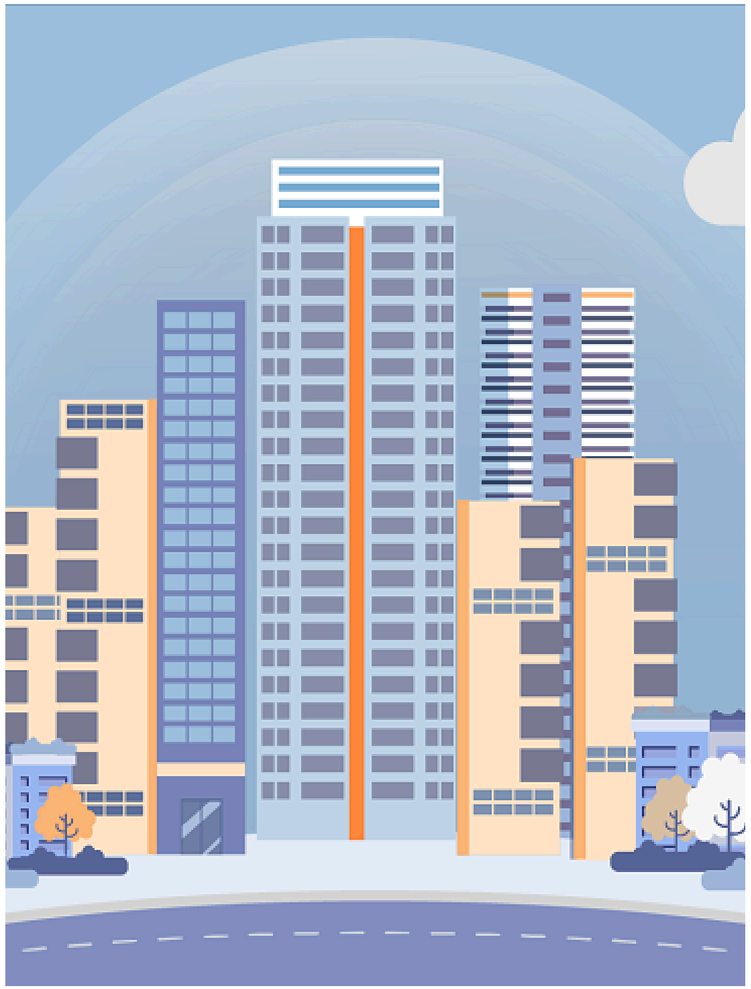
KM6312 Group Project

Data analysis of the HDB rental market in Singapore

Group 9

Presented By :
Cai Qinyi
Fan Hengtao
Li Yufei
Lim Yong Kang, Kenneth
Luo Kaiwen

Content



Introduction

Background: Increasing demand for HDB rentals due to migrant permit population growth and local improvement seekers

Complexity: Fluctuating prices, diverse amenities, and changing tenant preferences

- **Primary Objective**: Unravel underlying patterns and dynamics in Singapore's rental HDB market

- **Secondary Objectives**:
  - Provide nuanced perspective on rental price drivers
  - Foster deeper understanding among tenants, landlords, and estate agents

Related Works

Significant Body of Work

- Focuses on factors influencing rental prices
- Extraction and analytical methodologies to understand dynamic market forces

Text Mining and NLP

- Applied in the context of real estate descriptions and amenities
- Topic modeling and NLP techniques used to extract insights from textual data
- Identifying hidden topics in HDB rental listings

Machine Learning Models

- Used in predicting real estate prices
- Models like linear regression, logistic regression, SVM, random forests, decision trees, and deep learning models

Data Preparation

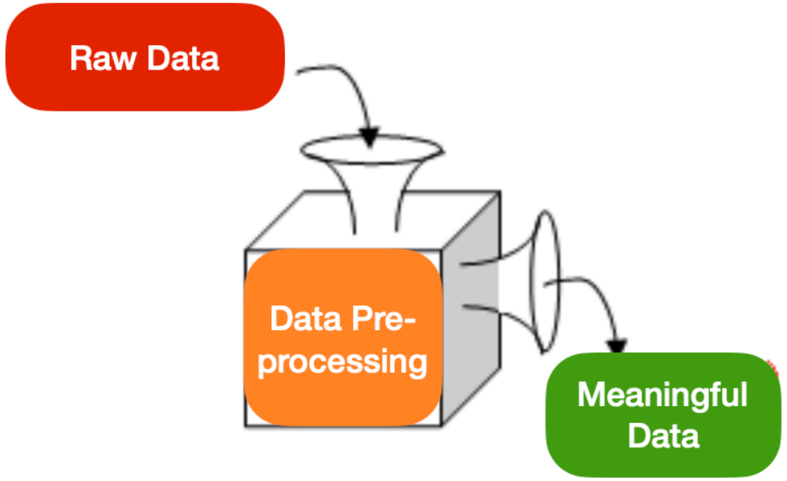
Data Source

- **Website**: 99.co
- **Data Extracted**: Property link, name, bedrooms, bathrooms, sqft, price_per_sqft, floor level, furnishing, year built, property type, amenities, and description
- **Tool Used**: Selenium library for Python with BeautifulSoup for parsing

Data Extraction Process

- Selenium Functionality: Mimics user actions like clicking, loading dynamic content
- HTML Parsing: Using find_all, find, and find_next functions
- Error Handling: Added tolerance for different page structures, checks for HTML elements before extraction
- Real-time Storage: Saving links and details to CSV for traceability

Data Preparation



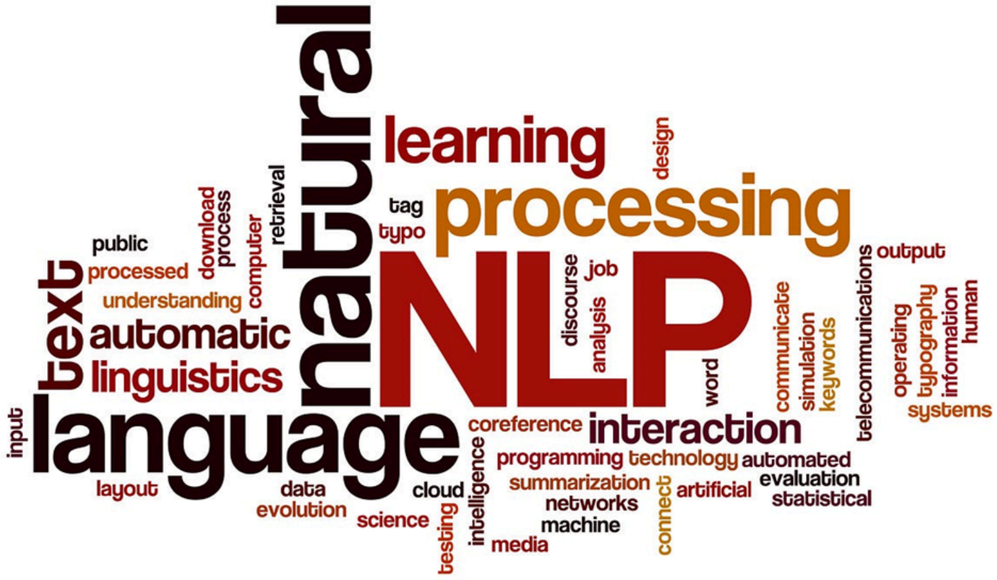
Data Pre-processing

- Target Variable Transformation: Except for linear regression. Continuous to categorical for classification models.
- Label Encoding: For categorical variables like floor level, furniture, and property type. Label Encoding distributes integer codes based on unique values for each category.
- Handling Missing Values: Remove the null values to reduce the noise caused by missing values.
- Outlier Removal: Using IQR method for numerical data

Text Pre-processing

Columns: Property description and amenities
Steps:

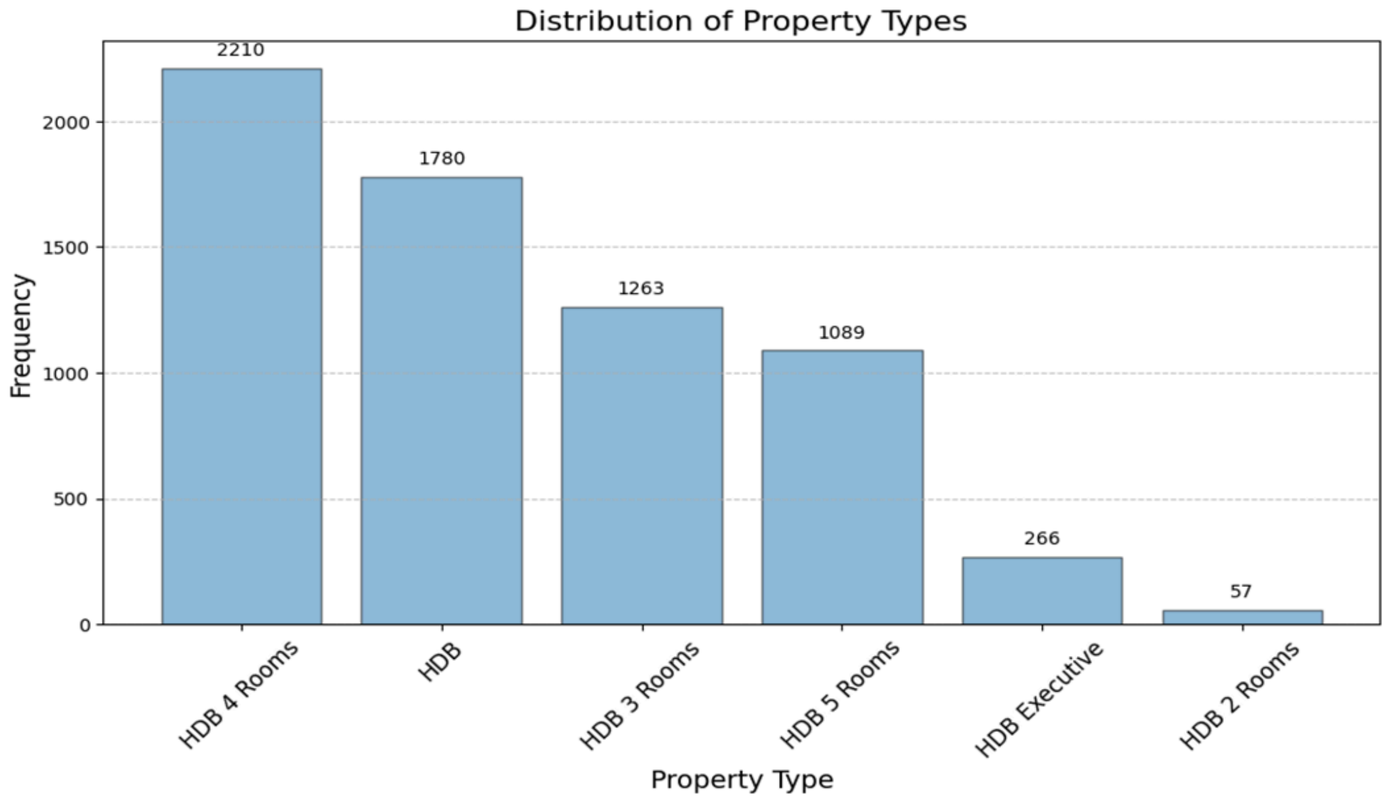
- Remove punctuation and symbols
- Convert to lowercase
- Remove non-alphanumeric characters
- Remove whitespace and normalize formatting
- Lemmatization and tokenization for NLP tasks

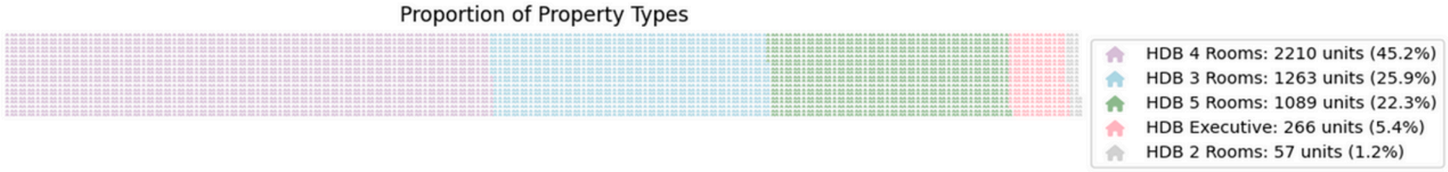


Data Structure

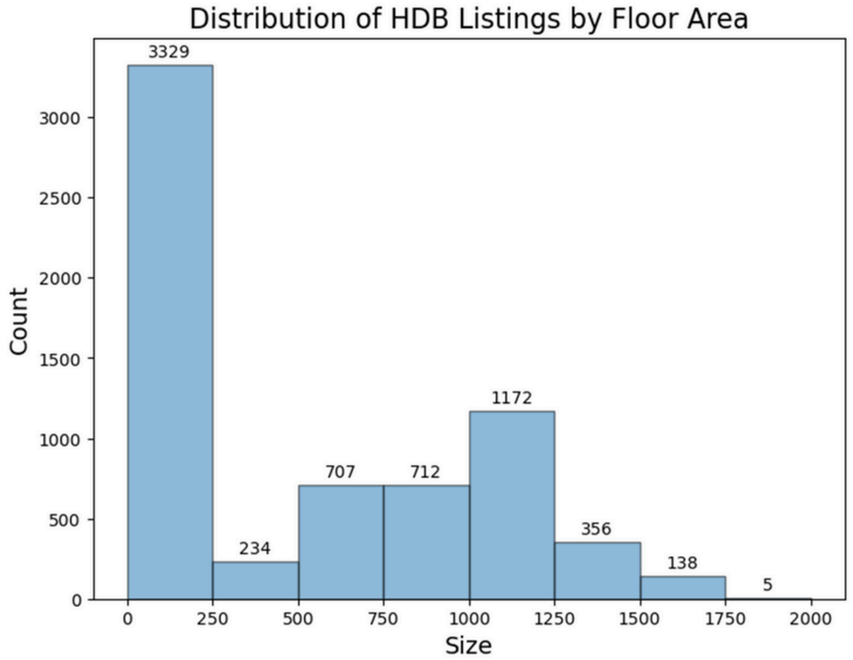
Attributes	Data Type	Example
Property_link	Text	https://www.99.co/singapore/rent/property/98-commonwealth-crescent-hdb-zjCdCY96KrCS2R22ukx6Nn
Name	Text	3 Room (3I) HDB for Rent in 98 Commonwealth Crescent
Beds	Numeric	2
Baths	Numeric	2
Size	Numeric	668
Price_per_sqft	Numeric	4.64
Floor_level	Text/Categorial	High
Furnishing	Text/Categorial	Partial
Built_year	Numeric	1970
Property_type	Text/Categorial	HDB 3 Rooms
Amenities	Text	Closet Washer Aircon Bed Dining room furniture Renovated Sofa Fridge
Description	Text	Best 3 room HDB flat whole unit for rent in Commonwealth Crescent - 3 room flat for rent - 2 spacious with aircons bedrooms - high floor - privacy unit - unblock view - renovated unit - partially furnished - spacious layout - around 10 mins walk to Commonwealth MRT - available in early January 2025 Pls call Ching at 9678XXXX for arrange viewing. Thank you.

Exploratory Data Analysis

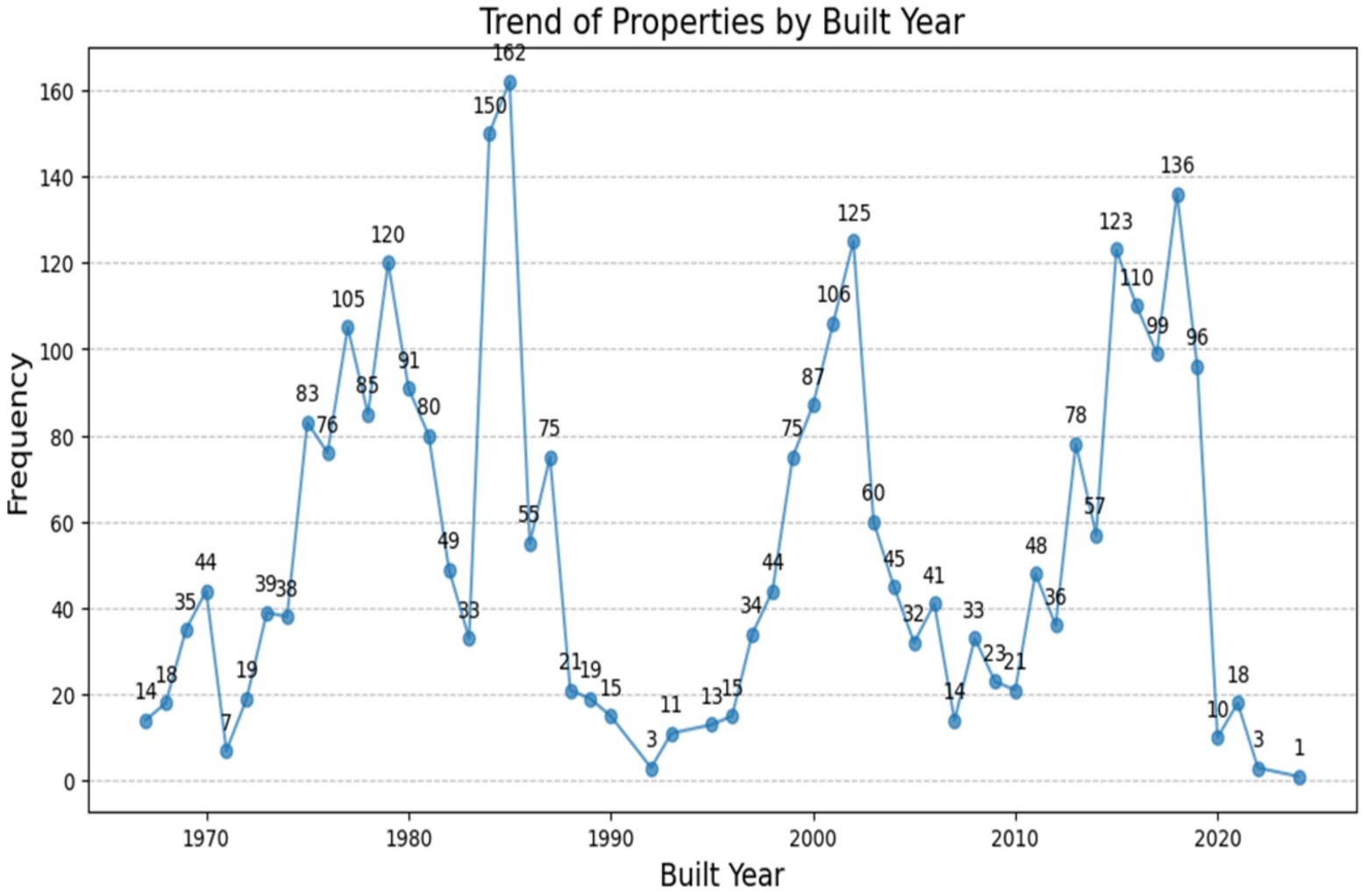




- There are a total of 6665 no. of rental listing
- Most common: HDB 4 Rooms, followed by HDB 5 Rooms.
- HDB 2 Rooms have the lowest number of units for rent.
- Majority of the listings are below the size of 250sf

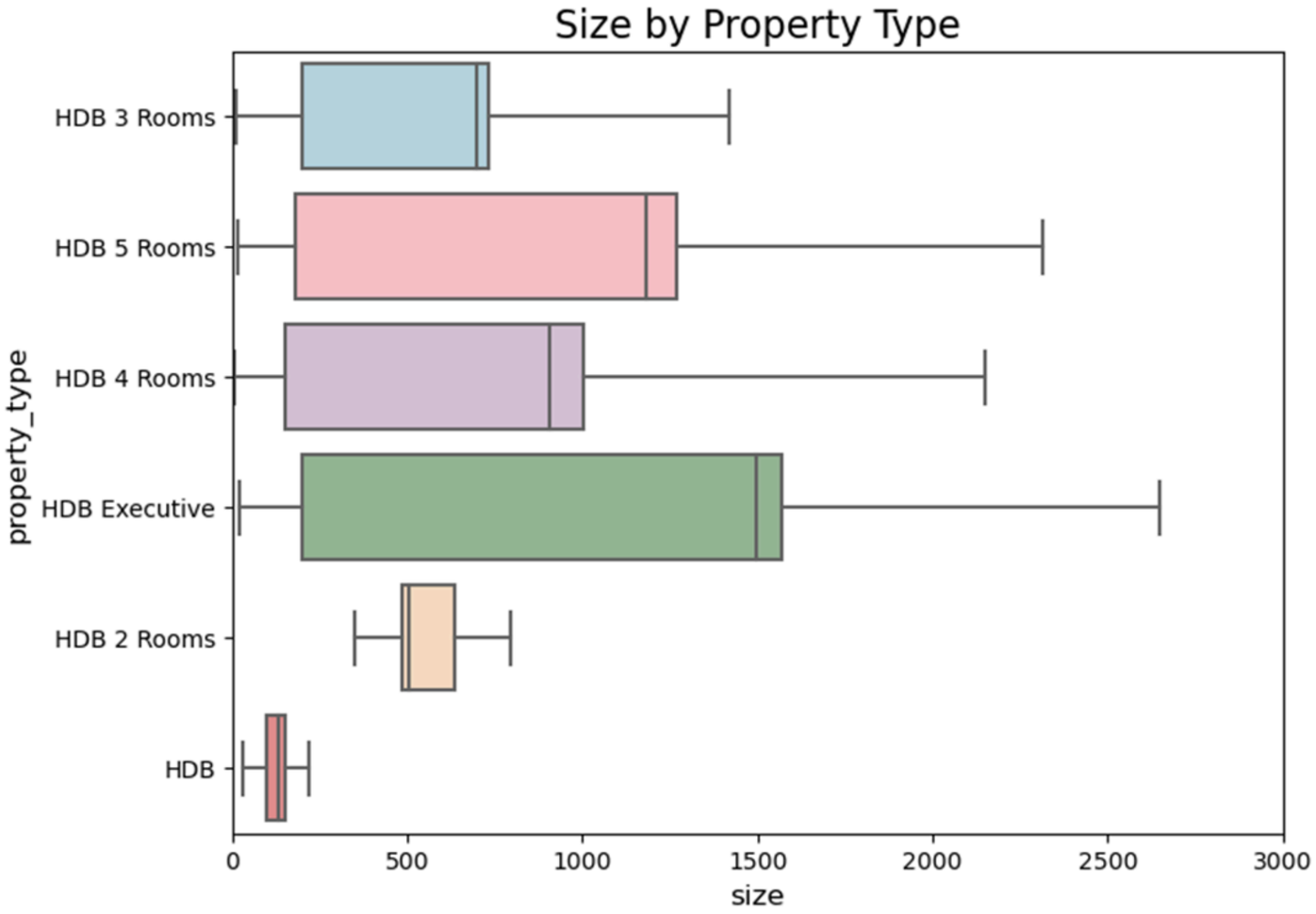


Exploratory Data Analysis



- Trend of properties by built year shows cyclical property development at various years
- This can be a good indicator on the type of properties that may be ageing, hence owners may be keen to sell.

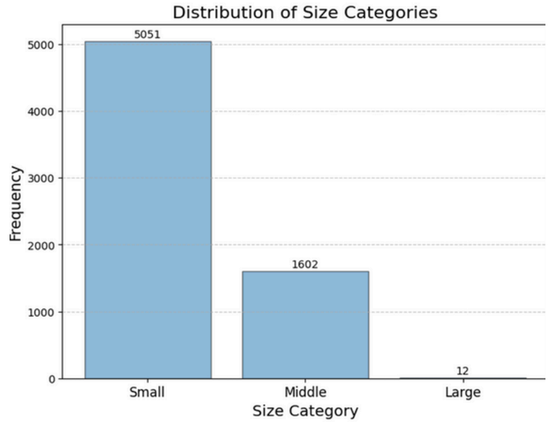
Exploratory Data Analysis



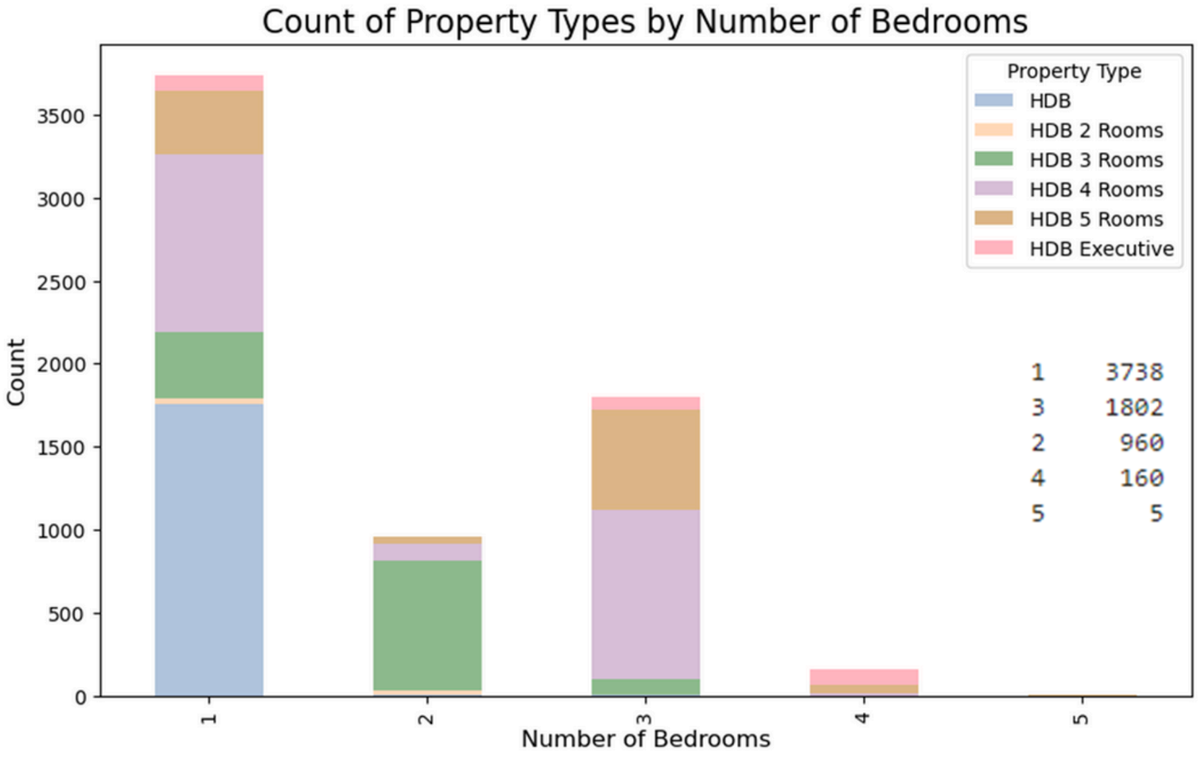
- Based on the Property type , we derive the various size range of the listings.
- The boxplot note that the median size across the different property types are approximately 200sf.
- We further classify the sizes of property by bins of 1000sf, as it may be a good size for families, hence we note that majority of the listings fell under the small bins

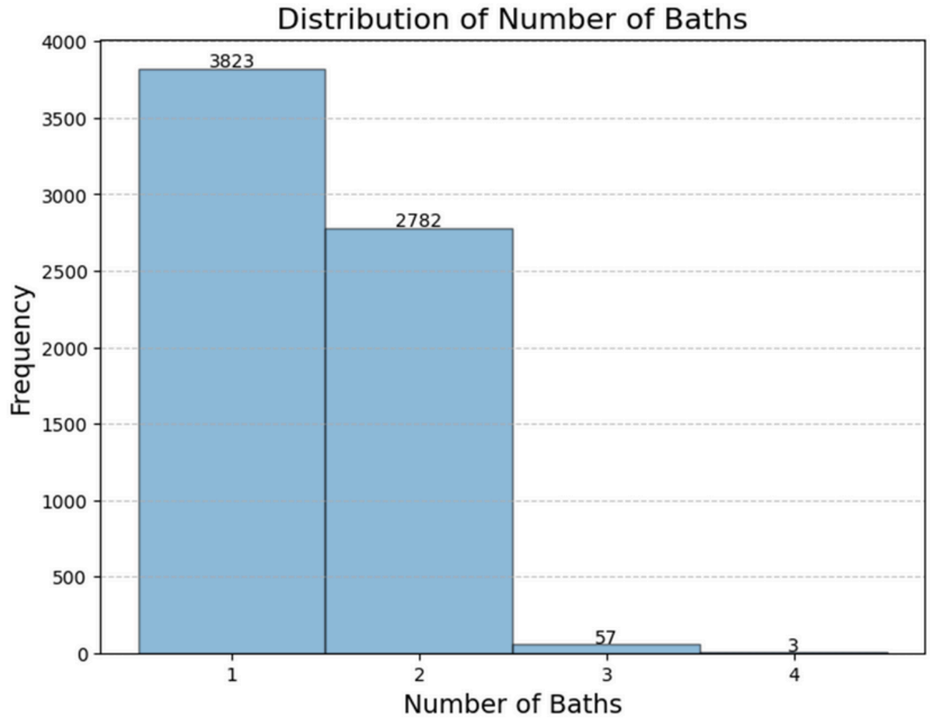
Median Size by Property Type:

	property_type	Median Size
0	HDB	130.0
1	HDB 2 Rooms	506.0
2	HDB 3 Rooms	700.0
3	HDB 4 Rooms	904.0
4	HDB 5 Rooms	1184.0
5	HDB Executive	1498.0



Exploratory Data Analysis



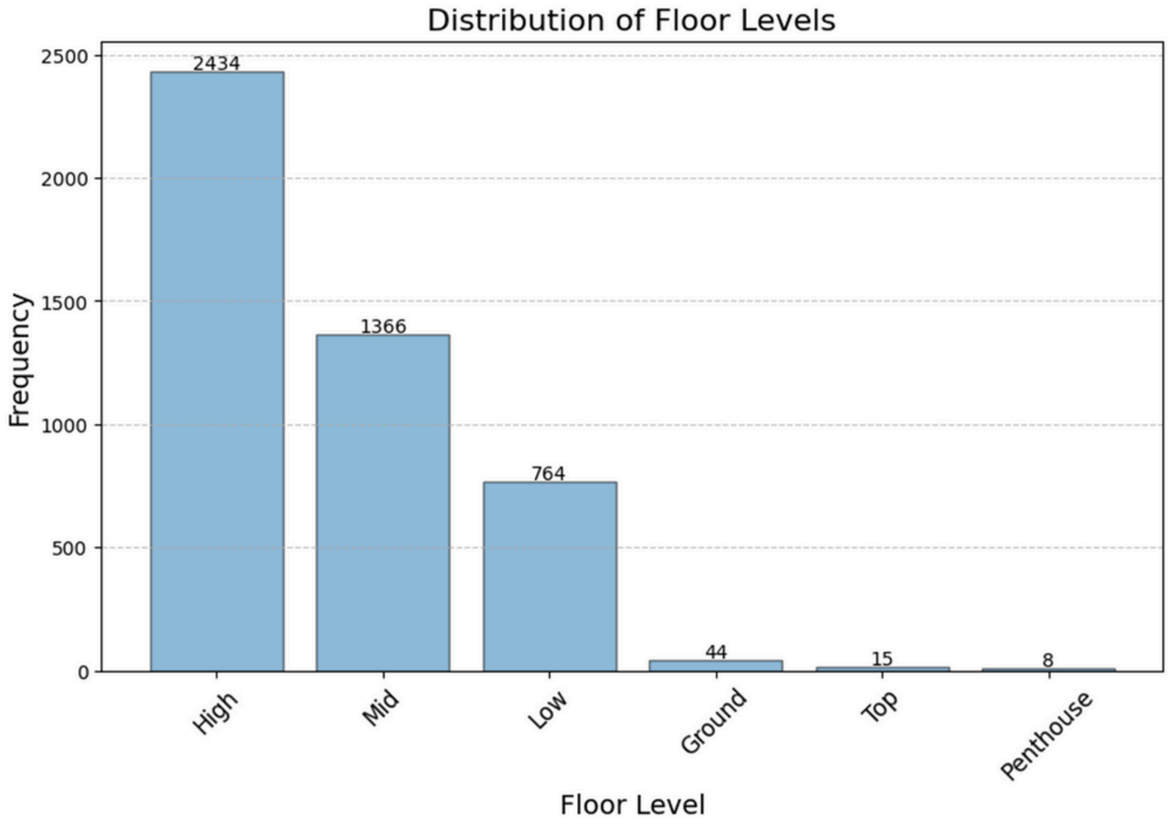


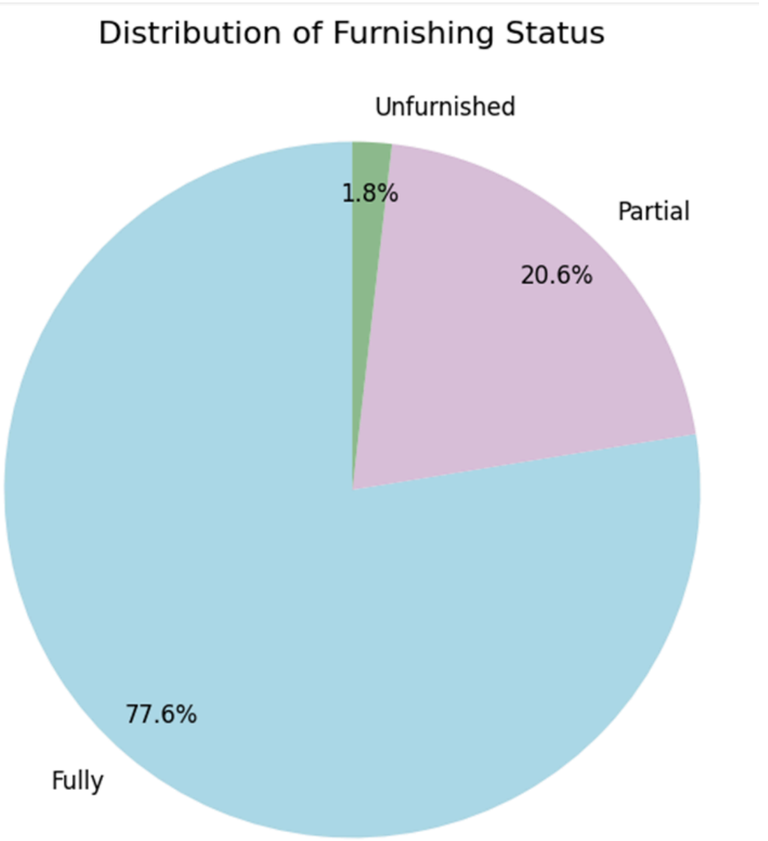
- Based on the charts, we can determine that majority of the listings are partial listings for 1 bedroom.

- We can also determine the proportion if full unit rental base on property type.

- 1 to 2 toilet per property is the common norm

Exploratory Data Analysis

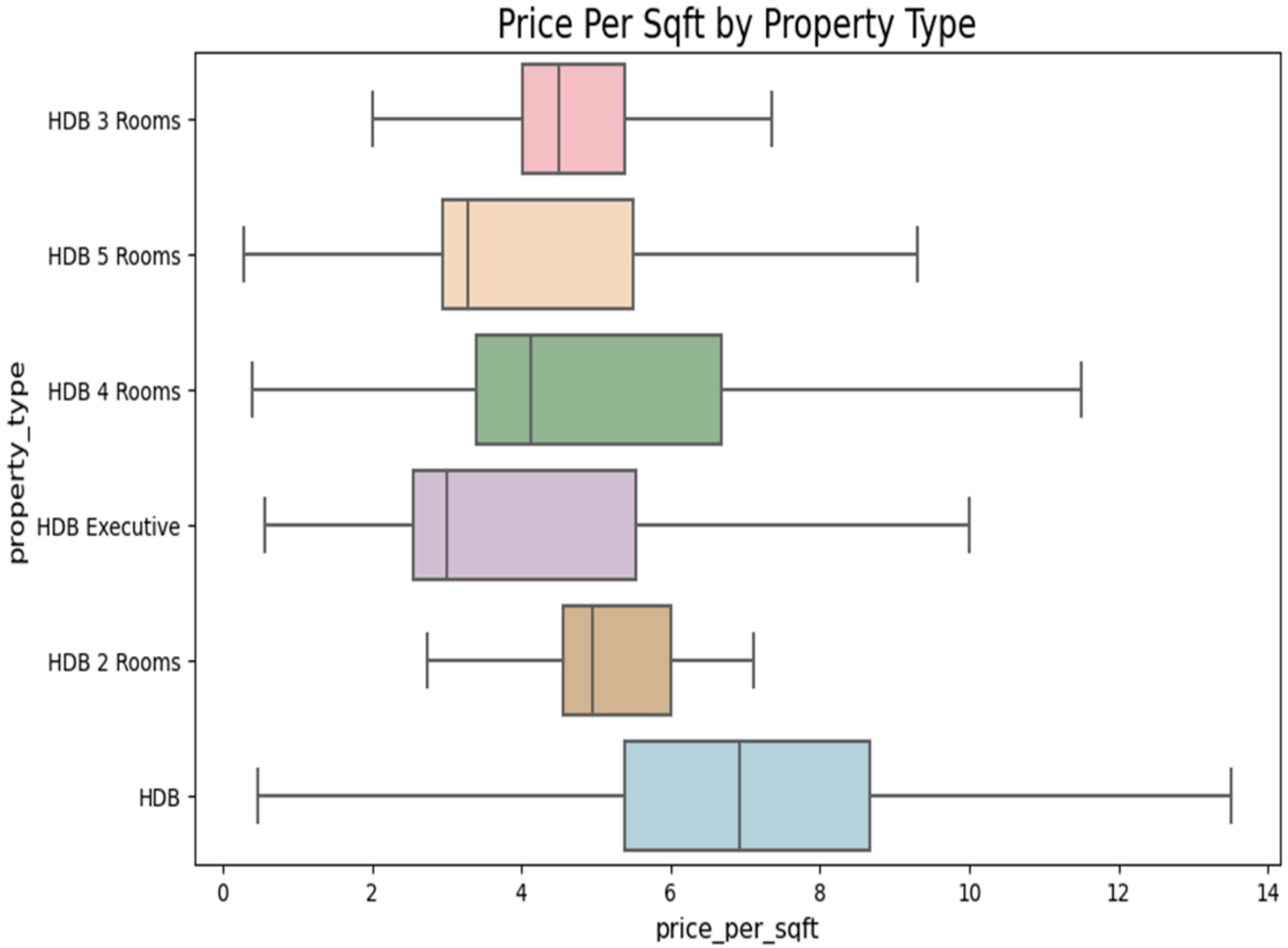




- Most of the current listing are for high floor units representing more than half the listings.

- 1 to 2 toilet per property is the common norm

Exploratory Data Analysis

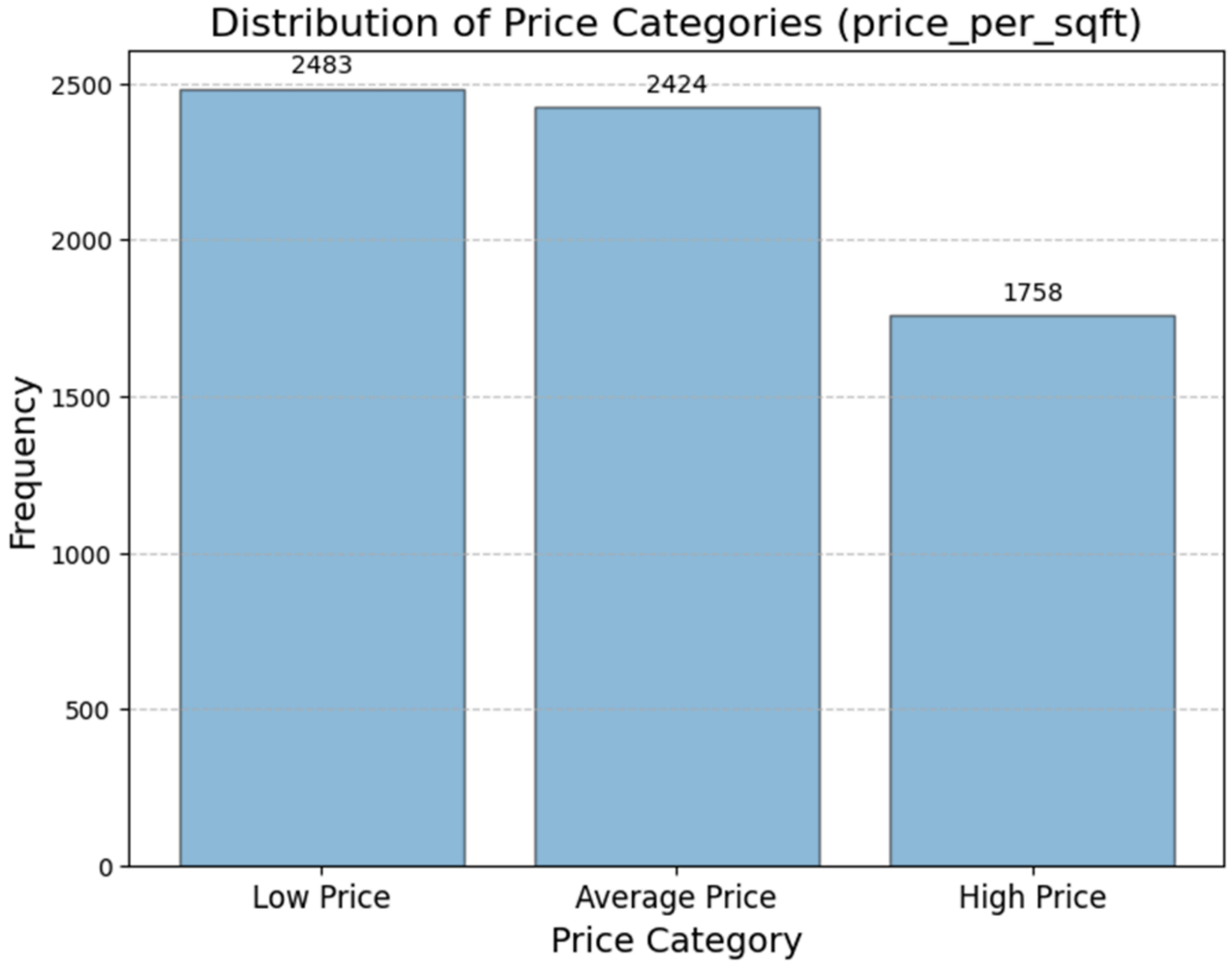


- The median decrease in price psf of the different unit varies approx. $0.50.
- The decrease is not proportional to the difference in median size.
- External factors i.e. location may be causing the difference

Median price_per_sqft by Property Type:

	property_type	Median price_per_sqft
0	HDB	6.930
1	HDB 2 Rooms	4.950
2	HDB 3 Rooms	4.500
3	HDB 4 Rooms	4.125
4	HDB 5 Rooms	3.280
5	HDB Executive	3.000

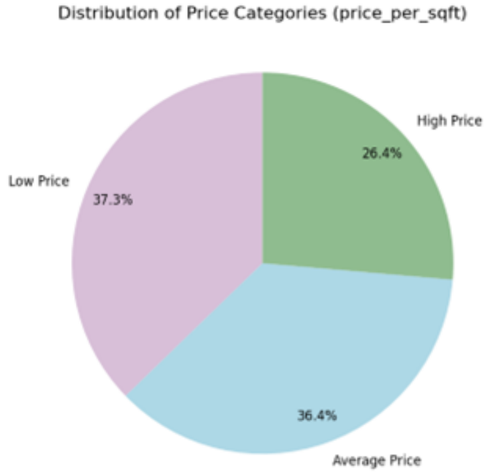
Exploratory Data Analysis



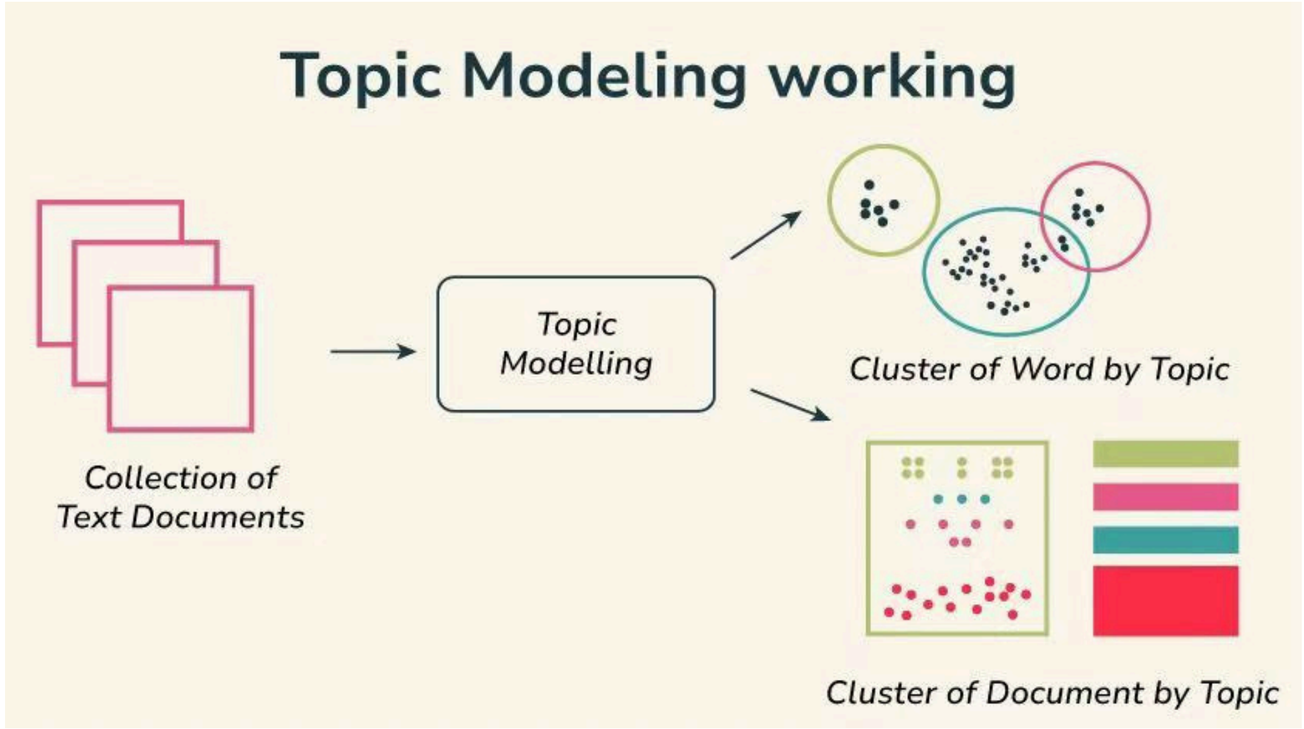
- We further classify the sizes of property by bins of $4 (avg), $7 (high)

- Based on the price classification, the prices are quite evenly spread, hence there are price options at all categories.

```
(   price_per_sqft price_category
 0            3.88      Low Price
 1            2.96      Low Price
 2            2.96      Low Price
 3            4.50  Average Price
 4            2.80      Low Price,
price_category
Low Price        2483
Average Price    2424
High Price       1758
Name: count, dtype: int64)
```



Topic Modeling



- **Method:** Applied LDA for topic modeling
- **Data Source:** Analyzed the amenities and description columns of HDB rental data
- **Objective:** Extract hidden topics and group semantically similar terms into meaningful themes

Research Questions

RQ 01

What topics are revealed by the rental amenities and descriptions of HDB listings?

RQ 02

Are there overlapping topics between topics extracted from amenities and descriptions?

RQ 03

Is reducing the number of topics improve the clarity of insights?

Machine Learning

Objective: Predict price per square foot for house prices.

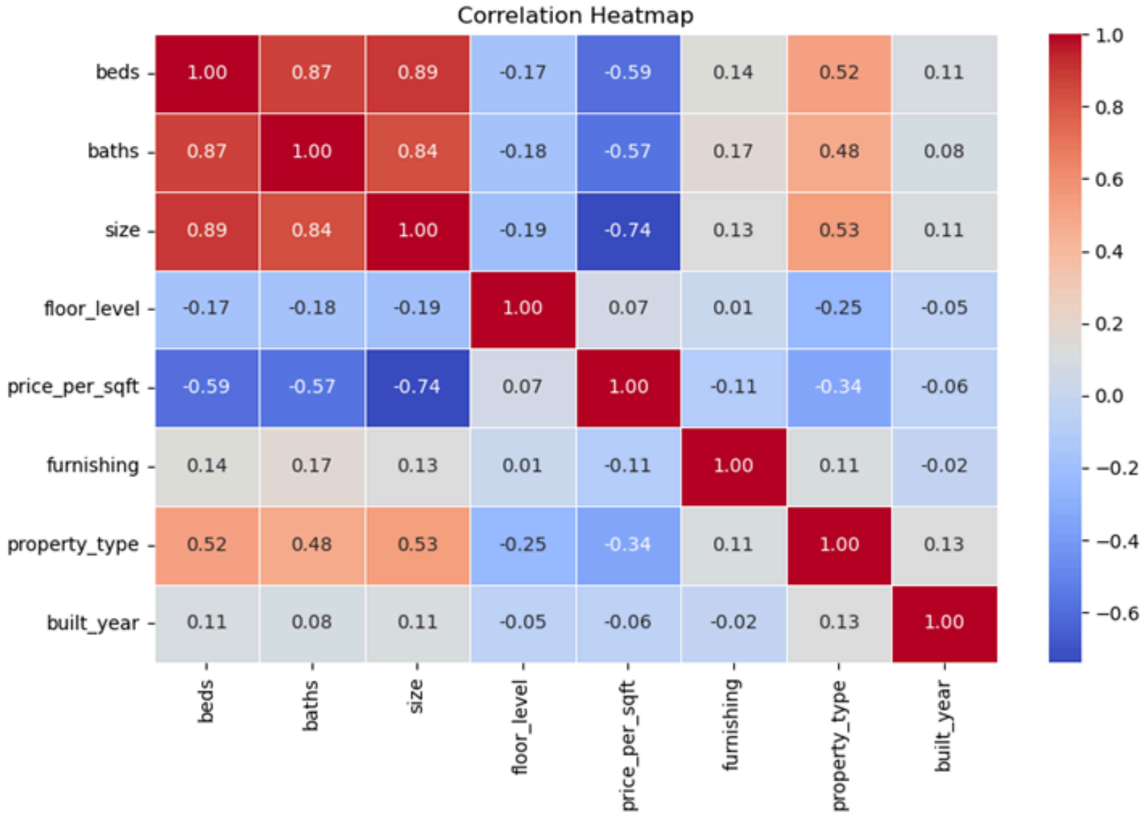
Models Used: Linear regression, logistic regression, SVM, random forests, decision trees, AdaBoost, XGBoost, gradient-boosted decision trees, 4 stacking model.

Issue: Initial models showed 100% accuracy, indicating overfitting due to using all variables.

Solution: Used a heatmap to identify key variables (beds, baths, size, property type, furnishing) to optimize the models.

Input variables: beds, baths, size, property type, furnishing

Output variable: The linear regression uses the continuous numerical variable ‘price_per_sqft’ before classification. The other models use the categorical variable ‘price_category’.




```python
def categorize_price(price):
    if price < 4:
        return 0  # Low price
    elif 4 <= price <= 7:
        return 1  # Median price
    else:
        return 2  # High price

df['price_category'] = df['price_per_sqft'].apply(categorize_price)
```

Deep Learning

Models Used

Multilayer Perceptrons, RNNs, LSTMs, GRUs, Bi-directional RNNs/GRUs/LSTMs, CNNs

Strengths

Effective for handling complex features and large datasets

Chanllenges

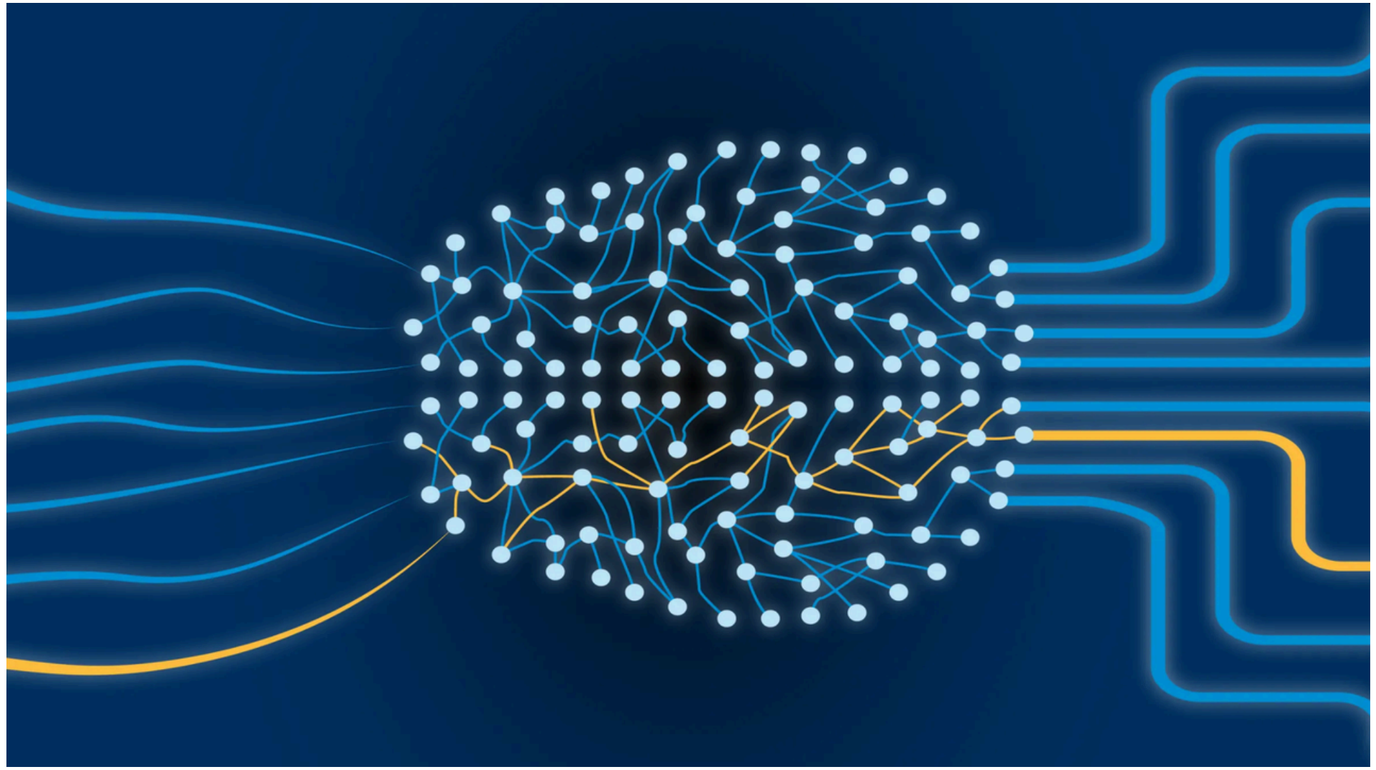
Dataset limitations: low feature dimensions and weak variable correlations

Result: Limited performance in this study context

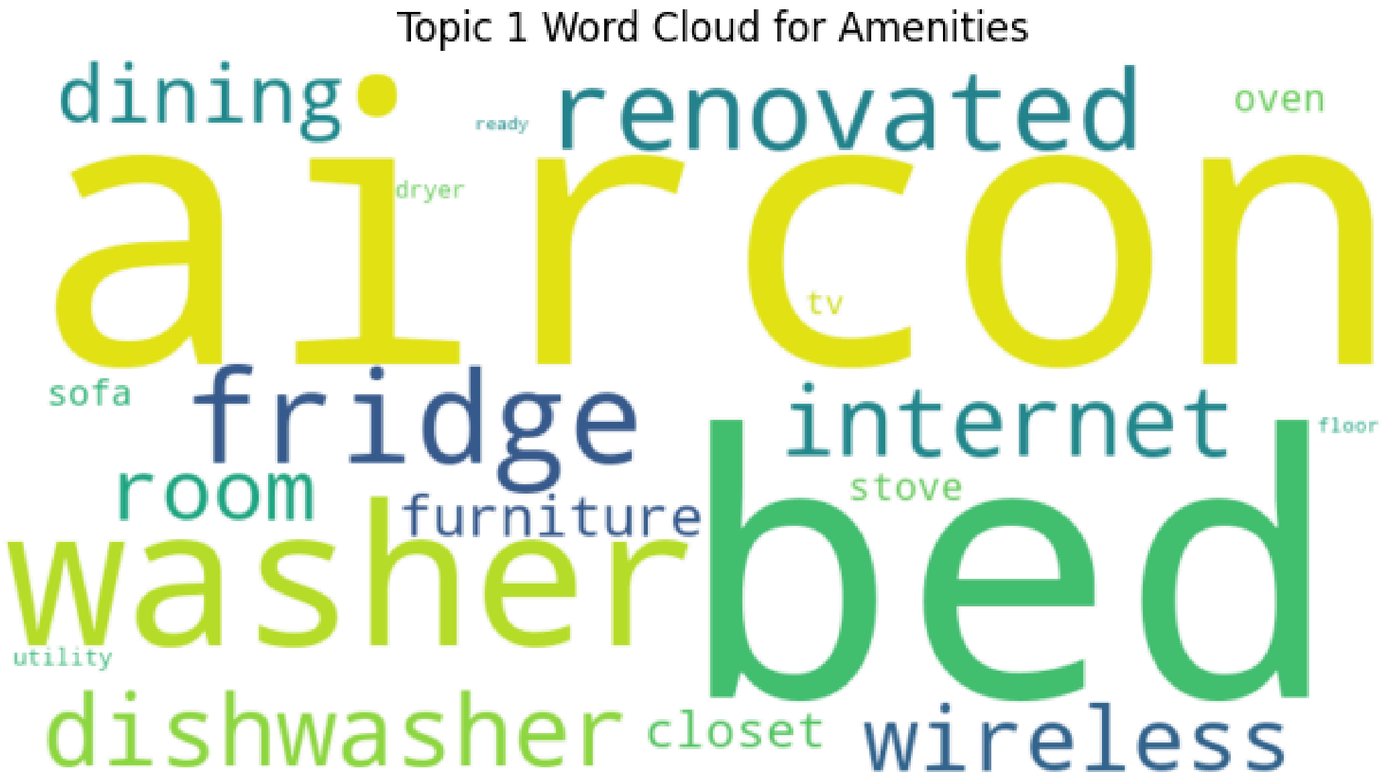
Variable Used

Input variables: beds, baths, size, property type, furnishing.

Output variable: Use categorical variable ‘price_category’ same as machine learning models except linear regression.



Result Interpretation of Topic Modeling



Explanation of Amenities Results

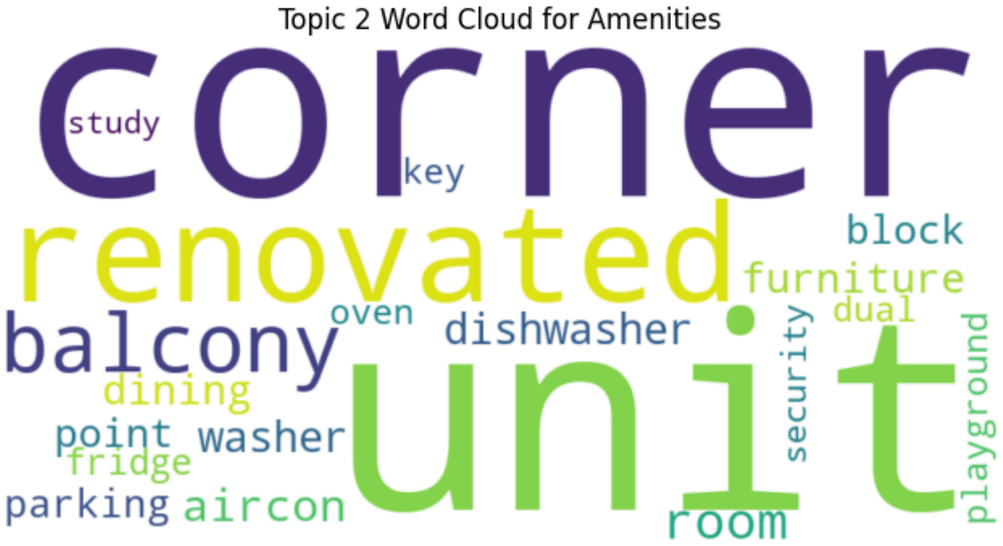
Topics focus on modern amenities enhancing tenant comfort and convenience.

Topic 1

This highlights a focus on modern amenities. Terms like “aircon,” “bed,” “internet,” and “dishwasher” emphasize tenants’ preferences for comfort, connectivity, and essential household appliances.

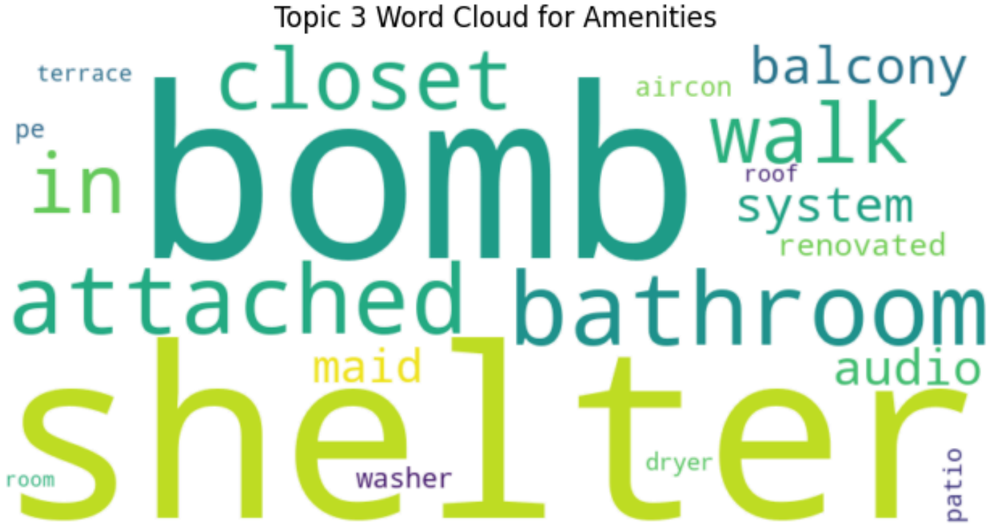
Result Interpretation of Topic Modeling

Topic Modeling - Topic 2 & 3



Topic 2

This highlights the appeal of structural design features. Terms like “corner,” “renovated,” and “balcony” emphasize the value of corner units, upgraded interiors, and outdoor spaces for tenants.



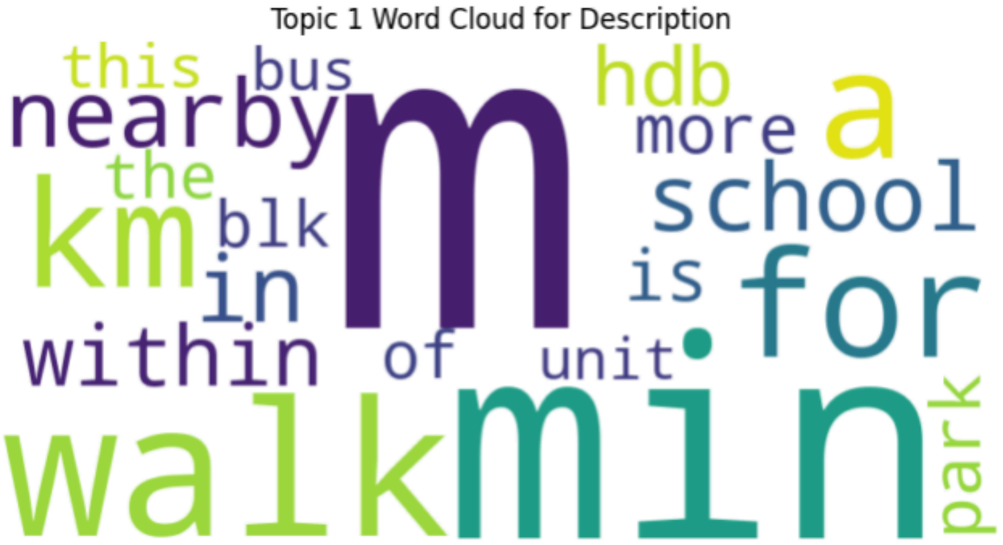
Topic 3

This highlights practicality and security aspects. Terms like “shelter,” “closet,” and “bathroom” underscore the value of bomb shelters as storage space and practical amenities that cater to tenants seeking functional living environments.

Result Interpretation of Topic Modeling

Explanation of Description Results

topics focus on rental details and location-based factors.



Topic 1

This underscores the importance of location. Terms like “walk,” “near,” and “school” highlight the appeal of properties close to schools, parks, and subway stations, reflecting tenants’ prioritization of convenience and accessibility.

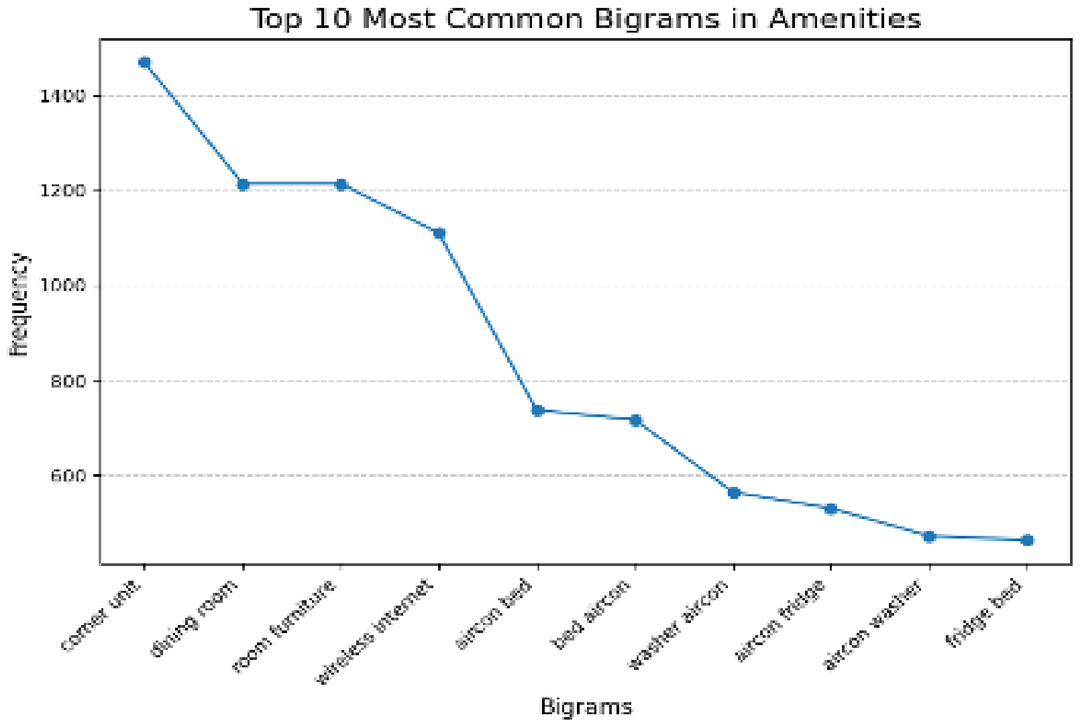
Topic 2

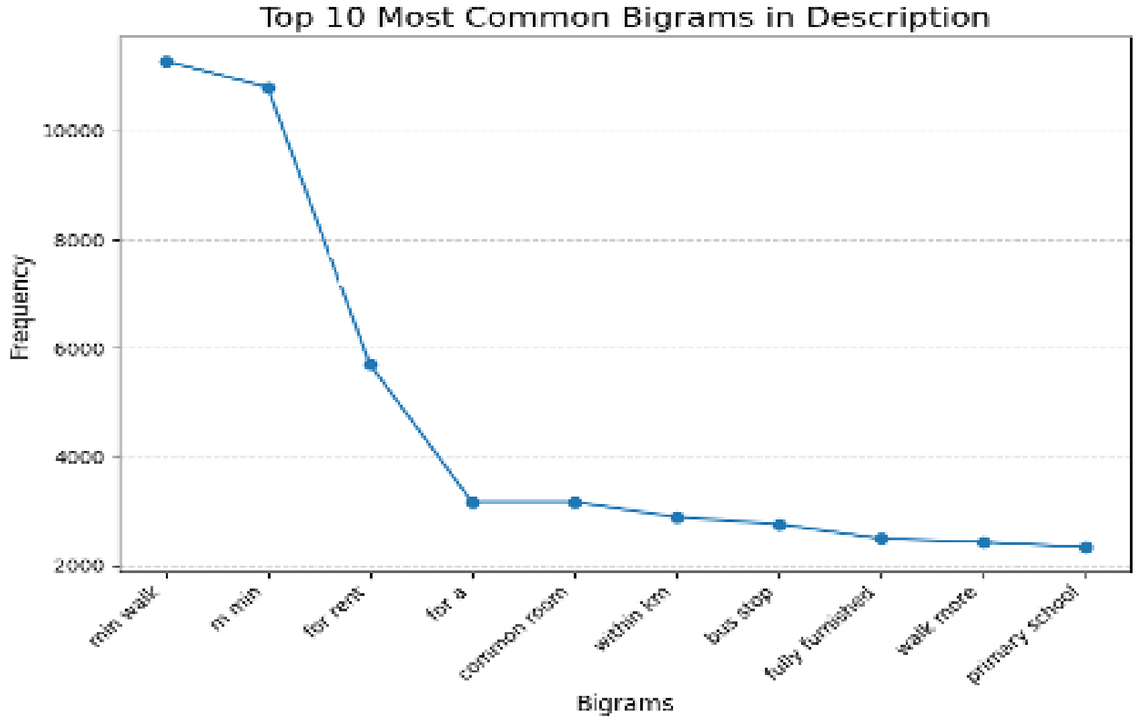
This highlights rental details. Terms like “room,” “rent,” and “MRT” emphasize room types, amenities, and transport options, while restrictions like “no pets allowed” clarify conditions.

Text Analysis

Analysis of Bigrams in Amenities and Description

- Bigram analysis reveals rental market preferences.
- Amenities: "Corner unit" highlights layout importance; "dining room" and "wireless internet" show demand for furniture and WiFi.
- Descriptions: "Min walk" emphasizes public amenity proximity; "for rent" and "common room" focus on rental use and room types.
- Further exploration: Analyzing relationships with size, price, floor level, and furnishing.





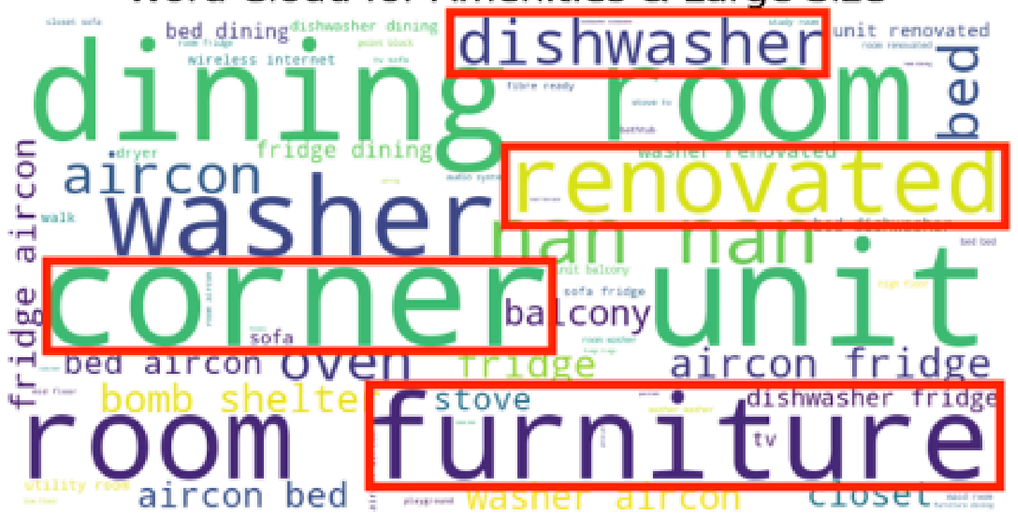
Text Analysis

Analysis of Amenities & Size

- Properties under 1,000 sqft emphasize convenience and basic functions .
- Larger properties (≥1,000 sqft) focus on luxury and space.
- Highlights differing tenant needs: small units prioritize affordability, large units prioritize comfort.

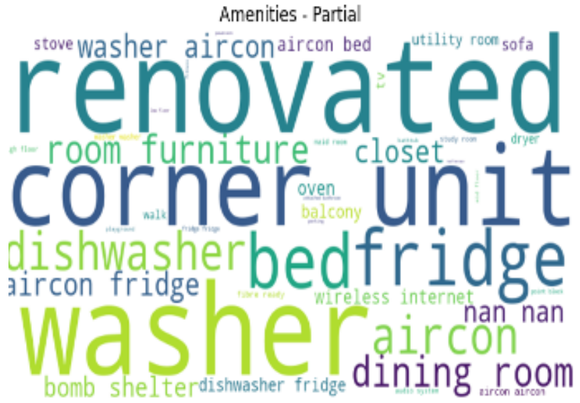
Word Cloud for Amenities & Small Size

Word Cloud for Amenities & Large Size



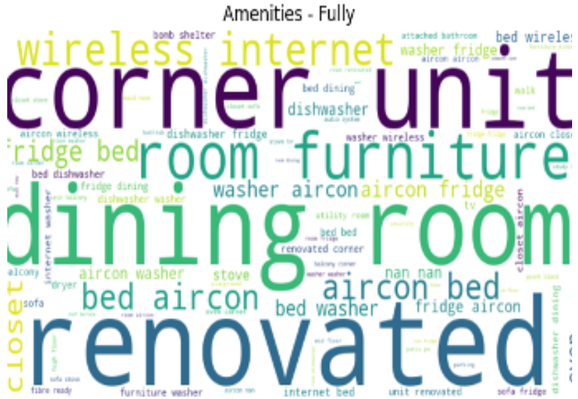
Text Analysis

Analysis of Amenities and Furnishing



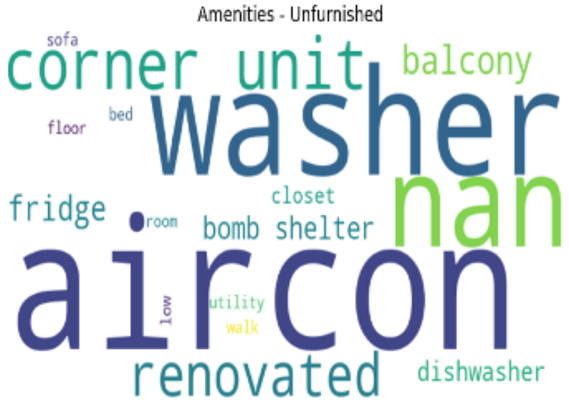
Partial Furnished

Focuses on necessities and upgraded features.



Fully Furnished

Emphasizes comfort, targeting convenience-seeking tenants.

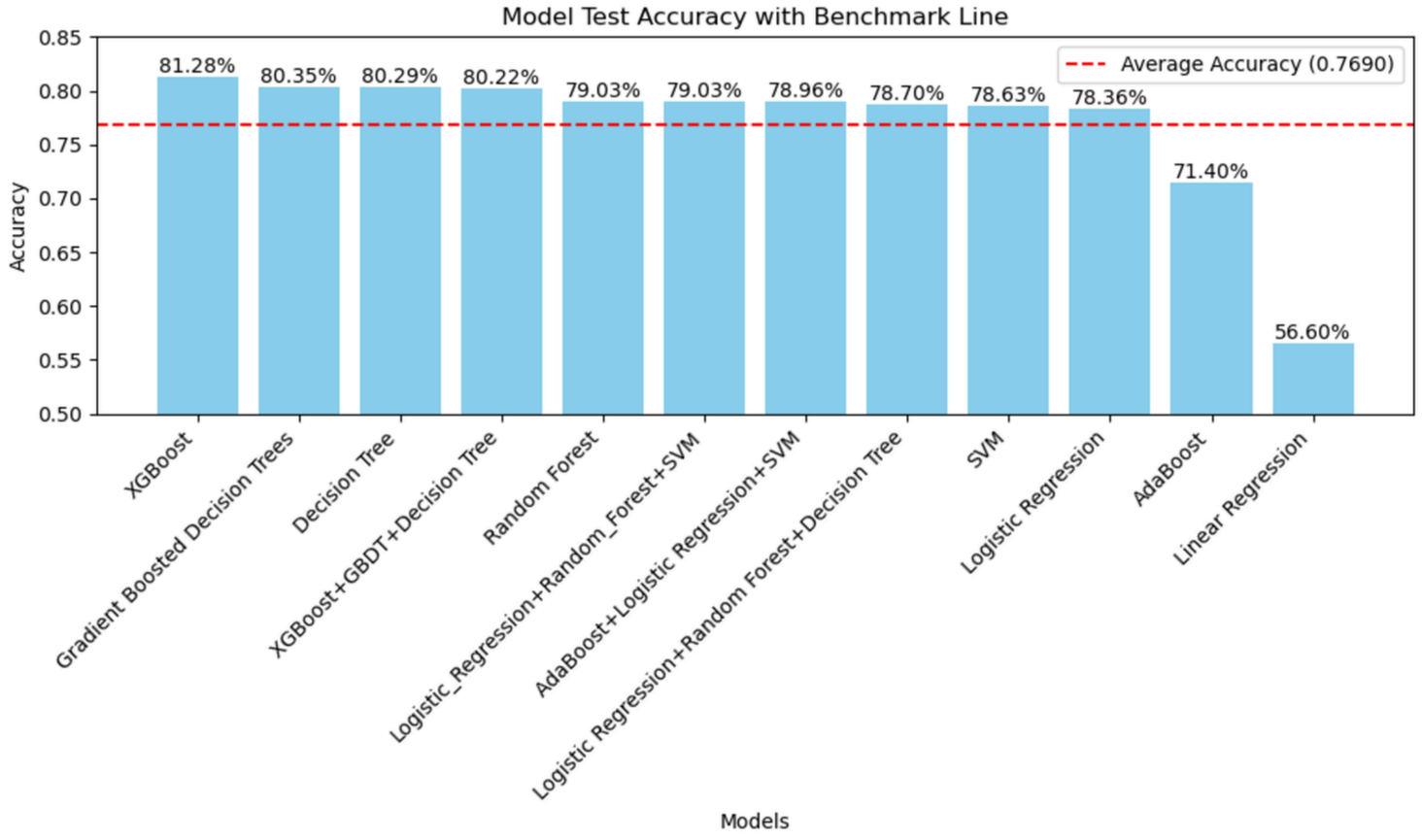


Unfurnished

Highlights structural features, catering to tenants who prefer flexible arrangements.

Common Keywords: “corner” and “unit” reflect consistent appeal across all categories.

Comparison of Machine Leaning Models



Parameter tuning : Random forest (n_estimators), XGBoost (tree depth, learning rate).

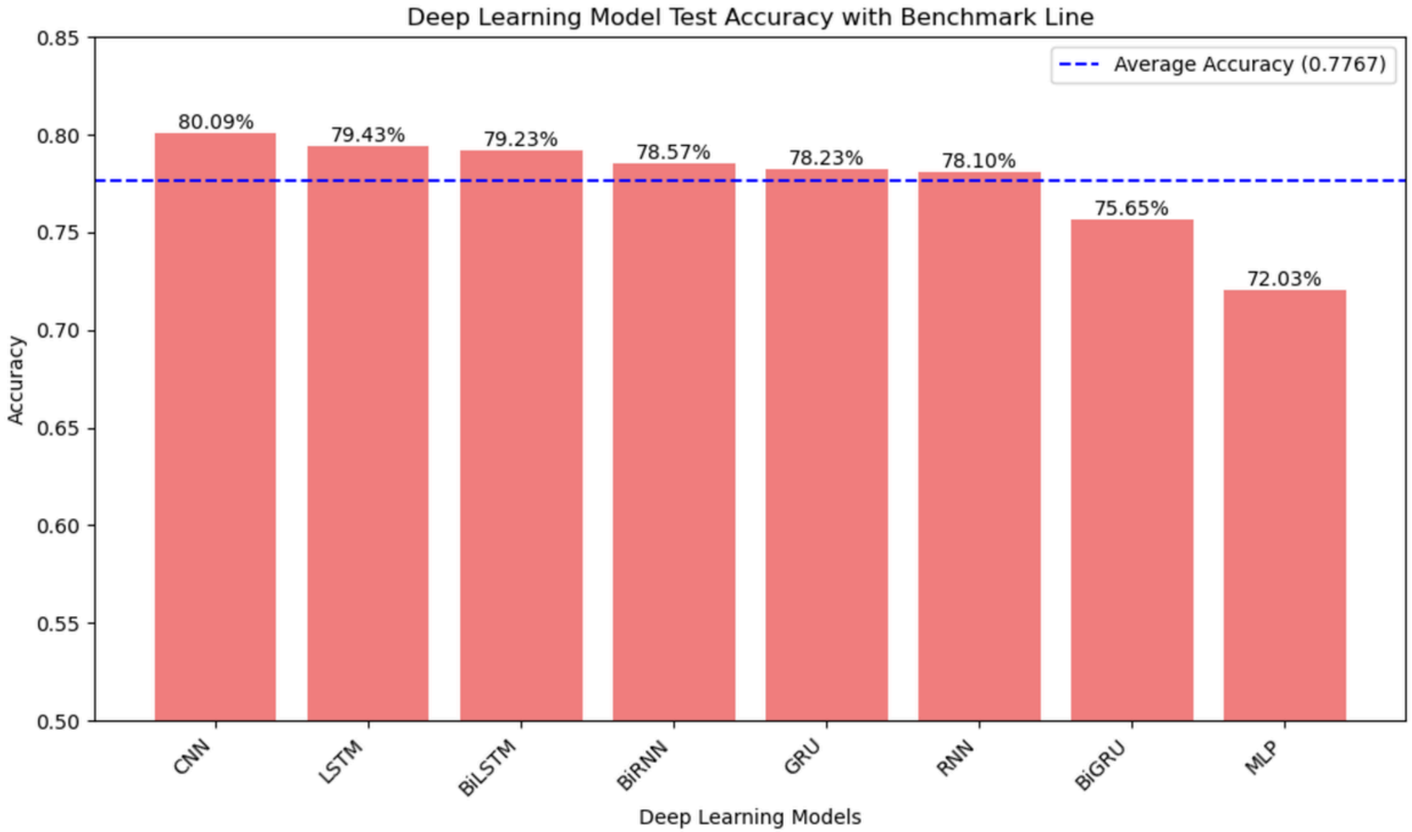
Linear regression: Performed worst (MSE=2.7499, R^2=0.5660).

Other models: Logistic regression, SVM, random forests and decision trees achieved similar accuracy (78%-79%).

Best-performing models:

- XGBoost: 81.28%
- Gradient Boosted Decision Tree: 80.36%

Comparison of Deep Learning Model



Hyperparameter Tuning

MLP: Adjusted hidden layer structure and activation function.
GRU: Tuned neurons and learning rate.

Results

The accuracy rates of CNN is over 80%.
Limitations: Sparse features, weak variable correlations, and small dataset size.

Conclusion

Deep learning models are less suitable for this dataset.

Conclusion

System Optimization

Improvement of Labeling Quality

Data Imbalance Problem

Academic and Practical Value

Moral and Ethical Statement

Reference

Pai, P. F., & Wang, W. C. (2020). Using machine learning models and actual transaction data for predicting real estate prices. Applied Sciences, 10(17), 5832-. https://doi.org/10.3390/app10175832

Singla, H. K., & Bendigiri, P. (2019). Factors affecting rentals of residential apartments in Pune, India: an empirical investigation. International Journal of Housing Markets and Analysis, 12(6), 1028–1054. https://doi.org/10.1108/IJHMA-12-2018-0097

Shen, L., Liu, Q., Chen, G., & Ji, S. (2020). Text-based price recommendation system for online rental houses. Big Data Mining and Analytics, 3(2), 143–152. https://doi.org/10.26599/BDMA.2019.9020023

Shaik, M. A., Praveen, P., Kumar, T. S., Parveen, M., Mucha, S., Mahender, K., & Reddy, I. R. (2024). Machine learning based approach for predicting house price in real estate. AIP Conference Proceedings, 2971(1). https://doi.org/10.1063/5.0196051

Wu, J., Cai, J., Luo, X. (Robert), & Benitez, J. (2021). How to increase customer repeated bookings in the short-term room rental market? A large-scale granular data investigation. DECISION SUPPORT SYSTEMS, 143, 113495-. https://doi.org/10.1016/j.dss.2021.113495

Zambrano-Monserrate, M. A., Ruano, M. A., Silva, C. A., Campoverde, R., Rosero, C., & Sanchez-Loor, D. A. (2023). Dynamism of the housing rental market in Guayaquil, Ecuador: an empirical analysis. Empirical Economics, 64(2), 747–764. https://doi.org/10.1007/s00181-022-02271-z

Thank You

Group 9