Data Analytics Mini Project

Real Estate Price Analysis and Prediction

COMP 333

Instructor: Dr. Yasser Esmaeili Salehani

By:

Imane Madda: 40208741

Uroosa Lakhani: 40227274

Concordia University

11 April, 2025

Abstract

This project analyzes and forecasts real estate prices in 45 Canadian cities using various machine learning algorithms to predict the most affordable housing cities. We investigate how attributes like the number of bedrooms, bathrooms, and median family income affect home values using a Kaggle dataset [1]. To learn more about market trends, we used clustering and linear regression algorithms. Our results provide a baseline model for predicting future property prices by indicating high relationships between listing prices, income, and population by city. Exploratory data analysis was performed, which included both descriptive data analysis and data wrangling. The refined dataset was used to train 4 models whose parameters, such as max_depth, were fine-tuned to achieve more accurate prediction models. These models were compared on the basis of 5 accuracy measures (R^2, MAE, RMSE, F1 Score, MCC Score). The comparison of these resulted in the Random Forest model being the most accurate, followed by Gradient Booster, Decision Tree and the least accurate being Linear Regression. Given the Random Forest predictions being the most accurate, we can predict that in the coming years, the top 3 most affordable cities will be Red Deer, Edmonton and Saskatoon, while the most unaffordable ones will be White Rock, Vancouver and Burnaby. Making Alberta and Saskatchewan amongst the most affordable cities and British Columbia amongst the most expensive.

Introduction

Housing affordability is a growing concern in Canada. With real estate prices fluctuating due to factors like economic shifts and varying supply and demand, understanding future affordability trends can help policymakers, investors, and homebuyers make more informed decisions. In order to identify trends and create accurate predictive models, the study is conducted on a housing dataset containing data from 45 different Canadian cities. The overall objective of the study is to identify the most affordable Canadian cities based on various factors present in the dataset such as property prices, median family income, etc.

1. Data Collection

The data analyzed in this study is the Kaggle dataset *Canadian House Prices for Top Cities* [1].

The dataset includes 35,768 (rows) property listings across 45 different Canadian cities.

Additionally, 9 columns of specifications are provided for each listing. Below are the descriptions provided for each, sourced directly from the dataset source [1].

Table 1: Column names with their descriptions

Column Name	Description
City	City or major metropolitan area within which listings were found. For example, Toronto may include listings from surrounding suburbs such as Markham, Oakville, etc.
Price	Listed price for the property in Canadian dollars.
Address	Street address and, where applicable, unit number for the listing.
Number_Beds	Number of bedrooms mentioned in the listing.
Number_Baths	Number of bathrooms mentioned in the listing.
Province	Province in which each city resides. Note, border towns such as Ottawa do not include listings from the surrounding out-of-province cities like Gatineau.
Population	City population. According to simplemaps (https://simplemaps.com/data/canada-cities)
Longitude / Latitude	Longitude and Latitude data for individual cities, taken from simpelmaps (https://simplemaps.com/data/canada-cities)
Median_Family_Income	Median household income for the city taken from the 2021 Canadian census.

2. Statistical Analyses & Modeling

2.1 Exploratory Data Analysis (EDA)

The exploratory data analysis of the dataset started off using the descriptive data analysis technique, which allowed the gathering of the following general findings found in the table below. Note that the address column was dropped, since specific listing addresses do not provide any significant value in this study.

Table 2: General descriptive insights for numeric columns in the dataset

	Price	Number_Beds	Number_Baths	Population	Latitude	Longitude	Median_Family_Income
count	3.576800e+04	35768.000000	35768.000000	3.576800e+04	35768.000000	35768.000000	35768.000000
mean	9.432963e+05	3.283661	2.532403	6.360151e+05	47.446556	-98.421636	89643.103416
std	1.020110e+06	1.730654	1.371910	1.120016e+06	3.333855	22.280935	12132.353510
min	2.150000e+04	0.000000	0.000000	6.338200e+04	42.283300	-123.936400	62400.000000
25%	4.599000e+05	2.000000	2.000000	1.091670e+05	43.866700	-122.316700	82000.000000
50%	6.990000e+05	3.000000	2.000000	2.424600e+05	49.025000	-104.606700	89000.000000
75%	1.095000e+06	4.000000	3.000000	5.228880e+05	49.888100	-79.866700	97000.000000
max	3.700000e+07	109.000000	59.000000	5.647656e+06	53.916900	63.100500	133000.000000

The above table provides key summary statistics of the dataset, showing that the average house price is approximately $943,296, with 3.28 bedrooms and 2.53 bathrooms per listing. Additionally, the most expensive house and the cheapest one are also provided by the max and min rows. The standard deviation highlights a wide variance in the range of population and median family income across cities, indicating strong variability in housing markets.

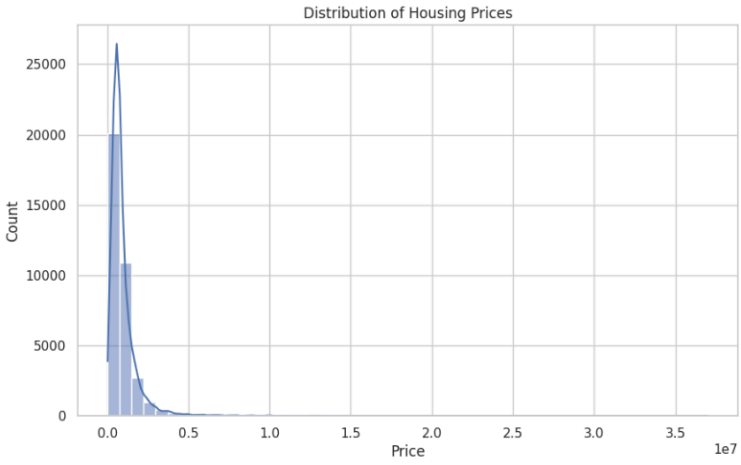

Figure 1: Housing Price Distribution Histogram

The distribution of housing prices is right-skewed, with most homes priced on the lower end. From the histogram and table 2, we can analyze that the lowest house price is $21,500, while the most expensive house is $37,000,000 the mean (average) price is $943,296 , the median price: $699,000 and the standard deviation is $1,020,110 which indicates that a small number of very expensive properties are pulling the average upward, making the median a more reliable indicator of typical housing prices.

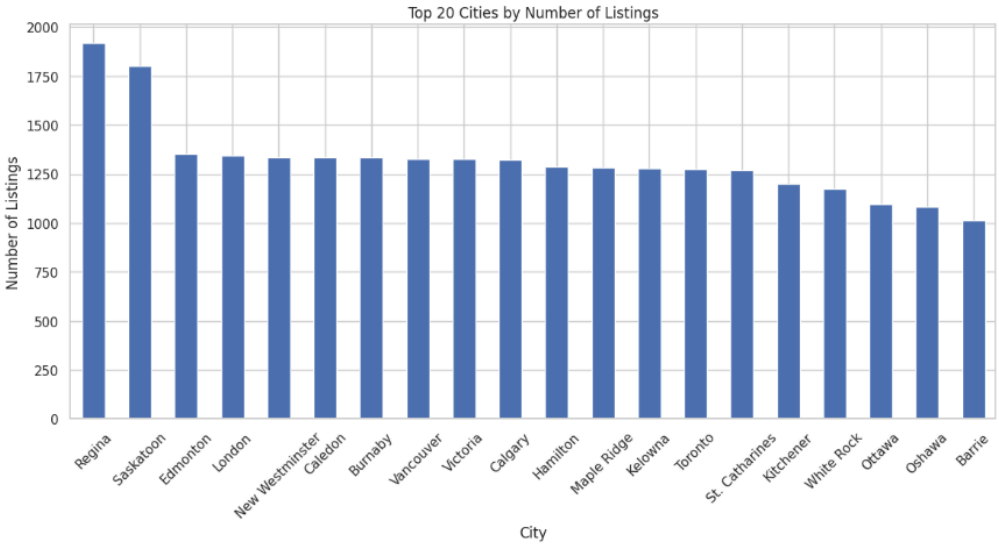

Figure 2: Top 20 Cities by Number of Listings

The above figure of top 20 Cities By number of listings shows Regina has the highest number of listings (1950), followed by Saskatoon approximately 1800. Most cities in the top 20 have fairly similar listing volumes, ranging from ~1000 to 1400. Bigger cities like Toronto, Ottawa, and Vancouver are not leading in listing counts, which suggests higher competition, stricter market or that the dataset includes relatively more listings from smaller cities. Also, cities like Toronto and Vancouver show significantly higher price averages.

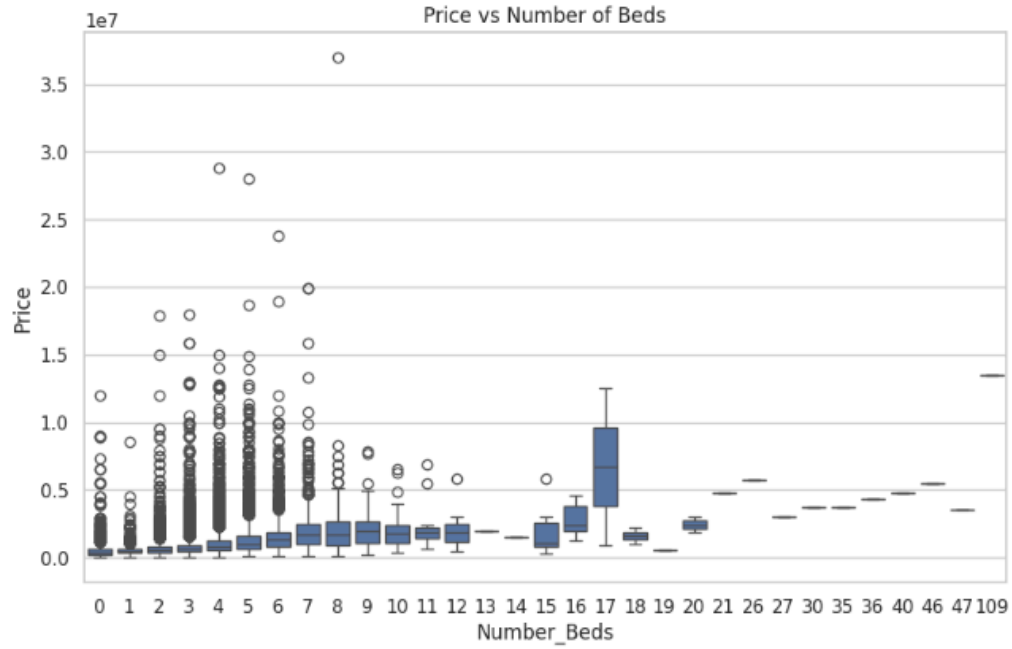

Figure 3: Price vs Number of Beds

The above graph of Price Vs Number of Beds, suggest that house prices generally increase with the number of bedrooms. Most 1–5 bedroom homes are priced below 1 million but some go beyond $10 million. Outliers with 46 and 109 bedrooms likely indicate anomalies or non-residential listings

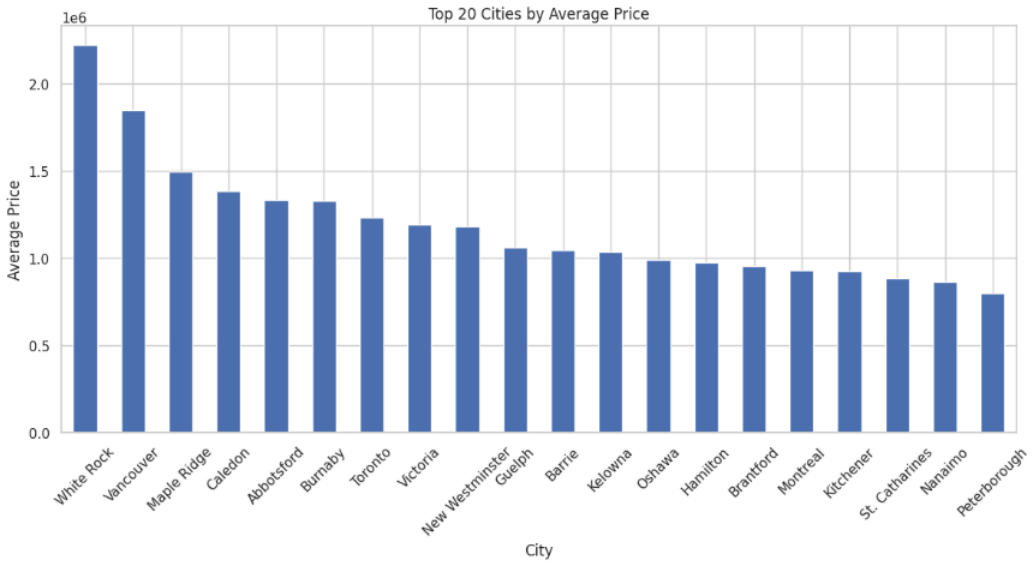

Figure 4: Top 20 Cities by Average Price

The above data suggests White Rock has the highest average house price (over 2 million), followed by Vancouver (1.8M). Cities like Toronto and Victoria also have high average prices, while others like Peterborough and Nanaimo are on the lower end of this top 20 list. This shows insights into which cities are the most expensive on average in the dataset.

2.2 Data Wrangling Findings

After the general EDA step, data wrangling was performed. This was performed in several steps, starting off with data loading using the pandas library. Verification and handling of missing data was performed by dropping rows that contain this defect. This was achieved using the pandas library as well as using built-in methods such as isnull().sum() and confirmed that there were no remaining missing values. The Address column was dropped as it does not provide any relevant information. Statistical summaries were explored using the describe() method. This allowed the identification of potential outliers using interquartile range (IQR) analysis. Additionally, we ensured numerical columns such as Price, Number_Beds, and Median_Family_Income were correctly typed as floats or integers, and standardized feature names for consistency in downstream machine learning tasks.

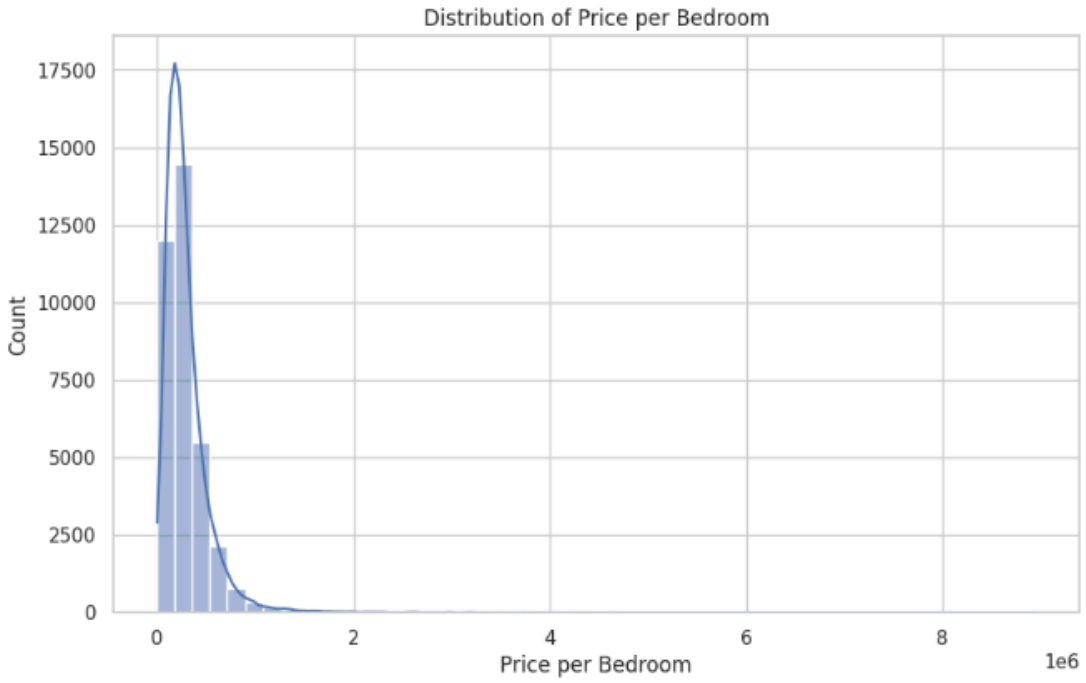

Figure 5: Distribution of Price per Bedroom

The distribution plot of Price per Bedroom shows a highly right-skewed distribution, indicating that the majority of homes are priced relatively low on a per-bedroom basis, with a long tail representing a small number of high-priced outliers. This suggests that while most properties are clustered within a typical price range, there are luxury or overpriced listings that significantly exceed the norm.

2.3 Regression Analysis :

In the regression analysis, we aimed to predict house prices using multiple linear regression. We selected key features such as Number_Beds, Number_Baths, Population, and Median_Family_Income as independent variables, and Price as the dependent variable. The dataset was split into training and testing sets using an 80-20 ratio with train_test_split() from sklearn. We then trained a Linear Regression model using LinearRegression() from Scikit-learn. After fitting the model to the training data, we evaluated its performance on the test set using Root Mean Squared Error (RMSE) and the R^2 score.

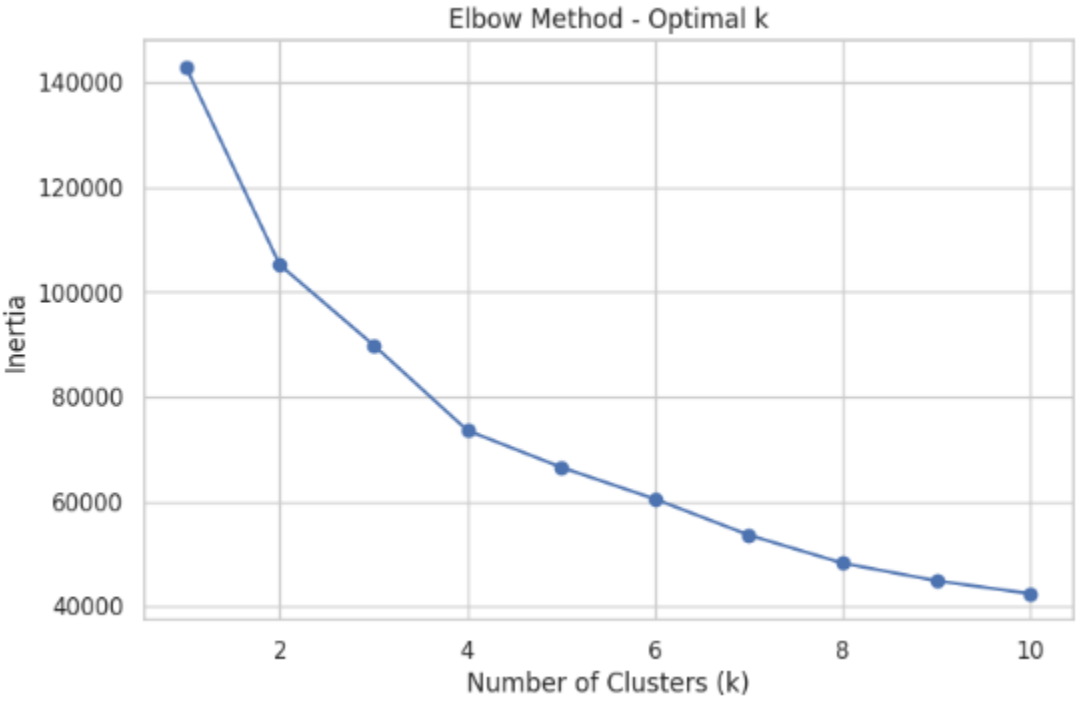

Figure 6: Elbow Method - Optimal K

The above Elbow Method plot is used to determine the optimal number of clusters (*k*) for K-Means clustering. The x-axis represents the number of clusters, while the y-axis shows the inertia—the sum of squared distances between each point and its assigned cluster center.

As *k* increases, inertia decreases because more clusters allow better fitting. However, after a certain point, the rate of improvement drops significantly. This "elbow" point, where the curve begins to flatten (around **k = 3**), suggests the most suitable number of clusters. Choosing *k = 3* balances clustering performance with model simplicity.

Figure 7: Clusters of Properties by Income vs Price

The scatter plot in Figure 7 displays three distinct clusters of properties based on Median Family Income (x-axis) and Price (y-axis).

- Cluster 2 represents high-priced properties, often considered luxury homes, and spans a wide range of income levels.

- Clusters 0 and 1 consist of lower to mid-range priced homes, showing overlap in price but slight differences in the income brackets they are associated with.

The clustering highlights how income and property prices group together, revealing patterns in housing affordability and market segmentation

3. Predictive Modeling

To identify the top 3 most affordable cities based on the various column variables in the dataset, such as the median family income, four models were trained. These are linear regression, decision tree, random forest and gradient boosting models.

In order to compare the categorical variables with the numeric ones, label encoding was performed on both the “City” and “Province” columns. Both encodings can be found below in Tables 3 and 4.

Table 3: City encoding

Encoding	City
0	Abbotsford
1	Airdrie
2	Barrie
3	Brantford
4	Burnaby
5	Caledon
6	Calgary
7	Edmonton
8	Guelph
9	Halifax
10	Hamilton
11	Kamloops
12	Kelowna
13	Kingston
14	Kitchener
15	Lethbridge
16	London
17	Maple Ridge
18	Medicine Hat
19	Moncton
20	Montreal
21	Nanaimo
22	New Westminster
23	Oshawa
24	Ottawa
25	Peterborough
26	Prince George
27	Quebec
28	Red Deer
29	Regina
30	Saint John
31	Saskatoon
32	Sault Ste. Marie
33	Sherbrooke
34	St. Catharines
35	St. John's

36	Sudbury
37	Thunder Bay
38	Toronto
39	Trois-Rivieres
40	Vancouver
41	Victoria
42	White Rock
43	Windsor

Table 5: Province encoding

Encoding	Province
0	Alberta
1	British Columbia
2	Manitoba
3	New Brunswick
4	Newfoundland and Labrador
5	Nova Scotia
6	Ontario
7	Quebec

City rankings were determined using an affordability index, which takes into account the median family income over the predicted house prices, calculated by each respective model.

A higher affordability index would correlate to a more affordable (cheaper) house, while a smaller index would represent a more expensive one.

Additionally, 5 accuracy measures were examined to compare each model and determine which one is the best.

3.1 Linear Regression

The Linear Regression model is a statistical model that is known to predict continuous target variables by finding a relationship between the target and input features. It does so by fitting lines to the best of its abilities, as can be seen in Figure 8. This model, however, is known to not be the most accurate, which has been proven as it's the model with the lowest accuracy

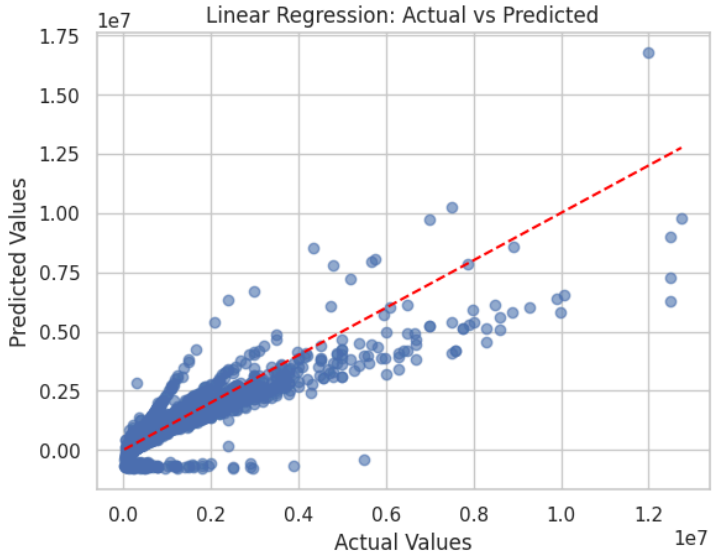

Figure 8: Linear Regression Predictive Model Plot

Table 5: Evaluation Metrics for Linear Regression

Metric	Value
R^2	0.77
MAE	233,396.54
RMSE	444,102.12
F1 Score	0.86
MCC Score	0.70

Table 6: Top 3 Most/Least Affordable Cities predicted by Linear Regression

Top 3 Most Affordable Cities	Affordability Index	Least 3 Affordable Cities	Affordability Index
St. John's (35.0)	0.3783	Moncton (19.0)	-4.2821
Thunder Bay (37.0)	0.3566	Edmonton (7.0)	-0.3557
Quebec (27.0)	0.3122	Saint John (30.0)	-0.0240

3.2 Decision Tree

The Decision Tree model is a non-linear model that splits data into subsets based on feature values to allow predictions in a tree-like structure of decisions, as can be seen in Figure 9. In the tree, each decision node is a feature test, while each leaf represents the predicted value

Note: To achieve a more accurate decision tree model, the maximum depth (7) of the decision regressor was increased until the top and bottom affordability cities remained the same.

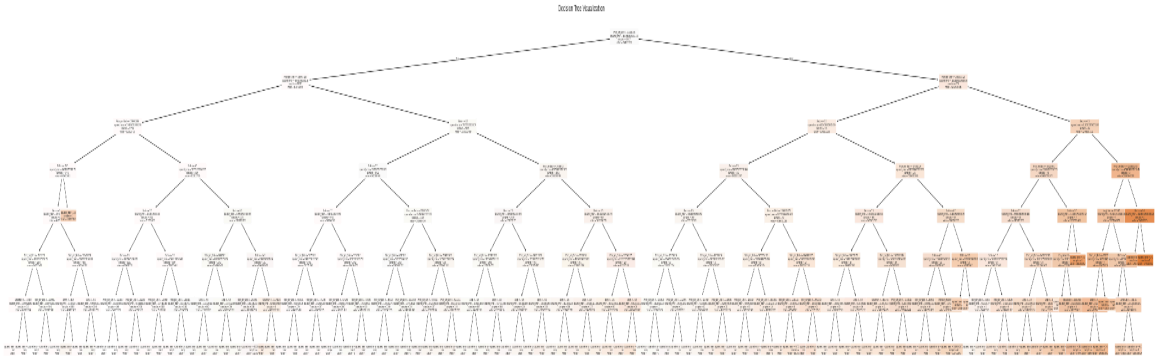

Figure 9: Decision Tree Predictive Model Plot

Table 7: Evaluation Metrics for Decision Tree

Metric	Value
R^2	0.95
MAE	96,645.90
RMSE	213,358.26
F1 Score	0.93
MCC Score	0.85

Table 8: Top 3 Most/Least Affordable Cities (Decision Tree)

Top 3 Most Affordable Cities	Affordability Index	Least 3 Affordable Cities	Affordability Index
Red Deer (29)	0.293993	White Rock(42)	0.057699
Edmonton (7)	0.274528	Vancouver (40)	0.080238
Saskatoon (31)	0.237477	Burnaby (4)	0.090136

3.3 Random Forest

Random Forest is a more complex version of the Decision Tree model. It creates multiple decision trees, which are each trained on random subsets of the data and combines their predictions to allow more accurate predictions. Overall, this technique allows the reduction of overfitting and variance which allows more robust and accurate predictions to be made, in comparison to a single decision tree.

Note: To achieve a more accurate random forest model, the maximum depth (70) of the random forest regressor was increased.

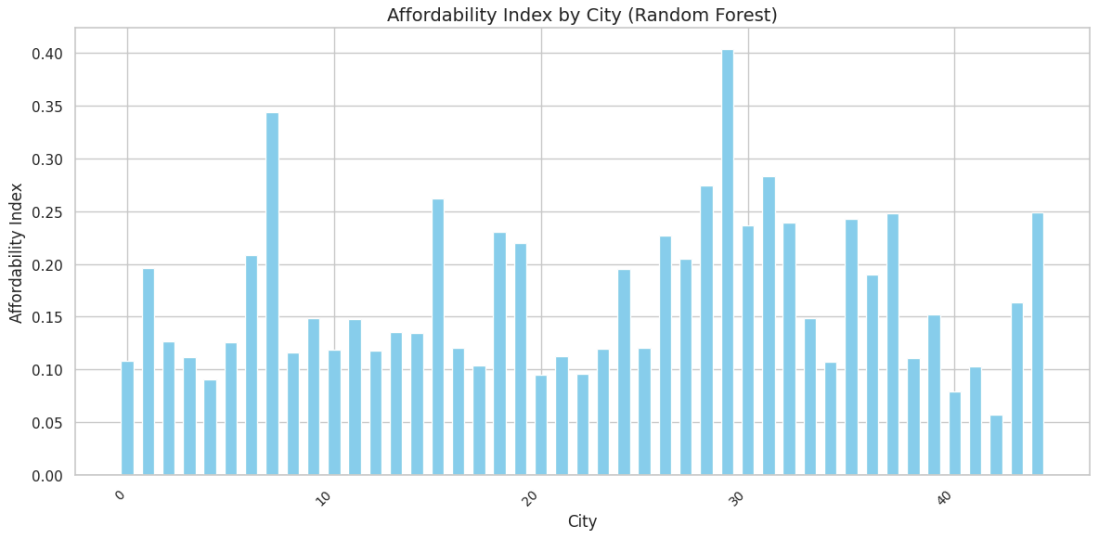

Figure 10: Random Forest Predictive Model Histogram

Table 9: Evaluation Metrics for Random Forest

Metric	Value
R^2	0.98
MAE	13254.03
RMSE	136 966.54
F1 Score	0.99
MCC Score	0.99

Table 10: Top 3 Most/Least Affordable Cities (Random Forest)

Top 3 Most Affordable Cities	Affordability Index	Least 3 Affordable Cities	Affordability Index
Red Deer (29)	0.4034	White Rock(42)	0.0570
Edmonton (7)	0.3434	Vancouver (40)	0.0792
Saskatoon (31)	0.2828	Burnaby (4)	0.0903

3.4 Gradient Boosting

Similar to the Random Forest Model, the Gradient Boosting model utilizes multiple trees, although they are built sequentially. Additionally, in this predictive model, each tree is expected to try to correct the errors made by the previous one. This can be understood as using simple decision tree models with iterative improvement to achieve a much more accurate prediction model.

Note: To achieve a more accurate gradient boosting model, the maximum depth (10) and learning rates (0.8) of the gradient boosting regressor were both increased.

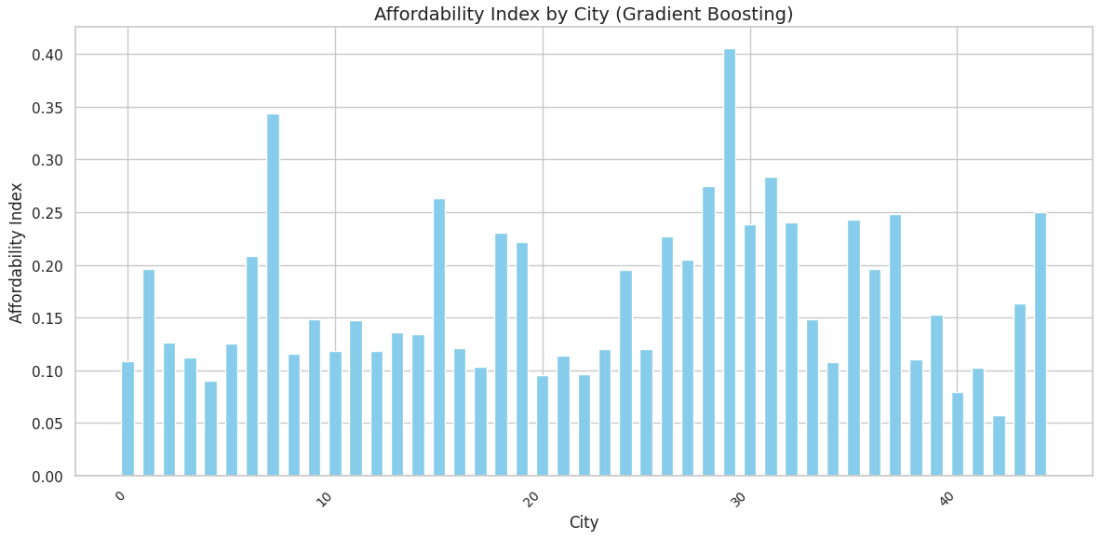

Figure 11: Random Forest Predictive Model Histogram

Table 11: Evaluation Metrics for Gradient Boosting

Metric	Value
R^2	0.97
MAE	16,018.51
RMSE	152,122.05
F1 Score	0.99
MCC Score	0.97

Table 12: Top 3 Most/Least Affordable Cities (Gradient Boosting)

Top 3 Most Affordable Cities	Affordability Index	Least 3 Affordable Cities	Affordability Index
Red Deer (29)	0.4052	White Rock(42)	0.0570
Edmonton (7)	0.3438	Vancouver (40)	0.0792
Saskatoon (31)	0.2834	Burnaby (4)	0.0903

Discussion of Results

Table 13: Table of Accuracy Metrics for All Four Models

Metric	Linear Regression	Decision Tree	Random Forest	Gradient Boosting
R^2	0.77	0.95	0.98	0.97
MAE	233396.54	96645.9	13254.03	16018.51
RMSE	444102.12	213358.26	136966.54	152122.05
F1 Score	0.86	0.93	0.99	0.99
MCC Score	0.7	0.85	0.99	0.97

Description of metrics analyzed:

- **R^2 (Coefficient of Determination)**: This metric allows the quantification of how well the predicted model match the actual data. Ranges from 0 to 1, where 1 is a perfect fit, while 0 would indicate no relevancy. Therefore, a Higher R^2 would indicate a better fit of the data.

- **MAE (Mean Absolute Error)**: This metric represents the prediction error by averaging the absolute difference between the predicted and actual values. In this case, a lower MAE value represents a more accurate model

- **RMSE (Root Mean Squared Error)**: This metric is also used to determine prediction errors, however it is better at penalizing larger errors as it squares the difference in comparison to MAE. Overall, this error metric takes the square root of the average squared differences between predicted and actual values. Similarly, a lower RMSE value represents a more accurate model.

- **F1 Score**: This metric is used to determine the performance for classification models, using both precision and recall into a single value. This score ranges from 0 to 1, where 1 would represent the perfect balance of precision and recall (best performance). Thus, a higher F1 score would suggest a better classification performance, which is significant in the case of a varying dataset.

- **MCC (Matthews Correlation Coefficient)**: A metric used for binary classification. It ranges from -1(worst) to 1 (best) where 0 would indicate random predictions. Therefore, a higher MCC would represent a better-performing model that is able to efficiently distinguish the different predicted classes

Model Comparison

After analyzing all 4 models, the ordering of the best performing model to the least is:

1. Random Forest (Best Model)
2. Gradient Boosting
3. Decision Tree

4. Linear Regression (Worst Model)

Random Forest has proven to be the best-performing model since it has the highest R^2 score (0.98), suggesting that the predictions align with the actual data. It has the lowest MAE (13,254.03) and RMSE (136,966.54) values, indicating that predictions are more accurate overall and that both small and larger errors are less present in the predictions. It has amongst the highest F1 scores (0.99), suggesting that it's good at handling precision and recall, allowing more accurate classification. Lastly, it had the highest MCC score (0.99), which indicates that there is a strong correlation between the predicted and actual values.

In second place for best performing model, we have Gradient Boosting. With a very close R^2 (0.97) to Random Forest, it also suggests an almost perfect match between the predicted and actual data. Although Gradient Booster has a slightly higher tendency of making errors as can be seen with both MAE (16,018.51) and RMSE (152,122.05). Despite having the same F1 score (0.99), Gradient Booster has a slightly lower MCC Score (0.97), making it a bit less accurate when is comes to binary classification.

In third place, we have the Decision Tree model. This model has a slightly lower R^2 (0.95) than the above two models, suggesting that the predictions are slightly less accurate than the actual data. Decision Tree tends to make a lot more smaller and larger errors by having significantly higher MAE (96,645.90) and RMSE (213,358.26) scores. This model has a slightly lower F1 score in comparison to both Random Forest and Gradient Boosting. Lastly, this model displays less accurate correlation between the predicted and actual data by having a lower MCC Score (0.85).

Lastly, in fourth place, the worst-performing prediction model is Linear Regression. It has the lowest R^2 (0.77) suggests that the predicted data doesn't match as well to the actual data. With the highest error scores for both MAE (233,396.54) and RMSE (444,102.12), this model tends to make a large number of errors, affecting the accuracy of the predictions. It has a moderate performance indicated by the F1 Score (0.86) and is the worst amongst all 4 when it comes to binary classification, as indicated by having the lowest MCC Score (0.70).

Note: It is important to note that the highest performing models were achieved by tuning parameters such as the max_depth (Random Forest, Gradient Boosting and Decision Tree) and learning rates (Gradient Boosting).

Considering the Random Forest model to be the most accurate, we can safely assume that in the coming years, the top 3 most affordable cities with be Red Deer, Edmonton and Saskatoon, while the most unaffordable ones will be White Rock, Vancouver and Burnaby, from Table 10.

Conclusion

To conclude, exploratory data analysis was conducted by performing descriptive data analysis and data wrangling. The purpose of performing these processes was to analyze and clean the dataset to be used in the predictive modeling step. In the predictive modeling step, 4 models were trained and compared using 5 accuracy metrics. The comparison of these yielded that the best performing model with the least amount of errors was the Random Forest model, followed by Gradient Boosting and Decision Tree, while the worst performing model was Linear regression. It is important to note that parameters such as max_depth and learning_rates were used to fine-tune the models to achieve better accuracy, which may have played a role in their performance ranking. Lastly, given that the Random Forest is the most accurate, based on its predictions, we can predict that in the coming years the top 3 most affordable cities will be Red Deer, Edmonton and Saskatoon while the most unaffordable ones will be White Rock, Vancouver and Burnaby.

References

[1] J. Larcher, Canadian House Prices for Top Cities, Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/jeremylarcher/canadian-house-prices-for-top-cities. [Accessed: Mar. 26, 2025].

[2] Scikit-learn developers, *scikit-learn: Machine Learning in Python*. [Online]. Available: https://scikit-learn.org/stable/. [Accessed: Apr. 5, 2025].

[3] M. Waskom *et al.*, *Seaborn: Statistical Data Visualization*. [Online]. Available: https://seaborn.pydata.org/tutorial/introduction.html. [Accessed: Apr. 5, 2025].