Dubai Real Estate Transactions Dataset – Detailed Introduction

Introduction

The real estate sector is one of the most dynamic and influential components of Dubai's economy, playing a central role in shaping the city's urban landscape and attracting global investment. Over the past two decades, Dubai has emerged as a leading hub for property development, offering a wide range of real estate opportunities that cater to both local residents and international investors. Understanding the structure, scale, and patterns of property transactions in Dubai is therefore essential for developers, investors, policymakers, and researchers who aim to gain a clear picture of market trends and future prospects.

The dataset under study provides a comprehensive record of 1.5 million property transactions carried out in Dubai. These transactions span a wide time frame and encompass multiple types of property dealings, including sales, mortgages, and gifts, thus capturing the diverse nature of Dubai's real estate activities. The dataset integrates rich details about each transaction, including property characteristics (such as type, usage, size, and parking availability), geographic attributes (such as area name, project, and nearby landmarks), and financial details (such as actual property worth and price per square meter). In addition, it records the number of parties involved in each transaction, providing insights into the structure of real estate deals.

By combining property features with location-based information, this dataset allows for a variety of analytical tasks. At a descriptive level, it supports exploration of transaction volumes, popular property types, and high-demand locations. At a predictive level, it can be used to estimate property prices based on multiple features, assess the effect of location and amenities on pricing, or forecast emerging real estate trends. From a strategic perspective, it enables researchers and stakeholders to identify investment hotspots, understand the relationship between infrastructure (such as proximity to metro stations or malls) and property value, and evaluate how Dubai's rapid development projects influence the housing market.

Moreover, this dataset is particularly valuable because of its granularity and scale. With over a million entries, it offers a large enough sample size to produce statistically significant results and uncover patterns that may not be visible in smaller samples. It also provides an opportunity to compare property types across different neighborhoods, study the evolution of real estate prices over time, and assess the impact of various economic factors on the market. For urban planners, the dataset sheds light on the distribution of residential and commercial properties, parking facilities, and property sizes across Dubai's neighborhoods.

In summary, this dataset is a rich resource for multi-dimensional analysis of Dubai's real estate sector. It captures the interplay between property characteristics, geographic location, and financial outcomes, providing actionable insights for multiple stakeholders. Whether the goal is to guide investment decisions, shape urban

policies, develop predictive models, or simply understand Dubai's housing market dynamics, this dataset offers a powerful foundation for research and practical applications.

Dataset Size and Structure

- Number of rows (records): ~1,548,772
- Number of columns (features): 24 (before cleaning)
- File size: ~283 MB in CSV format
- Time coverage: Transactions spanning multiple decades, from the late 1990s to recent years

Features Explained

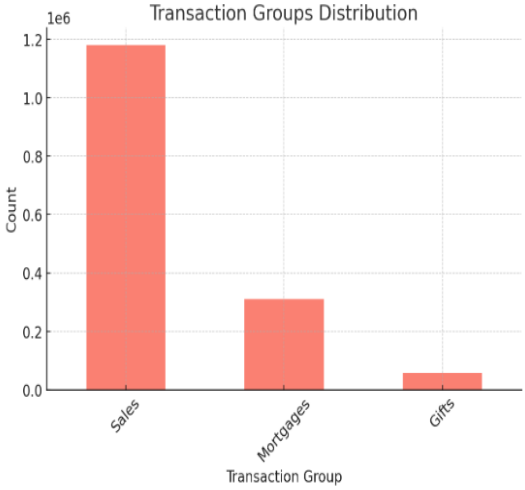
1. transaction_id → Unique identifier of each transaction.
2. trans_group_en → Transaction category (e.g., Sale, Mortgage, Gift).
3. procedure_name_en → Specific procedure (e.g., Grant, Mortgage Registration).
4. instance_date → Date of transaction.
5. property_type_en → High-level property type (e.g., Villa, Apartment, Land).
6. property_sub_type_en → More detailed type (e.g., Townhouse, Office, Retail Shop).
7. property_usage_en → Intended usage (e.g., Residential, Commercial, Mixed-use).
8. reg_type_en → Registration type (e.g., Existing Properties, Off-plan).
9. area_name_en → Geographic area in Dubai (e.g., Jumeirah, Marina).
10. building_name_en → Specific building (if available).
11. project_number → Identifier for the project.
12. project_name_en → Project name (e.g., Burj Khalifa Residences).
13. master_project_en → The master development/project (e.g., Downtown Dubai).
14. nearest_landmark_en → Notable landmark near the property.
15. nearest_metro_en → Closest metro station.
16. nearest_mall_en → Closest shopping mall.
17. rooms_en → Number of rooms (bedrooms).
18. has_parking → 1 if parking is available, 0 if not.
19. procedure_area → Property size in square meters.

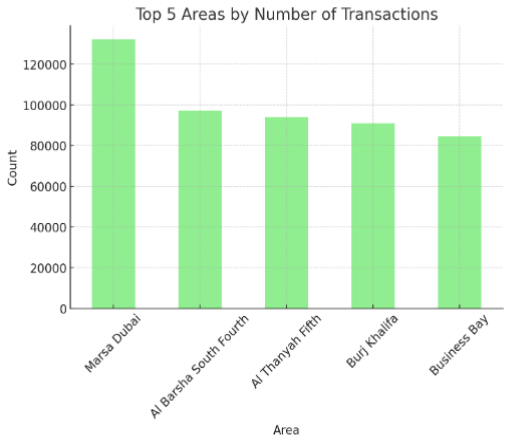
20. actual_worth → Total worth/price of the property in AED.
21. meter_sale_price → Price per square meter (AED/m²).
22. no_of_parties_role_1 → Number of parties in role 1 (e.g., buyers).
23. no_of_parties_role_2 → Number of parties in role 2 (e.g., sellers).
24. no_of_parties_role_3 → Number of parties in role 3 (e.g., agents/notaries).

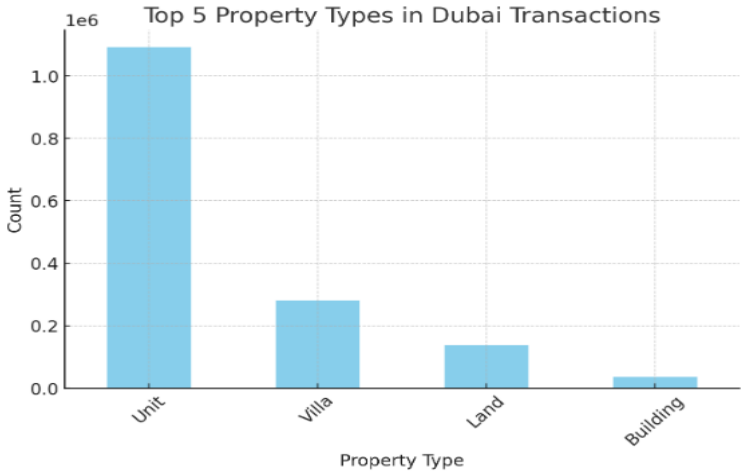
Potential Analytical and Predictive Tasks

1. **Price Prediction:** Building machine learning models to predict the actual worth (actual_worth) of a property using features like property type, location, area, and rooms.
2. **Trend Analysis:** Studying how property prices and transaction volumes have changed over time, across years and months.
3. **Spatial Analysis:** Comparing property prices across different areas of Dubai and analyzing hotspots.
4. **Infrastructure Impact:** Measuring the effect of proximity to metro stations and malls on property values.
5. **Clustering:** Grouping properties into clusters based on type, usage, and value to identify market segments.
6. **Outlier Detection:** Spotting unusually high or low transactions, which may signal luxury properties, undervalued deals, or data entry errors.
7. **Policy and Urban Studies:** Understanding how Dubai's urban planning and mega-projects have shaped property values over time.

visualize this Dataset







a. **Property Types** → Apartments (Units) dominate transactions, followed by Villas and Land.

b. **Top Areas** → Most deals are in **Marsa Dubai, Al Barsha South Fourth, Al Thanyah Fifth, Burj Khalifa, and Business Bay**.

c.**Transaction Groups** → Sales make up the majority, with Mortgages second, and Gifts much smaller.

d.**Parking Availability** → About **65% of properties include parking**, while 35% don't.

e.**Price Distribution** → Most properties are worth between **0.5M – 5M AED**, but a few very high-value transactions push the average up (hence the log scale).

Cleaning the dataset

Here are the features in your dataset that have missing values along with how many are missing:

Feature	Missing Values
nearest_mall_en	458,118
building_name_en	456,145
nearest_metro_en	449,856
project_number	442,675
project_name_en	442,675
rooms_en	342,275
property_sub_type_en	321,622
nearest_landmark_en	283,198
master_project_en	229,171
no_of_parties_role_1	968
no_of_parties_role_2	968
no_of_parties_role_3	968

Data Cleaning Procedure

The dataset contained over **1.5 million records** and 24 features describing real estate transactions in Dubai. Before analysis or modeling, the dataset required a comprehensive cleaning process to handle missing values, outliers, and inconsistencies. Below is a step-by-step procedure of the cleaning we applied, along with the reasoning behind each decision.

---

1. Creating a Clean Copy

The first step was to create a copy of the dataset (`df_clean`) to ensure the original data remained intact. This allows reproducibility and prevents accidental overwriting of the raw dataset.

---

2. Cleaning the `rooms_en` Column

The `rooms_en` column contained text values such as *"Studio"*, *"1BR"*, *"2BR"*.

- We replaced these with numeric equivalents (e.g., *Studio = 0, 1BR = 1*).
- We converted the column to numeric values to standardize the format.
- Missing values were filled with the **median number of rooms within each property type** (`property_type_en`).
    - For example, apartments tend to have fewer rooms than villas, so imputing by property type keeps the values realistic.
- If an entire property type had no room data (e.g., Land), we filled it with `0`.

This ensured that `rooms_en` became a fully numeric and reliable predictor.

```
                    room_mapping = {
  "Studio": 0,
  "1BR": 1,
"2BR": 2,
"3BR": 3,
"4BR": 4,
"5BR": 5
}
df_clean["rooms_en"] = df_clean["rooms_en"].replace(room_mapping)
```

---

3. Cleaning the `property_sub_type_en` Column

This column describes the subtype of the property (e.g., *Penthouse*, *Townhouse*). Many values were missing.

- We grouped the data by `property_type_en` and replaced missing subtypes with the **most frequent subtype (mode)** within each group.
- If no subtype existed at all for a property type, we filled with `"Missing"`.

This approach ensured consistency while maintaining realistic subtype distributions.

```
            df_clean["property_sub_type_en"] =
df_clean.groupby("property_type_en")["property_sub_type_en"].transform(
    lambda x: x.fillna(x.mode()[0] if not x.mode().empty else "Missing")
)
```

---

4. Cleaning Project and Building Columns

The columns `building_name_en`, `project_name_en`, and `master_project_en` often had missing values.

- We grouped by `area_name_en` and filled missing entries with the **most common value in that area**.
- If no value existed, we used `"Missing"`.

This leveraged the fact that certain areas are dominated by specific projects or buildings (e.g., Dubai Marina has Marina Towers).

```
                for col in ["building_name_en", "project_name_en",
"master_project_en"]:
                    df_clean[col] =
df_clean.groupby("area_name_en")[col].transform(
            lambda x: x.fillna(x.mode()[0] if not x.mode().empty
else "Missing")
        )
```

---

5. Cleaning Facility Proximity Columns

The columns `nearest_metro_en` and `nearest_mall_en` describe proximity to infrastructure. Many rows had missing values.

- We grouped by `area_name_en` and filled with the **most common facility in that area**.
- If no facility data existed for an area, we filled with `"None"`.

For `nearest_landmark_en`, the same approach was used: fill with the most common landmark per area, else `"None"`.

This ensured that every property has realistic neighborhood information.

```
                                for col in ["nearest_metro_en",
"nearest_mall_en"]:
    df_clean[col] = df_clean.groupby("area_name_en")[col].transform(
        lambda x: x.fillna(x.mode()[0] if not x.mode().empty else "None")
    )

      df_clean["nearest_landmark_en"] =
df_clean.groupby("area_name_en")["nearest_landmark_en"].transform(
    lambda x: x.fillna(x.mode()[0] if not x.mode().empty else "None")
)
```

---

6. Dropping the `project_number` Column

The `project_number` column was simply an internal ID with no predictive value. Keeping it could introduce noise or data leakage. Therefore, it was safely dropped.

```
                    df_clean = df_clean.drop(columns=["project_number"])
```

---

7. Cleaning the Party Roles

The columns `no_of_parties_role_1`, `no_of_parties_role_2`, and `no_of_parties_role_3` represent the number of buyers, sellers, and brokers in each transaction.

- Almost all rows had values, except for about 968 missing.
- Since most transactions involve **1 buyer and 1 seller**, we filled missing values with the **median (1)**.
- This preserves the typical structure of transactions without distortion.

```
          for col in ["no_of_parties_role_1", "no_of_parties_role_2",
"no_of_parties_role_3"]:
    df_clean[col] = df_clean[col].fillna(df_clean[col].median())
```

---

8. Handling Dates (`instance_date`)

The `instance_date` column contained transaction dates in **day-month-year format**.

- We converted it into proper datetime format using `dayfirst=True`.
- We extracted new features: `year`, `month`, and `day_of_week`.
- These features enable trend and seasonality analysis in future modeling.
- Four rows had invalid dates, which were dropped to remove the last missing values.

```
df_clean["year"] = df_clean["instance_date"].dt.year
df_clean["month"] = df_clean["instance_date"].dt.month
df_clean["day_of_week"] = df_clean["instance_date"].dt.dayofweek
```

---

9. Outlier Removal

Several columns contained extreme or unrealistic values. We applied filters to keep only valid ranges:

- **`actual_worth`**: kept between 50,000 and 1,000,000,000 AED.
- **`rooms_en`**: capped between 0 and 15.
- **`procedure_area`**: kept between 10 and 20,000 m².
- **`meter_sale_price`**: kept between 100 and 100,000 AED/m².
- **Party roles**: capped at 10 per role.

This removed data entry errors and extreme outliers without affecting normal records.

```
df_clean = df_clean[
    (df_clean["actual_worth"] >= 50_000) & (df_clean["actual_worth"] <=
1_000_000_000) &
    (df_clean["rooms_en"] >= 0) & (df_clean["rooms_en"] <= 15)
&
    (df_clean["procedure_area"] > 10) & (df_clean["procedure_area"] < 20_000)
&
    (df_clean["meter_sale_price"] > 100) & (df_clean["meter_sale_price"] <
100_000) &
    (df_clean["no_of_parties_role_1"] <= 10)
&
```

```
    (df_clean["no_of_parties_role_2"] <= 10)
&
    (df_clean["no_of_parties_role_3"] <=
10)
]
```

---

10. Final Export

After cleaning:

- The dataset had **1,535,582 records** and **26 features**.
- There were **no missing values**.
- The cleaned dataset was saved as `dubai_real_estate_clean.csv`, ready for analysis and modeling.

Conclusion

The cleaning process transformed the raw dataset into a structured, reliable, and analysis-ready dataset. By systematically addressing missing values, outliers, and inconsistencies, we ensured that the dataset reflects realistic real estate transactions in Dubai.

The inclusion of engineered features (year, month, day of week) increases the analytical value of the dataset, enabling richer insights into market dynamics. The result is a dataset that is:

- Free of missing values,
- Realistic in terms of property characteristics and prices,
- Enriched with new features for temporal analysis,
- Suitable for advanced exploratory data analysis, visualization, and predictive modeling.

This cleaned dataset now forms a solid foundation for understanding Dubai’s real estate market and building data-driven insights.

1.