Rental Market Analysis: Predictive and Clustering Insights

Problem Definition

The rental real estate market is characterized by its complexity, shaped by varying economic conditions, demographics, and local market characteristics across different regions. This project proposes an integrated approach combining predictive modeling for forecasting rental prices with clustering techniques for market segmentation. Also, clustering will be employed to segment the rental market into distinct groups based on similarities in property characteristics. The integration of predictive analytics with clustering will not only enhance the accuracy of rental price forecasts but also deepen the understanding of market segmentation. This comprehensive analysis will provide stakeholders with critical insights, enabling informed decisions on pricing, marketing, and investment strategies, thereby optimizing profitability and strategic engagement in the rental market.

Dataset

The data was downloaded from UCI Machine Learning Repository. (Link in the References Section). The dataset contains 99493 rows and 22 columns. (Data dictionary in Appendix)

Motivation

The project aims to empower stakeholders by enhancing pricing strategies and revealing market opportunities through advanced data analytics. Predictive modeling will provide robust price forecasts, while clustering uncovers market patterns, guiding targeted marketing and strategic planning. This dual approach will optimize revenue, improve marketing effectiveness, and support strategic decisions in the competitive rental market landscape.

Methodology

1. Data Preprocessing

The initial step was to read the data from its source and prepare it for analysis. Data preprocessing techniques included handling missing values, reconfiguring column data types, and removing duplicates. Missing values were replaced with arithmetic components (e.g., mean, median) for numerical data and relatable terms (e.g., 'Unknown') for categorical data.

2. Exploratory Data Analysis

This phase was primarily used to gain a basic understanding of the data at hand. Visualizations were created to better comprehend the data. For example, a frequency bar chart of the numerical columns was used to show that the most predominant types of apartments were 2-bedroom, 1-bathroom apartments.
(See fig. 1.1 in Appendix)

3. Clustering Analysis

Clustering techniques were applied to the data after basic preprocessing. K-means clustering identified three optimal clusters. One cluster represented high-end apartments averaging nearly $1,800, while the other two clusters, with average prices around $1,329, varied mainly by location. PCA plots were created, showing that the high-end apartment cluster was located in California, while the other two clusters were in Texas and Indiana. Apparently, California has higher rental prices compared to apartments in Texas and Indiana. (See fig. 1.2 in Appendix)

4. Clustering on Boston Metro Area

This analysis focused on apartments in the Boston Metro Area, filtered by the 'cityname' column, including nine cities like Boston, Cambridge, Malden etc. The same clustering process revealed three distinctive clusters: one for high-end apartments averaging $3,400, another for better-valued apartments at nearly $2,400, and a third for affordable units averaging $1,300. This cluster analysis varied apartments into three different price ranges.

5. Classification Analysis

The objective was to classify apartments by price, categorizing them as Affordable, Mid-Range, or High-End. These categories were formed by dividing the price range of Boston Metro apartments into three equal bins. SVM classifier yielded a low accuracy of 56%, but the random forest classifier yielded 75% accuracy, with Mid-Range and Affordable apartments achieving good F1 scores of 0.77 and 0.78, respectively, while High-End apartments had an F1 score of 0.59. The important features contributing to accurate classification were found to be `price_per_sqft`, calculated as `price`/`square_feet`, as well as `bathrooms` and `square_feet`. The lower F1 score for high-end apartments could be largely due to the limited availability of high-end apartments compared to affordable and mid-range apartments.

6. Regression Analysis

Various regression models, including Lasso, Ridge, and Decision Tree Regressor, were tested. The Decision Tree achieved the highest R^2 score of 95%, but had a high MSE of $25,506.61. Hyperparameter tuning reduced the MSE to $17,173.72 and improved the R^2 score to 97%. Lasso regularization identified the important features contributing to predicting apartment prices as `price_per_sqft`, `square_feet`, `longitude`, and `bathrooms`.

Conclusions

- Different clusters can be used to understand the requirements of various types of customers. From the previous clusters achieved in this analysis, it can be concluded that some customers prioritize the quality of the apartment, while others prioritize the worth of the apartment.

- This leads to the most important takeaway from the analysis: tailored marketing strategies are necessary to attract the greatest interest. High-end apartments could be marketed based on their amenities and spaciousness, while mid-range and affordable apartments could be promoted based on their unique property features, neighborhoods, or competitive pricing.
- Model predictions can be invaluable tools for investors and prospective buyers, enabling them to make informed decisions about high-yielding areas and properties. By analyzing these predictions, investors can strategically allocate their resources, ultimately increasing profitability and success in their investments.
- Cluster analysis helps renters understand their requirements and prioritize property types—high-end, mid-range, or affordable. By identifying distinct clusters, renters can make informed choices that align with their preferences and budget, leading to a more satisfying rental experience.
- This analysis also helps potential buyers and property owners understand the ever-changing real estate market. By identifying trends in what types of properties renters are interested in at any given moment, buyers can strategically invest in lucrative properties. For current owners, these insights enable the development of targeted marketing strategies to attract renters effectively, ensuring their properties remain competitive in the market.

Future Work

- Advanced machine learning models could be deployed to enhance predictive accuracy and reduce squared error values.
- More extensive location-based analyses may yield improved predictive results.
- Utilizing real-time data can help update and refine models, ensuring they remain relevant and accurate.
- Incorporating user preferences into predictive analysis could lead to better personalization techniques and tailored solutions.

References

1. UCI Machine Learning Repository. (n.d.). Apartment for Rent Classified. Retrieved from [UCI Machine Learning Repository] (https://archive.ics.uci.edu/dataset/555/apartment+for+rent+classified)

Appendix

Fig. 1.1 Histogram of Bedrooms and Bathrooms

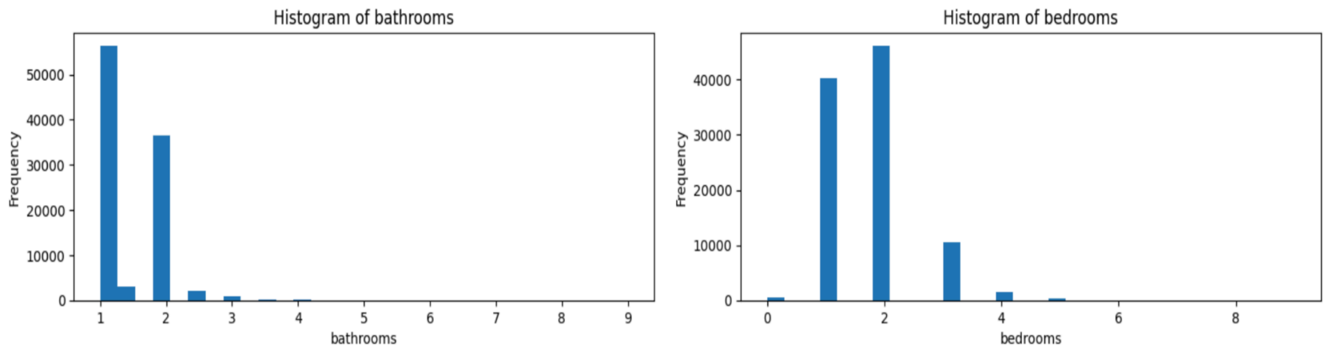
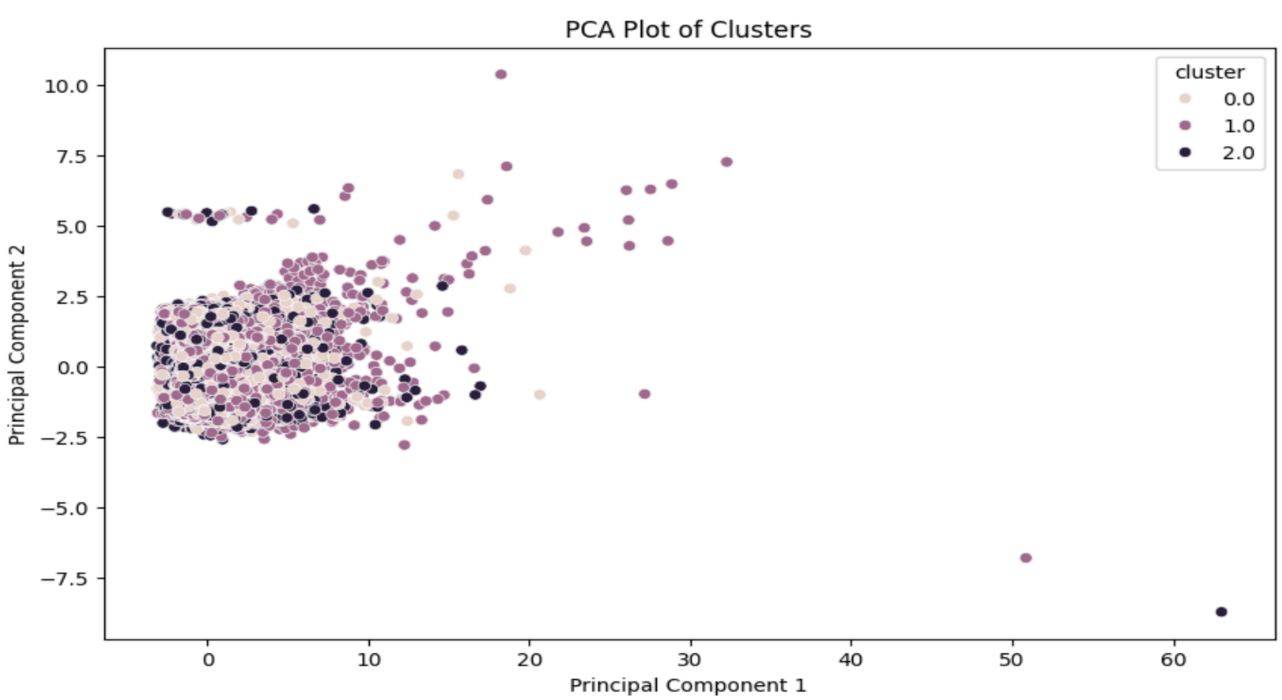


Fig 1.2 PCA of clusters



Data Dictionary

Column	Description	Variable Type
1. `id`	Unique identifier of the apartment	int64
2. `category`	Category of classified	object
3. `title`	Title text of apartment	object
4. `body`	Body text of apartment	object
5. `amenities`	Amenities included in the apartment	object
6. `bathrooms`	Number of bathrooms in the apartment	float64
7. `bedrooms`	Number of bedrooms in the apartment	float64
8. `currency`	Currency Type	object
9. `fee`	Fee	object
10. `has_photo`	Photo of the apartment	object
11. `pets_allowed`	What kind of pets are allowed in the apartment - dogs/cats etc.	object
12. `price`	Rental price of the apartment	float64
13. `price_display`	Price converted into display for reader	object
14. `price_type`	Type of rental price payment like monthly etc.	object
15. `square_feet`	Size of the apartment	int64
16. `address`	Address of the apartment	object
17. `cityname`	City in which the apartment is located	object
18. `state`	State in which the apartment is located	object
19. `latitude`	Latitude of the apartment location	float64
20. `longitude`	Longitude of the apartment location	float64
21. `source`	Origin of classified	object
22. `time`	Time when the classified was created	int64