Unlocking Shared Housing Supply in Canada: Happipad Capstone Project

Yuzhu Han, Foster Lockerbie, Jingran Zhao & Litao Zheng

Instructor: Scott Fazackerley

University of British Columbia Okanagan Campus

Master of Data Science

June 24, 2025

Executive Summary

Happipad exists to provide alternative avenues for access to housing opportunities for both renters and landlords, with a focus on affordability and accessibility. Our goal for this project was to utilize their data in two ways: to employ data modelling techniques to provide additional insights into their business, and use dashboarding to display geographical demand and temporal changes in rental activities on the Happipad platform.

We developed two dashboards: a Renter Overview and a Property Overview. To assist Happipad in understanding the housing supply, the Property Overview displays the geographical distribution of properties across Canada, and other important features such as the distribution of prices and property types and features in different locations. The Renter Overview helps give insight into housing demand and its changes over time by displaying the distribution of renters across Canada and their budget information.

We also employed machine learning modelling techniques in an attempt to predict the price of rental contracts based on the available data of the homes, such as the city, province, type of bed, and more factors. Happipad can use this information to understand the most important features of homes or listings on their prices, and to notify users if their listings have a price far from the predicted price in order to assist hosts in getting their room rented. Based on the success in previous research of the XGBoost, LightGBM, and random forest models of predicting home prices, we decided to use these models. The best performer was the random forest, with an RMSE of 133 (compared to the average rental price of $835 overall).

To assist the predictive modeling and explore the data further, we applied Ollama for natural language processing (NLP) for the free text property titles and descriptions provided in the dataset. A Mistral model was selected because of its light weight and relatively high efficiency. Highly detailed instructions, in combination with JSON-based outputs, and allowing for high flexibility with analysis of open-text responses led to better extraction of the data. Although this information did not improve model performance, this is likely due to the small sample size, and it has the potential to improve predictions as there is more data collected.

Taken together, we have provided Happipad with an effective pipeline to input their future data to get visual business insights through dashboards, as well as better understand the price distributions of their rental contracts.

Table of Contents

List of Figures

Introduction

Unaffordable housing is a well-known issue across Canada. Various strategies have been employed to tackle this problem, from the government introducing new legislation to private companies offering their own ideas. Happipad exists in an effort to provide a solution. Their goal is to provide an online platform where hosts and renters can find each other and have more confidence that they will be compatible, as Happipad requires users to create comprehensive personal profiles that include their living style preferences. In doing so, their hope is that hosts will be willing to rent out previously unused rooms in their homes, thus unlocking an additional source of housing. Through this project, we seek to help Happipad utilize their data to gain maximum insights, understand their users, and direct marketing actions in the future.

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. Difficulties in determining the appropriate price arise from the fact that there are many variables affecting home value: geographical location, proximity to amenities, various house features such as number of bedrooms and square footage, overall economic circumstances, and many more. Because estimating a reasonable price can be difficult, it can be useful to employ machine learning strategies to handle the complexity. Various methods have been developed to use machine learning to predict home prices, as this industry is lucrative. 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.

Approximating the appropriate rental price for a home based on its features would allow Happipad to notify hosts that their listing is within or outside of the appropriate price range, giving hosts an opportunity to reconsider their listed price if they choose to. Remaining within (or below) this estimated price range would ideally provide benefit to the hosts, as they would likely find renters more quickly and easily. This would also benefit the renters by encouraging affordable prices for them.

Happipad's role in unlocking housing and encouraging affordability is especially important given the high prices of homes compared to wages. According to a national study, the average asking price for shared accommodations in 2024 across Canada was $1010 (Urbanation, 2024). The average price for a studio apartment in 2024 was $2146 (Urbanation, 2024). This difference shows the potential for providing more affordable housing, as if more people rent out their unused bedrooms in their homes, this should provide housing at a lower cost than other options such as studio apartments.

Objectives

The following are the main research questions we seek to answer in this project:

- Can we effectively represent the housing demand and trends through dashboarding?
- Can we predict which home listings will be rented?
- Can we predict the appropriate price for rental listings?
- Can we use any large language models to improve our predictive modeling?

Data Sources

Our data was provided by Happipad in the form of CSV files, after being sourced from their API. All user data was anonymized, so no personal or identifiable information was included such as email addresses or names. These datasets contain information from January 2019 until April 2025. The following tables show the detailed variables in the four main datasets.

Table 1. Columns in the renters dataset.

Data type	**Column name**
Datetime	Registered At (Registration date), Looking For Start (Looking for start date)
Numerical	Budget (in CAD)
Categorical	Profile Status (Active, Inactive), Province, City, Country, Verification Status, Is Deleted, Gender, Prefer Live With, Looking In State, Lease Term
Other	ID, Name, Email, Phone Number, Client Name, Postal Code, Looking In Address, Looking In Postal Code, Number Adults, Number Children

Table 2. Columns in the hosts dataset.

Data type	**Column name**
Datetime	Host Registration date
Identifiers & Status	Host ID, Profile Status, Profile Deleted or not
Contact & Demographics	Host Name, Email, Phone Number, Gender, Self Describe Gender, Disabilities Description
Location	Country, Province, City, Postal Code

Hosting Preferences	Verification Status, Description, Cleaning Frequency, Visit Frequency, Cooking Frequency, Noise Tolerance, Schedules
Other	Client name (Happipad or other organizations)

Table 3. Columns in the properties dataset.

Data type	**Column name**
Datetime	Available from (Properties available date)
Numerical	Bathroom Occupants, Price, Utilities Fees, Total (total rental price),
Categorical	City, Province, Status (draft, deleted, or listed), Property Type
Other	Property ID, Property Title, Description (property description written by the host), Postal Code, Property details (Facilities, Household Items, Furnishings, Bed Type, Safety Features, Amenities, House Rules)

Table 4. Columns in the rental contracts dataset.

Data type	**Column name**
Datetime	Contract start date, end date, signed date, termination date, deadline
Numerical	Room rent (in CAD), room utilities (in CAD)
Categorical	Status (signed, active, expired, terminated), Was accepted (yes, no), city, province, Rental client name (Happipad, Canada Homeshare, Canada Connections, Refugee Housing, Matthew House Ottawa)
Other	Room title (open text response), postal code, home details (contains lists of information about property type, home furnishings and facilities)

Background

To develop our strategies to predict home rental prices, we drew upon previous research on predicting home sale 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.

To explore and understand complex data, data visualization also plays an important role. Today, people can use diverse tools for this such as Season, Echart, Plotly Dash, Tableau, Power BI. Each visualization has its own strengths. One of the most popular data visualization tools is Plotly, which is a free and open-source module in Python. It is commonly used for data visualization across various disciplines, including buildings, public health and statistical analysis (Addepalli, Sindhuja, Gaurav, & Ali, 2023). This indicates that Plotly Dash can be suitable for our dashboard including properties related data. Therefore, we used it to explore property characteristics and market trends in our dashboard.

Another efficient tool is Tableau, because it is easy to use, and has a large number of users (Patel, 2021). In particular, it allows users without development experience to perform basic operations, which means everyone can use it to create dashboards. It is friendly for our group to master it quickly, and it is easy for Happipad to navigate. More importantly, Tableau can easily synthesize large amounts of unstructured data, which is very beneficial for large or messy datasets. Tableau can also draw complex charts with different styles, which can help us make our dashboards more beautiful (Bhombe, Walukar, Thakare, & Kamble, 2019).

For natural language processing tasks, Ollama with the Mistral model was selected because it enables fully local execution, ensuring data privacy and compliance while also supporting GPU-accelerated inference and backend flexibility (Bendi-Ouis, Dutartre & Hinaut, 2024). A recent study shows that the open-source large language models (LLMs) deployed locally can achieve competitive performance compared to closed-source models like GPT-4, while also promoting broader access, enabling collaboration, and supporting diverse applications through techniques such as instruction tuning (Manchanda et al., 2024).

Methods

Data Cleaning

All data cleaning and analysis was conducted using Python. As there are four datasets in this project, each group member was responsible for cleaning and completing the exploratory data analysis (EDA) for one dataset.

1. Renters Dataset:

To effectively monitor renters' information with a dashboard, we standardized the columns for "country," "province," "city," "looking in state," and "looking in address." The country_converter package in Python was used to verify whether a given country name is valid. For province and city standardization, we use a predefined dictionary. Given the highly diverse structure of address formats, each address string is parsed in reverse, matching individual words against a city dataset. This approach helps ensure that the correct city is identified. For example, in the address "Kelowna, BC, Canada," the reverse matching starts with "Canada" (not a city), then "BC" (also not a city), and finally "Kelowna," which is correctly recognized as a valid city and recorded in the "looking in city cleaned" column.

2. Hosts Dataset:

There are many free-text fields in this dataset (e.g. Self Describe Gender, Disabilities Description, Host Preference) exhibiting very high rates of missingness (over 75% missing in some cases), while location fields (Province, Country) also have substantial nulls (~30% of rows).

First, all nulls were filled with empty strings to simplify subsequent text processing. Country names were then standardized using the country_converter (coco) library and any unrecognized entries ("not found") were recorded to "Unknown." For provinces, each entry was lowercased and stripped of punctuation. Direct matches to standard two-letter Canadian provinces abbreviations (e.g. BC, AB, ON) were retained, and fuzzy matching (with an 80-point cutoff) mapped the rest to canonical province names (e.g. "British Columbia" to "BC"), with any unmatched or blank entries set to "Unknown." Whitespace and punctuation cleaning were applied on the city column. Embedded province or country terms were removed by extracting the last word from the single city token. Then the cleaned names were validated against a GeonamesCache list to replace any non-recognized names with "Unknown." Finally, personally identifiable or unnecessary columns (Name, Email, Phone Number, Postal Code, Verification

Status, Self Describe Gender, Schedules) were dropped, and all remaining “Unknown” or placeholder values (‘NA’, ‘NaN’, ‘’, ‘None’) were converted back to true nulls for further analysis.

3. Properties Dataset:

A key part of cleaning the properties dataset involved standardizing the location information, including province, city, and postal code. To ensure consistency, province names were converted to their official two-letter Canadian abbreviations. City names were matched against a Canadian city name dataset from SimpleMaps to correct spelling errors and updated manually to reflect the most recent administrative names. Postal codes were standardized to the official Canadian format (e.g., *V1V 1V8*).

Another important data cleaning step involved handling columns containing multiple values. These columns initially stored comma-separated strings, which were first converted into Python lists. For exploratory data analysis and visualization purposes, these list-based columns were exploded into separate rows. Later, for predictive modelling, they were transformed into binary format to enable effective use in machine learning algorithms.

4. Rental Contracts Dataset:

The rental contracts dataset was cleaned with the goal of making it as accessible and usable as possible for the predictive analyses of room rent prices. The Home details column was expanded into its multiple lists for the furnishings, safety features, amenities, and house rules, and each feature was either binarized or made into a categorical variable. For example, a home either contains a “0” or “1” in the private entry column created. Conversely, for the “Bed type” column, a home contains one of a set number of options, such as “double” or “queen” size. This dataset was relatively clean compared to the hosts and renters data, so many of the remaining columns merely had to be converted to the correct data type (such as datetime object or float). The province names were all standardized to the two-letter acronym to retain consistency with the other datasets. The postal codes were also standardized to all capital letters with no spaces.

Statistical analysis

1. Exploratory Data Analysis (EDA)

We conducted a comprehensive EDA to understand the distribution, variability, and relationships within the dataset, including descriptive statistics and visualization of feature distributions. The results can help identify potential outliers or patterns and guide the modelling and dashboard design. This EDA can also provide inherent value for Happipad as it can reveal insights about their user behaviour.

2. Feature analysis and modelling

We developed a model to predict the price of rental contracts depending on the various features of contracts provided in the rental contracts dataset, such as location, lease term, and bed type. This model was created using algorithms known to be successful in this type of prediction, such as XGBoost, random forest and LightGBM. The best model was selected after searching for its best parameters using a grid search. The model performance was judged by root mean-squared error (RMSE), with a lower RMSE indicating more accurate predictions and thus better performance. After determining the best performer, we extracted feature importances from the model to understand which features of a home are most impactful on its price.

We then applied Ollama with a Mistral model to extract important information from the open-text fields ("Property Title" and "Description"), and added these additional columns to the dataset to enhance the model's predictive ability.

Lastly, we developed a predictive classifier model using a random forest to predict whether each home listed on the properties dataset will become rented or not. This can provide Happipad insight into which types of homes or hosts are deemed desirable by the renters.

Dashboard Development and Visualization

We created a dashboard to visualize the trends of main indicators such as rental price, budget and property status. It captured some insights about property features and renter preference, helping Happipad monitor a daily rental activity and drive decision-making. We used Dash and Tableau to develop the dashboards since each tool has its unique advantages. The first dashboard was created in Dash, as Happipad is mainly familiar with Python-based interfaces. We also created another dashboard in Tableau to provide another option that is more user-friendly.

Results

Exploratory Data Analysis

1. Renters Dataset:

To explore the features of the renter dataset, we generated charts for the distribution of country, city, looking in state, looking in city, budget, and lease term. Among these, the distributions of renter location, looking in province, budget, and lease term are the most insightful.

The dataset contains renters from a total of 82 countries. As shown in the bar chart of country distribution (Figure 1), the majority of renters are from Canada, accounting for 7,155 entries. India ranks second with 187 renters, followed by the United States with 67.

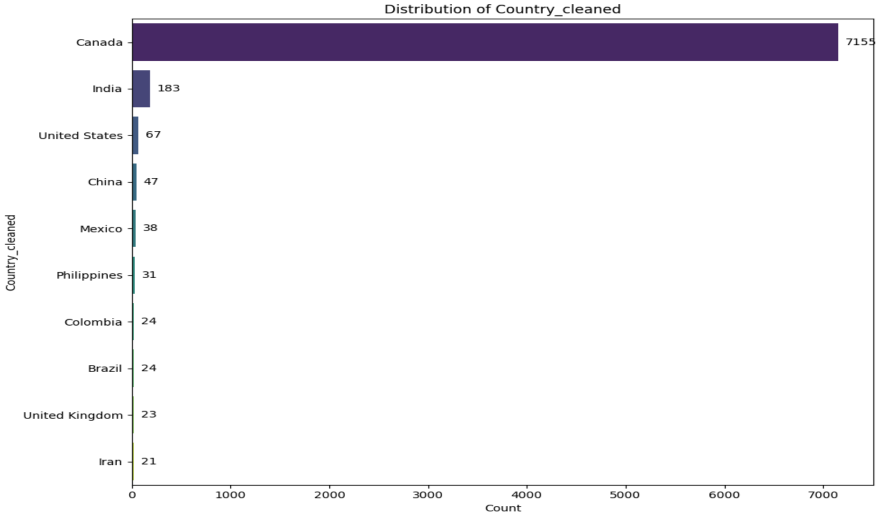


Figure 1. Country distribution of renters.

Regarding the distribution of looking in province (Figure 2), British Columbia has the highest demand, with 7,856 renters seeking properties there. Ontario is the second most popular province, with 2,301 renters. Nova Scotia and Alberta also show relatively high levels of demand.

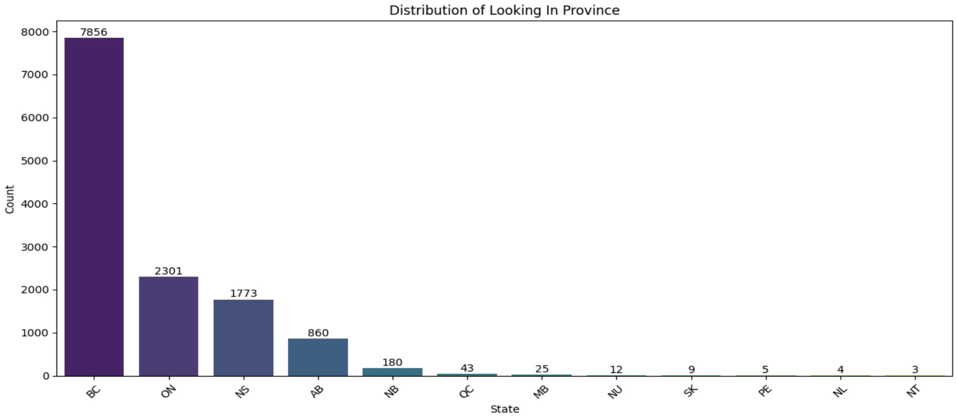


Figure 2. Distribution of locations renters seek to live in.

The budget distribution chart below (Figure 3) excludes outliers, specifically those budgets that are less than $50 per month and those that exceed $2,000 per month. The resulting distribution resembles a normal curve, with a mean of $1,043. The highest frequency occurs around the $1,000 mark, indicating that most renters are willing to pay approximately $1,000 per month for a property.

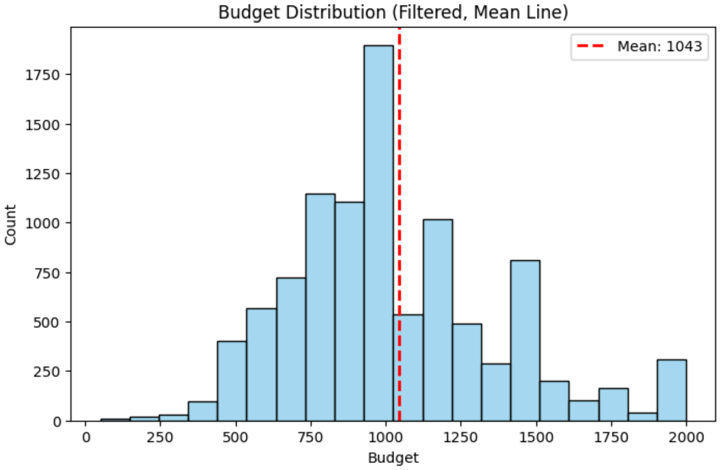


Figure 3. Budget distribution of renters.

The trend of lease terms over the years is illustrated in the line chart (Figure 4). Generally, 1-month leases are the most common. However, in 2021 and 2025, 6-month leases were the most common option. Long-term leases have consistently been the least common option over the past six years.

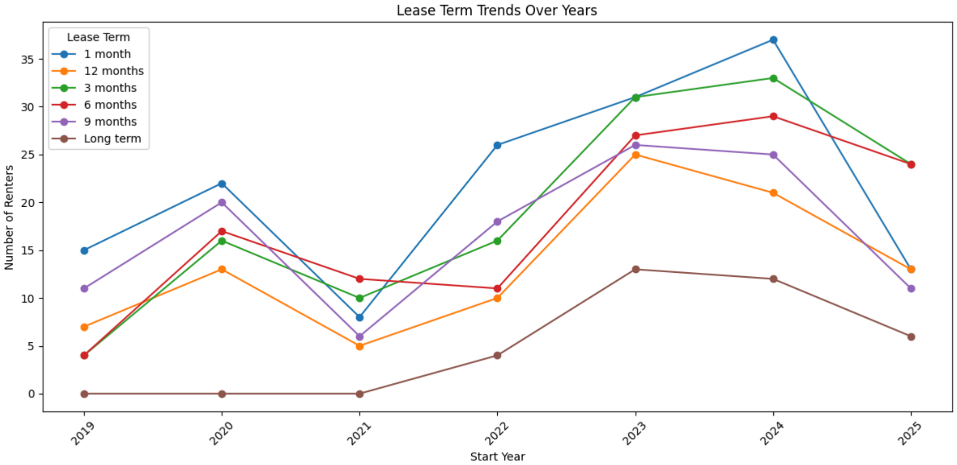


Figure 4. Lease length trends by year.

2. Hosts Dataset:

Since the host dataset only contains useful categorical values, a for loop was used to produce the bar charts of the distribution of each categorical variable, such as the city, roommate gender preference, cooking frequency, and visitor frequency. The following plots are essential in helping us understand the geographical distribution and preferences of hosts. Figure 5 shows the cities with the most hosts. Kelowna stands out with the highest number of hosts, at around 700. This is likely because Kelowna is where Happipad's headquarters are located and initially launched. Halifax comes next with just over 300 hosts, followed by Toronto, Calgary, and Dartmouth. Figure 6 explores hosts' gender preferences for who they are comfortable living with. The majority of hosts selected 'no preference'. Around 1,300 hosts prefer to live with females, while the others selected males or non-binary. There are two more lifestyle variables that could impact host–renter compatibility: visit frequency and cooking frequency. The visit frequency chart (Figure 7) shows that the most common response is 'once a month', followed by every week and never. This suggests that most hosts expect occasional visitors. The cook frequency chart (Figure 8) on the right shows that the majority of hosts allow renters to cook daily.

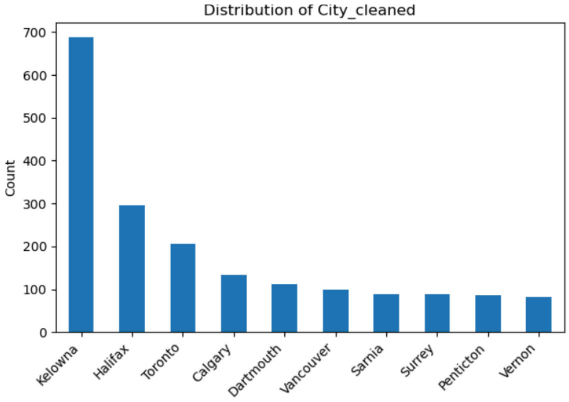


Figure 5. Host city distribution.

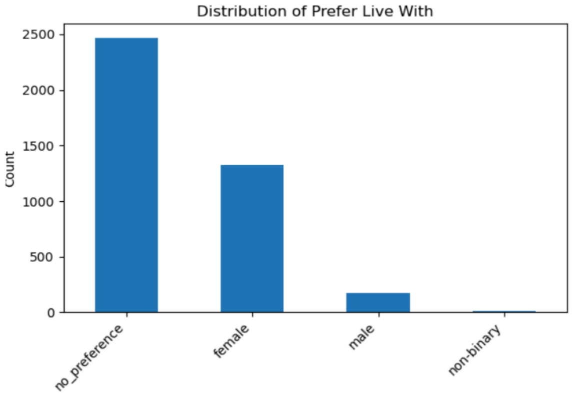


Figure 6. Living preference distribution.

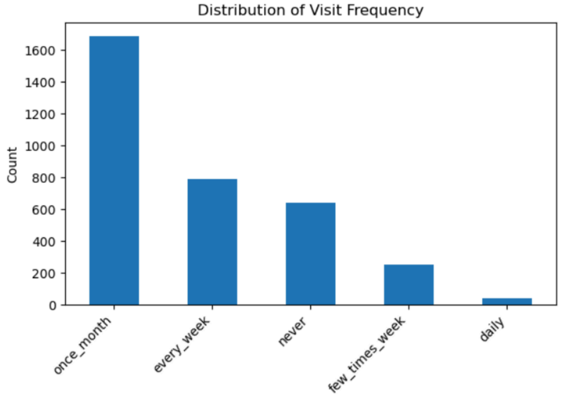


Figure 7. Visit frequency distribution.

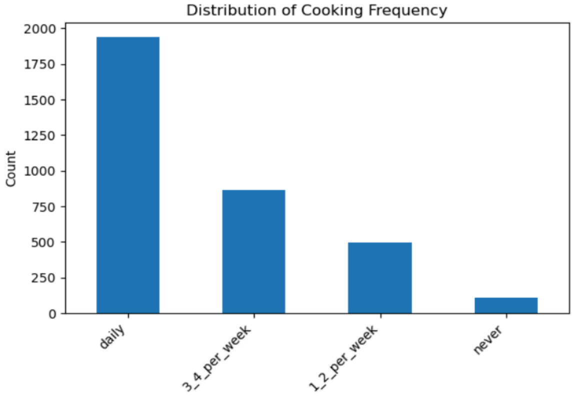


Figure 8. Cooking frequency distribution.

3. Properties Dataset:

The properties dataset was analyzed to better understand the types of rental listings available on Happipad and how they are distributed across Canada. This helped our group to identify key trends in property types, rental pricing, and geographic coverage.

Looking at rental prices, we observed that the minimum listed price is $0 and the maximum is $118,800, which are likely unrealistic and were treated as outliers. To address this, we filtered out listings with prices below $50 and above $3,000 before plotting the distribution shown in Figure 9. Most listings fall between $500 and $1,500, with the distribution skewed to the right. A small number of higher-priced listings push the average rental price slightly above $1,000. This distribution suggests that Happipad currently supports a wide range of rental options, but the majority of listings remain within a relatively affordable range for most renters.

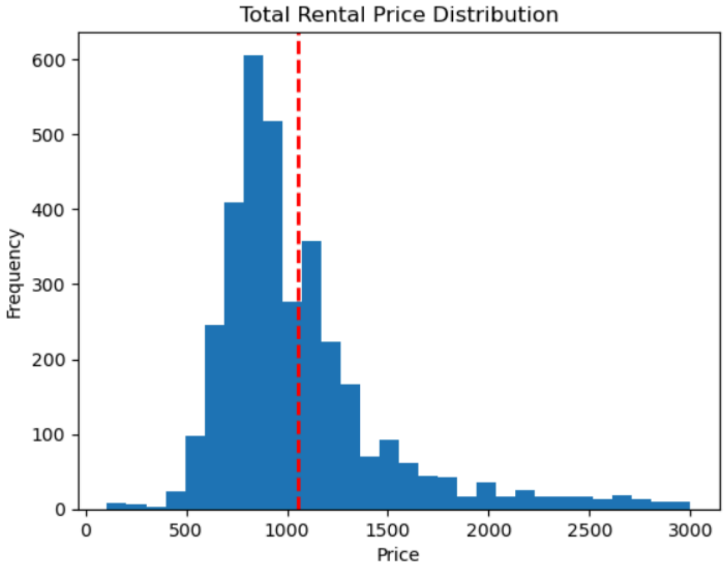


Figure 9. Distribution of rental listing prices.

According to Figure 10, most listings on the platform are categorized as houses, followed by the "other" category, apartments, and townhouses. In contrast, suites and mobile homes appear far less frequently. This suggests that Happipad primarily features smaller or shared accommodations, with fewer listings for larger, independent living spaces.

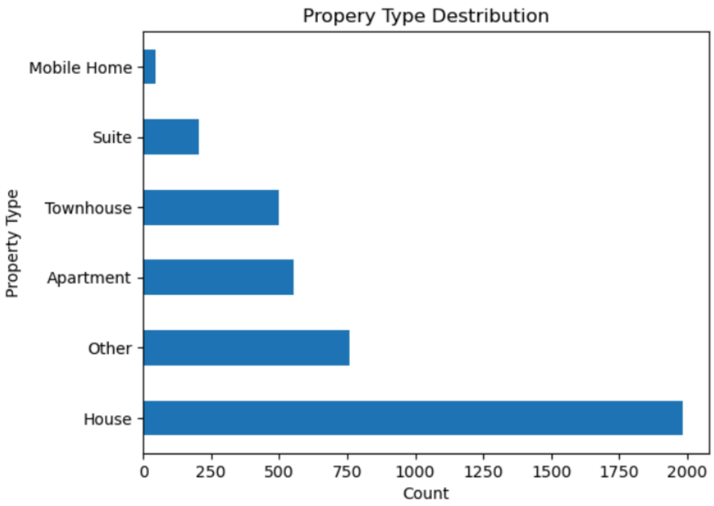


Figure 10. Distribution of home listing property types.

In terms of location, Figure 11 shows that the majority of properties are located in British Columbia (BC), followed by Nova Scotia (NS), Ontario (ON), and Alberta (AB). Listings from other provinces are much less common. British Columbia and Ontario having high numbers of listings likely reflects the high populations in these provinces, but the high numbers of listings in places such as Nova Scotia may indicate that Happipad is providing a new source of housing.

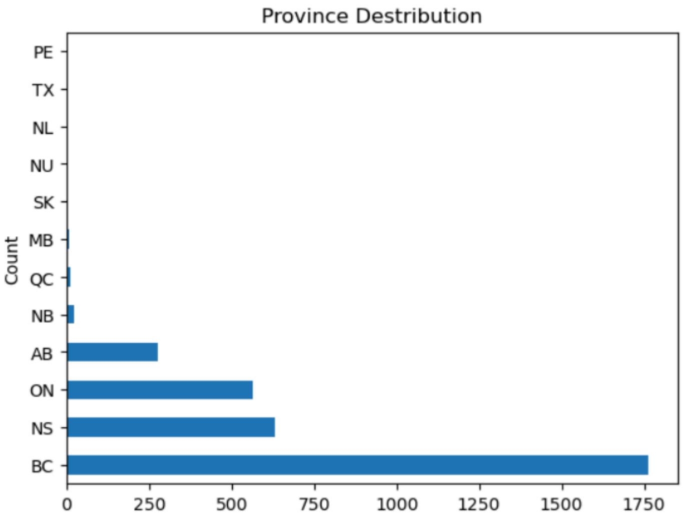


Figure 11. Distribution of home listings across Canadian provinces.

Building on the analysis of geographic distribution, we further examined the types of amenities offered in the top four provinces by listing volume: British Columbia, Nova Scotia, Ontario, and Alberta (Figure 12). Basic amenities, such as a fridge, stove and oven, Wi-Fi, and washer, are provided in nearly all properties, indicating a consistent baseline standard across regions. However, some regional differences were observed. Properties in Alberta are more likely to offer street parking but less likely to include air conditioning, whereas air conditioning is more commonly available in Ontario. Swimming pools are the least common amenity across all four provinces, suggesting that such features remain relatively rare in the Happipad current listings.

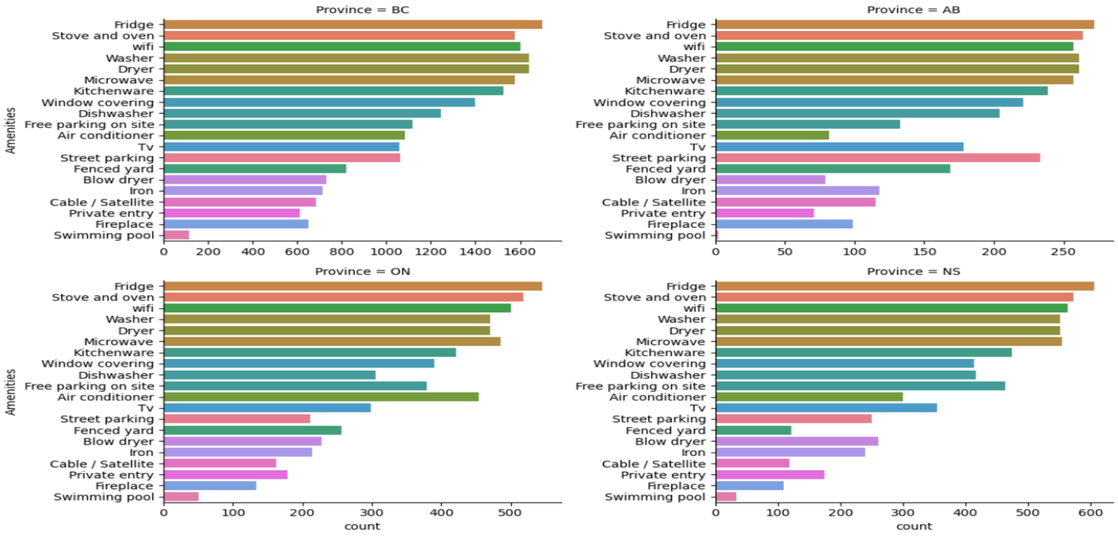


Figure 12. Rental property amenities included in the top four provinces.

4. Rental Contracts Dataset:

We then explored the rental contracts data in order to understand more about the subset of the property listings that have had a rental contract created for them between a renter and host. The aim was to understand more about Happipad rentals and how they compare to the overall rental market in Canada. First, we explored the price distribution of all contracts in the dataset. Figure 13 shows that the average price of a rental contract on Happipad in the last five years has been $835.

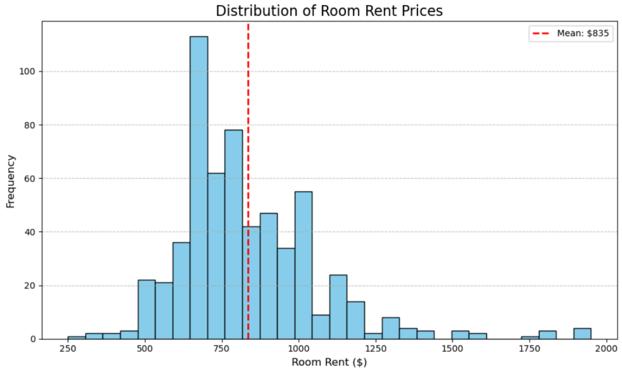


Figure 13. Distribution of room rent prices in contracts.

In Figure 14, we can see that in 2024, the average rental price for a room was $883. Comparing this to the overall national average of $1010 for shared accommodations referenced earlier, we can see that Happipad is indeed providing a source for relatively affordable homes. This is especially apparent if we compare this price to the national average of $2146 for studio apartments.

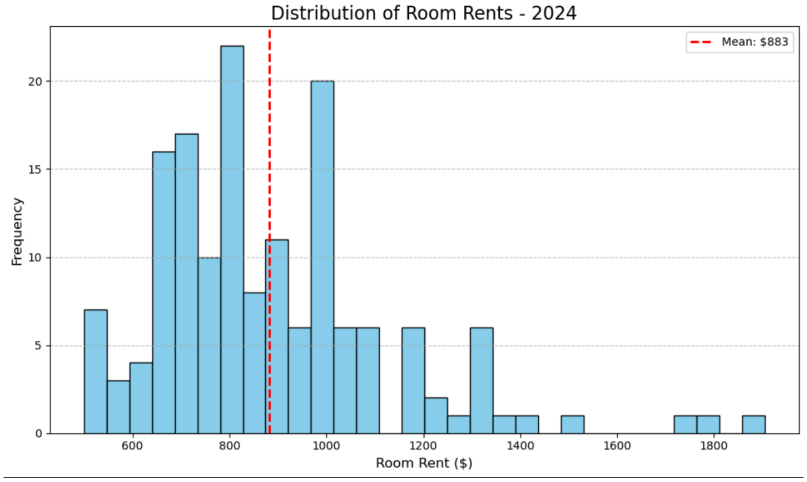


Figure 14. Distribution of room rent prices in Happipad contracts in 2024.

Then, we determined the cities with the highest number of rental contracts from 2019 to 2025 on Happipad. Figure 15 shows the most popular cities for rental contracts, with the highest number starting on the left and descending toward the right. The presence of certain small cities is surprising, such as Sarnia or Drumheller. This could indicate that because there are not as many purpose-built rental buildings in these smaller communities, Happipad is encouraging homeowners to rent out bedrooms within their homes, thus opening up a new avenue of supply.

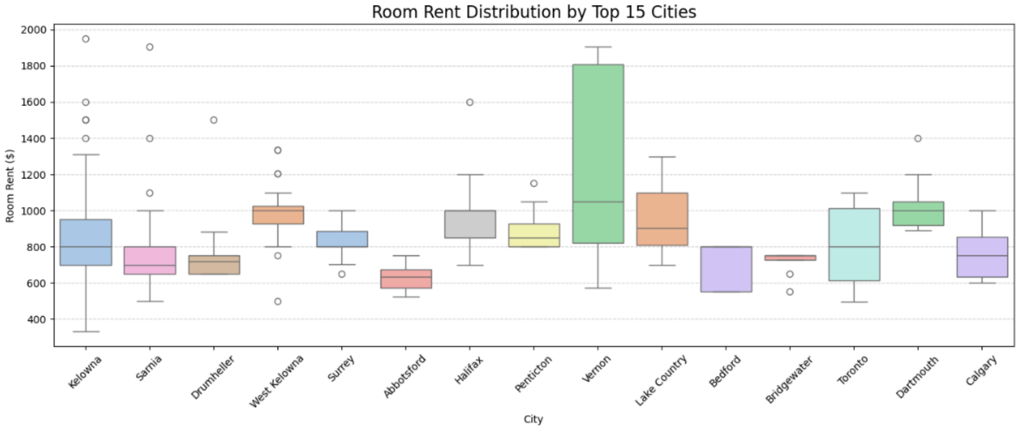


Figure 15. Room rent distributions for each of the top 15 most popular cities on Happipad.

Modeling

We used the rental properties dataset to better understand the rental contracts, and then used modeling to predict the appropriate price for listings based on the other characteristics in the dataset. To accomplish this, we first used Scikit-learn Python packages to split the dataset into a training and a test set, using an 80%-20% split. We then one-hot encoded the necessary features of the data, such as province and property type. We dropped the columns in the training and test data that would be unnecessary or hinder our analyses, such as deadline date and host name.

As we had planned, we first implemented an XGBoost algorithm due to its known high performance in predicting home prices. We performed two different editions of the XGBoost model. For the first model, we applied one-hot encoding to the "city" variable. Then, because there are over 70 cities within the dataset, we created another XGBoost model incorporating target encoding on the city variable instead. This was done with the intention of reducing model dimensionality and potential for overfitting. The XGBoost model with one-hot encoding of the city variable achieved a root mean-squared error (RMSE) of 141.89, while the comparative model with target encoding of the city variable achieved an RMSE of 137.41.

Next, we utilized a random forest algorithm from Scikit-learn to predict the rental contract prices, sticking with target encoding of the city variable because that had the best performance so far. The initial model used common default parameters, with 100 trees. The RMSE of this model was 133.60. After performing parameter grid search with various different numbers of trees, maximum tree depth, minimum number of samples per leaf, and maximum number of features per tree, the highest performing model still had an RMSE of 133.60. From this model, we determined the importance of each feature in the dataset. Figure 16 shows that the year of the contract is the most important feature, which makes sense as rental prices have increased significantly in the past five years in Canada. Other important features include the contract length, city (Vernon, Kelowna, and Sarnia), and wifi inclusion.

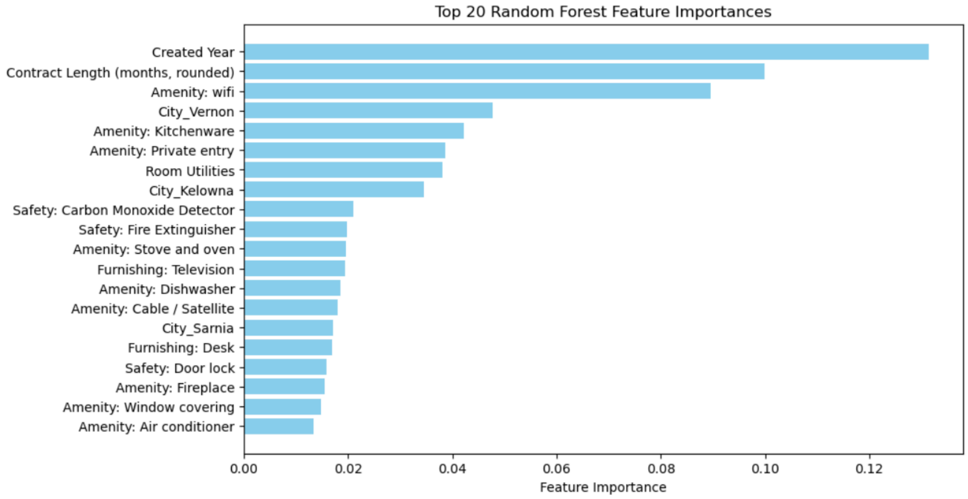


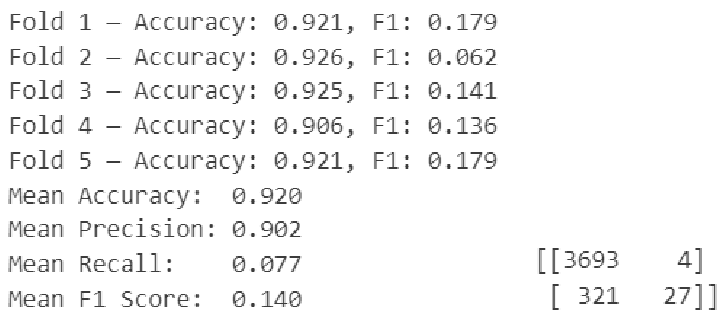
Figure 16. Top 20 most important home features for determining price in rental contracts.

We then fitted a LightGBM model to the dataset to investigate if this model would perform better than a random forest. Similarly to the random forest, a grid search was performed to find the best parameters for the model. After testing various different numbers of trees, leaves, maximum depth of trees, learning rate, and bagging fraction values, the best RMSE achieved was 137.42, which was slightly worse than our best performing random forest model.

We also developed a rental outcome prediction model to estimate whether a listed property is likely to be rented based on its features. The model was trained using the properties dataset, and the target variable was a binary indicator specifying whether the property has an associated rental record in the contracts dataset.

To train and evaluate the model, we implemented a random forest classifier with 5-fold cross-validation. This approach involved splitting the dataset into five equal folds, using four folds for training and one for testing in each iteration, ensuring that every listing was used once for validation. The default parameter with 100 trees and no maximum depth of tree were used. This method helped us assess the model's generalizability and reduce the risk of overfitting.

While the model achieved relatively high average accuracy and precision, the F1 score remained low, suggesting poor overall classification performance (Figure 17). This outcome is primarily driven by class imbalance in the dataset, where out of over 4,000 property listings, only around 360 have been rented in the past. As shown in the confusion matrix, the model produces a large number of false negatives, meaning the majority of rented properties are incorrectly predicted as not rented.



```
Fold 1 – Accuracy: 0.921, F1: 0.179
Fold 2 – Accuracy: 0.926, F1: 0.062
Fold 3 – Accuracy: 0.925, F1: 0.141
Fold 4 – Accuracy: 0.906, F1: 0.136
Fold 5 – Accuracy: 0.921, F1: 0.179
Mean Accuracy:  0.920
Mean Precision: 0.902
Mean Recall:    0.077            [[3693    4]
Mean F1 Score:  0.140             [ 321   27]]
```

Figure 17. Model result and the confusion matrix of rental likelihood classifier model.

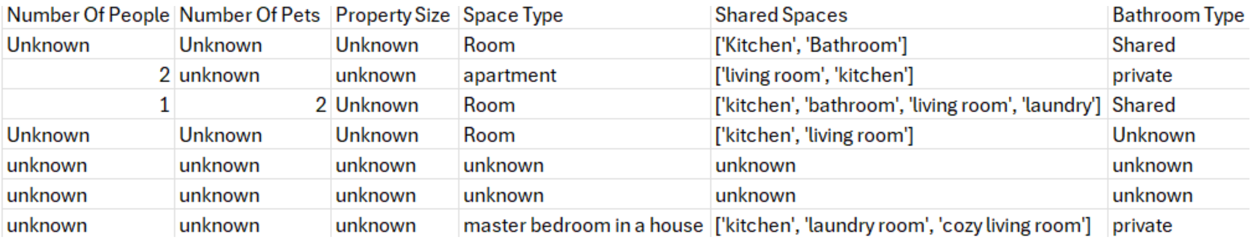
Natural Language Processing

Building on the predictive modelling work, we explored ways to enrich the dataset by extracting additional information from property titles and descriptions using natural language processing (NLP). The aim was to capture unstructured details users entered about their properties, such as shared space availability or nearby amenities, that were not directly available in the provided dataset but could enhance predictive performance and make the dataset more comprehensive.

To achieve this, we used Ollama, a local language model platform that offers a flexible and private environment for tasks such as summarization, classification, and information extraction. Specifically, we implemented the Mistral model, an efficient and open-source LLM known for its strong performance and ability to run on local hardware. We used it to extract structured fields from the unstructured open-text descriptions of properties. The extracted columns include some newly introduced variables, such as number of people, property size, and shared spaces, which are variables that Happipad would like to collect in the future. We also added two more variables, nearby amenities and unique features, as these could provide valuable context for understanding rental listings and could enhance downstream analysis.

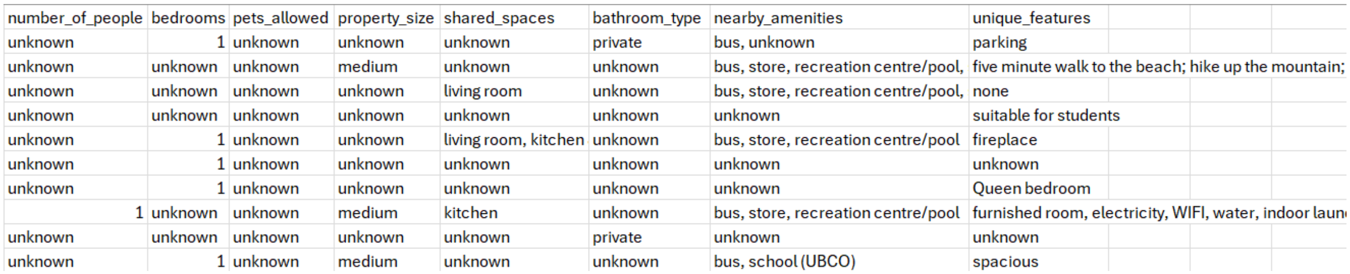
One of the key factors influencing the quality of extraction was the design of the input prompt. In early attempts, we used a CSV-style output format, which led to inconsistent and error-prone results. Switching to a JSON-based output format significantly improved reliability and readability. More importantly, by making the instructions more specific, clearly defining the expected field formats (data types, categorical domains), and allowing flexibility for open-text fields like unique features, we observed a noticeable improvement in both the accuracy and completeness of the extracted data (Figure 18 & 19).

This result highlighted that prompt design plays a critical role in the effectiveness of language model outputs. Well-crafted prompts not only lead to more accurate and structured output, but also reduce the need for manual post-processing, making the extraction pipeline more efficient. These findings offer valuable guidance on how Happipad can implement LLMs like Mistral for future information extraction tasks that enable scalable use of unstructured data to support more informed decision-making.



Number Of People	Number Of Pets	Property Size	Space Type	Shared Spaces	Bathroom Type
Unknown	Unknown	Unknown	Room	['Kitchen', 'Bathroom']	Shared
2	unknown	unknown	apartment	['living room', 'kitchen']	private
1	2	Unknown	Room	['kitchen', 'bathroom', 'living room', 'laundry']	Shared
Unknown	Unknown	Unknown	Room	['kitchen', 'living room']	Unknown
unknown	unknown	unknown	unknown	unknown	unknown
unknown	unknown	unknown	unknown	unknown	unknown
unknown	unknown	unknown	master bedroom in a house	['kitchen', 'laundry room', 'cozy living room']	private

Figure 18. Example extraction output of the original input prompt.



number_of_people	bedrooms	pets_allowed	property_size	shared_spaces	bathroom_type	nearby_amenities	unique_features
unknown	1	unknown	unknown	unknown	private	bus, unknown	parking
unknown	unknown	unknown	medium	unknown	unknown	bus, store, recreation centre/pool,	five minute walk to the beach; hike up the mountain;
unknown	unknown	unknown	unknown	living room	unknown	bus, store, recreation centre/pool,	none
unknown	unknown	unknown	unknown	unknown	unknown	unknown	suitable for students
unknown	1	unknown	unknown	living room, kitchen	unknown	bus, store, recreation centre/pool	fireplace
unknown	1	unknown	unknown	unknown	unknown	unknown	unknown
unknown	1	unknown	unknown	unknown	unknown	unknown	Queen bedroom
1	unknown	unknown	medium	kitchen	unknown	bus, store, recreation centre/pool	furnished room, electricity, WIFI, water, indoor laun
unknown	unknown	unknown	unknown	unknown	private	unknown	unknown
unknown	1	unknown	medium	unknown	unknown	bus, school (UBCO)	spacious

Figure 19. Example extraction output of the best tuned final input prompt.

After finalizing the Ollama extraction method, we attempted to use it to enhance our aforementioned predictive modeling of the rental prices. The bathroom type, property size, and nearby amenities were deemed to be a large enough sample size to include. We then joined the newly created dataset with the existing dataframe used for our modeling to see if these extra features would improve model accuracy. As the random forest model performed the best with our data initially, we applied this model again with the added features. The resulting best RMSE achieved was 140.44, which is slightly worse than the random forest model without these new features. This is likely due to the small sample size of these new features, which could have led to overfitting or poor learning in the model. For example, nearly one third of the bathroom type column responses were unknown after the modeling. Despite this, there is still potential for this analysis pipeline to be useful for Happipad, because as their user base grows, this sample size will increase and it is likely the results will improve over time.

Dashboard

We initially developed the dashboards using Dash, because Happipad is more familiar with Python based frameworks. To provide greater flexibility, we also recreated the dashboards in Tableau, as it can provide more stylistic options and is more user-friendly. Both versions of the dashboard include two main components: a property overview and a renter overview.

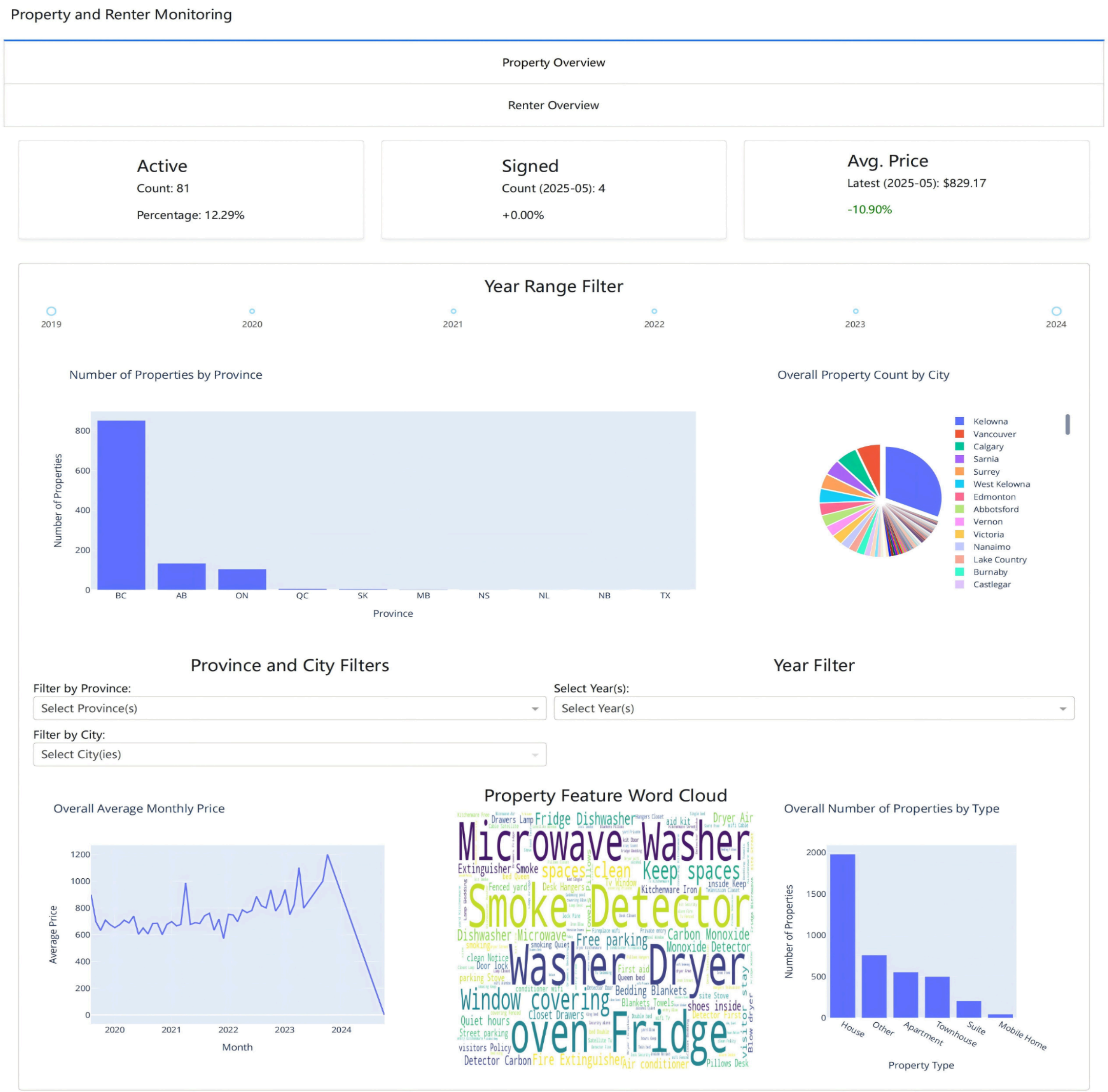
The figures below show the property overview page on Dash and Tableau respectively. The property overview (Figure 20 & 21) consists of three main parts: property status at the top, geographical distribution of properties in the middle, and other property features at the bottom.

Property status includes the number of active properties, the number of signed contracts, and the average monthly rent price. These three indicators reflect the current status of properties on the platform, helping assess whether the property performance is within a normal range. For example, if the number of active properties is significantly lower than in the previous month, it may indicate an anomaly that requires further investigation.

In the middle section, the dashboard displays the distribution of properties by province and by city. A year filter enables us to view property counts for a specific year or a selected range of years. For example, Figure 20 shows the geographical distribution of properties from 2019 to 2025. When we click on a bar in the bar chart, the pie chart automatically updates to show the number of properties in each city within the selected province. From these two charts, we can identify specific areas with the highest and lowest property counts. Combined with the renter overview, this insight helps to determine whether the supply of properties meets the needs of renters.

The bottom section of the dashboard presents the price trend, key property amenities, and the distribution of property types. By default, the data displayed includes information for all properties. The three charts can also be controlled by filters for province, city, and year. Since some city names appear in multiple provinces, the city can only be selected after choosing the province. In addition, the line chart, word cloud, and bar chart operate along a single dimension. This means that when multiple provinces are selected, the charts compare data across those provinces; similarly, when multiple cities are selected, the comparison is made across cities.

There is also an interaction between the property type distribution and the price trend. When we click on a bar in the property type distribution chart (in the right corner), the price trend chart will be updated correspondingly to reflect data of the selected property type. With these three charts, we can monitor the price trend over years, identify the most commonly provided amenities, and understand the distribution of property types. These visualizations help detect abnormal price fluctuations, assess whether the available amenities meet renters' needs, and observe any significant changes in property type distribution.



Property and Renter Monitoring

Property Overview

Renter Overview

Active
Count: 81
Percentage: 12.29%

Signed
Count (2025-05): 4
+0.00%

Avg. Price
Latest (2025-05): $829.17
-10.90%

Year Range Filter

Province and City Filters

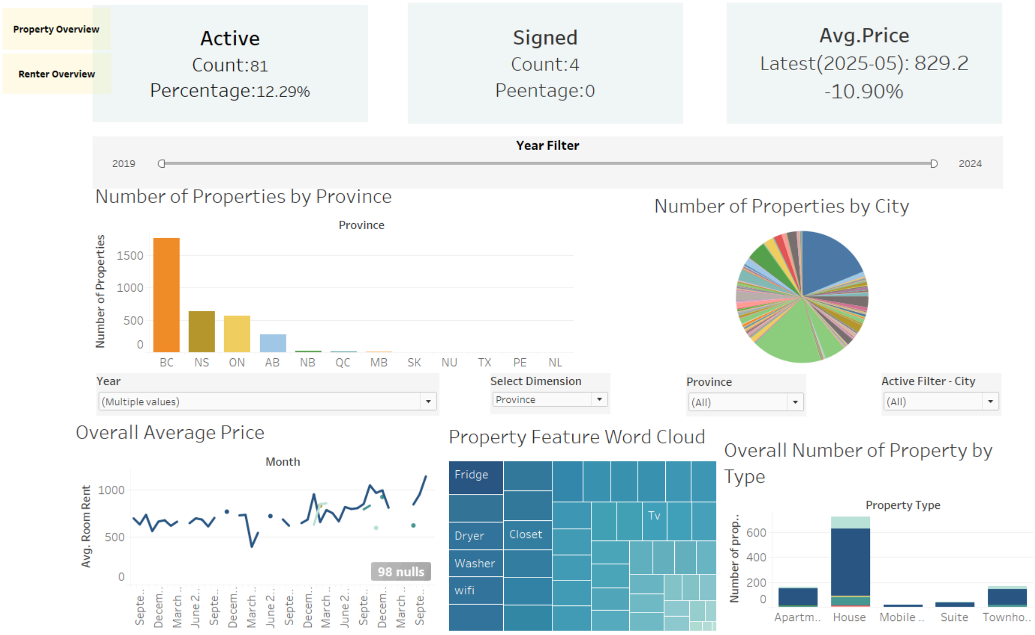
Filter by Province:
Select Province(s)

Filter by City:
Select City(ies)

Year Filter

Select Year(s):
Select Year(s)

Figure 20. Property Overview (Dash Version)


Active
Count:81
Percentage:12.29%

Signed
Count:4
Peentage:0

Avg.Price
Latest(2025-05): 829.2
-10.90%

Year Filter
2019
2024

Year
(Multiple values)

Select Dimension
Province

Province
(All)

Active Filter - City
(All)

Figure 21. Property Overview (Tableau Version)

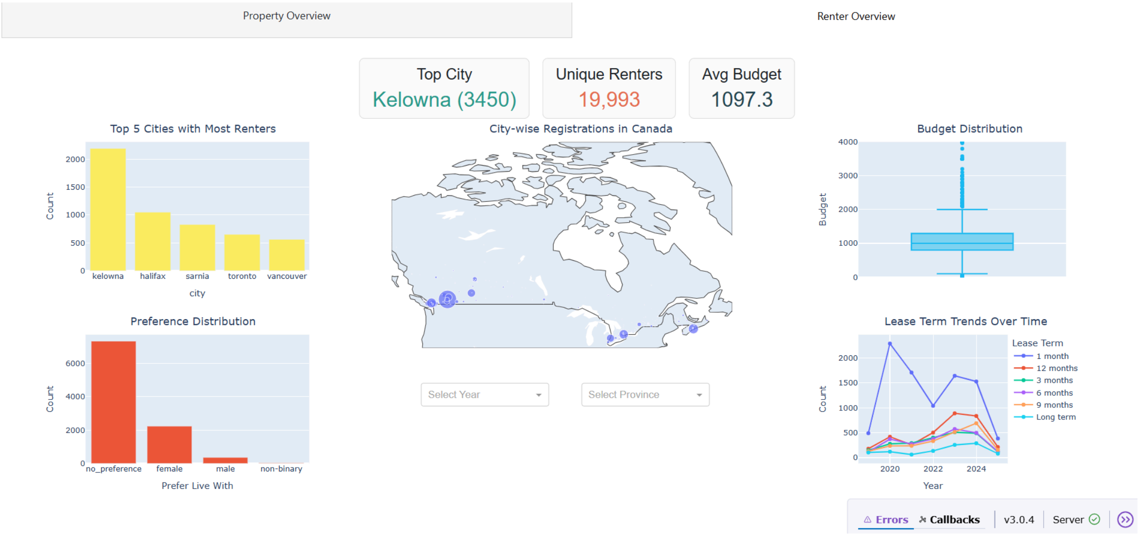
The second page of our interactive dashboard visualizes renter trends and market dynamics across Canada using data from the Happipad platform as shown in Figure 22 and 23. This page provides a comprehensive overview of where renters want to live, how much they are willing to spend, and their rental preferences over time. At the top of the dashboard, three key information cards summarize essential statistics. The total number of unique renters registered on the Happipad platform from 2019 to 2025 is 19,993. The top city where renters want to live is Kelowna, with 3,448 distinct renters. This popularity may be influenced by the fact that Happipad is headquartered in Kelowna. Finally, the average monthly budget across all renters is approximately $1,097 dollars per month.

At the center of the dashboard, a city-level map provides a geographic visualization of renter distribution across Canada. Each dot on the map represents a city, with the size of the dot corresponding to the number of renters. The main hotspots are found in Kelowna, Halifax, Vancouver, Toronto, and Calgary. To the left of the map, a bar chart titled “Top 5 Cities with Most Renters” confirms that Kelowna is the top city that renters wish to live in, followed by Halifax and Vancouver. Below this, another bar chart displays the preference distribution among renters. The majority of users selected “no preference” regarding their desired gender of roommates, followed by females. These preferences and location data can help inform

Happipad's future marketing campaigns as they will understand more of what their users want and where they reside.

To the right of the map, a histogram of budget distribution shows that most renters have a monthly budget below $2,000, with a strong peak around the $1,000 mark. The distribution is right-skewed, including the presence of a few high-budget outliers reaching beyond $4,000. Below this, a line chart illustrates lease term preferences over time, broken down by the year of registration from 2019 to 2025. One-month and one-year leases are the most commonly selected durations across all time ranges, while other lease terms such as 3, 6 months, and long-term are comparatively less selected.

The bottom section of the dashboard includes interactive filter controls that allow users to explore the data in greater detail. These filters include the ability to select specific provinces, years of registration, and lease term durations. By adjusting these parameters, users can gain deeper insights into regional and temporal trends in renter behavior. Overall, the renters overview page provides an informative and interactive summary of renter preferences across Canada, which Happipad can use to monitor their user base and trends over time.

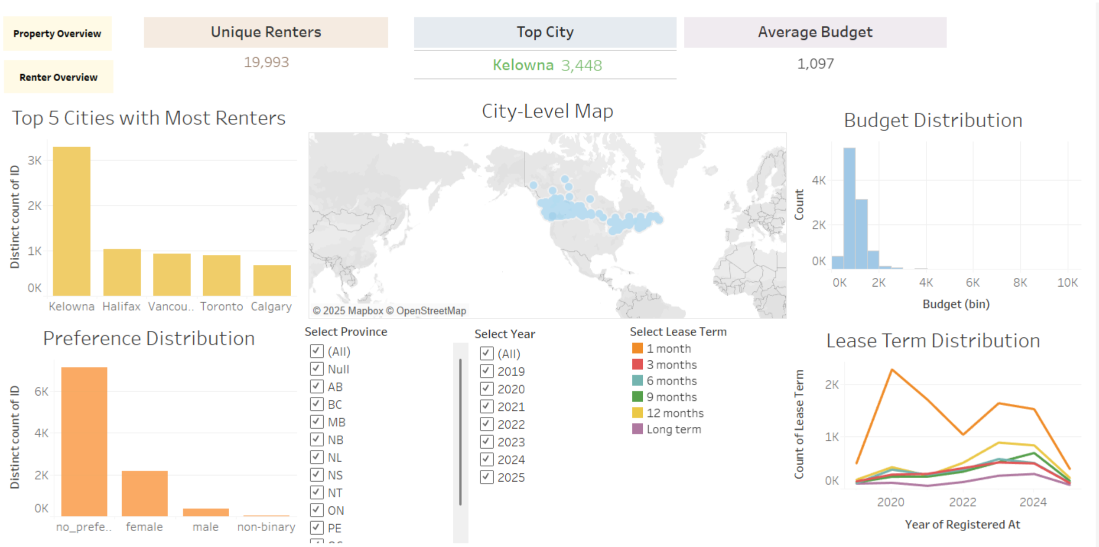


Top City
Kelowna (3450)

Unique Renters
19,993

Avg Budget
1097.3

Figure 22. Renters Overview (Dash Version)



Property Overview

Renter Overview

Unique Renters 19,993

Top City Kelowna 3,448

Average Budget 1,097

Select Province
☑ (All)
☑ Null
☑ AB
☑ BC
☑ MB
☑ NB
☑ NL
☑ NS
☑ NT
☑ ON
☑ PE

Select Year
☑ (All)
☑ 2019
☑ 2020
☑ 2021
☑ 2022
☑ 2023
☑ 2024
☑ 2025

Select Lease Term
🟧 1 month
🟥 3 months
🟦 6 months
🟩 9 months
🟨 12 months
🟪 Long term

Figure 23. Renters Overview (Tableau Version)

Conclusion

To support Happipad in exploring trends in key metrics of renters and properties, we developed interactive dashboards using both Dash and Tableau. The design process began with initial sketches, followed by iterative refinement based on feedback collected from Happipad to ensure the dashboards were aligned with their operational needs. Among various options, the monitoring of renters and properties emerged as Happipad's top priority. The property overview displayed contracts status, properties location and features of properties, while the renter overview showed the geographical distribution, budgets and lease term preferences of renters. These dashboards enabled Happipad to comprehensively understand their platform data, effectively identify key trends in indicators, and detect any anomalies.

Our dashboard's future work will follow two main directions. First, to make our dashboard operate more quickly, we will consider embedding it into a website instead of running it locally, which would require a server. Although free servers are available, using them may violate client confidentiality. As a result, an internal website is better. Moreover, we aim to build a pipeline between data preprocessing and visualization. This would only be possible by understanding more about the Happipad website and data collection structure.

The predictive modeling of the contract prices yielded moderately good results, and provides a solid starting point if Happipad would like to utilize the modeling for identifying if listed homes fall within a reasonable price window. The model's accuracy would likely improve significantly as the dataset grows. Likewise, although the NLP techniques applied did not improve the model at this stage, it is likely due to the small sample size and they could prove to be useful in the future. Future work for modelling could include monitoring the model performance as the dataset grows, and ensuring that performance increases over time, or adjusting model parameters as necessary if performance begins to falter. Lastly, if data collection methods were made more comprehensive, we could develop an effective model to automatically suggest renters to landlords based on their profile compatibility.

References

Addepalli, L., Sindhuja, S., Gaurav, L., & Ali, W. (2023). A comprehensive review of data visualization tools: Features, strengths, and weaknesses. *International Journal of Computer Engineering in Research Trends*, 10(1), 10–20. https://doi.org/10.22362/ijcert/2023/v10/i01/v10i0102

Bhombe, A., Walukar, K., Thakare, Y., & Kamble, S. (2019). Comparative analysis of two BI tools: MicroStrategy and Tableau. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3462539

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

Jiya Manchanda, Laura Boettcher, Matheus Westphalen & Jasser Jasser. (2024). *The Open Source Advantage in Large Language Models*. arXiv:2412.12004 https://arxiv.org/abs/2412.12004

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

Patel, A. (2021). Data visualization using Tableau (Master's thesis, Metropolia University of Applied Sciences). *Theseus Digital Repository*. https://www.theseus.fi/bitstream/handle/10024/652129/Patel_Ashwin.pdf?sequence=4

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

SimpleMaps. (2023). *Canada Cities Database*. SimpleMaps.com. Retrieved May 29, 2025, from https://simplemaps.com/data/canada-cities

Urbanation. (2024). *March 2024 Rentals.ca report.* Rentals.ca https://rentals.ca/blog/march-2024-rentals-ca-report

Yannis Bendi-Ouis, Dan Dutartre & Xavier Hinaut. (2024). *Deploying Open-Source Large Language Models: A Performance Analysis*. arXiv:2409.14887 https://arxiv.org/abs/2409.14887