Projected Price Prediction in 5 years

Team Members: Urvashi Dube, Sudhish Subramaniam

Business Understanding

One of the finest ways to diversify net worth is to invest in land, even if it is undeveloped raw property. The primary goal is to identify the relationship between the locality of the house, previous land value, previous improvement value, current improvement value and current land value to predict the future land value in 5 years of land in Vancouver City.

The benchmark price of homes in Metro Vancouver stands at $1,155,300, representing a 2.1% monthly decline but a 3.9% yearly increase in September 2022 [1]. A B.C. real estate expert is downplaying a possible decline in average home prices nationally by the end of next year [2]. The rise and fall of Vancouver houses are predicted in these articles, but accurate predictions can be achieved only through machine learning and analysis of the most recent data.

The information will be useful to the public, who plans to make an investment or plan to buy a house for living.

The model will not only predict the future rate but also the growth of the area in which the land is developed. Builders will be helped by this, as with the result of the model great insights into locality, tax rate, land quality (from the age of the building) and land value can be drawn. The model will also predict next big improvement year.

Who are the stakeholders?

Builders: people who wish to develop building projects in Vancouver. It will help them to identify the most developing and profitable area.

Public: People who wish to purchase land of their own in Vancouver either for their stay or as an investment.

Shop Owners, entrepreneurs, and Companies: They can identify a suitable area of Vancouver for their business.

An organization can use the results of the model to decide the location of its upcoming project. The model will help them with the zoning classification and the street name for the most profitable land in 5 years.

The business goal of this project is "identify profitable lands of Vancouver in 5 years". Outputs from the regression analysis to predict the land's future value and future improvement values are the key business success criteria.

"Predicting the future land value and future improvement value on analysis" is the data mining goal of the project. The data mining would be considered a success if the accuracy of the model developed is above 80 % and all the essential columns are taken into consideration.

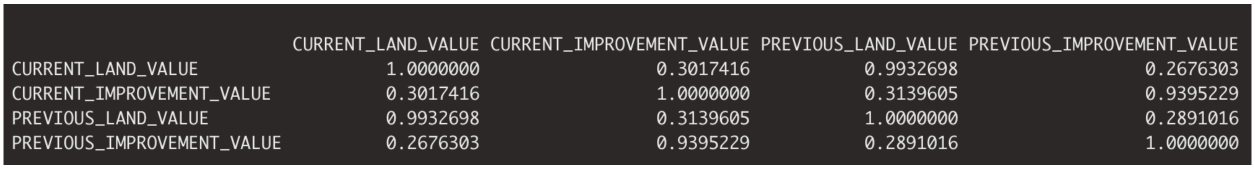
Data Understanding & Preparation

The dataset is a CSV file which has 29 columns and 218563 rows of data points (features). The dataset was taken from the City of Vancouver's open data portal. The City of Vancouver, BC Assessment, Finance Risk, and Supply Chain Management - Revenue Services is the owner of the data. The public can access data via the open data portal for data analysis. The dataset needs a lot of cleaning, narrative lines are discarded since they are not necessary for processing. All the null values in the numeric columns of the dataset are replaced with the column's mean, and the null values in the year columns are replaced with the mode (column's most prevalent element).

The prediction subject of the dataset is future land value and future improvement value. The domain concepts of the project are to identify the most relevant columns and categories to predict the future land value and improvement value of properties of Vancouver. Features such as PID, Legal Type of land, Zoning Classification, Street name, current land value, current improvement value, previous land value, previous improvement value, year built, and big improvement year.

The target attribute here is the future land value and future improvement value. This can be predicted only using the columns mentioned above. Different types of analysis and machine learning models must be applied to the dataset to achieve maximum accuracy.

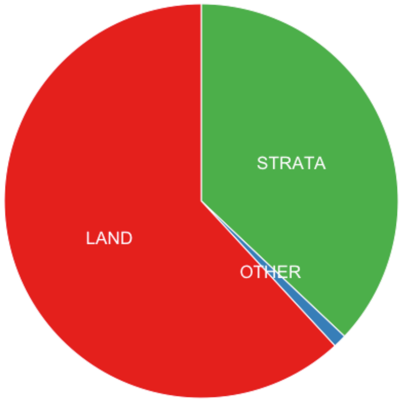
We have found the correlation matrix between the numeric columns in the dataset.

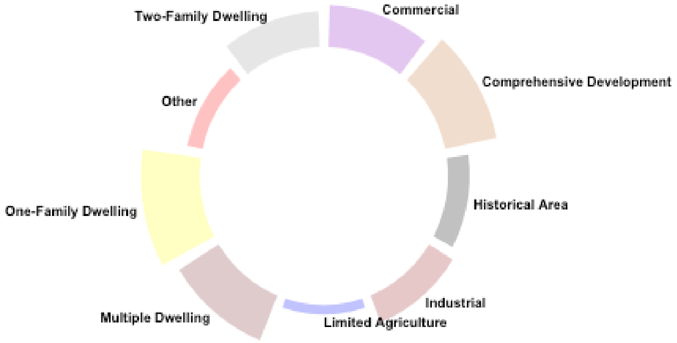


```
                           CURRENT_LAND_VALUE CURRENT_IMPROVEMENT_VALUE PREVIOUS_LAND_VALUE PREVIOUS_IMPROVEMENT_VALUE
CURRENT_LAND_VALUE                  1.0000000                 0.3017416           0.9932698                  0.2676303
CURRENT_IMPROVEMENT_VALUE           0.3017416                 1.0000000           0.3139605                  0.9395229
PREVIOUS_LAND_VALUE                 0.9932698                 0.3139605           1.0000000                  0.2891016
PREVIOUS_IMPROVEMENT_VALUE          0.2676303                 0.9395229           0.2891016                  1.0000000
```

The matrix indicates that current land value, previous land values, current improvement value and previous improvement value are highly correlated with each other.

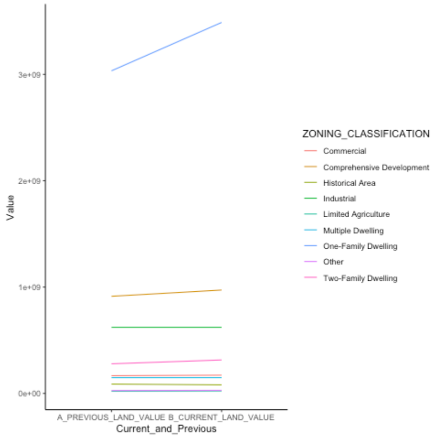
Current Land Value of Each Legal Type and Each Zone





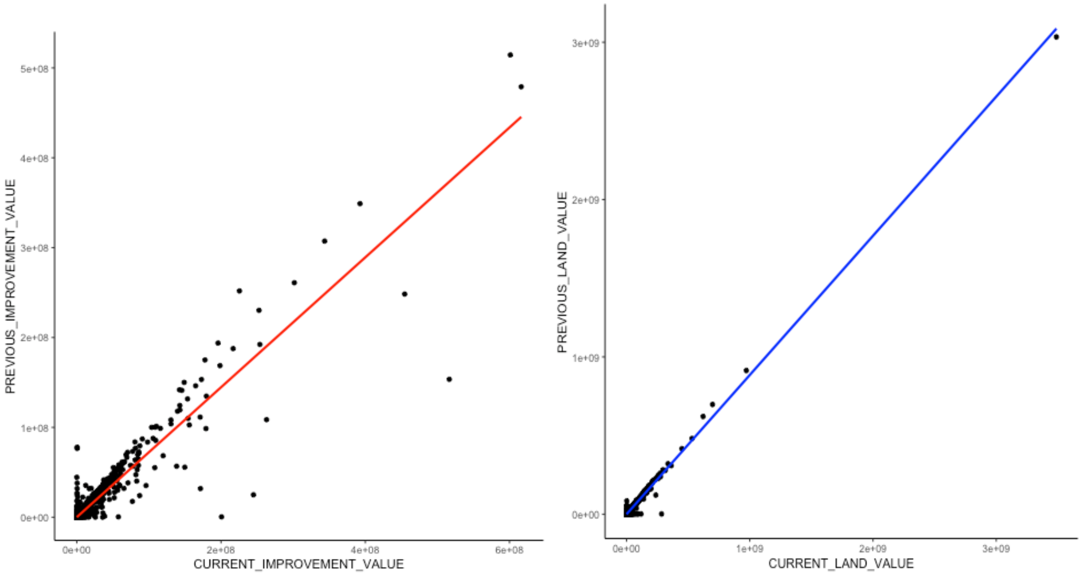
The land has the maximum current land value as compared to the other property types. About the zones, one-family dwelling is the most expensive property in Vancouver, followed by comprehensive development and multiple dwelling. The least expensive is limited agriculture.

Zone-wise Current Land Value Vs Previous Land Value



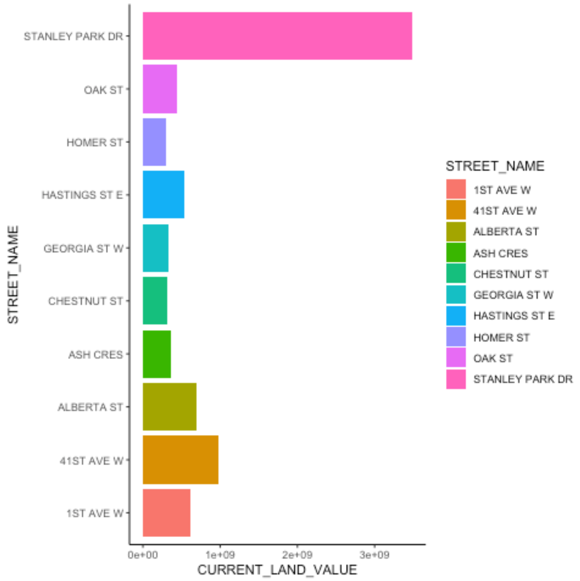
With the highest Land Value, multiple dwelling has shown an increase in land value in the past years. Comprehensive Developments and two-family dwellings have also shown a significant increase in land value. Apart from them, there is not much change in the land. Values of other properties in the past years.

Current Land Value Vs Previous Land Value of Each Property and Zone



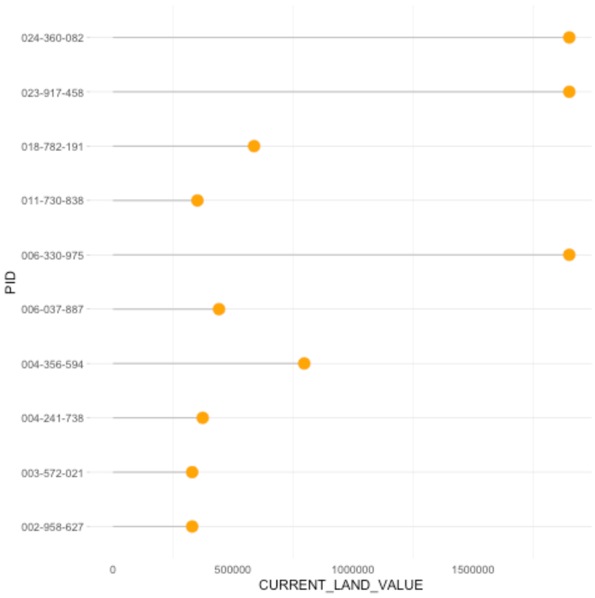
There is a linear relationship between the current improvement value and the previous improvement value as well as between the current land value and the previous land value.

Top 10 Expensive Streets of Vancouver



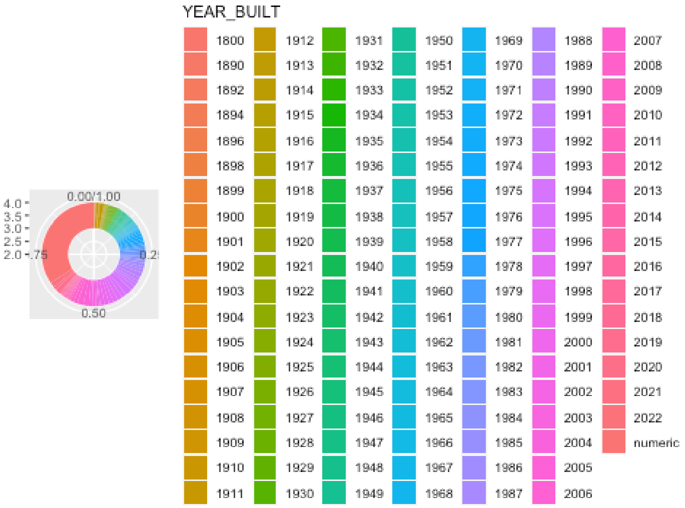
Properties in Stanley Park are the most expensive And properties at Chestnut Street are the least expensive in Vancouver. 41st Avenue and Alberta Ave W also follows.

Top 10 Expensive Properties of Vancouver



Out of the top 10 most expensive properties in Vancouver, 024-360-082 property is the most expensive property and 003-572-021 is the least expensive. These properties are in the prime location of the city with the highest current improvement value.

The recency of Properties in Vancouver



This graph tells us more than 40 % of the properties in Vancouver are recently built (2007-2022). There are not many properties in Vancouver built between 1920 – 1950. Most of the properties in Vancouver are recently built.

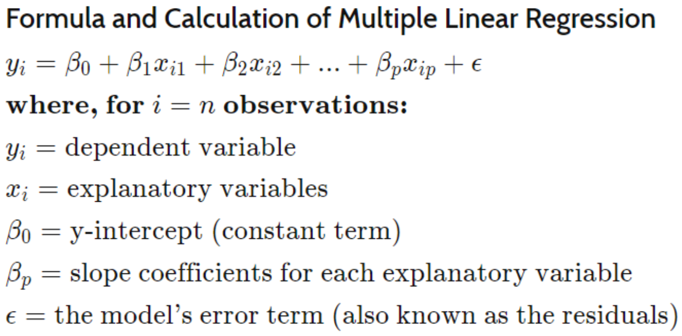
Modelling

For the regression analysis and predictions, we used two different models. Considering the correlation between the columns in the dataset we implemented Multiple Linear Regression and Random Forest Regression.

Multiple Linear Regression

A single dependent variable and several independent variables can be analysed using the statistical technique known as multiple regression. To forecast the value of the single dependent value, multiple regression analysis uses independent variables whose values are known.

It is used as a predictive approach to describe the relationship between a continuous dependent variable and two or more independent variables.



Formula and Calculation of Multiple Linear Regression

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$

where, for $i = n$ observations:

$y_i =$ dependent variable

$x_i =$ explanatory variables

$\beta_0 =$ y-intercept (constant term)

$\beta_p =$ slope coefficients for each explanatory variable

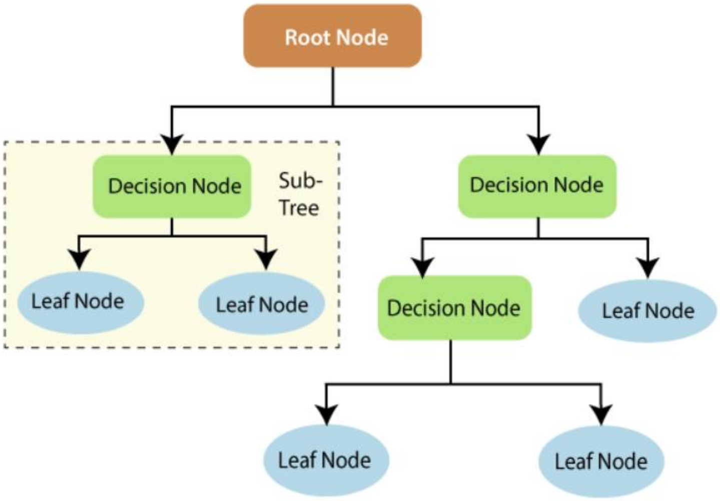
$\epsilon =$ the model's error term (also known as the residuals)

A statistician or analyst can use the simple linear regression function to forecast the value of one variable based on the knowledge of another one. Only two continuous variables—an independent variable and a dependent variable—can be utilised in a linear regression. The parameter that is utilised to calculate the dependent variable or result is known as the independent variable. A multivariate regression model incorporates several explanatory factors.

Random Forest

A supervised learning technique called Random Forest Regression leverages the ensemble learning approach for regression. The ensemble learning method combines predictions from various machine learning algorithms to provide predictions that are more accurate than those from a single model. It is capable of both classification and regression tasks.

A random forest generates accurate predictions that are simple to comprehend. Large datasets can be handled effectively. In comparison to the decision tree method, the random forest algorithm offers a higher level of accuracy in outcome prediction.

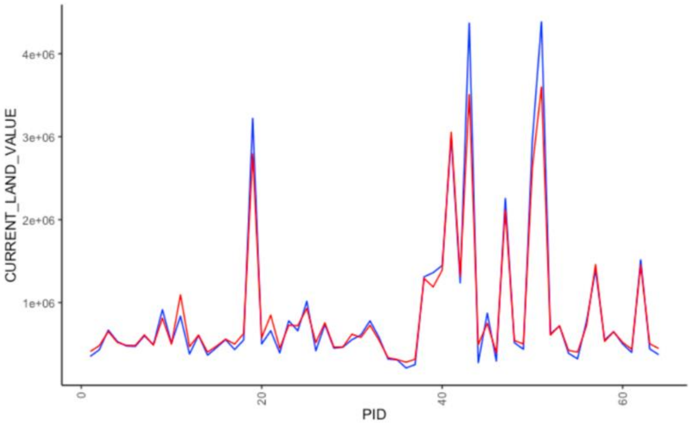


- Compared to the decision tree algorithm, it is more accurate.
- It offers a practical method for dealing with missing data.
- Without hyper-parameter adjustment, it can generate a reasonable prediction.
- It fixes the overfitting problem with decision trees.
- At the node's splitting point in every random forest tree, a subset of features is chosen at random.

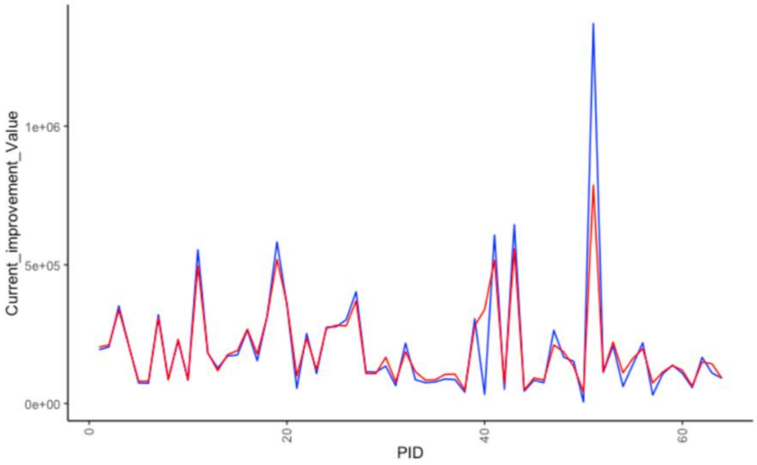
For both the regression models, we used PID, previous improvement value, previous land value, year built and current land values to train and test the model. We used the above parameters to predict the current land values. On obtaining the values we tested our models by comparing the predicted current land values and the original current land values. For random forest regression, we chose proximity = TRUE, to get confined results from the dataset.

Using Random Forest regression, we have obtained the following results:

We predicted the future land value of 90 random properties in Vancouver after 5 years.



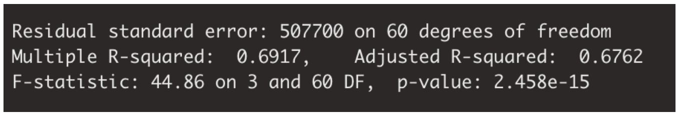
The blue line indicates the current land values and the red line shows the predicted land values. The graph indicates that for the PID taken into consideration, for some properties there will be a significant drop in the land value, for some there will be a slight increase but for the majority ones, there will be no change in the value. This is because no new constructions or developments are predicted in those areas. A significant drop in the land value, happens if there is an increase in the land value of another property in a street. The hike in the land value can be explained by the construction of new buildings, increased transportation routes and more schools, colleges, and supermarkets nearby. This graph can be used by people to identify the streets and zones where properties can be built or bought to increase their profits. The properties that have significant hikes in the graph from the previous land values will be the ones which would be ideal to be bought.



This graph shows the predicted and current improvement values. The blue line shows the current improvement value, and the red line shows the predicted improvement value. Considering the graph we can say that, there has been a significant decrease in the improvement value in the upcoming years. The main reason behind this downfall can be the recession in countries, the downfall of the labour market, the lack of skilled labour, the unavailability of fruitful properties and the impact of the COVID-19 pandemic. The world is still opening, and construction workers and property owners are deeply affected by the covid pandemic. Incomes of many had been cut down. Due to this people have been refraining from investing in properties.

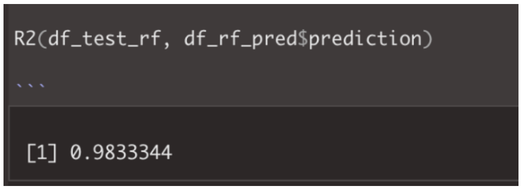
Results & Evaluation

For the performance metrics we have used R2 score and RMSE score to compare the two models.



```
Residual standard error: 507700 on 60 degrees of freedom
Multiple R-squared:  0.6917,    Adjusted R-squared:  0.6762
F-statistic: 44.86 on 3 and 60 DF,  p-value: 2.458e-15
```

This shows the summary of the linear regression model.



~~~
R2(df_test_rf, df_rf_pred$prediction)
```

 [1] 0.9833344
~~~

This shows the R2 score of the random forest model. R2 Scores of :

Multiple linear regression model: 0.6917

Random Forest Regression: 0.9833

It is evident from both R2 scores of both models, that Random Forest regression has better results as compared to the linear regression model.

Hence, we use random forest regression to predict the future land value and future improvement value. The results from random forest regression can be used to analyze different properties in Vancouver. City planners in Vancouver can decide upon the weak areas where predicted land improvement value and land value are less, and ways to increase the property values. Areas which can be worked on to increase the land value are transportation, schools, colleges, offices, supermarkets, etc. By working on these areas, there will be an overall development of the city and the economy of Vancouver. This will increase the urge to invest in property. Property dealers and estate agents will also be benefited from the predictions. It will help them to decide to choose the property to invest in. This will help them increase their business. Students of universities and the public would be benefited from the predictions as it would be easy for them to analyze the properties with more land values, which will be close to colleges, universities, supermarkets and transport.

Hence the predictions will not help to invest in properties but also in the overall development of the city.

Recommendations

To increase the efficiency and range of the project we provide the following recommendations-

- Sentiment Analysis could be applied to the narrative lines to get more insights and features.
- Advanced Machine learning algorithms could be applied to get more deep insights into the problem.
- More parameters could be considered for evaluation. Other datasets could be combined to get more insights into the dataset. For further analysis, tax levy could also be considered, and the tax levy after 5 years could also be predicted.
- More data could be incorporated to increase the accuracy of the machine-learning model.

References

1. https://wowa.ca/vancouver-housing-market

2. https://vancouver.citynews.ca/2022/08/12/bc-real-estate-price-decline-questions/

3. Dataset: https://opendata.vancouver.ca/explore/embed/dataset/property-tax-report/table/?refine.report_year=2022&dataChart=eyJxdWVyaWVzIjpbeyJjb25maWciOnsiZGF0YXNldCI6InByb3BlcnR5LXRheC1yZXBvcnQiLCJvcHRpb25zIjp7fX0sImNoYXJ0cyI6W3siYWxpZ25Nb250aCI6dHJ1ZSwidHlwZSI6ImNvbHVtbiIsImZ1bmMiOiJBVkciLCJ5QXhpcyI6ImN1cnJlbnRfbGFuZF92YWx1ZSIsInNjaWVudGlmaWNEaXNwbGF5Ijp0cnVlLCJjb2xvciI6IiMwMjc5QjEifV0sInhBeGlzIjoibGVnYWxfdHlwZSIsIm1heHBvaW50cyI6NTAsInNvcnQiOiIifV0sInRpbWVzY2FsZSI6IiIsImRpc3BsYXlMZWdlbmQiOnRydWUsImFsaWduTW9udGgiOnRydWV9

4. Dataset Information: https://opendata.vancouver.ca/explore/dataset/property-tax-report/information/?refine.report_year=2022&dataChart=eyJxdWVyaWVzIjpbeyJjb25maWciOnsiZGF0YXNldCI6InByb3BlcnR5LXRheC1yZXBvcnQiLCJvcHRpb25zIjp7fX0sImNoYXJ0cyI6W3siYWxpZ25Nb250aCI6dHJ1ZSwidHlwZSI6ImNvbHVtbiIsImZ1bmMiOiJBVkciLCJ5QXhpcyI6ImN1cnJlbnRfbGFuZF92YWx1ZSIsInNjaWVudGlmaWNEaXNwbGF5Ijp0cnVlLCJjb2xvciI6IiMwMjc5QjEifV0sInhBeGlzIjoibGVnYWxfdHlwZSIsIm1heHBvaW50cyI6NTAsInNvcnQiOiIifV0sInRpbWVzY2FsZSI6IiIsImRpc3BsYXlMZWdlbmQiOnRydWUsImFsaWduTW9udGgiOnRydWV9

5. https://www.rdocumentation.org/

6. https://www.datacamp.com/tutorial/linear-regression-R

7. https://www.geeksforgeeks.org/random-forest-approach-for-regression-in-r-programming/

8. http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r

9. https://vancouversun.com/news/local-news/home-sales-decline-bcrea-forecast#:~:text=In%202023%2C%20home%20sales%20are,cent%20in%202023%20to%20%24939%2C500.

Appendix

Linear Regression	Regression analysis model
Random Forest Regression	Regression analysis model
R2 Score	Performance metric for regression model
RMSE Score	Performance metric for regression model
Correlation	Evaluation of association between two or more variables