Boston Rental Property Recommendation
(COMP3125-01 Individual Project)

Krish Ruchir Patel
School of Computing and Data Science

***Abstract*— This project develops a recommendation system for Boston rental properties tailored for WIT students. Using data analysis and a Random Forest Regressor, it predicts rental prices by balancing affordability and travel convenience. Key findings highlight the impact of travel times and property features on pricing. The system provides tailored recommendations while offering scope for future enhancements with advanced modeling and real-time data integration.**

Keywords—Rental properties, machine learning, random forest, Boston, property recommendation system.

I. Introduction

The rising demand for rental housing in urban areas necessitates innovative solutions to identify affordable and convenient housing options tailored to specific user needs. This study focuses on developing a data-driven rental property recommendation system for students attending Wentworth Institute of Technology (WIT) in Boston. By analyzing critical factors such as rental prices, travel times, and property features, the project aims to address key challenges in affordability and accessibility. The system leverages machine learning techniques, such as Random Forest Regressor, to make accurate predictions and offer actionable recommendations based on user preferences.

The importance of this research lies in bridging the gap between affordable housing and transportation accessibility, which are often disconnected in urban planning. Previous studies, such as Hamidi and Ewing's work on the affordability of HUD housing, highlight the significant impact of housing costs and transit accessibility on urban living standards [7]. This project extends this concept by integrating real-time travel data and detailed property attributes to provide an effective recommendation framework. By addressing affordability and convenience through a machine learning lens, the project demonstrates a scalable methodology applicable to other urban areas and user demographics.

II. Datasets

A. Source of dataset

The dataset for this project was obtained from two primary sources:

1. **HomeHarvest Library:** The HomeHarvest Library, accessible via a GitHub repository, provided the foundational data on rental properties in Boston, including attributes like rental prices, property type, number of bedrooms and bathrooms, and availability. This library is recognized as a credible and trusted resource within the data science community due to its reliance on real-time data scraping from multiple verified property listing websites. The datasets were generated by the library through automated data collection processes that ensure up-to-date and accurate property details. The data was accessed and downloaded on October 30, 2024, for this project [1].
2. **Google Distance Matrix API:** To enhance the dataset with travel-related metrics, the Google Distance Matrix API was utilized. The API calculates walking, transit, and driving times from each rental property to a predefined destination, in this case, the Wentworth Institute of Technology (WIT). This API is a reliable and industry-standard tool for travel data, frequently used in real-world transportation and logistics applications. For this project, API calls were made during October 2024, ensuring the inclusion of accurate and timely travel data [2]. each rental property to Wentworth Institute of Technology. This integration ensures accurate travel distance data, critical for evaluating accessibility and convenience [2].

B. Character of the datasets

The final dataset contains the following columns and their characteristics:

Column Name	Units
property_url	String (URL)
property_id	Alphanumeric
text	String
style	String
full_street_line	String
street	String
unit	String or Integer
zip_code	Alphanumeric
beds	Integer
full_baths	Integer
half_baths	Integer
sqft	Square Feet (sqft)
days_on_mls	Integer (days)
list_price	USD ($)
list_date	Date (YYYY-MM-DD)
new_construction	Boolean (True/False)
latitude	Decimal Degrees
longitude	Decimal Degrees
neighborhoods	String
county	String
fips_code	Integer
parking_garage	Boolean (True/False)
primary_photo	String (URL)
alt_photos	List of Strings (URLs)
walking_time_min	Minutes
transit_time_min	Minutes
driving_time_min	Minutes
price_per_min_walk	USD/min
price_per_min_transit	USD/min
price_per_min_drive	USD/min
price_per_sqft	USD/sqft
bed_bath_ratio	Decimal
avg_zipcode_price	USD ($)
neighborhood_travel_score	Integer (scaled 1-10)

Table 1

Preprocessing and Cleaning: The dataset underwent comprehensive preprocessing to ensure accuracy and uniformity. Missing values in numerical columns, such as sqft and list_price, were imputed using the median, while categorical columns like style were filled using the mode. Duplicate entries, identified based on the property_id, were removed to maintain data consistency. Furthermore, travel times (walking_time_min, transit_time_min, driving_time_min) were standardized into minutes, and monetary values, including list_price, were converted to USD for consistent analysis. To ensure proper scaling, all values were normalized to adjust for discrepancies across data ranges.

Combining and Adjusting: Data from multiple sources, including the HomeHarvest Library and Google Distance Matrix API, were merged to provide a comprehensive dataset. This involved aligning property details from the HomeHarvest Library with travel times derived from the Google Distance Matrix API. ZIP codes served as the primary key for combining these datasets. Careful adjustments were made to reconcile discrepancies, such as matching property IDs and validating the consistency of geographic coordinates.

Feature Engineering: Several derived features were created to enhance the dataset's analytical depth. Travel cost-efficiency metrics, including price_per_min_walk, price_per_min_transit, and price_per_min_drive, were generated to capture affordability based on accessibility. A neighborhood_travel_score was engineered to normalize travel times across different modes of transport, providing a comprehensive measure of accessibility on a 1-10 scale. Property-specific insights, such as the bed_bath_ratio and price_per_sqft, were calculated to represent layout efficiency and price competitiveness. These features added granularity and enabled more effective predictions.

Size: The dataset consists of **2,504 rows** and **34 columns**, encompassing detailed property attributes and travel metrics.

III. Methodology

Exploratory Data Analysis (EDA): Before implementing the model, an extensive EDA was performed to uncover trends and patterns within the dataset. Key features such as price_per_sqft, walking_time_min, and avg_zipcode_price was analyzed for their impact on rental prices. Visualization techniques, including scatter plots, heatmaps, and bar charts, were employed to explore relationships between property characteristics and travel convenience metrics. For example, it was observed that shorter walking times were weakly correlated with higher rental prices, highlighting the influence of proximity on affordability. These insights guided feature selection and informed the model-building process.

Model Overview: The project utilized a Random Forest Regressor (RFR) to predict rental property prices based on a wide array of features, including travel times, property size, and neighborhood scores. Random Forest is an ensemble method that builds multiple decision trees, averaging their results to produce robust predictions. This approach effectively captures non-linear relationships and interactions between variables.

Assumptions of Random Forest: The model assumes that each decision tree learns independently through random sampling and feature selection, ensuring diversity among trees. This diversity reduces overfitting and enhances generalization.

Advantages of Random Forest: Random Forest excels in handling high-dimensional data and identifying important features, such as travel convenience metrics, while being robust to outliers. Its ensemble structure reduces the impact of overfitting and improves accuracy.

Challenges and Mitigation: While Random Forest can handle complex datasets effectively, it can be computationally intensive. The model parameters, such as the number of trees (n_estimators), were kept at default or practical levels to balance performance and computational efficiency. Missing data was imputed, and features were scaled where necessary to maintain model consistency.

Implementation Details: The model was implemented using Python's Scikit-Learn library. Feature engineering introduced additional variables like price_per_min_transit to enhance predictions. Model evaluation relied on metrics such as Root Mean Squared Error (RMSE), which provided insights into prediction accuracy.

By integrating EDA insights and employing Random Forest, the methodology combines analytical rigor and robust machine learning techniques, enabling actionable recommendations for renters based on key property and travel attributes.

IV. Results

A. *Result 1: How do travel times by walking, transit, and driving influence rental prices?*

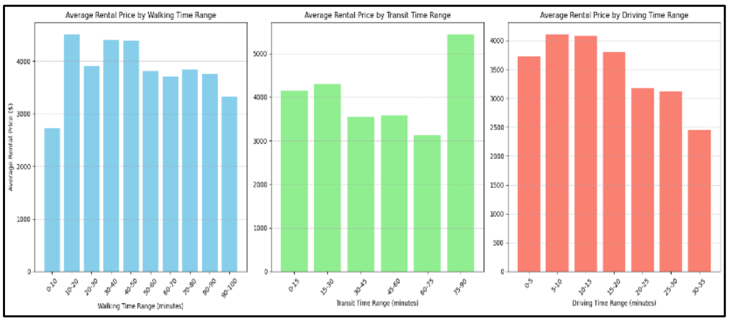

Image 1

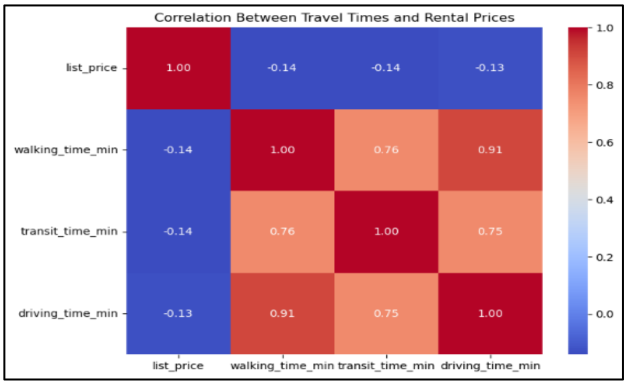

Image 2

The analysis highlights the relationship between travel times to Wentworth Institute of Technology (WIT) and rental prices. **Image 1** reveals that properties with shorter travel times, especially walking and transit times, exhibit higher average rental prices. This demonstrates the premium associated with proximity to WIT, as shorter commuting times are highly valued by renters. Conversely, properties with longer travel times, particularly driving times, tend to have lower rental prices, reflecting decreased desirability for extended commutes.

Image 2, the correlation heatmap, provides further evidence of this relationship. The weak negative correlations observed between travel times and rental prices (-0.14 for walking time, -0.14 for transit time, and -0.13 for driving time) indicate that shorter travel times are associated with slightly higher rental prices. However, the weak strength of these correlations suggests that while proximity is a factor, other property and neighborhood characteristics, such as size, amenities, and local conveniences, have a greater impact on determining rental prices.

Travel times by walking, transit, and driving have a weak but notable influence on rental prices, with shorter travel times, particularly walking and transit times, resulting in slightly higher prices. This indicates that proximity to WIT is a consideration for renters, though not the sole determinant, emphasizing the need for a holistic evaluation of property features and affordability when making recommendations.

B. *Results 2: What is the impact of property characteristics on rental prices?*

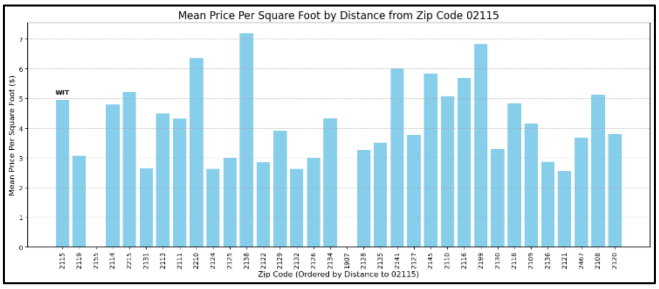

Image 3

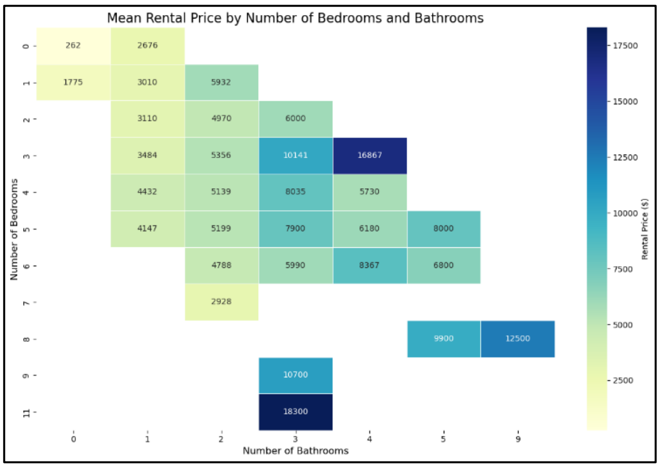

Image 4

Image 3 illustrates the mean price per square foot across zip codes, organized by their distance from zip code 02115, which corresponds to the Wentworth Institute of Technology (WIT). Properties located closer to WIT tend to have higher rental prices per square foot, indicating the premium placed on accessibility and convenience. However, certain distant zip codes also exhibit competitive pricing, potentially due to luxury developments, neighborhood demand, or enhanced amenities. This finding highlights that while proximity to WIT is a strong factor, localized economic and social factors significantly affect property valuation in specific regions.

Image 4 examines the relationship between bedroom and bathroom configurations and their impact on rental prices. Balanced configurations, such as one bedroom to one bathroom, consistently command higher rental prices, reflecting their practicality and desirability. The configuration of three bedrooms and two bathrooms is particularly valuable, offering a sweet spot of space and functionality. Conversely, imbalanced configurations, such as many bedrooms with fewer bathrooms, result in lower rental prices, suggesting inefficiency in layout. Properties with three or more bathrooms typically attract higher prices, catering to larger families or shared living arrangements.

The analysis reveals that both location and property characteristics significantly impact rental prices. Proximity to WIT drives up costs due to increased demand for accessibility. However, distant zip codes with high prices indicate the influence of localized amenities and neighborhood quality. Structurally, well-configured properties with balanced bedroom-to-bathroom ratios and sufficient space are highly valued, while imbalanced layouts are less desirable. This insight underscores that both locational and structural features are critical determinants of rental pricing, influencing affordability and desirability for prospective renters.

C. *Results 3: How Do Rental Prices Vary by Location?*

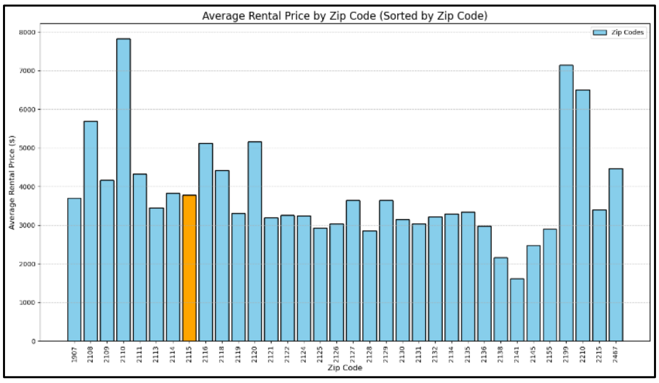

Image 5

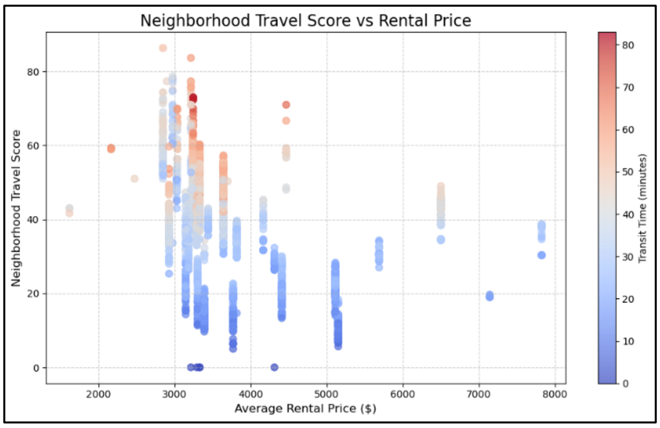

Image 6

Image 5 presents the average rental prices by zip codes, sorted by proximity to Wentworth Institute of Technology (WIT). Zip codes closer to WIT, such as 02115, exhibit moderately higher rental prices, likely due to their convenience for students and professionals. However, zip codes further away, such as 02120 and 02122, demonstrate competitive prices due to factors like luxury developments or other location-specific attributes. The variance highlights that location near educational or professional hubs influences pricing dynamics. Additionally, premium neighborhoods further from WIT show a spike in prices, reflecting demand for luxury housing.

Image 6 visualizes the relationship between the Neighborhood Travel Score and rental prices, with transit time represented by color intensity. Neighborhoods with high travel scores (indicating accessibility) tend to cluster in the higher price ranges, showing a direct relationship between convenience and rental costs. Areas with longer transit times generally exhibit lower rental prices, as indicated by the blue-colored points at the lower rental price range. This suggests that travel convenience, represented by shorter transit times and higher accessibility, is a significant driver of higher rental costs.

Rental prices vary significantly by location, influenced by proximity to WIT and neighborhood accessibility. Properties near WIT offer a balance between affordability and convenience, making them attractive to students. Luxury housing further away tends to command higher prices due to its exclusivity. This insight emphasizes the role of both geographical proximity and travel convenience in determining rental pricing.

D. Result 4: How can affordability and convenience be balanced?

Feature	Importance
price_per_min_transit	0.2533
walking_time_min	0.1436
price_per_sqft	0.1230
price_per_min_drive	0.1048
neighborhood_travel_score	0.0952
driving_time_min	0.0787
beds	0.0694
price_per_min_walk	0.0657
transit_time_min	0.0583
full_baths	0.0080

Table 2

The Random Forest Regressor model was created to predict rental prices while balancing affordability and convenience. This model incorporated several critical features such as price per minute transit, walking time, price per square foot, and neighborhood travel scores to assess the trade-offs between cost and accessibility. The model's preferences were designed to align with user needs, especially those of students at Wentworth Institute of Technology (WIT). For instance, the model prioritizes properties with lower average zip code prices and shorter travel times, ensuring affordability without compromising accessibility to WIT.

As shown the **Table 2** which is output of the model, The feature price_per_min_transit ranks as the most influential variable, emphasizing the importance of cost-efficiency in transit for balancing affordability and convenience. Properties with shorter and affordable transit times to Wentworth Institute are highly favored, as they reduce the time and monetary burden of commuting.

Other critical factors include walking_time_min and price_per_sqft, which reflect the desirability of proximity and affordability of space, respectively. Short walking times to Wentworth Institute, while typically commanding higher rental costs, provide significant value for students prioritizing convenience. Additionally, variables like neighborhood_travel_score offer a broader perspective on accessibility by combining multiple transportation modes into a single metric.

The model also highlights the trade-offs between travel times (walking_time_min, price_per_min_drive) and cost per unit area (price_per_sqft). While closer properties often come at a premium, the combined insights allow renters to optimize their decisions by balancing these competing priorities.

Affordability and convenience can be balanced by focusing on properties that optimize price_per_min_transit and walking_time_min, as these factors provide the best value for renters who prioritize both cost and accessibility to Wentworth Institute. The Random Forest model effectively identifies such properties by incorporating user preferences for maximum price and acceptable travel times, enabling personalized and informed recommendations. This approach ensures that renters can achieve the best compromise between affordability and convenience.

E. Result 5: Can a machine learning-based recommendation system identify the best properties based on user preferences and current listings?

The machine learning-based recommendation system, built on the Random Forest model, demonstrates its ability to identify optimal rental properties based on user preferences such as maximum price, acceptable travel time, and other key metrics. The system achieves moderate prediction accuracy with a Root Mean Squared Error (RMSE) of 555.94, indicating its effectiveness in estimating rental prices. While its performance is reasonable, it could benefit from further hyperparameter tuning and data expansion to enhance its accuracy.

A key strength of the model is its ability to prioritize user-centric metrics like affordability and accessibility. The importance of features such as price_per_min_transit and walking_time_min reflects its alignment with practical preferences, particularly for renters seeking proximity to Wentworth Institute. The system filters properties based on constraints, such as maximum price and travel time, and ranks options effectively by balancing cost and convenience.

The recommendations generated by the system integrate user-defined constraints with property-specific attributes like avg_zipcode_price and neighborhood_travel_score. This ensures that renters receive tailored suggestions for properties that meet their specific needs. Despite its strengths, the system does have limitations, including a reliance on a static dataset and challenges with missing data in critical metrics. Real-time updates to property listings and additional data inputs could improve the system's scalability and reliability.

The machine learning-based recommendation system successfully identifies the best properties by leveraging data-driven insights and user-defined preferences. Its ability to balance affordability and accessibility makes it a valuable tool for renters, particularly in urban settings where trade-offs between cost and convenience are critical. The system highlights the power of integrating machine learning with real-world property data to deliver actionable recommendations.

V. Discussion

The discussion section critically evaluates the limitations of the project while proposing improvements and future directions. The primary limitation lies in the accuracy of the model, as indicated by a Root Mean Squared Error (RMSE) of 555.94. Although this level of error is moderate, fine-tuning the model through hyperparameter optimization, such as increasing the number of trees or adjusting maximum features in the Random Forest Regressor, could enhance its predictive capabilities.

Another key limitation is the dataset itself, which relies heavily on static data and API-generated travel distances. This may not fully capture temporal variations or real-time factors influencing rental property preferences, such as fluctuating demand or seasonal variations. Extending the dataset to include temporal data or dynamically updated property listings could better align the system with real-world scenarios.

The scalability of the project is also constrained by its focus on Wentworth Institute of Technology students in Boston. While the system performs well in this localized context, applying it to other cities or broader user groups would require substantial adjustments, such as integrating additional property attributes or recalibrating the travel-time metrics for new geographic areas.

To address these challenges, future work could focus on enhancing feature engineering by incorporating additional variables, such as amenities scores, real-time travel data, or user reviews. This would provide a more comprehensive view of property characteristics and further personalize recommendations. Advanced modeling techniques, including deep learning frameworks or hybrid models, could also be explored to improve prediction accuracy.

Lastly, developing a user-centric application, such as a mobile app, would make the system accessible to a broader audience. By leveraging OpenAI's APIs and the current recommendation system, such an application could provide dynamic and personalized property suggestions to renters nationwide, ensuring the scalability and usability of the project in diverse contexts.

VI. Conclusion

This project developed a data-driven recommendation system to assist Wentworth Institute of Technology (WIT) students in finding rental properties that balance affordability and convenience. By integrating property attributes, travel times, and rental prices, the study highlighted the importance of location and travel accessibility in determining rental costs. Properties closer to WIT, especially within a 15–30-minute travel range, were more desirable, reflecting the trade-off between proximity and affordability.

The Random Forest Regressor model effectively predicted rental prices, with price per minute of transit and walking time identified as the most influential factors. The system enabled personalized recommendations based on user preferences, streamlining the property selection process.

This framework not only provides valuable tools for students but also demonstrates the potential of data analysis and machine learning in solving real-world housing challenges. Future improvements could include real-time data integration and additional features like neighborhood amenities to enhance its accuracy and scope.

Acknowledgment

I would like to express my gratitude to Professor Weijie Pang for her invaluable guidance and support throughout this project. Her insights and encouragement were instrumental in shaping the methodology and achieving the objectives. Additionally, I extend my thanks to Wentworth Institute of Technology for providing the resources necessary for this study.

References

[1] Bunsly, "HomeHarvest Library: Collects real-time rental property data in Boston, including details such as price, property type, and availability," GitHub Repository, accessed Dec. 7, 2024. Available: https://github.com/Bunsly/HomeHarvest

[2] Google Developers, "Google Distance Matrix API: Calculates walking, transit, and car travel times from each rental property to Wentworth Institute of Technology," API Documentation, accessed Dec. 7, 2024. Available:https://developers.google.com/maps/documentation/distance-matrix/overview

[3] Scikit-Learn, "Python library for machine learning model development, used for creating and evaluating the Random Forest Regressor model," Documentation, accessed Dec. 7, 2024. Available: https://pypi.org/project/scikit-learn/

[4] Matplotlib & Seaborn, "Python visualization libraries used for creating graphs and analyzing relationships in the dataset," Matplotlib Documentation, Seaborn Documentation, accessed Dec. 7, 2024. Available: https://mode.com/blog/python-data-visualization-libraries

[5] Pandas, "Python library for data manipulation and preparation, extensively used for cleaning and feature engineering," Documentation, accessed Dec. 7, 2024. Available: https://pandas.pydata.org/

[6] Python Requests Library, "Used for fetching travel time data via API calls from the Google Distance Matrix API," Documentation, accessed Dec. 7, 2024. Available: https://pypi.org/project/requests/

[7] Hamidi, S., & Ewing, R. "How Affordable Is HUD Affordable Housing?" Journal of Housing and Urban Development, vol. 8, no. 2, pp. 1-15, 2015. Available: https://www.tandfonline.com/doi/full/10.1080/10511482.2015.1123753