A Proposal for Data-Driven Insights for Renter-Landlord Matching

Foster Lockerbie, Yuzhu Han, Jingran Zhao, Litao Zheng

May 9, 2025

Master of Data Science Capstone

University of British Columbia

Introduction

Happipad exists to provide alternative avenues for access to housing opportunities for both renters and landlords, with a focus on affordability and accessibility. Part of the challenge for equitable housing is to provide a reasonable price. Additionally, the price must be high enough that a landlord is satisfied with their investment and continues the rental relationship.

Because estimating a reasonable price can be difficult, it can be useful to employ machine learning strategies. Various methods have been developed to use machine learning to predict home prices, as this industry is lucrative. Other difficulties also arise from the fact that there are many variables affecting home prices: geographical location, proximity to amenities, various house features such as number of bedrooms and square footage, overall economic circumstances, and many more. Although most existing research is focused on prices of home sales, we believe that using similar strategies would work effectively on predicting rental prices as similar factors contribute to rental prices as selling prices.

A previous study by Ho, Tang and Wong (2021) tested the effectiveness of support vector machine (SVM), random forest (RF), and gradient boosting machine (GBM) in solving this problem. In testing these three methods on a sample of 40,000 housing transactions over 18 years in Hong Kong, they found that RF and GBM had lower mean squared error and mean absolute percentage error than SVM, suggesting these two strategies can make more accurate predictions.

RF algorithms have also been compared to more simplistic methods such as linear regression and ridge regression, and RF has displayed lower mean absolute error (Koktashev et al., 2019). This suggests that RF methods are better equipped to handle home pricing problems, possibly due to the fact that these problems are highly complex and methods such as linear regression are too simple to handle them effectively.

Another study compared the predictive capabilities of XGBoost, support vector, RF, multilayer perceptron, and multiple linear regression algorithms on house prices (Sharma, Harsora & Ogunleye, 2024). They found that XGBoost had the lowest mean squared error. This suggests there may be some value in us investigating XGBoost algorithms along with RF algorithms.

A second component of our project is dashboarding. Dashboards are an intuitive way for us (data experts) to present data to company members who wish to gain insights from the data but are not necessarily experts in handling data. There are various methods for creating dashboards, but our group has chosen to use PowerBI. This is because it is a free program, and provides ample flexibility in how to represent the data and is intuitive to use to create attractive dashboards. Happipad would like to receive a dashboard that helps them understand the trends within their data regarding regional rental markets. This includes average price, proportion of listing in each type of home (apartment, house, etc.) and more descriptive statistics. We will provide dashboards that provide this information across time at both a provincial level and within specific cities in the dataset.

Overall, we aim to help Happipad utilize their data to understand user preferences and market trends across cities, and how this translates to renter and landlord outcomes, using visual aids like dashboards and predictive modelling algorithms. This will help landlords list their properties with optimal pricing, and it will help Happipad understand where user demand lies within their platform.

Objectives

The following outlines the main objectives of this project:

- Identify the key factors that drive successful renter-landlord matches using historical booking data.
- Predict the best-matching properties for each renter based on their stated preferences.
- Develop predictive models to predict the optimal housing price for landlords based on property features and location, supporting data-driven pricing strategies.
- Uncover meaningful insights into daily rental activity and renter preferences through a dashboard.
- Provide insights into property characteristics and user behaviour on the platform using the interactive dashboard, supporting better business decisions and platform improvements.

Data Sources

At this stage, we have not yet been provided with the actual dataset. Therefore, we base our understanding on the project proposal shared by Happipad and the information available on its official website.

The expected dataset is anticipated to include approximately 238,700 users, 3,500 host listings, and associated booking records. Potential variables may include property listing details, user preferences, booking information, registration dates, seasonal trends, postal codes for both renters and hosts, browser and device types, and rental prices.

Each property listing is expected to provide categorical details such as property type (e.g., house, apartment), location, available amenities, pet allowance, and number of rooms. To ensure confidentiality, sensitive information such as usernames will be replaced with synthetic data while maintaining the utility of the dataset. According to the provided proposal, approximately 10% of the dataset is expected to contain missing values, which will be addressed during the data preparation phase.

Methods

1. **Data preparation**

The preparation phase will involve two key steps:

- Normalization of Variables:

Some continuous variables can be normalized to a standard scale to ensure comparability across features and to avoid dominance of features with large numeric ranges in subsequent analyses.

- Weighting of Variables:

Since certain features (e.g., location match, price compatibility) may be more critical for matching than others, we can apply a weighting scheme to reflect their relative importance. We will either use feature importance scores from models to define these weights or ask the partner for instructions.

2. **Statistical analysis**

- Exploratory Data Analysis (EDA)

We will conduct a comprehensive EDA to understand the distribution, variability, and relationships within the dataset, including descriptive statistics and visualization of feature distributions. The result can help identify potential mismatches or patterns and guide the modelling and dashboard design.

- Feature analysis and modelling

We will analyze how renters' stated preferences align with the properties they ultimately book, focusing on which features show high match rates and which renters are more willing to compromise on. This will guide our predictive modelling, where we will develop models to estimate rental price ranges and identify key factors influencing booking decisions. By interpreting model outputs, we aim to highlight the most important features shaping rental outcomes.

3. **Dashboard Development and Visualization**

We will develop a Power BI dashboard that can be used to monitor daily rental activity and uncover insights into property features and user behaviour. The dashboard will present key metrics and trends, enabling interactive exploration and supporting data-driven decisions.

Deliverables

1. **Dashboard:**

Due to the uncertainty of business requirements and the pending delivery of the dataset, we have considered the following two directions for dashboard design:

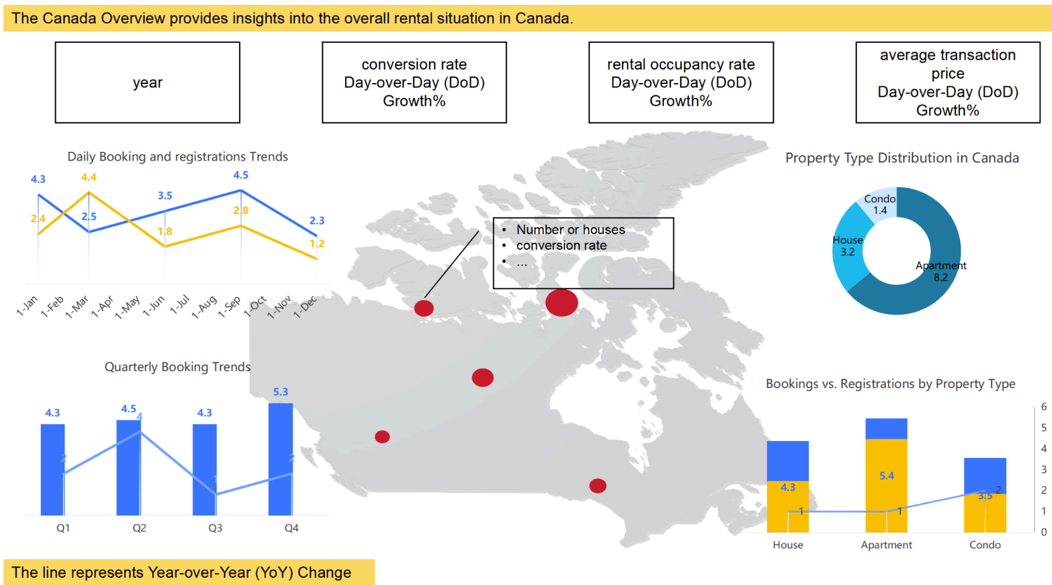
- **Direction 1: Rental Situation Monitoring**

Purpose:
To monitor rental activities across Canada from both a national and provincial perspective.

Structure:

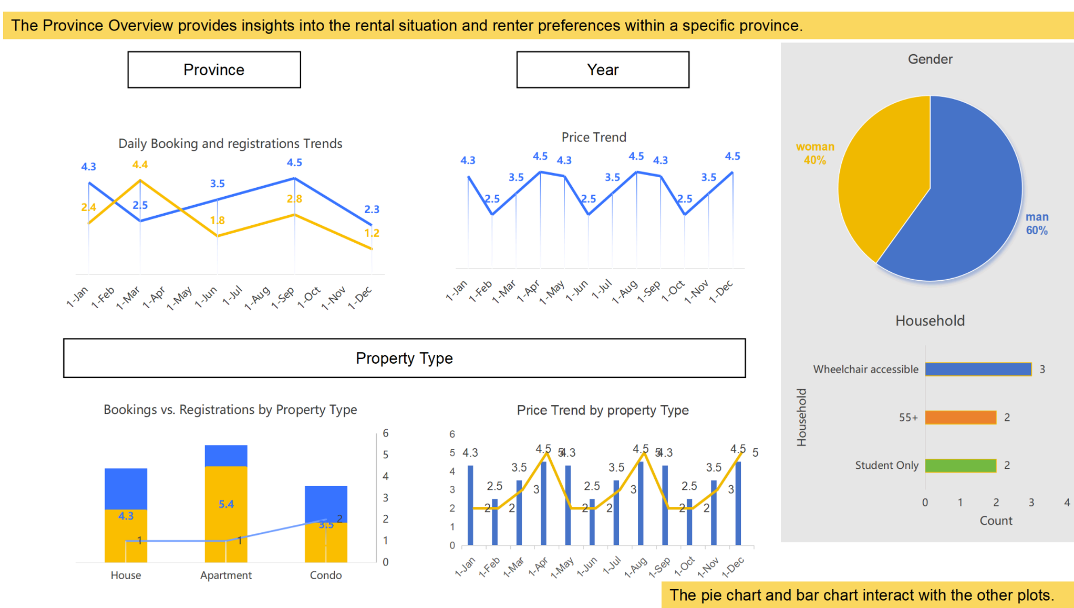
- Canada Overview:

Analyze rental trends at the national level on a daily basis. Detect anomalies, such as sudden drops in rental prices or booking volumes during peak rental seasons. For example, identify if there is a sharp decline in daily bookings despite seasonal demand.



- Provincial View:

Investigate provinces that show unusual patterns. If a province experiences a decrease in rental activity, explore potential local reasons, such as a decline in the student population or changes in regional demand.



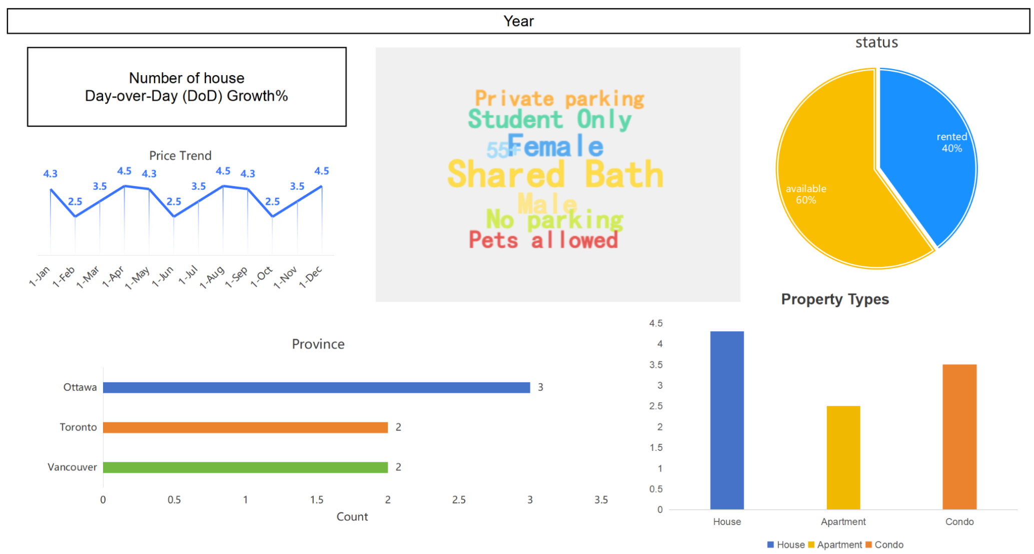
- **Direction 2: Property and Renter Monitoring**

Purpose:
To track the status of listed properties and renter behaviour on the platform.

Structure:

- Property Overview:

Monitor property-related metrics, including: total number of listings, distribution of property types, rental performance indicators, and price trends over time. The interactive function can allow filtering by province, property type, and year to support comparative and segmented analysis.



The Property Overview displays the distribution and types of properties. The province and property type charts are clickable and interact with the other

- Renter Overview:

Analyze renter-related data, including: demographic characteristics (e.g., age, gender), preference behaviour (e.g., preferred property types, locations), booking behaviour (e.g., frequency, duration, conversion).

2. **Final Project Report and Presentation**

A comprehensive document and slide deck summarizing the entire project, including the data preparation process, exploratory data analysis findings, insights into renter preferences and compromises, predictive modelling results, and key interpretations. This report will integrate all analyses into a clear, cohesive narrative for the partner.

3. **Supporting Materials**

If required, we will provide the cleaned dataset and Python code scripts used for the analyses and modelling to ensure reproducibility.

Project Plan

The project will run from May 5, 2025, to June 24, 2025. The midterm presentation is scheduled for June 3, and the final presentation will be on June 24. Group members are expected to work five days a week, with a weekly meeting held every Tuesday for progress update. The general project timeline is as follows:

WBS NUMBER	TASK TITLE	TASK OWNER	START DATE	DUE DATE
1	**Project Initiation**			
	Group Charter	Foster	05/05/25	05/06/25
	Project Proposal	Yuzhu	05/07/25	05/09/25
2	**Analytical Development**			
	Data cleaning	Litao	12/05/25	14/05/25
	Exploratory Data Anlysis	Foster	14/05/25	16/05/25
	Dashboard Creation	Jingran	15/05/25	11/06/25
	Predictive Modelling	Yuzhu	19/05/25	11/06/25
3	**Project Deliverables**			
	Midterm Presentation	Litao	28/05/25	03/06/25
	Final Presentation	Jingran	16/06/25	24/06/25
	Final Report	Yuzhu	12/06/25	24/06/25

Limitations

- The data only includes rented properties and completed bookings, which may exclude properties that were listed but never rented. This can lead to a bias in modelling renter preferences and predictions of matching the renters and houses.
- Not all user preferences may be explicitly captured in the dataset. Implicit preferences or preferences not recorded. For example, the renters may not list all of their requirements. Many completed bookings are the result after negotiation.
- The dataset contains missing information. This may affect the model's performance and lead to a bias in predictions.
- Rental trends may change over time due to external factors such as economic shifts, seasonal effects, or changes in platform policies. Our models may not fully generalize to future data without frequent updates.

Conclusion

This project aims to enhance the housing rental experience and match renters with the appropriate housing for rent by predictive models to support both renters and landlords. Moreover, the dashboard is made for internal use to visualize the status of the rental activities. Through predictive modeling, such as support vector machine (SVM), random forest (RF), XGBoost, and gradient boosting machine (GBM) listed above from previous research papers, and an interactive dashboard, we strive to help renters find listings that align with their preferences, assist landlords in setting competitive and data-informed prices, understand occupancy and pricing trends across locations and seasons, and explore the dynamics of successful renter-landlord matching.

Despite limitations, such as missing data and data only containing completed bookings, this project provides a strong foundation for future development and implementation. The final deliverables, including a dashboard, predictive models, and analysis, will offer actionable insights and demonstrate how machine learning and statistical methods can optimize outcomes in the many-to-many rental housing market.

References

Ho, W. K. O., Tang, B.-S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research, 38*(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558

Koktashev, V., Makee, V., Shchepin, E., Peresunko, P., & Tynchenko, V. V. (2019). Pricing modeling in the housing market with urban infrastructure effect. *Journal of Physics: Conference Series, 1353*(1), 012139. https://doi.org/10.1088/1742-6596/1353/1/012139

Sharma, H., Harsora, H., & Ogunleye, B. (2024). An optimal house price prediction algorithm: XGBoost. *Analytics, 3*(1), 30–45. https://doi.org/10.3390/analytics3010003