*Machine Learning-Fall 2024*
*GROUP 13-Calingo, Francis, J.F. De Assis, J. Guo, A. Mohammed, B. Olugbenjo, Y. Wai*

# Canadian Rental Prices and Immigration

Predicting how immigration policy changes will impact rental accommodation costs in Canada for 2025

## Objective

In October 2024, Canadian Prime Minister Justin Trudeau announced a significant reduction in immigration targets previously set for 2025 - 2027, including a 21% decrease in the number of permanent residents (PR) being admitted into the country.[1] For a nation with historically liberal policies and attitudes towards immigration[2] - this pivot in policy was made in the face of ongoing pressure from political counterparts and public sentiment.[3]

As a result of the macroeconomic policy changes during the COVID-19 pandemic to stimulate and preserve the economy, many countries faced increased inflationary pressure resulting in unprecedented increases to cost-of-living.[4] With downstream impacts to the supply chain and labour preventing growth in housing, many Canadians were faced with soaring prices for owning or renting their homes. For rentals specifically, Desjardins bank reported an 8.3% inflation in rented accommodation during Q3 2024, the fastest pace since the 1980s.[5]

As cited by the federal government, in order to relieve major sectors of Canada's infrastructure, such as the rental and housing markets, permanent resident targets will be reduced from half a million admissions to 395,000. The objective of this report was to determine whether or not reductions in immigration targets set out by the Government of Canada will in fact reduce prices in the rented accommodation market. By using supervised learning, and specifically regression to predict rental prices in Canada, this analysis can determine the degree of impact that changes in PR admissions will have on the average cost of renting a home in Canada.

**$H_0$ = "New PR admissions targets will not reduce rental accommodation costs in Canada"**

## Data Preparation

*Data Preparation: What was your data source (e.g., web scraping, corporate data, a standard machine learning data set, open data, etc.)? How good was the data quality? What did you need to do to procure it? What tools or code did you need to use to prepare it for analysis? What challenges did you face?*

[1] https://www.cbc.ca/news/politics/immigration-changes-announcement-1.7360827
[2] https://www.cfr.org/backgrounder/what-canadas-immigration-policy
[3] https://www.environicsinstitute.org/projects/project-details/canadian-public-opinion-about-immigration-and-refugees---fall-2024
[4] https://www.imf.org/-/media/Files/Publications/WP/2023/English/wpiea2023010-print-pdf
[5] https://www.ctvnews.ca/business/rent-inflation-to-slow-in-the-next-few-years-desjardins-predicts-1.7109777

# Sources of Data

There were two major components to data required for this analysis. In order to leverage a supervised machine learning model and regression to predict rental prices with PR admissions as a feature, historical immigration and rental data was required. The final data used in this project can be downloaded here : data_sources.›

## Immigration data

Immigration data rigorously tracked by Federal and Provincial authorities, and is used to forecast future immigration trends. Open Canada, the Government of Canada's portal for open source data stores a historical pivoted dataset with PR admissions results from 2015 to Sept 2024, monthly and by major cities across Canada.

Monthly IRCC Updates - Canada - Permanent Residents by Province/Territory and Census Metropolitan Area (CMA) - Open Government Portal

While the source was comprehensive and high quality, effort needed to be taken in order to unpivot data into a usable tabular format, with one record corresponding to each city for each month.

| Province | City | Date | Admissions | Month | Year |
|---|---|---|---|---|---|
| Alberta | Brooks | 2015-01-01 00:00:00 | 10 | Jan | 2015 |
| Alberta | Calgary | 2015-01-01 00:00:00 | 990 | Jan | 2015 |

*Figure - sample two rows from unpivoted PR admissions dataset*

## Rental data

Rental data was required with a similar geographic granularity to the immigration data. As rental prices can vary greatly across cities, data must be representative of at least major urban areas in Canada. Sourcing high quality rental data was difficult to do from open sources, with most data being kept as proprietary for major real estate companies and organizations like MLS, HouseSigma, and Zillow.

In industry, a common practice has developed to leverage synthetic data[6] to train machine learning models. Synthetic data uses fictitious but often representative data to develop ML models in the absence of real data due to limited availability. This is often done to prevent violations of data privacy, or proprietary knowledge. Additionally, with the advent of publicly accessible LLMs such as ChatGPT, business users can often feed their requirements to such tools and receive an output that is representative of real world data. By using prompt engineering[7] to develop a structured prompt that encompasses the need for realistic data, with a minimum of 10,000 samples across major urban areas, including multiple relevant features often included in rental listings, and representing monthly results from 2025 to 2023, ChatGPT was able to create the following prompt.

[6] https://mostly.ai/what-is-synthetic-data
[7] https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-prompt-engineering

*Generate a realistic dataset of rental prices for major Canadian cities, including Vancouver, Toronto, Montreal, Calgary, Ottawa, Edmonton, and Halifax. The dataset should include:*

1. ***Data Columns**:*
    - ***City**: Major cities like Toronto, Vancouver, Montreal, Calgary, etc.*
    - ***Province**: Corresponding provinces (e.g., Ontario, British Columbia).*
    - ***Year**: From 2019 to 2023.*
    - ***Month**: January to December.*
    - ***Rental Type**: Apartment, Condo, Detached House, Townhouse.*
    - ***Number of Bedrooms**: 1, 2, 3, 4, etc.*
    - ***Number of Bathrooms**: 1, 2, 3, etc.*
    - ***Square Footage**: Ranges for different rental types.*
    - ***Furnished**: Yes/No.*
    - ***Pet Friendly**: Yes/No.*
    - ***Parking Included**: Yes/No.*
    - ***Distance to City Center (km)**: Numeric value.*
    - ***Monthly Rent (Target)**: Dependent variable, with realistic pricing trends.*
    - ***Walk Score: A score between 0 and 100 indicating walkability.***
    - ***Transit Score: A score between 0 and 100 indicating access to public transit.***
    - ***Age of Building: Number of years since the building was constructed.***
    - ***Energy Efficiency Rating: Numeric score (e.g., 0–10).***
    - ***Lease Term: Length of the lease in months (e.g., 6, 12, 24).***
    - ***Noise Level: Numeric score (e.g., 1–10, with 10 being very noisy).***
    - ***Nearby Schools Rating: Average rating of schools in the area (1–10).***
    - ***Internet Availability: Yes/No indicating high-speed internet availability.***
    - ***Crime Rate Index: A score representing the area's safety.***
    - ***Annual Property Tax: Approximation based on rent and location.***
2. ***Realism**:*
    - *Average monthly rent should reflect the general cost of living in each city. For example, Vancouver and Toronto should have higher average rents compared to Edmonton or Halifax.*
    - *Include a range of rental prices within cities to capture variability (e.g., downtown areas vs. suburban neighborhoods).*
    - *Use realistic distributions for rental prices, square footage, and proximity to transit. For instance, apartments should generally be smaller and less expensive than single-family homes.*
3. ***Additional Notes**:*
    - *Include 10,000 rows of data distributed proportionally across cities.*
    - *Reflect seasonality and trends where applicable (e.g., higher prices in Toronto and Vancouver for smaller units due to demand).*
    - *Ensure property types align with city norms (e.g., more condos in downtown Toronto, more single-family homes in Calgary suburbs).*

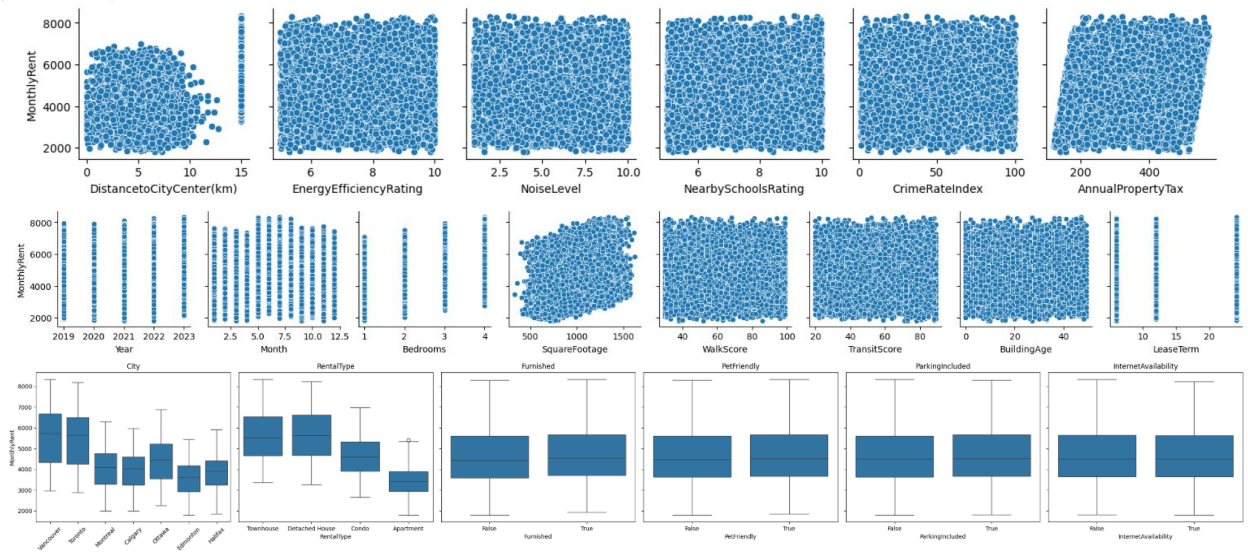
*Figure. Prompt used to generate synthetic data from ChatGPT*

The final prompt was then fed back into the ChatGPT to generate a synthetic dataset for our analysis. Initial inspection of the data showed that the data was compliant to the prompt, with 10,000 records. EDA was still required to see if the synthetic data was usable for the purposes of this analysis, and the data was also meant to be merged with PR admissions data for a completely comprehensive dataset.

## Data Cleaning & Exploratory Data Analysis

Given the need to analyze data from two different sources, we performed several preprocessing steps, including: Standardizing value representations within the same column across both dataframes, and aligning the data types of columns to ensure compatibility.

After these preparations, we explored how various columns impact MonthlyRent. We categorized the columns into three types: Categorical Variables, Discrete Variables and Continuous Variables. We then plotted pairwise comparisons to identify patterns and relationships. Based on our analysis, we selected the following columns for further investigation, as they significantly influence MonthlyRent: City, RentalType, Year, Month, Bedrooms, SquareFootage, and AnnualPropertyTax. These selected columns will be the focus of subsequent steps in the analysis.




## Pre-processing Pipeline

The EDA process identified columns with a high degree of correlation to the dependent variable of 'MonthlyRent'. Once the relevant columns were merged from PR admissions data and rental data, the data was ready to be pre-processed as part of a pipeline in order to feed a machine learning model.

### One hot encoding

The categorical variables of City and Rental_Type were one-hot encoded to create multiple columns with binary numerical values representing the city and type of rental for each record in our dataset. Spare_output was set to FALSE in order to create a condensed array of value that would aide in saving computational memory when training the model.

### Scaling

Standardized Scaler was used on discrete and continuous variables to prevent large scale features from over-influencing the model. This was selected instead of min-max scaling to avoid any influence from outliers in the dataset.

# Model Design

To make our model selection varied, we opted to go for one model that worked best for determining linear relationships, and for the other model that worked best for handling non-

linear relationships. We used the following two supervised learning regression models: Linear Regression and Random Forest Regressor. Other models that we considered include Decision Tree, Support Vector Regression, and Lasso Regression.

However, as these models are either beyond the scope of this course, have a higher risk of overfitting or underfitting, or are not complex enough to have the predicting power to predict our target variable, we did not choose those models.

As we are trying to predict rental prices, that will be our target variable.

## Linear Regression

Following feature engineering, we split our data into training and testing sets (test size=20%). Then, we trained the linear regression model, then made predictions on the training set.

## Random Forest Regressor

The algorithm is virtually the same as for the linear regression model, this time setting model=RandomForestRegressor() with n_estimators=100 and random_state=0.

## Hyperparameter Tuning

Here are the steps for our hyperparameter tuning:

1) We ensured that our categorical columns were consistent (i.e., datatype set as string)
2) Set target variable name (i.e., target_variable='MonthlyRent')
3) Identify and set categorical and numerical columns
4) Exclude target variable from features
5) Separate target and features
6) Define preprocessing for categorical data
7) Define the Random Forest Model
8) Create a Pipeline, then define the parameter grid for tuning:

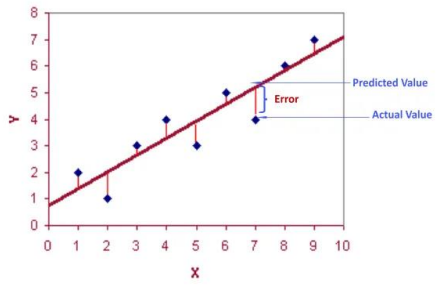
```
# Create a pipeline
pipeline = Pipeline(steps=[('preprocessor', preprocessor), ('model', model)])

# Define the parameter grid for tuning
param_grid = {
    'model__n_estimators': [50, 100, 200],
    'model__max_depth': [10, 20, None],
    'model__min_samples_split': [2, 5, 10],
    'model__min_samples_leaf': [1, 2, 4]
}
```

9) Split the dataset into training and test sets
10) Perform grid search with Cross-Validation
11) Print the best parameters and score, then train the model with said parameters
12) Results:

```
Best Parameters: {'model__max_depth': None, 'model__min_samples_leaf': 1, 'model__min_samples_split': 5, 'model__n_estimators': 200}
Best Negative Mean Squared Error: -32626.113915237056
Test Set Score (R^2): 0.9846805817641215
```

# Model Evaluation

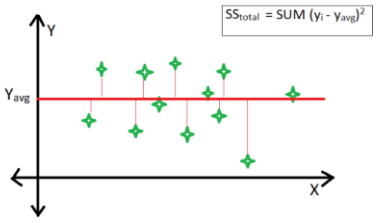
For both models, we focused our attention to two metrics: **Root Mean Squared Error (RMSE)** and the **R-squared ($R^2$) Score**. RMSE measures the average magnitude of the errors between the predicted values and the actual values.

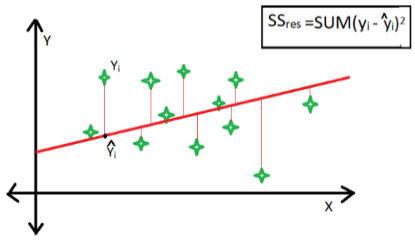



$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

*LEFT IMAGE: A simple example visualizing how RMSE is measured, with the line representing the regression line of a model (source: https://medium.com/@mygreatlearning/rmse-what-does-it-mean-2d446c0b1d0e )*
*RIGHT IMAGE: Formula for RMSE*

$R^2$ Score (also known as the Coefficient of Determination) indicates how well our regression model's predictions fit the actual data (i.e., the proportion of variance in our target variable that can be explained by the features in the model).




$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

*LEFT TO RIGHT IMAGES: Visualization of the $R^2$ Score formula. (source: https://www.geeksforgeeks.org/ml-r-squared-in-regression-analysis/ )*

Therefore, it follows that the model's predictive power will increase as the RMSE value decreases and the $R^2$ Score increases.

*$R^2$ Score and RMSE for Linear Regression*

```
Mean Squared Error: 114799.00900506652
R² Score: 0.9353817159270535
```

*$R^2$ Score and RMSE for Random Forest Regressor*

```
Mean Squared Error: 34611.56304064926
R² Score: 0.9805177777042425
```

It appears that the Random Forest Regressor's $R^2$ score is slightly higher than the Linear Regression Model's $R^2$ score. However, its RMSE is significantly lower than the RMSE for the Linear Regression Model. This leads us to conclude at this stage that the Random Forest Regressors does a better job in predicting our target variable than the linear regression model, implying that the relationships between the target and feature variables are non-linear.

We then carried out cross-validation using k-fold cross validation technique to assess model performance and generalization.

*Cross-Validation Results for Linear Regression*

After encoding categorical features, initializing the linear regression, defining a custom scorer for MSE, performing k-fold cross-validation with k=5, and converting negative MSE to positive for interpretability, here are the results:

```
Linear Regression Cross-Validation Results:
Mean Squared Errors for each fold: [2619180.59845891  153294.52588975  122173.16363917   99712.61513702
  140622.86745715]
Average MSE: 626996.7541164018
```

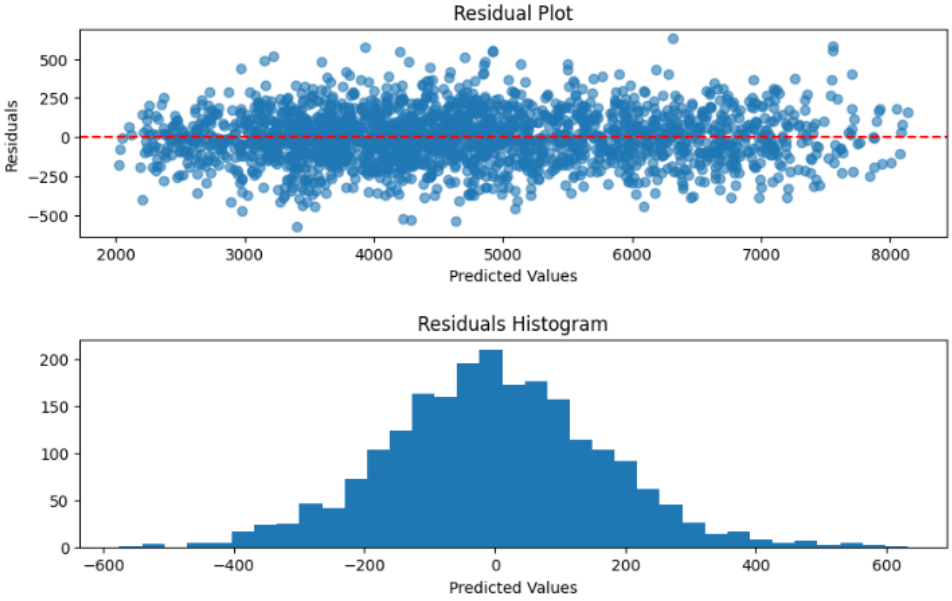
## *Cross-Validation Results for Random Forest Regressor*

The same procedure, but with the Random Forest Regressor, was executed. Here are the results:

```
Random Forest Regressor Cross-Validation Results:
Mean Squared Errors for each fold: [2406430.86870896   99202.60377874  934899.47851753   64722.21715475
   76471.10794809]
Average MSE: 716345.2552216125
```

## *Testing the Best Model*

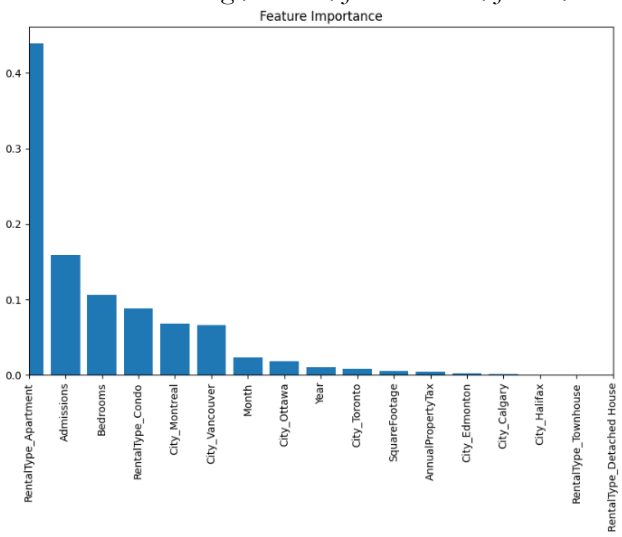
When we plotted the residuals, they seem to be well-distributed along the predicted value.




Using the best_model.score() function, we checked for overfitting. Results show negligible difference between the training and test $R^2$ score, meaning that there is a lack of evidence of overfitting, and therefore the model performs well and the $R^2$ score is exceptional.

```
Training Set Score (R^2): 0.996224777357276
Test Set Score (R^2): 0.9846805817641215
```

After checking for feature importance, we get Rental Type=Apartment being by far the most important feature, even surpassing Admissions, which still remained a very important feature.

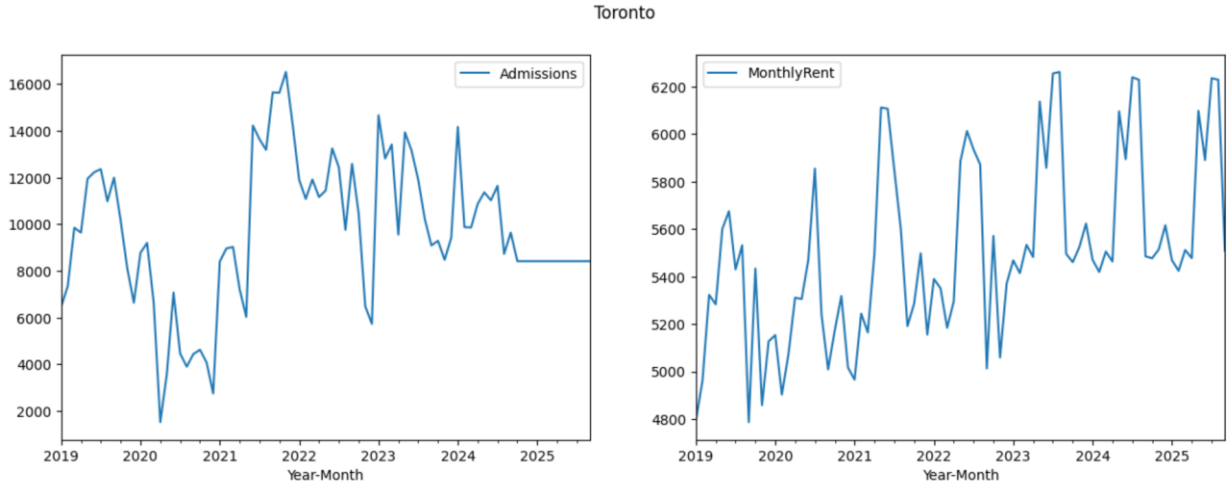



# Model Prediction

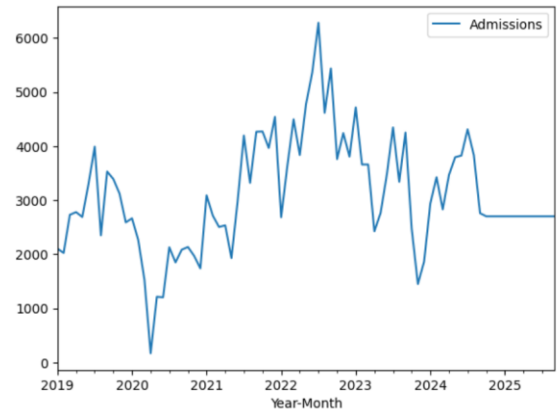
We used our best model to predict rental prices for 2024 and 2025.

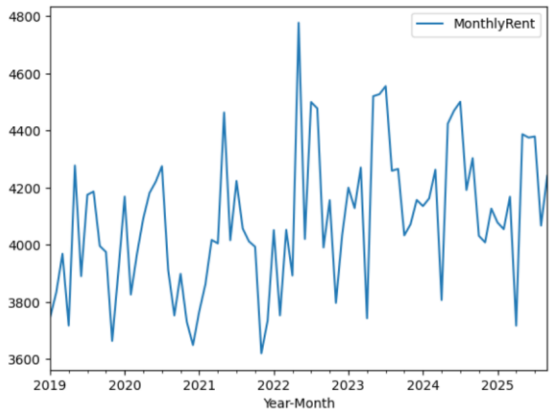
1) Use the best_model.predict() function on the combined dataset (with 'MonthlyRent' column dropped)
2) Then the prediction dataset was combined with the historic rent dataset via concatenation.
3) Admissions data was prepared for plotting (for the purposes of this project, we assumed that admissions levels will remain the same for 2024 and 2025).
4) Admissions data and combined rent data for Canada's largest cities (i.e., Toronto, Montreal, Vancouver, Calgary, Edmonton, Ottawa) were plotted side-by-side

SAMPLE RESULTS:

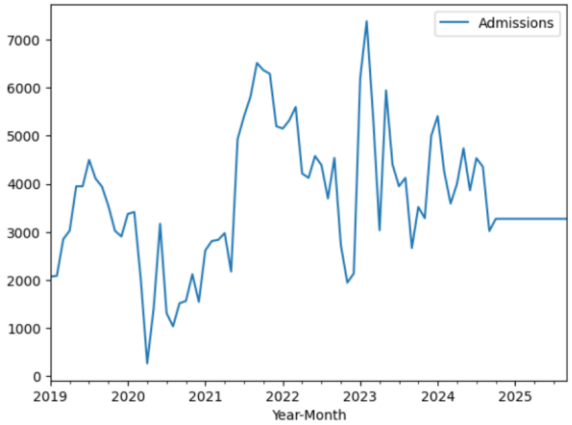


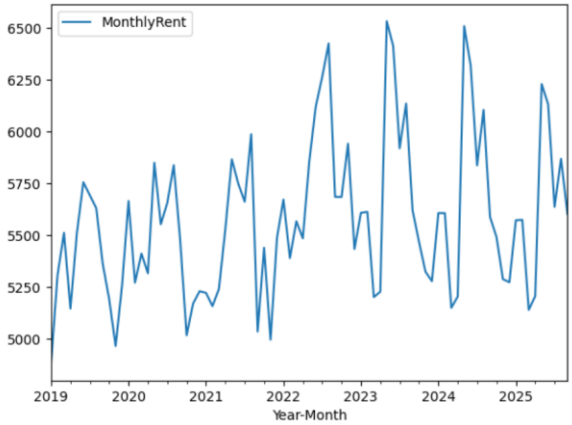
Montreal



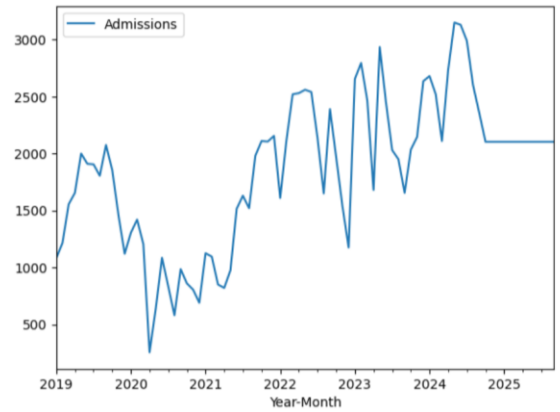



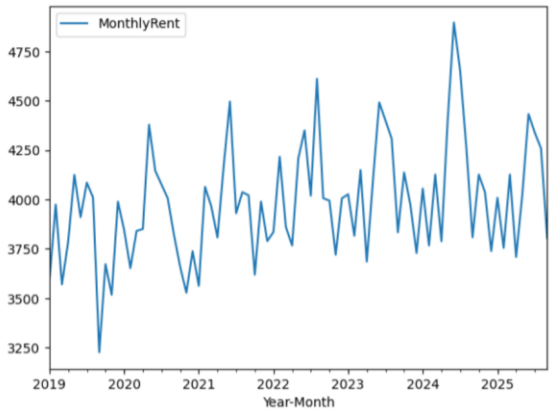

Vancouver




Calgary



# Conclusion

After completing our prediction using Random Forest Regressor, we can observe that 2025 rental prices appear lower in many cities (i.e. Vancouver, Montreal and Calgary). Additionally, by observing feature importances, it is clear that PR admissions, while not the most important feature, does meaningfully contribute to the model's ability to predict rental prices. As such, we can reject the null hypothesis that the new PR admission targets will not reduce rental accommodation costs in Canada.

By using synthetic data for rental prices, we were able to build a well performing, generalized model that can be used to predict economic impacts relative to immigration policies. Synthetic data is a useful tool businesses and policy makers can leverage to demonstrate technical capabilities, and serve as a proof of concept in order to justify any capital expenses required to gather real-world, proprietary data. While the prompt required ChatGPT to provide realistic data, future EDA could be done with real-world proprietary data to compare how realistic synthetic data was.

Visualizing rental prices side-by-side with admissions data does show that rental prices in 2025 would decrease, but not by a significant margin. While the scope of this report was aimed to examine the impact of permanent resident admissions, future studies could leverage temporary resident (TR) immigration targets. TR admissions and other immigration categories could be included to develop a more holistic analysis to understand the relationship between immigration policy and Canada's economic outlook.