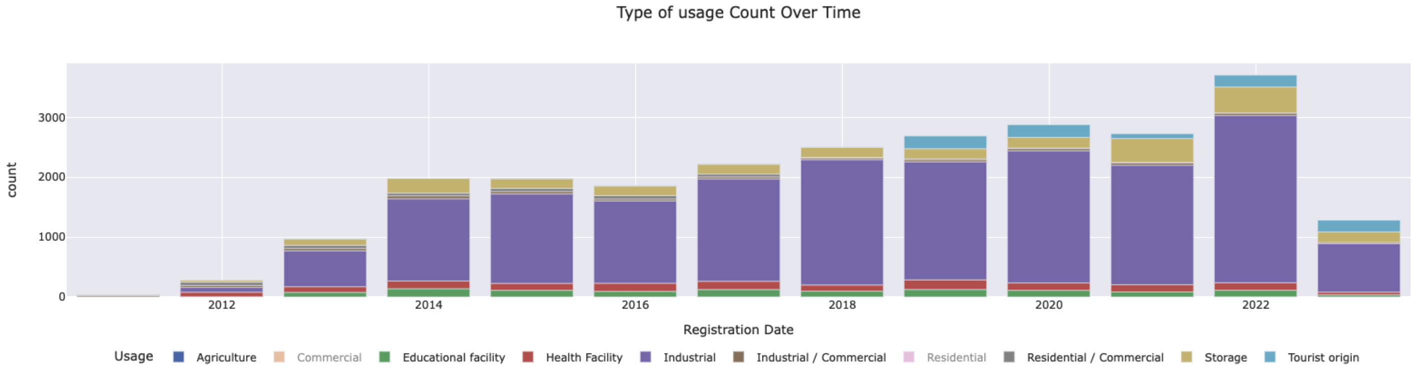




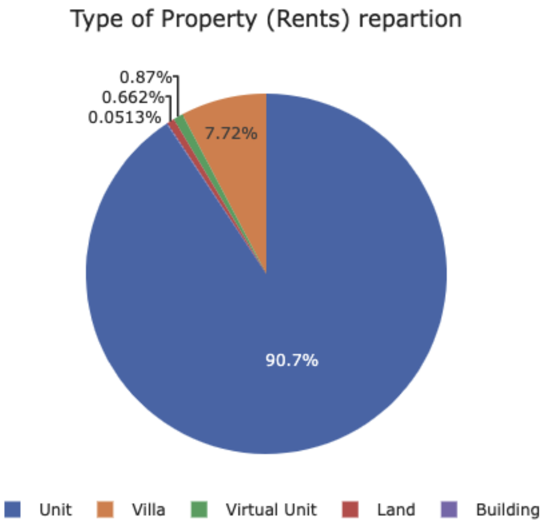
From a property usage perspective, the majority falls under residential (73%) and commercial (25%) categories. Interestingly, beyond these categories, there's a marked emergence of properties earmarked for tourism starting from 2019. This could be indicative of an evolving real estate landscape adapting to Dubai's growing status as a global tourist destination.

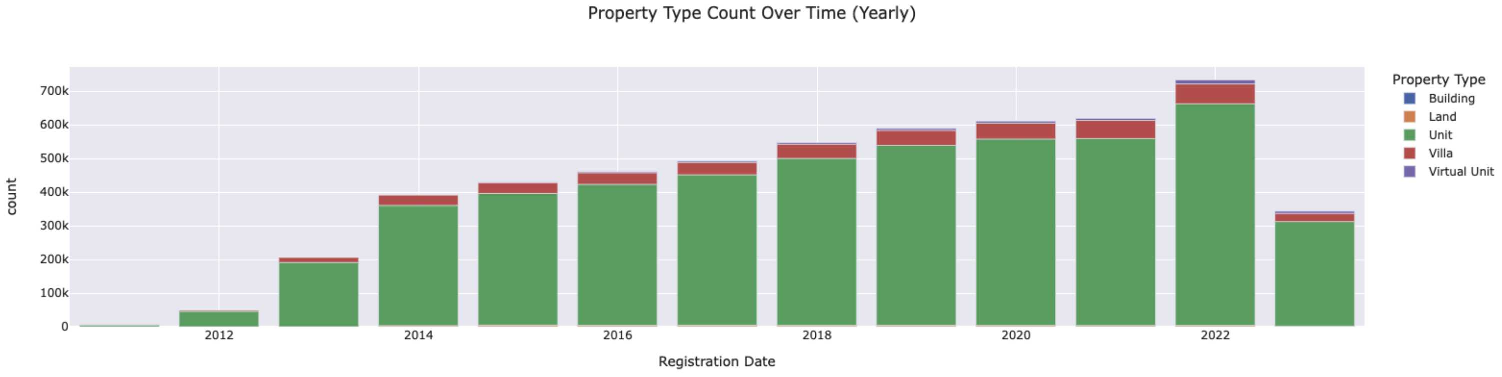


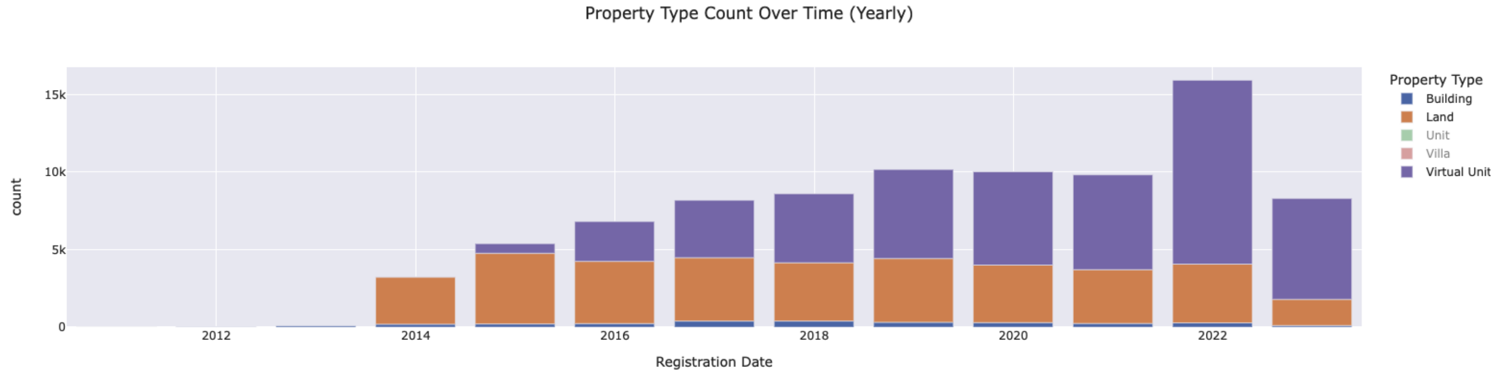




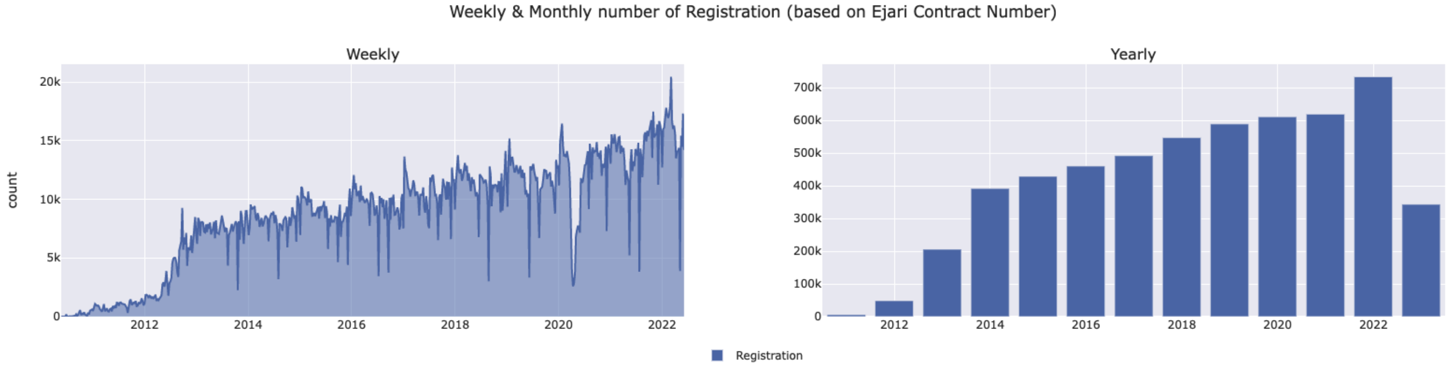
As for property types, our data predominantly featured Units (90%) and Villas (7%). However, a noteworthy trend is the significant increase in the number of virtual units, demonstrating the digital transformation influencing the real estate industry.



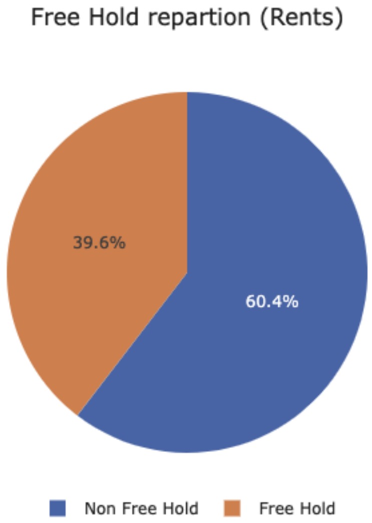


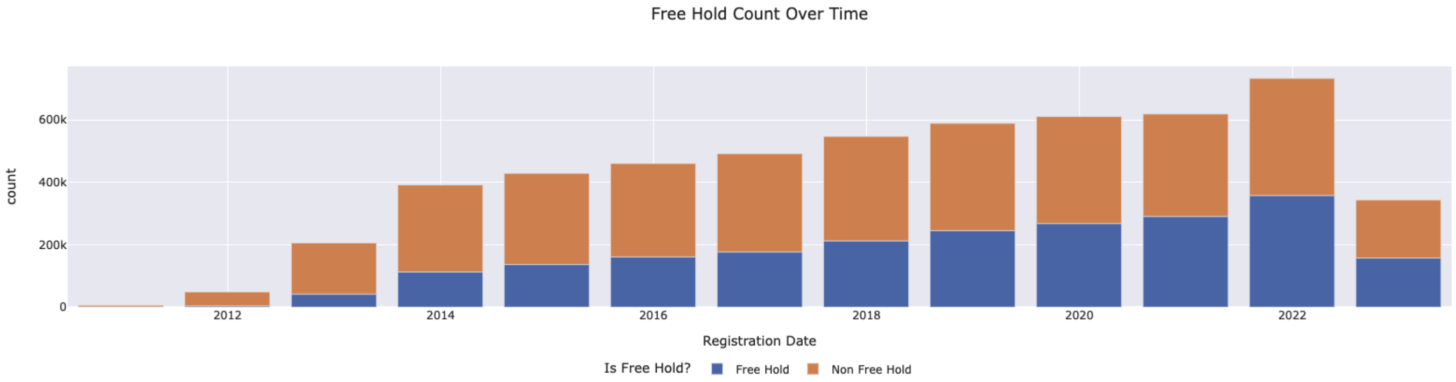


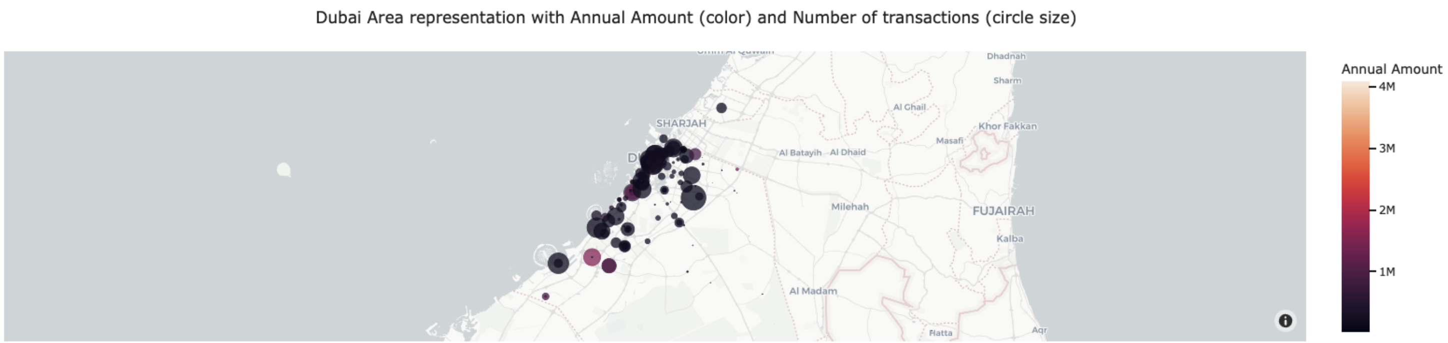
The number of registrations showed a notable spike in 2014, only to experience a dip in 2020, likely attributable to the COVID-19 crisis. Despite this setback, the overall trajectory of property registrations has been positive and constantly ascending, hinting at a robust market.



Looking at the concept of Freehold properties, the data revealed a balanced split, with Non-Freehold properties forming the majority (60%), while Freehold properties constituted 40%. A rise in Freehold registrations was observed starting from 2014. This could possibly be linked to regulatory changes introduced in Dubai's property laws during that period, broadening the opportunities for expatriates to own properties and significantly impacting the dynamics of the market.



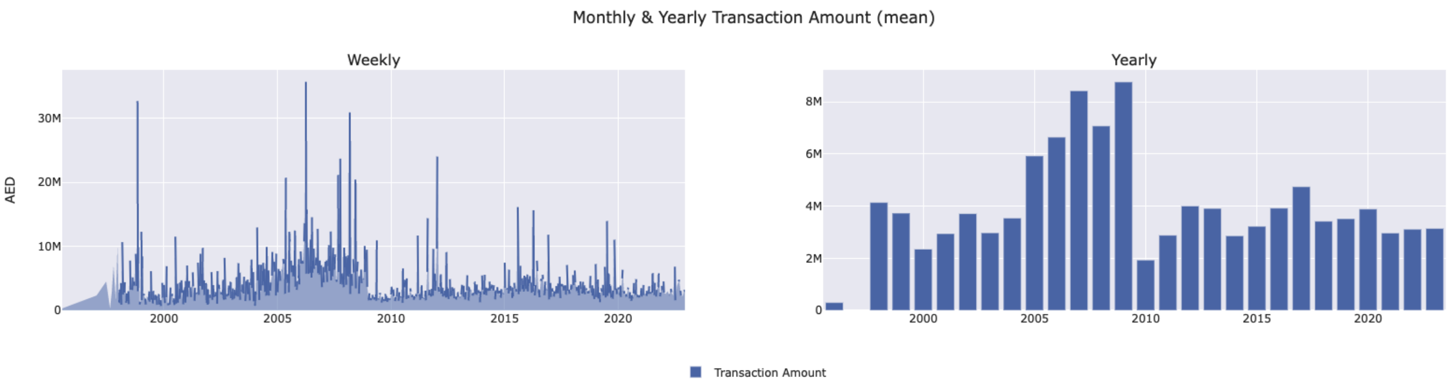




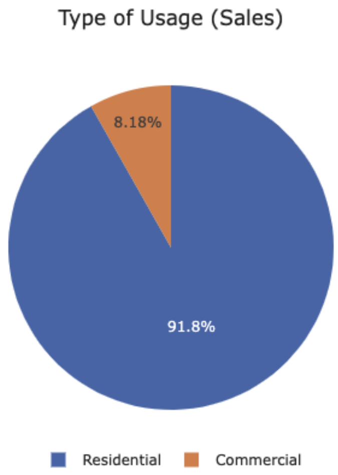
Sales

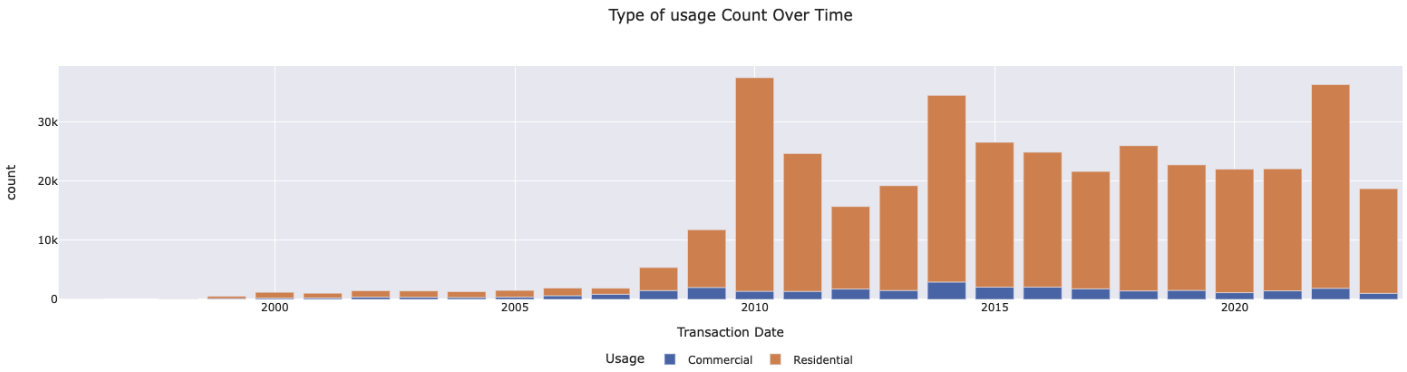
Transitioning our exploration from rentals to property sales, we attempted to unearth trends, patterns, and correlations across various dimensions in the dataset.

The volume of transactions presents a cyclical pattern, which is a common trait in real estate markets reflecting periods of expansion and contraction.

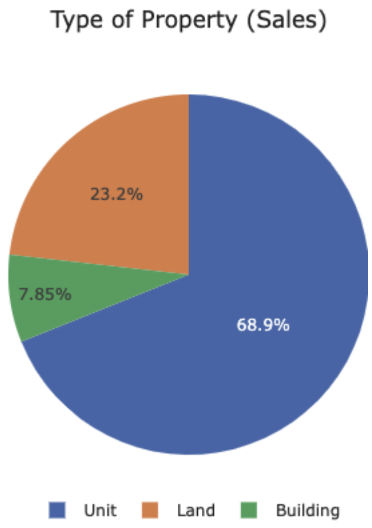


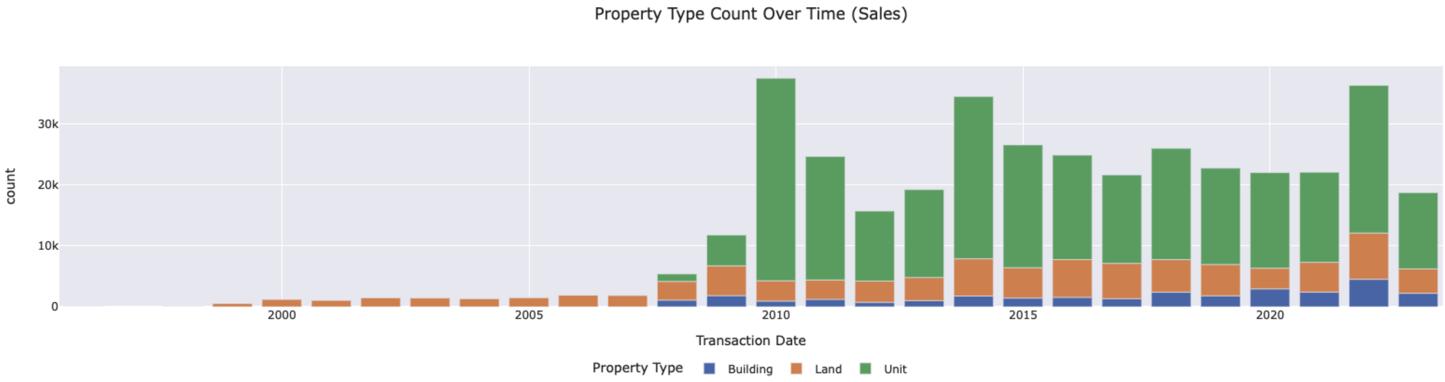
Analyzing the property type, our data primarily encompassed Units (69%), followed by Land (23%), and Buildings (8%). A historical view revealed that until 2007, land was the principal commodity in sales. Subsequently, we observed an influx of units and buildings entering the market.



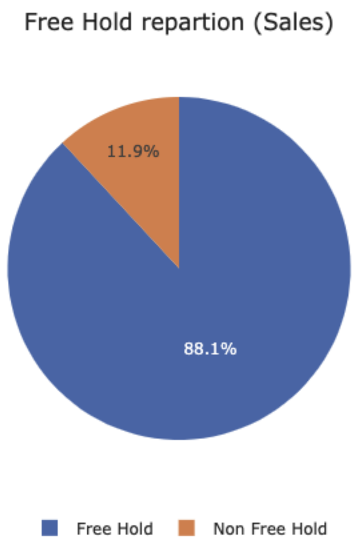


Breaking down the usage of properties, an overwhelming majority was residential (92%), with commercial properties forming a smaller fraction (8%). In terms of sales amount, the market was thriving until 2009, when a steep fall occurred due to the global financial crisis. However, the Dubai real estate market has since recovered and stabilized, with sales hovering around 3 to 3.5 million AED.

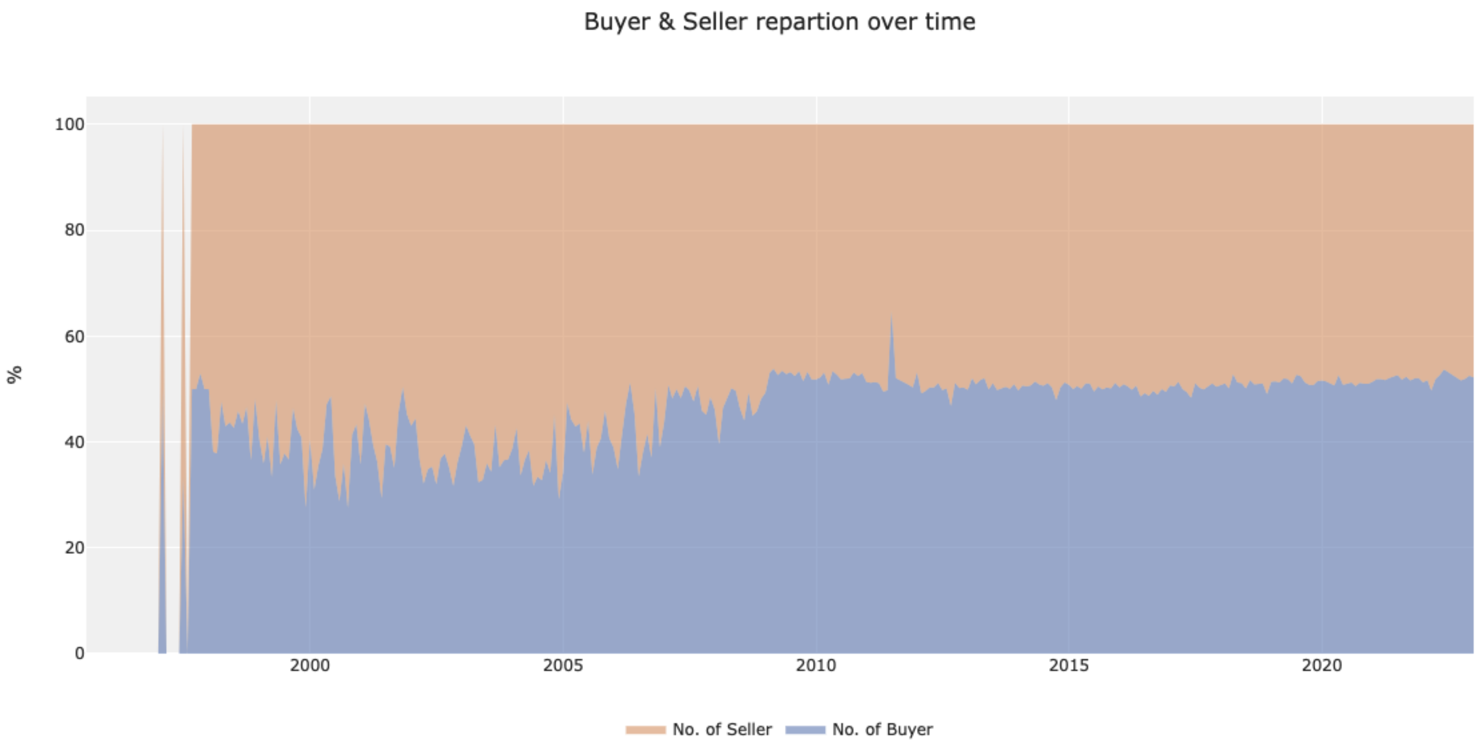




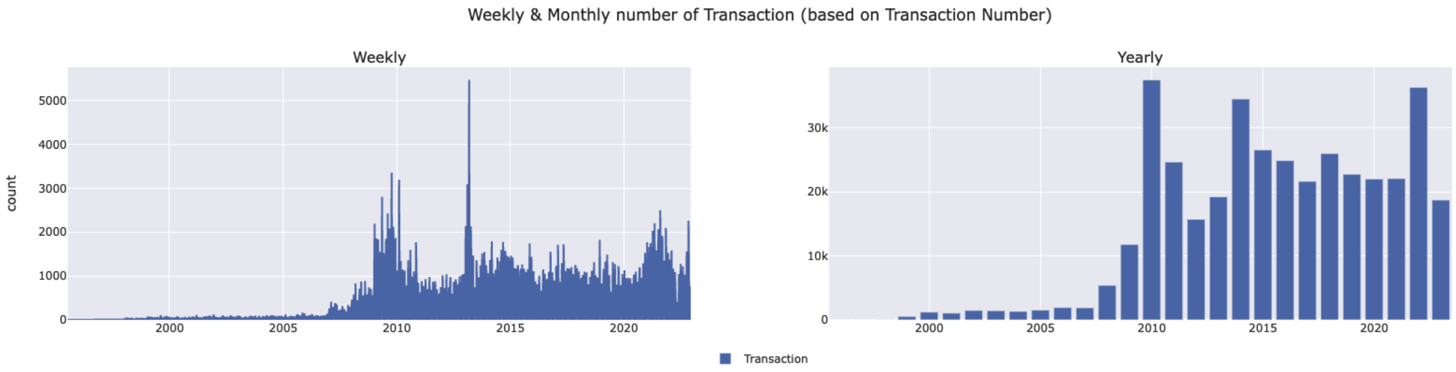
A striking contrast was found when comparing the Freehold status between rentals and sales. Sales had a much higher proportion of Freehold properties at 88%, leaving Non-Freehold properties at a mere 12%. This dominance of Freehold property sales emerged post-2008, which could likely be a result of revised laws that made it easier for expatriates to purchase properties.

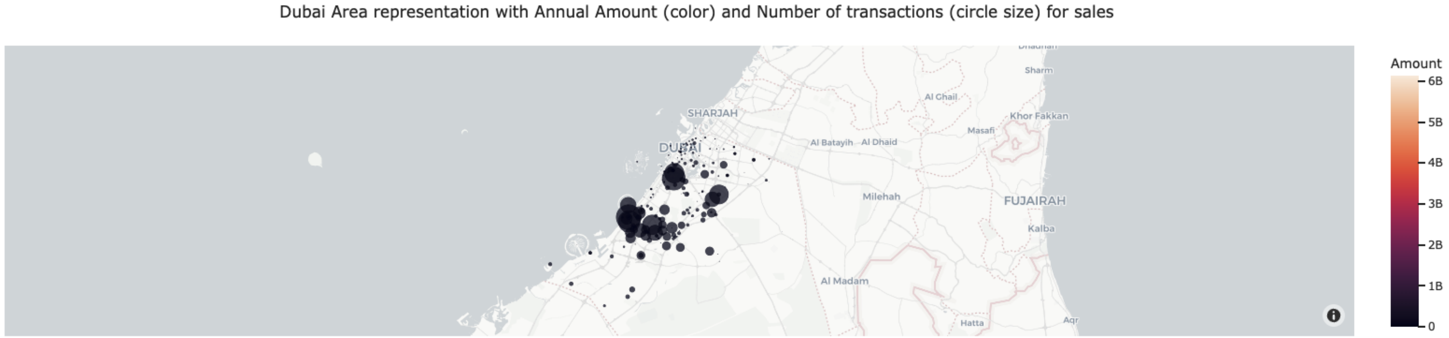


Another interesting trend is the shift in the buyer-seller dynamics. Until 2009, there was a buyer-seller imbalance, with more sellers (around 60%) than buyers (40%), likely due to price volatility and elevated prices. In contrast, the current scenario depicts a balanced market, with an equal number of sellers and buyers, indicating stability and maturity in the market.



The number of registrations showed a notable spike in 2014, only to experience a dip in 2020, likely attributable to the COVID-19 crisis. Despite this setback, the overall trajectory of property registrations has been positive and constantly ascending, hinting at a robust market.



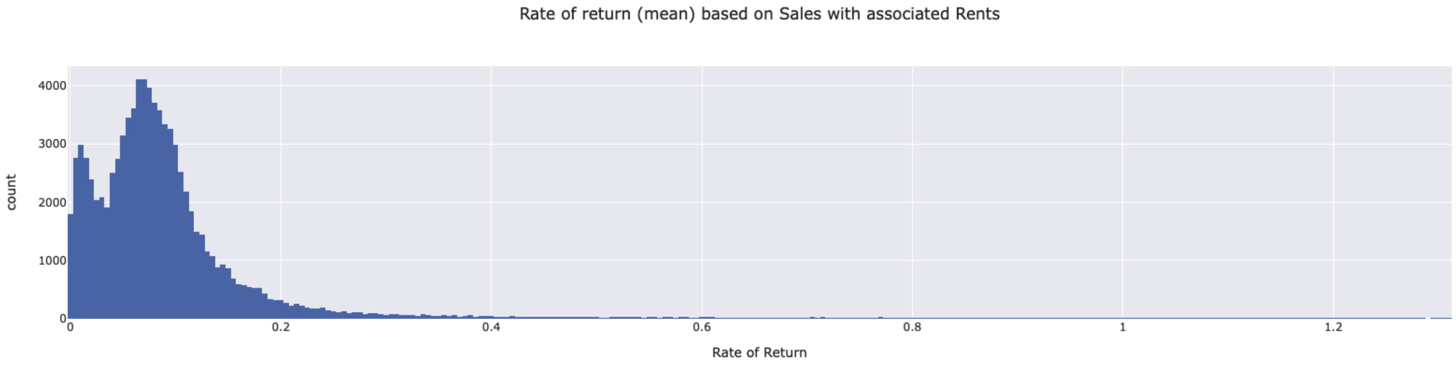


In our analysis of real estate properties, we employed an approach to calculate and analyze the rate of return, defined as the annual rental income divided by the property sale price. This key metric allows us to quantify the performance of properties from an investment perspective.

Our process began by merging two datasets, one representing rental contracts and the other property transactions. Using the merge_asof function in pandas, we were able to align each property sale with the most recent corresponding rental agreement based on their dates. This ensured a time-consistent and accurate comparison.

Subsequently, we calculated the rate of return for each property and investigated its distribution and relationship with other variables. The rate of return predominantly ranged between 0.06 (6%) and 0.08 (8%), indicating a decent return on investment for most properties. This is within a reasonable expectation for the real estate market in Dubai.

However, we also identified a few outliers in the data with abnormally high rates of return. These cases were examined in detail to identify potential data errors or anomalies. After validation, certain outliers that were concluded to be data entry errors were removed from the dataset to prevent skewing of our analysis.



0.1.3 Surface area (m^2) and sale/rental price corr

Our analysis of the correlation between the surface area (m^2) and the rental/sale price in the Dubai real estate market reveals interesting insights. Generally, we would expect a stronger correlation between property size and price - the larger the property, the higher the price. However, our calculations show a weak correlation when examining individual transactions, both for rentals and sales, with values of 0.09 and 0.35 respectively. This suggests that factors beyond surface area are influencing price to a significant degree when we look at transactions in isolation.

Nevertheless, the correlation becomes notably stronger when transactions are grouped by different factors such as Area, Ejari Contract Number, Property Type, Usage, and Project for rental prices and Area and Project for sales prices. For rental transactions, the Property Type shows the strongest correlation of 0.756, implying that the type of the property (Building, Land, Unit, Villa, Virtual Unit) significantly influences how the size of the property correlates with the rental price. Project, or the name of the real estate project, also shows a fairly strong correlation (0.37), indicating that the specific project a property is a part of can impact how size relates to price.

On the other hand, for sales transactions, the Project and Area have correlations of 0.55 and 0.46, respectively. This indicates that where a property is located and which project it is part of, play a significant role in how the property's size affects its sale price.

Interestingly, the Usage factor shows a negative correlation (-0.18) with rental prices. This suggests that the purpose of the property (Commercial, Industrial, Residential) inversely affects the correlation between the size and rental price - perhaps larger commercial or industrial properties aren't rented at a proportionally higher price compared to smaller ones.

These findings underscore the complexity of real estate pricing and the role various factors play in determining price. Property size is certainly an influential factor, but its impact is often modulated by aspects like location (Area), specific project, property type, and usage.

Grouping Factor	Rental Price Correlation	Sale Price Correlation
Individual Transactions	0.092	0.354
Area	0.246	0.456
Ejari Contract Number	0.209	-
Property Type	0.756	-
Usage	-0.180	-
Project	0.367	0.550

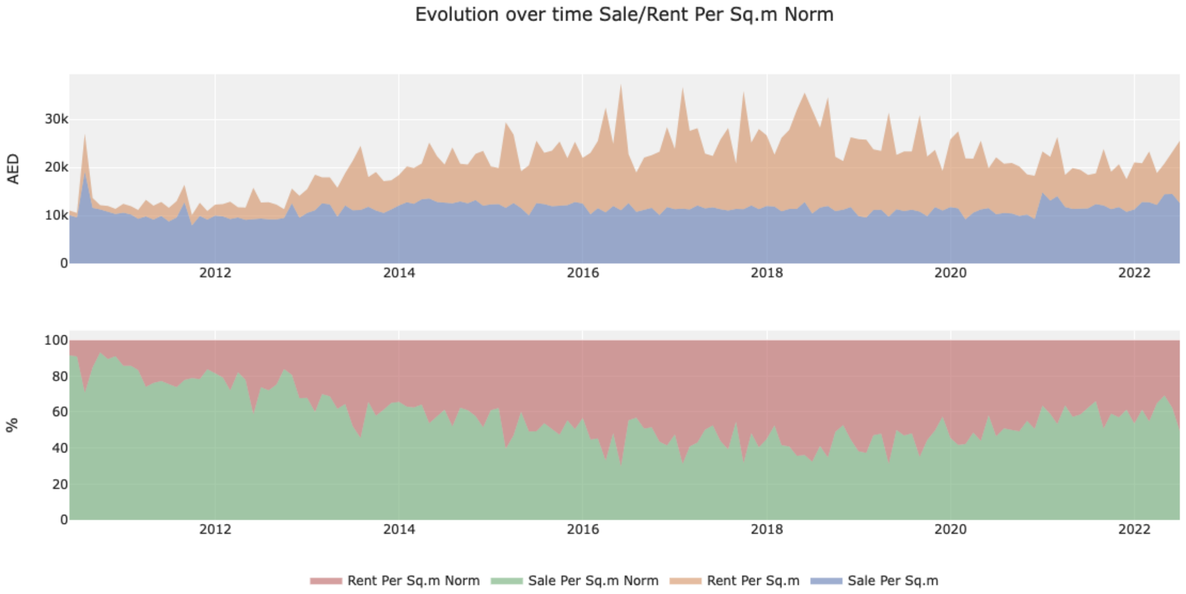
Table 1: Correlations between surface area and rental/sale prices grouped by different factors

0.1.4 Rental and Sales real estate markets comparison

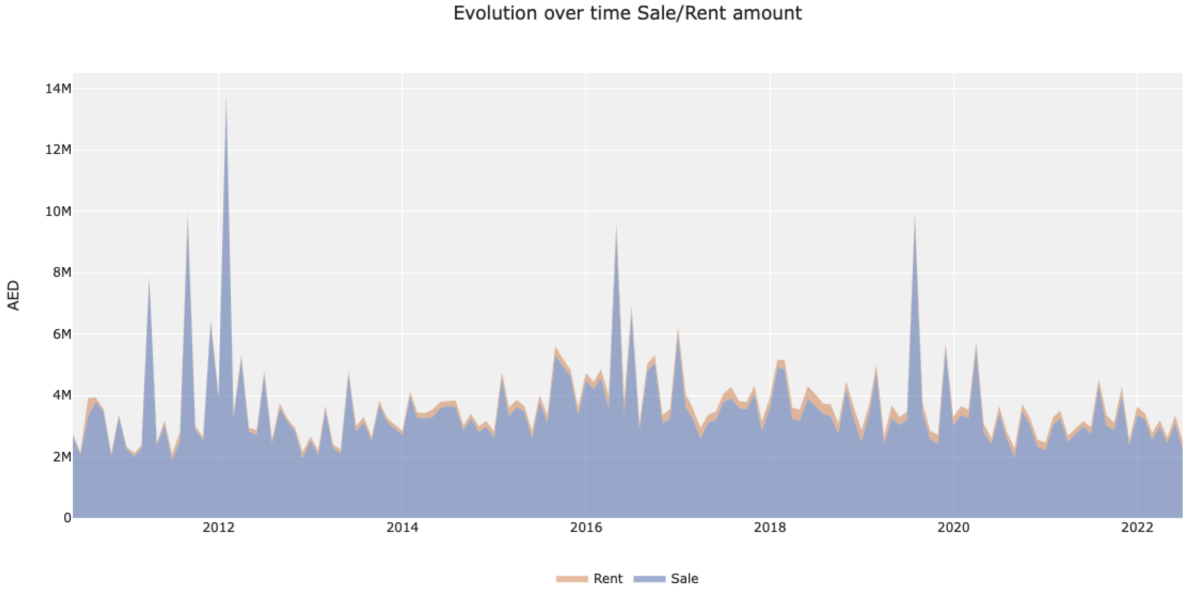
In the course of our study of Dubai's real estate landscape, we've delved into the nuanced complexities of both the rental and sales markets. With a meticulous comparison of these two distinct segments, we observed a array of trends, correlations, and contrasts.

A pivotal area of investigation was the relationship between sales and rental prices. Our analysis indicated a moderate positive correlation of 0.21 between Sale Per Sq.m and

Rent Per Sq.m. This suggests that as the sales price per square meter tends to increase, the rent per square meter often follows a similar upward trend, albeit not as strongly. This could be attributed to the fact that both values are typically driven by similar market forces, such as location, amenities, and property type.



However, when it comes to correlating the overall Sale and Rent amounts, the relationship was significantly weaker, with a correlation of only 0.005. This divergence could be a result of various factors that affect sale prices and rental prices differently. For example, the decision to buy is usually influenced by long-term considerations, such as potential for appreciation, mortgage rates, and changes in real estate laws. Conversely, rentals are more sensitive to short-term factors like shifts in population, employment rates, and immediate demand-supply dynamics.



Furthermore, the Freehold property status, which allows owners full and indefinite ownership, exhibited contrasting trends in the two markets. While Freehold properties made up a significant portion of sales (88%), they were less prominent in the rental market (40%).

This disparity likely reflects policy changes post-2008 that made it easier for expatriates to buy properties in Dubai, encouraging more sales of Freehold properties.

0.1.5 Impacting factors on rents over time

In the analysis of the factors impacting the sale/rental price of properties in Dubai over time, the two most consistently significant factors were found to be ’feature_property_sub_type’ and ’feature_units’. These factors remained as the top two most important features from 2013 to 2022, although their relative importances changed.

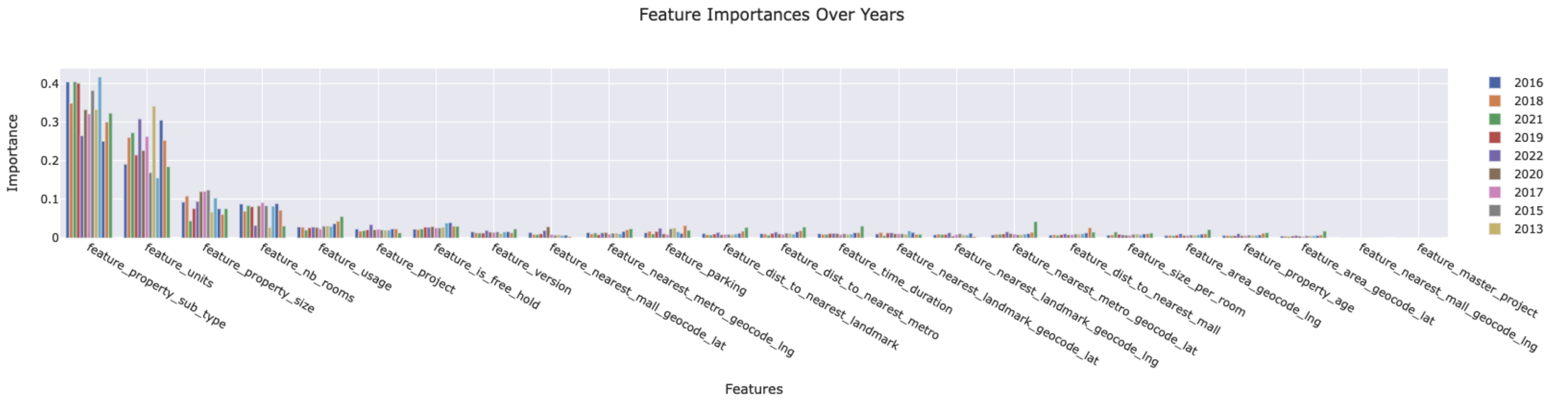
The ’feature_property_sub_type’ had the highest impact on property prices in 2013, with an importance value of 0.3818, but saw a slight decrease over the years. However, it still remained one of the most significant factors, showcasing an importance value of 0.2648 in 2022.

The factor ’feature_units’ was ranked as the second most important feature in 2013 with an importance value of 0.1693. Over the years, its significance has been increasing, and by 2022, it held the most influential position with an importance value of 0.3085.

’Feature_property_size’ and ’feature_nb_rooms’ also played a crucial role in determining the property prices. Although their relative importance varied slightly over the years, they remained consistently in the top four positions from 2015 to 2022. It suggests that larger properties with more rooms tend to command higher prices in Dubai’s real estate market.

Interestingly, we also noticed a marked increase in the importance of ’feature_project’ between 2015 and 2022, growing from an importance value of 0.0209 to 0.0346, indicating that the specific project the property is part of has become increasingly significant in determining its price.

These results suggest that over time, the nature of the property (its type and number of units), its size and the number of rooms, and the project it is a part of, have become increasingly significant in driving the sale/rental prices of properties in Dubai.



0.2 Prediction Models

0.2.1 Data Cleaning

Prior to the development of the machine learning models, additional data preprocessing was carried out to enhance the quality and interpretability of the datasets. The 'Parking' feature was transformed into a binary variable, with the presence of a 'NaN' value interpreted as the absence of parking, thus being replaced with 'False', while properties with parking were indicated with 'True'. Moreover, any missing values in the 'Number of Rooms' column were filled with '0', under the assumption that 'NaN' could denote studio properties without specific rooms.

Further manipulation was conducted to manage missing data in the 'Property Size (sq.m)' and 'Number of Rooms' columns. A more sophisticated approach was taken here, where missing values were filled with the mean value of similar properties. This was achieved by grouping the properties based on 'Property ID', 'Property Type', and 'Property Sub Type', and then applying a lambda function to replace 'NaN' values with the mean size or room number of that group. The Python code for this operation used the Pandas 'groupby' and 'transform' functions, with 'fillna(x.mean())' used to replace the missing values with the group mean.

Lastly, for the 'Master Project' and 'Project' columns, any 'NaN' values were replaced with 'None'. The reasoning behind this choice was that not all properties are associated with a specific project, so 'NaN' values in these cases were treated as meaningful and represented as 'None'.

0.2.2 Feature Engineering

Feature creation

As part of our advanced data manipulation process, we introduced three new features into our dataset to enrich its context and enhance the predictive performance of our machine learning models.

The first derived feature, "Rent Time," was obtained by calculating the difference between the end and start dates of the rental contracts. This provided a numerical measure of the duration of rent, allowing us to potentially identify correlations between rental duration and rental price, and thus contribute to our model's predictive power.

Another critical aspect of real estate pricing lies in the location of the property. To capture this information, geocoding was employed to convert the categorical location data from the 'Area', 'Nearest Metro', 'Nearest Mall', and 'Nearest Landmark' features into numerical data that could be used by the machine learning models. The Google API was utilized to obtain longitude and latitude values corresponding to these location descriptions. This transformation resulted in a more granular and continuous representation of location, which is advantageous for machine learning algorithms as it allows them to find more intricate relationships and make more accurate predictions. Moreover, we computed the distance between the Area of the property and Nearest Metro, Nearest Mall, and Nearest Landmark giving other potential information to the model.

The final feature created, 'Property Size Per Room' represents the average size of each room in a property, derived by dividing the total property size by the number of rooms. This feature provides nuanced insights into the spatial distribution within a property, potentially enhancing the model's accuracy by differentiating properties not only by total size but also by room spaciousness.

Feature encoding

To prepare the dataset for machine learning modeling, we needed to encode the categorical variables to convert them into a format that can be interpreted by the algorithms. Given that most machine learning models cannot operate directly on categorical data, this step was crucial to enable meaningful analysis and accurate predictions.

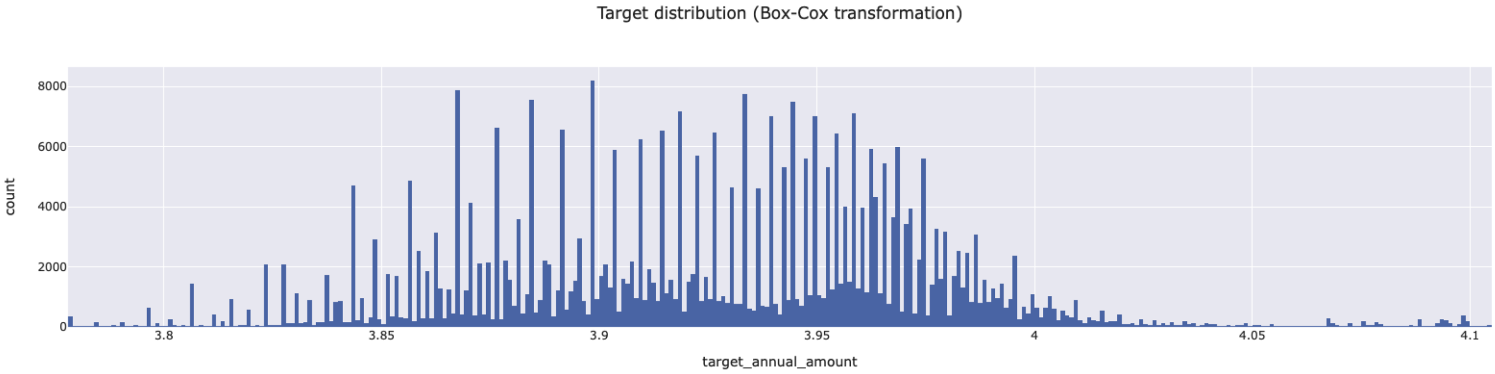
First, the 'Property Sub Type', 'Usage', 'Master Project', and 'Project' columns, which had more than 10 unique values, were processed using the LabelEncoder method. LabelEncoder transforms each unique category in these features into an integer. This approach was beneficial as it retained the necessary categorical information while converting it into a more digestible numerical format for the machine learning model. However, it should be noted that LabelEncoder implies an order in the categories, which may not always be accurate or beneficial, especially when dealing with nominal data.

For the 'Property Type' feature, which had 5 unique values, we opted for the pandas get_dummies method. This method creates a new binary column for each category in the feature, thereby eliminating any potential ordinal relationship that does not exist in the actual data. This technique, also known as one-hot encoding, is particularly useful for handling categorical variables with a reasonably low number of unique categories.

Lastly, for features with only two unique values, such as 'Parking', 'Version', and 'Free Hold', we encoded them simply as 0 and 1. This binary encoding method is straightforward and efficient for representing dichotomous data in a machine learning-friendly manner.

These encoding strategies were carefully selected based on the nature and characteristics of each feature, ensuring that our dataset was well-optimized for the ensuing regression model training.

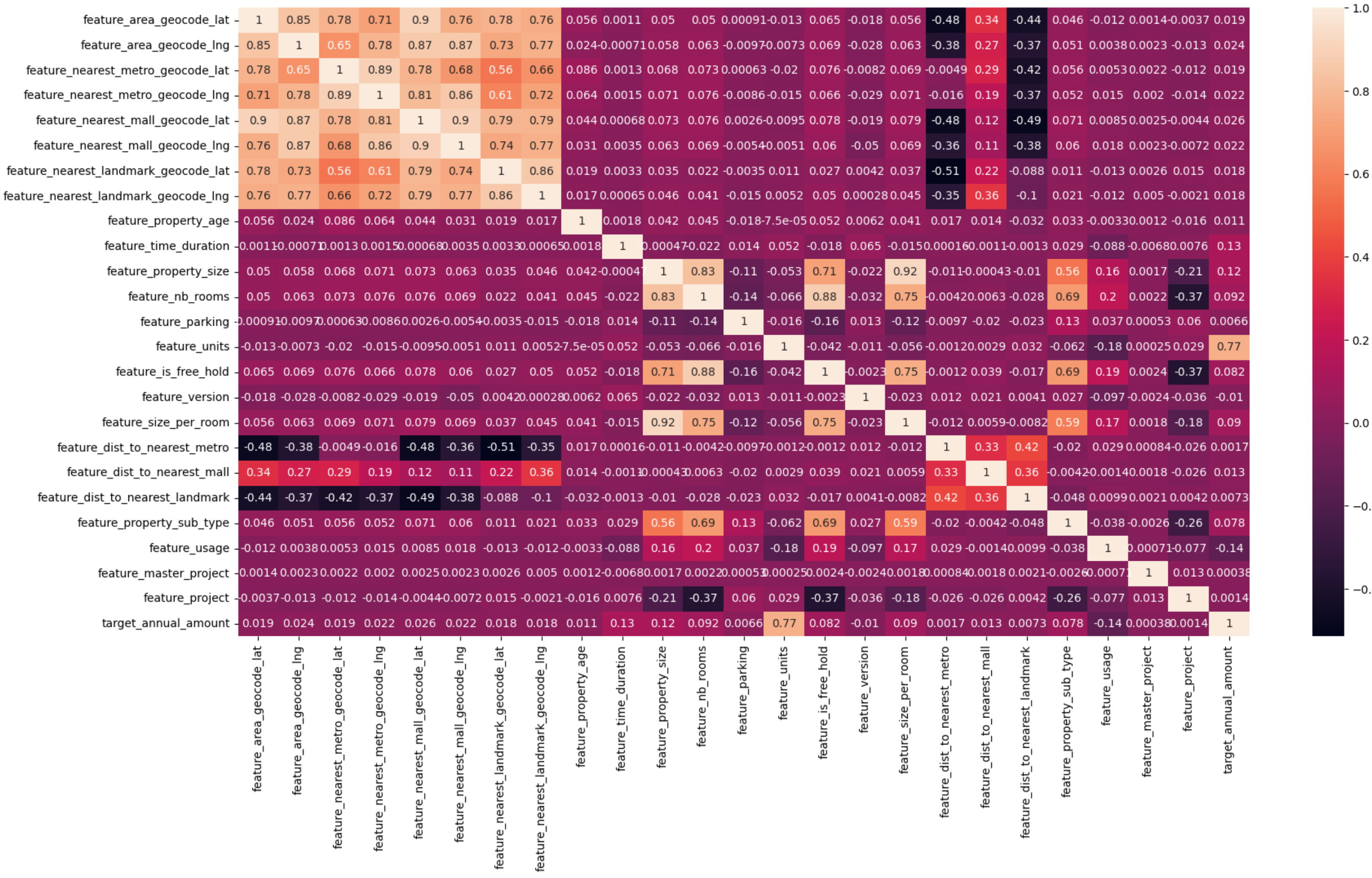
In addition to encoding the categorical features, we applied a Box-Cox transformation to our target variable, 'Annual Amount'. This transformation was adopted as an essential preprocessing step to adjust the distribution of this variable, aiming to shape it into a more Gaussian or 'normal' distribution.



The Box-Cox transformation, which requires input data to be positive, is a power transform method that can effectively stabilize variance and make the data more closely align with the key assumptions of linear regression models. In our case, this transformation was especially advantageous as the distribution of 'Annual Amount' might have been skewed, with possible presence of outliers or heavy tails.

Normalizing the target variable distribution can result in improved model performance, because many machine learning algorithms, including regression-based models like LightGBM, assume that the input data is normally distributed. This assumption, if met, can help the algorithm to learn the underlying patterns more effectively and to produce more accurate predictions.

After all this steps, you have all the features to train our Regression model (while ensuring that there is no price leakage in the model's target):



	feature_area_geocode_lat	feature_area_geocode_lng	feature_nearest_metro_geocode_lat	feature_nearest_metro_geocode_lng	feature_nearest_mall_geocode_lat	feature_nearest_mall_geocode_lng	feature_nearest_landmark_geocode_lat	feature_nearest_landmark_geocode_lng	feature_property_age	feature_time_duration	feature_property_size	feature_nb_rooms	feature_parking	feature_units	feature_is_free_hold	feature_version	feature_size_per_room	feature_dist_to_nearest_metro	feature_dist_to_nearest_mall	feature_dist_to_nearest_landmark	feature_property_sub_type	feature_usage	feature_master_project	feature_project	target_annual_amount
feature_area_geocode_lat	1	0.85	0.78	0.71	0.9	0.76	0.78	0.76	0.056	0.0011	0.05	0.05	0.00091	-0.013	0.065	-0.018	0.056	-0.48	0.34	-0.44	0.046	-0.012	0.0014	-0.0037	0.019
feature_area_geocode_lng	0.85	1	0.65	0.78	0.87	0.87	0.73	0.77	0.024	-0.00071	0.058	0.063	-0.0097	-0.0073	0.069	-0.028	0.063	-0.38	0.27	-0.37	0.051	0.0038	0.0023	-0.013	0.024
feature_nearest_metro_geocode_lat	0.78	0.65	1	0.89	0.78	0.68	0.56	0.66	0.086	0.0013	0.068	0.073	0.00063	-0.02	0.076	-0.0082	0.069	-0.0049	0.29	-0.42	0.056	0.0053	0.0022	-0.012	0.019
feature_nearest_metro_geocode_lng	0.71	0.78	0.89	1	0.81	0.86	0.61	0.72	0.064	0.0015	0.071	0.076	-0.0086	-0.015	0.066	-0.029	0.071	-0.016	0.19	-0.37	0.052	0.015	0.002	-0.014	0.022
feature_nearest_mall_geocode_lat	0.9	0.87	0.78	0.81	1	0.9	0.79	0.79	0.044	0.00068	0.073	0.076	0.0026	-0.0095	0.078	-0.019	0.079	-0.48	0.12	-0.49	0.071	0.0085	0.0025	-0.0044	0.026
feature_nearest_mall_geocode_lng	0.76	0.87	0.68	0.86	0.9	1	0.74	0.77	0.031	0.0035	0.063	0.069	-0.0054	-0.0051	0.06	-0.05	0.069	-0.36	0.11	-0.38	0.06	0.018	0.0023	-0.0072	0.022
feature_nearest_landmark_geocode_lat	0.78	0.73	0.56	0.61	0.79	0.74	1	0.86	0.019	0.0033	0.035	0.022	-0.0035	0.011	0.027	0.0042	0.037	-0.51	0.22	-0.088	0.011	-0.013	0.0026	0.015	0.018
feature_nearest_landmark_geocode_lng	0.76	0.77	0.66	0.72	0.79	0.77	0.86	1	0.017	0.00065	0.046	0.041	-0.015	0.0052	0.05	0.00028	0.045	-0.35	0.36	-0.1	0.021	-0.012	0.005	-0.0021	0.018
feature_property_age	0.056	0.024	0.086	0.064	0.044	0.031	0.019	0.017	1	0.0018	0.042	0.045	-0.018	-7.5e-05	0.052	0.0062	0.041	0.017	0.014	-0.032	0.033	-0.0033	0.0012	-0.016	0.011
feature_time_duration	0.0011	-0.00071	0.0013	0.0015	0.00068	0.0035	0.0033	0.00065	0.0018	1	-0.00047	-0.022	0.014	0.052	-0.018	0.065	-0.015	0.00016	-0.0011	-0.0013	0.029	-0.088	-0.0068	0.0076	0.13
feature_property_size	0.05	0.058	0.068	0.071	0.073	0.063	0.035	0.046	0.042	-0.00047	1	0.83	-0.11	-0.053	0.71	-0.022	0.92	-0.011	-0.00043	-0.01	0.56	0.16	0.0017	-0.21	0.12
feature_nb_rooms	0.05	0.063	0.073	0.076	0.076	0.069	0.022	0.041	0.045	-0.022	0.83	1	-0.14	-0.066	0.88	-0.032	0.75	-0.0042	0.0063	-0.028	0.69	0.2	0.0022	-0.37	0.092
feature_parking	0.00091	-0.0097	0.00063	-0.0086	0.0026	-0.0054	-0.0035	-0.015	-0.018	0.014	-0.11	-0.14	1	-0.016	-0.16	0.013	-0.12	-0.0097	-0.02	-0.023	0.13	0.037	0.00053	0.06	0.0066
feature_units	-0.013	-0.0073	-0.02	-0.015	-0.0095	-0.0051	0.011	0.0052	-7.5e-05	0.052	-0.053	-0.066	-0.016	1	-0.042	-0.011	-0.056	-0.0012	0.0029	0.032	-0.062	-0.18	0.00025	0.029	0.77
feature_is_free_hold	0.065	0.069	0.076	0.066	0.078	0.06	0.027	0.05	0.052	-0.018	0.71	0.88	-0.16	-0.042	1	-0.0023	0.75	-0.0012	0.039	-0.017	0.69	0.19	0.0024	-0.37	0.082
feature_version	-0.018	-0.028	-0.0082	-0.029	-0.019	-0.05	0.0042	0.00028	0.0062	0.065	-0.022	-0.032	0.013	-0.011	-0.0023	1	-0.023	0.012	0.021	0.0041	0.027	-0.097	-0.0024	-0.036	-0.01
feature_size_per_room	0.056	0.063	0.069	0.071	0.079	0.069	0.037	0.045	0.041	-0.015	0.92	0.75	-0.12	-0.056	0.75	-0.023	1	-0.012	0.0059	-0.0082	0.59	0.17	0.0018	-0.18	0.09
feature_dist_to_nearest_metro	-0.48	-0.38	-0.0049	-0.016	-0.48	-0.36	-0.51	-0.35	0.017	0.00016	-0.011	-0.0042	-0.0097	-0.0012	-0.0012	0.012	-0.012	1	0.33	0.42	-0.02	0.029	0.00084	-0.026	0.0017
feature_dist_to_nearest_mall	0.34	0.27	0.29	0.19	0.12	0.11	0.22	0.36	0.014	-0.0011	-0.00043	0.0063	-0.02	0.0029	0.039	0.021	0.0059	0.33	1	0.36	-0.0042	-0.0014	0.0018	-0.026	0.013
feature_dist_to_nearest_landmark	-0.44	-0.37	-0.42	-0.37	-0.49	-0.38	-0.088	-0.1	-0.032	-0.0013	-0.01	-0.028	-0.023	0.032	-0.017	0.0041	-0.0082	0.42	0.36	1	-0.048	0.0099	0.0021	0.0042	0.0073
feature_property_sub_type	0.046	0.051	0.056	0.052	0.071	0.06	0.011	0.021	0.033	0.029	0.56	0.69	0.13	-0.062	0.69	0.027	0.59	-0.02	-0.0042	-0.048	1	-0.038	-0.0026	-0.26	0.078
feature_usage	-0.012	0.0038	0.0053	0.015	0.0085	0.018	-0.013	-0.012	-0.0033	-0.088	0.16	0.2	0.037	-0.18	0.19	-0.097	0.17	0.029	-0.0014	0.0099	-0.038	1	-0.00071	-0.077	-0.14
feature_master_project	0.0014	0.0023	0.0022	0.002	0.0025	0.0023	0.0026	0.005	0.0012	-0.0068	0.0017	0.0022	0.00053	0.00025	0.0024	-0.0024	0.0018	0.00084	0.0018	0.0021	-0.0026	-0.00071	1	0.013	0.00038
feature_project	-0.0037	-0.013	-0.012	-0.014	-0.0044	-0.0072	0.015	-0.0021	-0.016	0.0076	-0.21	-0.37	0.06	0.029	-0.37	-0.036	-0.18	-0.026	-0.026	0.0042	-0.26	-0.077	0.013	1	0.0014
target_annual_amount	0.019	0.024	0.019	0.022	0.026	0.022	0.018	0.018	0.011	0.13	0.12	0.092	0.0066	0.77	0.082	-0.01	0.09	0.0017	0.013	0.0073	0.078	-0.14	0.00038	0.0014	1

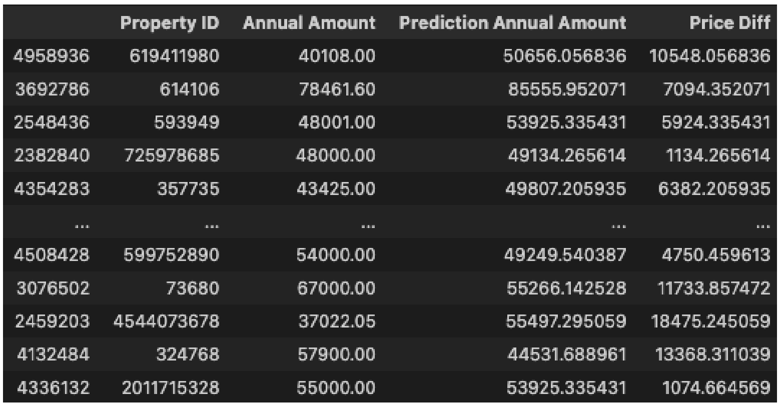
0.2.3 Rental price prediction

In our endeavor to accurately predict the rental price of a property based on its defining characteristics, we employed a machine learning approach using the LightGBM regression model. Our dataset was comprised of diverse property features such as geographical location, proximity to landmarks, property size, the number of units, and more. To account for skewness in the distribution of the rental prices (our target variable), a Box-Cox transformation was applied, rendering the data more suitable for analysis. This transformation is particularly useful for stabilizing the variance and rendering the data as normally distributed as possible, enhancing the accuracy of our model.

Our predictive model was validated using a k-fold cross-validation method with 5 folds, a widely accepted approach for assessing the robustness and generalizability of machine learning models. The scoring metric used was the negative mean squared error, which offers insights into the average squared differences between our predicted and actual values; this score is then square-rooted to bring it back to the same unit as the target variable, giving us the root mean square error (RMSE). Our model achieved an encouraging average RMSE score of 0.0301 across the folds, with a relatively low variance of 0.0035, indicating consistent performance.

The performance of our model is particularly impressive when considering the scale of our target variable, which, after the Box-Cox transformation, ranged between 3.77 and 4.10. The RMSE of 0.0301 represents an average error that is approximately 1% of this range, signifying a high level of accuracy in predicting rental prices.

Here are some predictions of the model:



	Property ID	Annual Amount	Prediction Annual Amount	Price Diff
4958936	619411980	40108.00	50656.056836	10548.056836
3692786	614106	78461.60	85555.952071	7094.352071
2548436	593949	48001.00	53925.335431	5924.335431
2382840	725978685	48000.00	49134.265614	1134.265614
4354283	357735	43425.00	49807.205935	6382.205935
...	...	...	...	...
4508428	599752890	54000.00	49249.540387	4750.459613
3076502	73680	67000.00	55266.142528	11733.857472
2459203	4544073678	37022.05	55497.295059	18475.245059
4132484	324768	57900.00	44531.688961	13368.311039
4336132	2011715328	55000.00	53925.335431	1074.664569

0.2.4 Rental price prediction X months ahead

In our endeavor to forecast the rental price of a property x months into the future, we adopted a two-pronged modeling approach involving a SARIMAX model and a regression model. The primary objective was to capture the temporal dependencies in the data along with the impact of property-specific characteristics on its rental price.

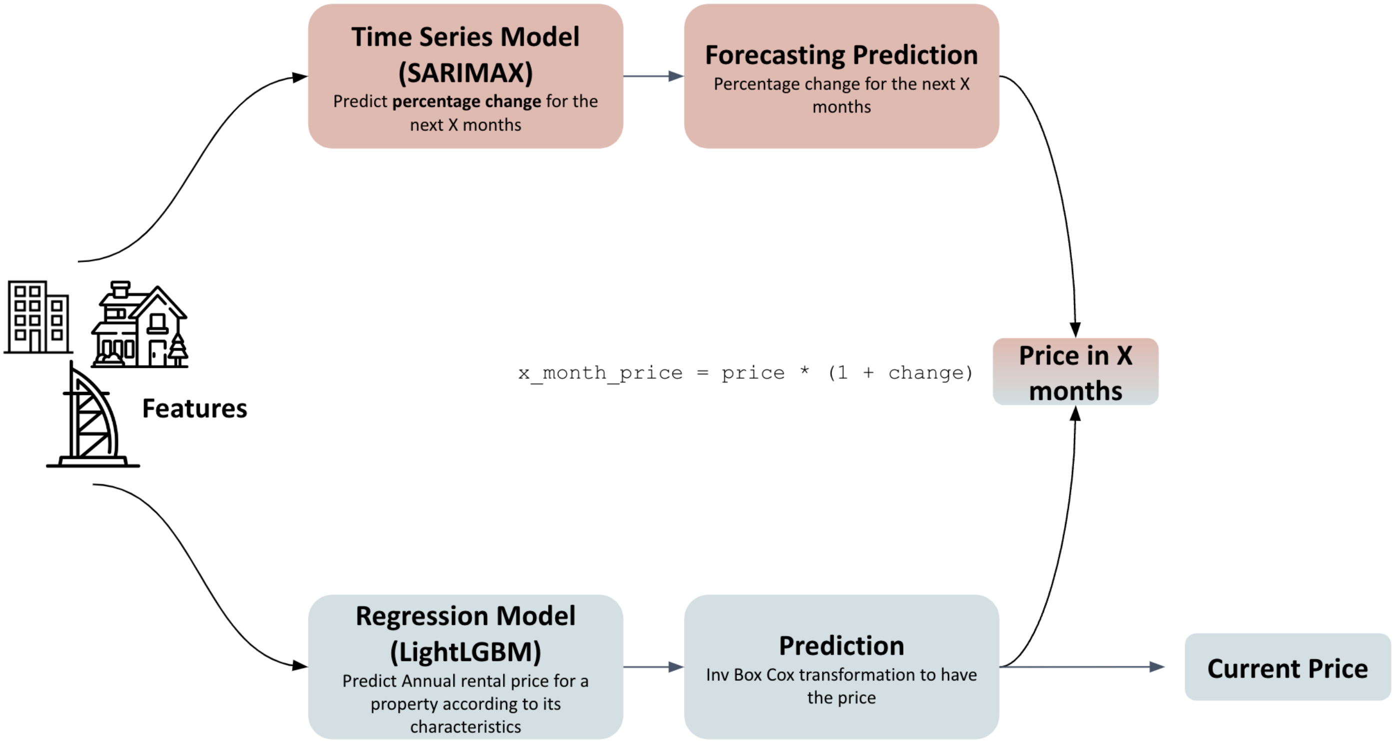
For the time series component, we used a SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables) model, a powerful and versatile time series analysis technique that handles seasonality by integrating differencing, autoregression, and moving average into a composite model. Our target variable for this model was the percentage change in rental price over one month, as this reflects the rate of change in the real estate market, which is often more informative than the absolute price itself.

The choice of target as the percentage change rather than absolute price allows us to capture relative movements and trends in the market, while also making the model more adaptable to different scales of prices. This way, we can quantify monthly market movements in a normalized way that's independent of the specific price level, offering better generalizability across different property price ranges.

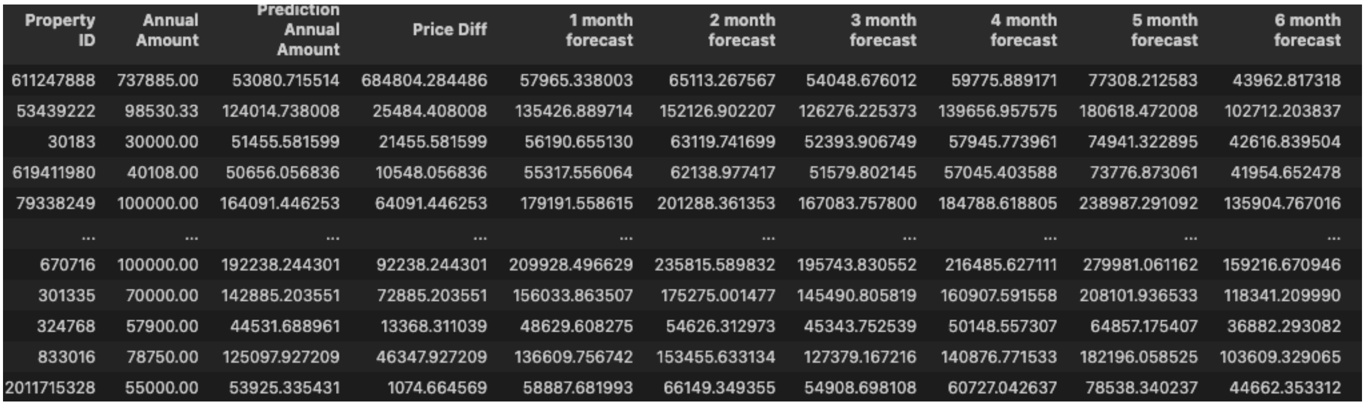
Along with this, the SARIMAX model incorporated exogenous variables that were aggregated on a monthly basis (mean value). These variables—Property Size, Number of Rooms, Units, and Number of Contracts—provided additional information about the real estate market dynamics for each month. Exogenous variables are crucial in our model because they represent external factors that can influence our target variable, providing our model with valuable contextual information that simple time series models might miss.

The output of this SARIMAX model gave us a projected percentage change for the Dubai Real Estate Market for the next x months. These projections were then incorporated into our regression model, which was previously trained to predict property prices based on property-specific attributes. By adjusting our original price predictions with the projected market movements, we managed to generate future predictions that take into account both individual property characteristics and expected market trends.

This combined approach, leveraging both time series and regression modeling, offered a comprehensive solution that integrates micro-level property specifics with macro-level market dynamics, enabling robust and informed predictions of future property prices.


Time Series Model (SARIMAX)
Predict **percentage change** for the next X months
Forecasting Prediction
Percentage change for the next X months
Features
x_month_price = price * (1 + change)
Price in X months
Regression Model (LightLGBM)
Predict Annual rental price for a property according to its characteristics
Prediction
Inv Box Cox transformation to have the price
Current Price

Here are some result on the future price of a property:



Property ID	Annual Amount	Prediction Annual Amount	Price Diff	1 month forecast	2 month forecast	3 month forecast	4 month forecast	5 month forecast	6 month forecast
611247888	737885.00	53080.715514	684804.284486	57965.338003	65113.267567	54048.676012	59775.889171	77308.212583	43962.817318
53439222	98530.33	124014.738008	25484.408008	135426.889714	152126.902207	126276.225373	139656.957575	180618.472008	102712.203837
30183	30000.00	51455.581599	21455.581599	56190.655130	63119.741699	52393.906749	57945.773961	74941.322895	42616.839504
619411980	40108.00	50656.056836	10548.056836	55317.556064	62138.977417	51579.802145	57045.403588	73776.873061	41954.652478
79338249	100000.00	164091.446253	64091.446253	179191.558615	201288.361353	167083.757800	184788.618805	238987.291092	135904.767016
...	...	...	...	...	...	...	...	...	...
670716	100000.00	192238.244301	92238.244301	209928.496629	235815.589832	195743.830552	216485.627111	279981.061162	159216.670946
301335	70000.00	142885.203551	72885.203551	156033.863507	175275.001477	145490.805819	160907.591558	208101.936533	118341.209990
324768	57900.00	44531.688961	13368.311039	48629.608275	54626.312973	45343.752539	50148.557307	64857.175407	36882.293082
833016	78750.00	125097.927209	46347.927209	136609.756742	153455.633134	127379.167216	140876.771533	182196.058525	103609.329065
2011715328	55000.00	53925.335431	1074.664569	58887.681993	66149.349355	54908.698108	60727.042637	78538.340237	44662.353312

Conclusion

In conclusion, this comprehensive study employed advanced data analytics and machine learning to understand and predict trends in the Dubai real estate market. Through a robust data preprocessing strategy, we transformed and enriched our datasets for enhanced interpretability by machine learning algorithms. A LightGBM regression model and a SARIMAX model were effectively integrated, resulting in highly accurate predictions of rental prices and informed forecasts of future property prices. This integrated approach underscores the power of data-driven decision making, combining micro-level property specifics with macro-level market dynamics. The models built offer valuable insights to stakeholders, helping them navigate the complex landscape of real estate pricing. The study also highlights the significant impact of regulatory changes and market maturity on real estate dynamics.

Future advancements could enhance this model by deploying ensemble methods that stack multiple models for better predictive power and incorporating additional influential data sources for a more comprehensive analysis.