

**CAPSTONE PROJECT**

Assignment 3

**Analyzing and predicting rent prices in Canada**

*Professor: Samer Al-Obaidi*

*Submitted on 18th July 2024*

**Group: Ottawa**

**Team Members**

|  |  |
| --- | --- |
| **NAME** | **Student ID** |
| Tanya Nimesh | N01579781 |
| Yshika Shrikhande | N01545656 |
| Apeksha Hipparagi | N01579432 |
| Shrutika Desai | N01580471 |
| Gurleen Banga | N01579137 |

|  |  |
| --- | --- |
| **Table of Contents** | |
| **Topics** | **Page Number** |
| Overview | 3 |
| Process Data Diagram | 4 |
| Existing IT Architecture | 7 |
| Solution Design | 11 |
| Visualizing our Predictions | 15 |
| The fit of the new solution into the existing IT architecture | 20 |
| Impact Analysis | 22 |
| Suggestions to further enhance the IT architecture | 23 |
| Challenges and future enhancement opportunities | 24 |
| Group Contribution | 26 |
| References | 27 |

**Overview**

The main objective of this capstone project was to forecast rent prices in areas of Canada using the data obtained from the Government of Canada. The project itself was oriented around cleaning up a big dataset so the trends in rent and their causes could be analysed and understood properly.

Each of the sources used in data gathering was a master dataset that was obtained from the Government of Canada website and contained data entities such as rent prices, types of houses, regions, economic characteristics, and population data. Run processing and cleaning of the data is notable for several stages, which are aimed at improving the quality and usability of the data collected. The options of the data frame were narrowed to the years 2019-2023, and unnecessary columns with data, such as “index adjustment factor”, were deleted.

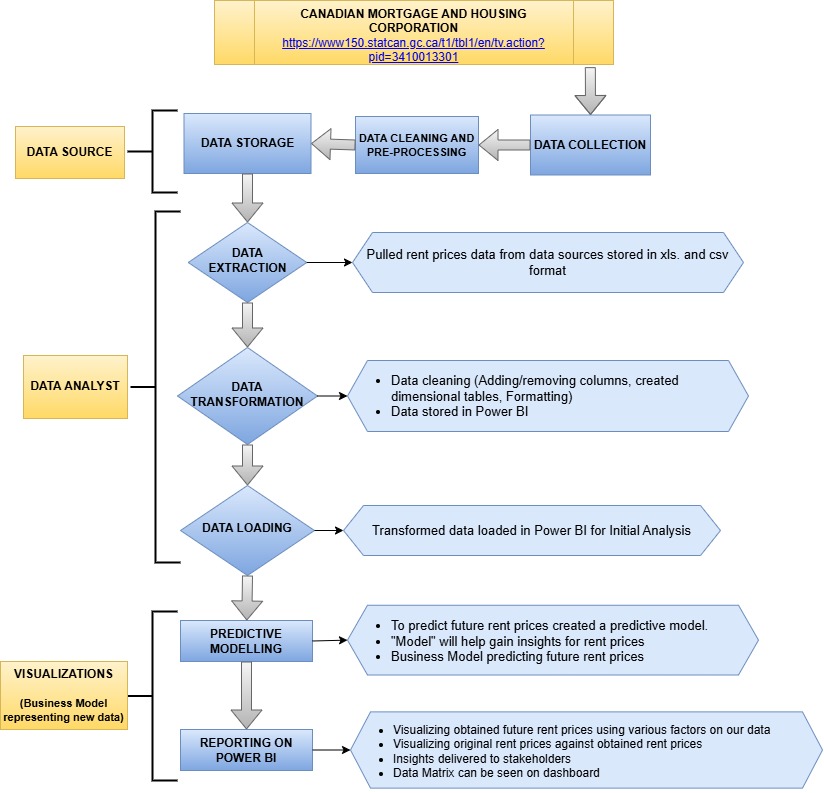
Data storage primarily made use of were Power BI and Excel files. Data was temporarily stored in the memory or exported in excel form in any of the processing stages. If data was not being processed, it was either stored in Excel files or datasets available on Power BI which made the storage very safe and convenient. Several issues were encountered during the project like determining genuine data source, handling with ineffective and repetition data, handling with unavailable data and constructing right kind of data presentation with new dimension tables. Some of the basic preprocessing steps that had to be followed included data transformation to ensure homogeneity across different data sets, normalisation, and outliers’ treatment had to be followed with a lot of care for them to have the expected impact. Furthermore, errors resulting from Data Source Paths into Power BI files were common when opening and using files from other system units or other devices.

Nevertheless, the accomplishment of the project was to rectify and formalise the dataset by means of harmonising and normalising it and relations within it.

This formed a good platform in analysing the real estate industry and determining the rents to be charged in Canada. They were able to gather all the necessary data to elaborate the trends in rental and prepare the ground for more qualitative analysis and visualisation.

**Process Data diagram**

The process flow below shows the data flow and the process followed in from historical data to predictive modelling which shows the prediction of Future rent prices.

**Fig 1.1 - Process Data Diagram**

1. **Data Source**

The raw data collected from data source CANADIAN MORTGAGE AND HOUSING CORPORATION. The data collected in Excel format (.xlsx). The URL provided in the chart likely points to a specific dataset on the Statistics Canada website. Statistics (Canada. (n.d.). *Table 34-10-0133-01)*

* **Data Collection** – The data collected from the website was transferred to Power BI database to facilitate easy access and management. The data stored transferred to Power BI stored in excel format.
* **Data Cleaning and Pre-processing** – Dealing with inaccurate data, duplicate records, and missing information was one of the main obstacles. Datasets were having gaps that needed to be filled up or removed to avoid interfering with the research. Eliminate duplicate records to keep the data accurate and clean. Identified and fixed any erroneous data inputs, including outliers, to ensure the study was reliable. All these processes took time, but they were necessary to maintain the data's integrity.
* **Data Storage** - Data storage in Power BI involves Excel files and datasets during processing, while at rest data is stored in Excel files or Power BI datasets. Power BI files have a storage capacity of 312 KB and Excel files 1113 KB. Accessing source and output data from different devices can cause errors due to changes in data source routes. To prevent issues, ensure data sources are accessible from all available devices or use relative routes.

1. **Data Analyst**

In this part of process will include the analysis of the historical data to delve into the prediction of future rent prices by transforming and cleaning the data. The crucial part will be the extracting and loading the data in Power BI to initialize the data analysis.

* **Data Extraction –** Inthis step, pulled the data stored in Power BI in .xls And .csv format for further analysis. The goal is to collect the necessary data that will be used for further processing and analysis.
* **Data Transformation –** This step is very crucial to identify errors in the dataset, for example unnecessary rows/columns, data redundancy, irrelevant data. This procedure ensures that the information being used is accurate, well arranged and relevant to trustworthy analysis.
* **Data Loading -** This process involves transferring cleansed and converted data into Power BI. Key tasks include importing the data into the Power Bi and ensuring data integrity throughout the process.

1. **Visualizations (Business Model representing new data)**

In this part of the process comprehensive reports are prepared to provide an overview of the research and insights, and interactive dashboards are developed to graphically depict the project's findings.

* **Predictive Modelling -** By understanding past trends and patterns, the model can provide valuable insights into future rent price movements. The use of tools like Power BI enhances the accessibility and visualization of these predictions, making the insights actionable and understandable for all stakeholders.
* **Reporting on Power BI** – Data Matrix can be seen on the dashboard once created. Visualization of rent prices across various provinces can be seen in which original rent price can be compared to future rent prices based on various factors on our dataset. This will help stakeholders to derive insights on future predictions of rent prices.

**Existing IT Architecture**

High level details of all major components of the IT system:

* **IT architecture:** While Shared Services Canada (SSC) oversees network and data center operations, Statistics Canada maintains security over Sensitive Statistical Information (SSI). Together, they constitute the foundation of Statistics Canada's IT infrastructure. To better address data security, risk management, and business continuity, the agency has amended its IT policies. Managing massive data sets and enabling intricate statistical analysis to depend heavily on these policies. (Statistics Canada. (2017, February)  
  In order to handle emerging risks, risk management entails creating detailed strategies that are periodically evaluated. A focus area for development is ensuring that all service providers fulfill security and operational needs by keeping an eye on third parties' adherence to IT standards. Plans for maintaining vital IT services—such as data backup and system recovery—in the event of outages have been put in place. (Statistics Canada. (2017, February)

With a focus on filling in any gaps in risk management and compliance, Statistics Canada is dedicated to ongoing development. Delivering high-quality statistical information is the agency's objective, and it is supported by this continuous review and improvement. (Statistics Canada. (2017, February)

* **Software tools:** Data is gathered, stored, and analyzed by Statistics Canada using a variety of computing tools and applications. Among them are: (Statistics Canada. (n.d.). *Statistical methods)*

1. Spreadsheets: Used for doing descriptive statistical analyses, adding columns and rows of numbers, and computing means. They produce pie charts, line charts, and bar charts, among other visualizations and summaries. (Statistics Canada. (n.d.). *Statistical methods)*
2. Databases: Help in the creation of summaries, aggregate data, and reports by storing, accessing, sorting, and analyzing data. (Statistics Canada. (n.d.). *Statistical methods)*
3. Specialized Programs: Designed to prepare data for publication by editing, cleaning, imputing, and processing the final tabular output. (Statistics Canada. (n.d.). *Statistical methods)*
4. Statistical software: Programs such as R studio are used for complex statistical studies like modeling, data processing, and the creation of summaries and infographics. (Statistics Canada. (n.d.). *Statistical methods)*

* **Data sources:** To acquire data, Statistics Canada uses both direct data collection from people and businesses and the utilization of pre-existing data from public and private sector entities. This strategy guarantees thorough coverage and superior data, offering crucial insights into sociological and economic trends. (Statistics Canada. (n.d.), *Where do our data come from?*)

1. **Direct Data Collection:** Direct data collection is done using the following sources:
2. Paper surveys: Standard forms that are mailed in.
3. Phone surveys: Using phone calls to gather data.
4. In-person interviews: Collecting information in person.
5. Online surveys: Web-based questionnaires used to collect digital data.
6. Crowdsourcing: Utilizing the public to provide data.

(Statistics Canada. (n.d.), *Where do our data come from?*)

1. **Using Existing Data Sources:** Statistics Canada has been using information gathered by other governmental bodies and businesses for more than a century. This comprises of:
2. Administrative Data: Information gathered during operational and administrative procedures.
3. Microdata Linkage: Adding information from many sources to improve the datasets.
4. Open Data: Making use of publicly accessible datasets.
5. Web Scraping: Data extraction from webpages and websites.

(Statistics Canada. (n.d.), *Where do our data come from?*)

* **APIs:** The APIs (Application Programmer Interfaces) of the Canadian government website, particularly those provided by Statistics Canada, play a crucial role in enabling access to aggregate data and metadata. These APIs allow data users to connect directly to public-facing databases, facilitating the consumption of a discrete amount of data points. The primary API service provided is the Web Data Service (WDS), which offers access to data and metadata released daily. Web Data Service (WDS) provides access to data and metadata released daily by Statistics Canada. The output of these is in JSON and XML formats. (Statistics Canada. (n.d.). *Application Programming Interface (API))*  
  The JSON API includes the following methods: (Statistics Canada. (n.d.). *Application Programming Interface (API))*

1. Product Change Listings:
   1. getChangedSeriesList
   2. getChangedCubeList
2. Cube Metadata and Series Information:
   1. getCubeMetadata
   2. getSeriesInfoFromCubePidCoord
   3. getSeriesInfoFromVector
3. Data Access; data changes for today, over time and full table:
   1. getChangedSeriesDataFromCubePidCoord
   2. getChangedSeriesDataFromVector
   3. getDataFromCubePidCoordAndLatestNPeriods
   4. getDataFromVectorsAndLatestNPeriods
   5. getBulkVectorDataByRange
   6. getDataFromVectorByReferencePeriodRange
   7. getFullTableDownloadCSV
   8. GetFullTableDownloadSDMX

The APIs are smoothly integrated into the overall IT ecosystem by offering standardized methods for accessing and retrieving data. They promote interoperability and data sharing across various components of the IT infrastructure, allowing data users to efficiently consume and utilize Statistics Canada's data. Comprehensive API documentation and user guides are available to assist developers and data users in effectively interacting with these services, ensuring seamless integration and utilization. (Statistics Canada. (n.d.). *Application Programming Interface (API))*

* **Servers:** Statistics Canada's Web Data Service (WDS) offers daily access to data and metadata, making it perfect for users requiring specific data point updates. It features 15 methods for developers to retrieve data, including getChangedSeriesList, getCubeMetadata, and getFullTableDownloadCSV. Comprehensive instructions for using WDS and SDMX REST web services can be found in their user guides. CANSIM table numbers have been renamed to Product IDs (PID), with a lookup table available for reference. The Delta File service provides extensive daily data and metadata updates. (Statistics Canada. (n.d.). *Web data services)*
* **Network integration – ports, drivers, firewalls**: A variety of technologies are used by Statistics Canada to integrate its network and provide strong security and effective data management. This comprises: (Shared Services Canada. (n.d.). *Network security strategy)*

1. Ports:Uses a variety of TCP and UDP ports, including TCP 22 for SSH and TCP 443 for HTTPS and secure connections. (Shared Services Canada. (n.d.). *Network security strategy)*
2. Drivers:Advanced drivers are included to control data flow and device communications. (Shared Services Canada. (n.d.). *Network security strategy)*
3. Firewall: Uses various firewall technologies, such as stateful inspection, packet-filtering, and next-generation firewalls (NGFWs) with deep packet inspection and intrusion prevention features. (Shared Services Canada. (n.d.). *Network security strategy)*

**Solution Design**

**Solution to solve the business problem**

Amid the vibrant and ever-changing landscape of Canada's real estate market, stakeholders sought a sophisticated tool to help them understand the intricate web of rent prices across various regions and housing types. The mission of this project was straightforward: delve into historical rent prices in regions with populations exceeding 10,000, predict future rent trends, and deliver actionable insights for real estate professionals, policymakers, and business owners alike.

To tackle the pressing issues of rent affordability and the dynamics of market demand and supply, we envisioned a robust solution. The solution contains a relational databases, time series model, and an interactive dashboard. It equips stakeholders with practical, data-driven insights into affordability, market conditions, and future rent price predictions.

The solution is a **Regression model** along with an interactive **dashboard** on PowerBI. The interactive data dashboard will be an assessment and visualization interface of various rental housing results for the stakeholders.

**Key Features of our Business Solution:**

* Affordability Metrics Visualization: Rent prices compared with households’ average income across the various geographical areas.
* Regional Comparison: Highlight regions with rent prices significantly higher than the national average.
* Impact Assessment: Visualize the impact of rising rent prices on various demographic groups.
* Vacancy Rates Analysis: Show the vacancy rates and job availability in different regions.
* New Constructions Tracking: Monitor the number of new rental housing units being produced.

**Detailed Solution Design steps**

1. **Data Collection and Pre-processing:**

We gathered the historical rent price data (2020-2023) from The Canada Mortgage and Housing Corporation (CMHC) website. For cleaning and pre-processing the dataset, we used PowerBI. Power BI is great because it can connect to various data sources like databases, spreadsheets, and cloud services, making it easy to bring in raw data for cleaning and transformation. Inside Power BI, there is a tool called Power Query that lets you clean, reshape, remove duplicates, split columns, and apply transformations using a visual interface, so you don’t need to be a coding expert. One of the best parts is that once you have set up, you are cleaning and preprocessing steps in Power BI, you can automate them. This means your data will automatically refresh at regular intervals or whenever new data is added. This ensures that every time you analyze your data, you are working with the most up-to-date and cleanest information available.

1. **Model Development:**

* Categorical Features: Used one-hot encoding to convert categorical variables into a suitable format for modeling.
* Numerical Features: Used standard scaling to normalize numerical data.
* Model Selection: We looked at several different models to figure out which one worked best with our data. We tested out the **Gradient Boosting Regressor, K-Nearest Neighbors (KNN) Regressor, Random Forest Regressor, CatBoost Regressor, and LightGBM Regressor**. Each model went through thorough training and testing, where we split our data into 70% for training and 30% for testing to make sure the models were tested on new data they had not seen before. To see how well they did, we used Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) metrics, which helped us see how accurate and dependable our predictions were.
* **Comparing the Root Mean Square Error of the predictive models used**

The square root of the average of the squared error between the expected and actual values is determined by RMSE. Since the model's predictions are closer to the actual values, a low RMSE value is suitable to choose. This depicts that the predictive model is accomplished in analyzing the trends in the historical data.

Interpretation: RMSE enables easy interpretation since the outputs are displayed in unit of the response variable. We tried predicting Rent values for the years 2025, 2026, 2027 by the following 4 predictive models and found the following RMSE:

|  |  |
| --- | --- |
| Model | RMSE |
| Random Forest Regressor | 89.259 |
| K-Nearest Neighbor | 153.938 |
| CatBoost Regressor | 99.157 |
| LightGBM Regressor | 101.721 |

Taking all the values above into consideration, the RMSE is the lowest at **89.259** among all the models where the **Random Forest Regressor** was applied.

* Accuracy: In the examined models, the Random Forest Regressor has the lowest value of RMSE therefore this model is least deviant from actual values and better determines patterns of data set than others.
* Robustness: From the above explanation, one can deduce that Random Forest is robust and can work effectively with almost any input data type and dimensions. Moreover, it minimizes overfitting and increases the theory’s generalization because it includes several decision trees.
* Interpretability: Random Forests also provide feature importance making it easier to identify which features have the most significant impact to the predictions even though the model is not very interpretable as the linear model.
* Scalability: Random Forests do not have problems with scalability, especially if they are designed to work on parallel computations as training is computationally intensive.  
    
  However, Random Forests are slightly easier to use than some of the other complicated models like Gradient Boosting because it involves lesser parameter optimization.

**Benefits of chosen methodology:**

It appeared that the Random Forest Regressor was one of the best ones to select due to its group of trees that provide reliable predictions. We utilized 100 estimators, and as such, its accuracy was favorable as was its speed. We maintained a fixed random state of 42, so if it ran with the same seed number, we would get the same results as before, which was useful to check on our solutions and compare with each other.  
  
Once our chosen model was ready, we proceeded to make some predictions. Hereby, we created the new data for the years 2025, 2026, and 2027 and made different combinations of features for the mentioned years. To prepare this new data with the same sequences for making the predictions we used the same steps. This acts like an experienced guide and with the help of our model, it provided a prediction for the rent prices in these years. To enhance understanding, visuals were created to enable people to easily comprehend the various data. These visualizations highlighted trends and predicted changes, offering a roadmap for policymakers and business owners. These images gave cognizance of where things are headed and showed what was coming for people making big choices.

**Comparing our methodology with others:**

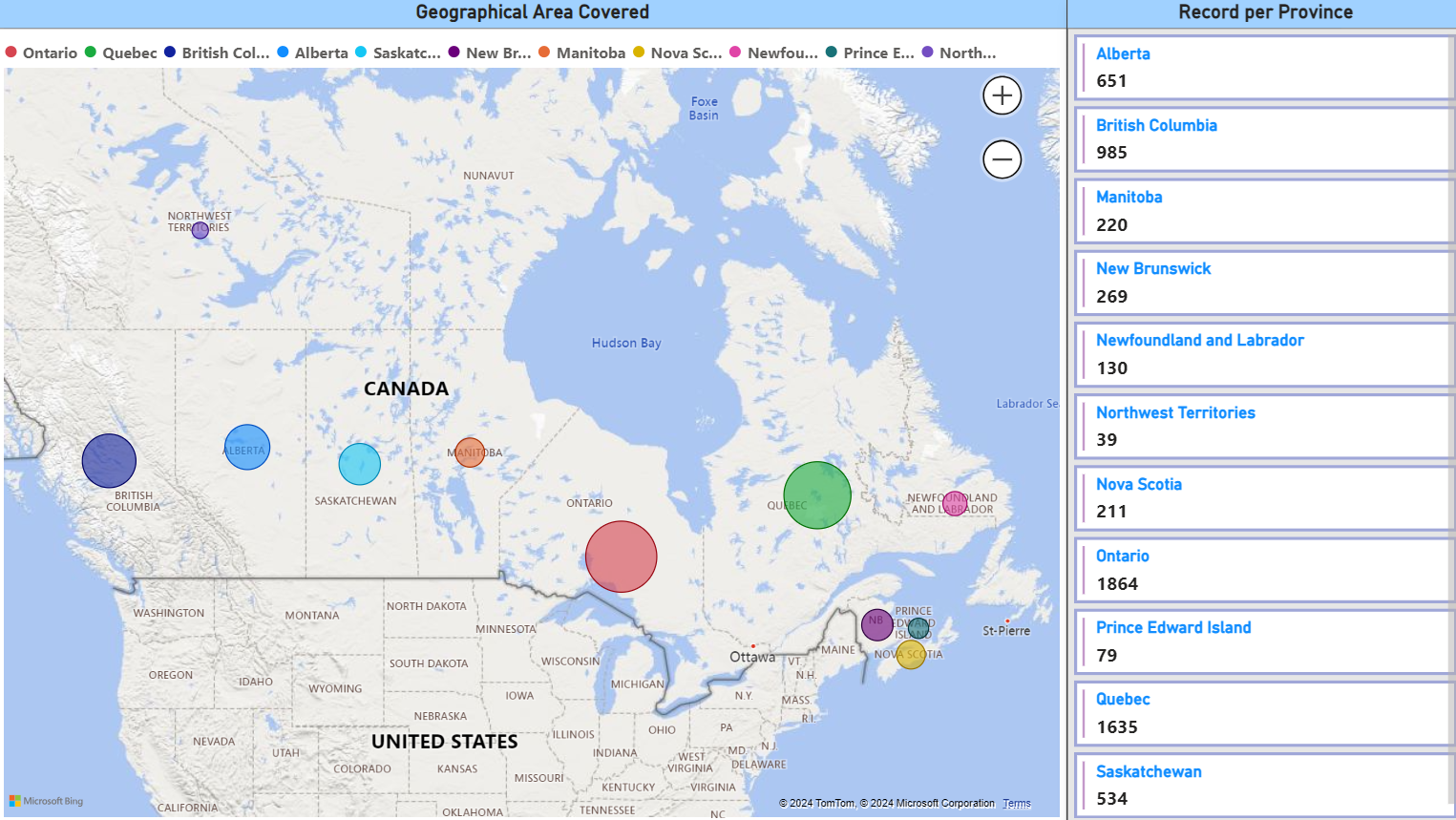
This ensured that our forecasts were exact and did not allow for variations apart from the level of complexity that was understood during the preprocessing phase and evaluation of several models. It provided at least a good structure, which was open for suggestion and further enhancement if better information was obtainable. Thus, our approach gave more detailed information because more complicated forms of relations within a dataset are considered in contrast to simpler forms like linear ones. The model's performance was improved by automated feature preprocessing, which decreased human error and guaranteed consistent transformations.

* Simple Linear Regression: Compared to ensemble techniques like Random Forest and Gradient Boosting, it may be more difficult to construct and may not be able to capture complicated relationships in the data.
* Single Model Approach: Performance may be less than ideal if a single model is used without comparison to other options. Robustness is ensured by our approach of analyzing several models.
* Manual Feature Engineering: Automated feature preprocessing using pipelines reduces human error and guarantees consistent application of transformations compared to manual feature engineering.

**Visualization out Predictions**

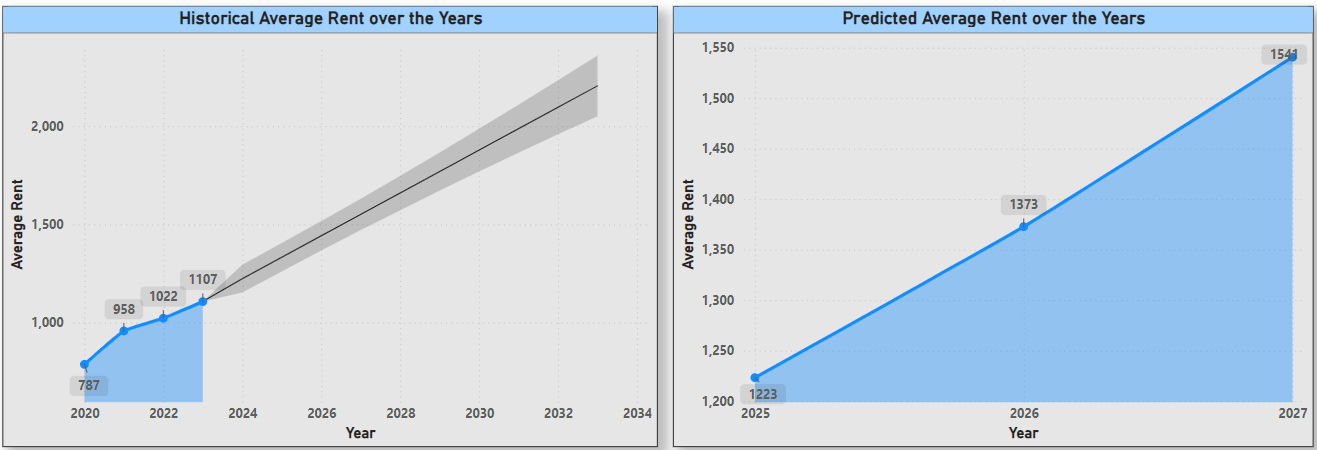
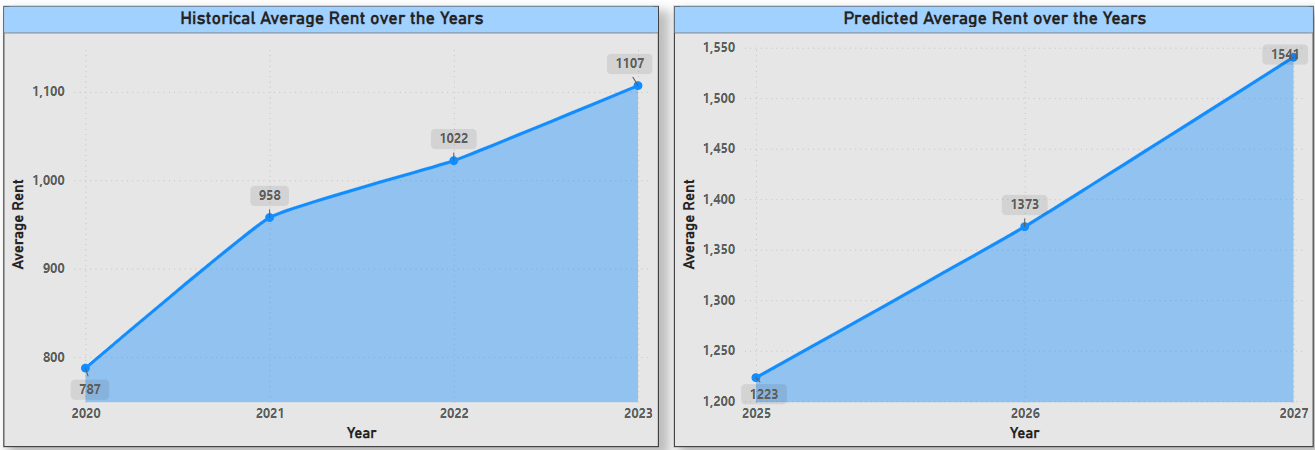
Our Dataset covers the Geographical data across Canada indicating the regions with population exceeding 10,000. Below is the visualization presented based on our predicted data which includes the rental market value of Canada.

**Geographical area covered**

**Fig 1.2 – The Geographical area that our dataset covers**

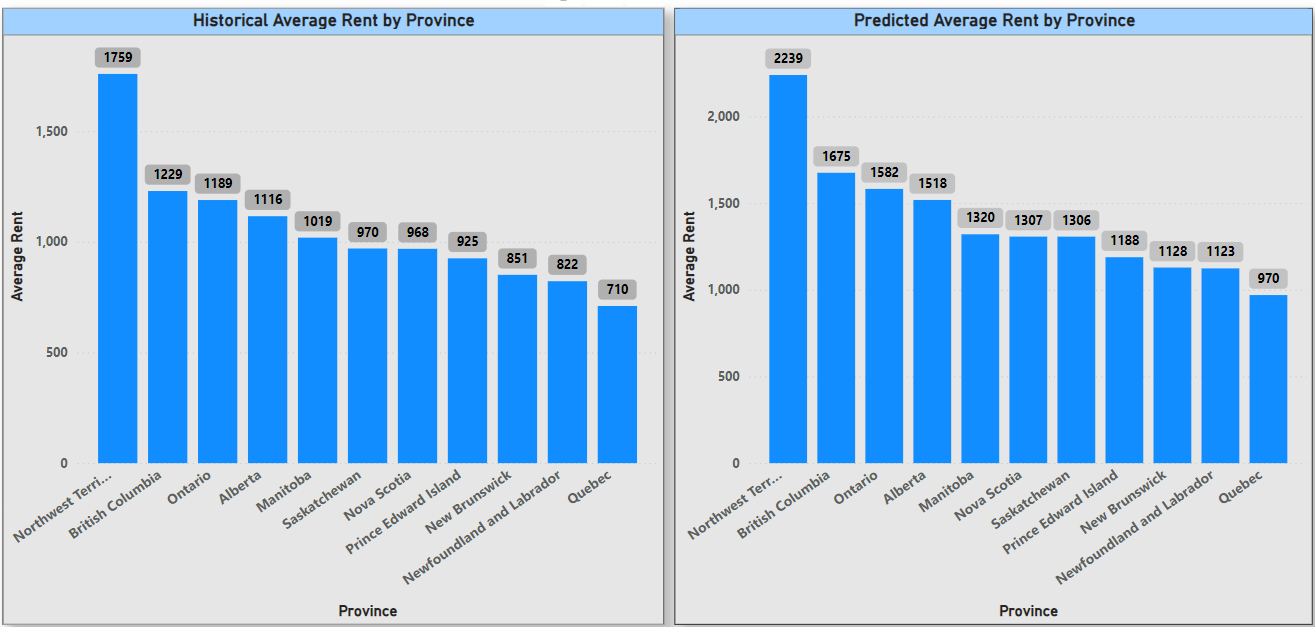
As per the above fig 1.2, the Geographical chart represents the Visualization of the rent prices of provinces in Canada exceeding with population 10,000 based on our historical data. This helps us to identify the rental market value over all provinces. Ontario has the highest value of rental market following with Quebec, British Columbia, Alberta, Saskatchewan, and other provinces.

**Comparing Historical and Predicted Average Rent price values over the years**

**Fig. 1.3 - Historical Average vs Predicted average rent over the years**

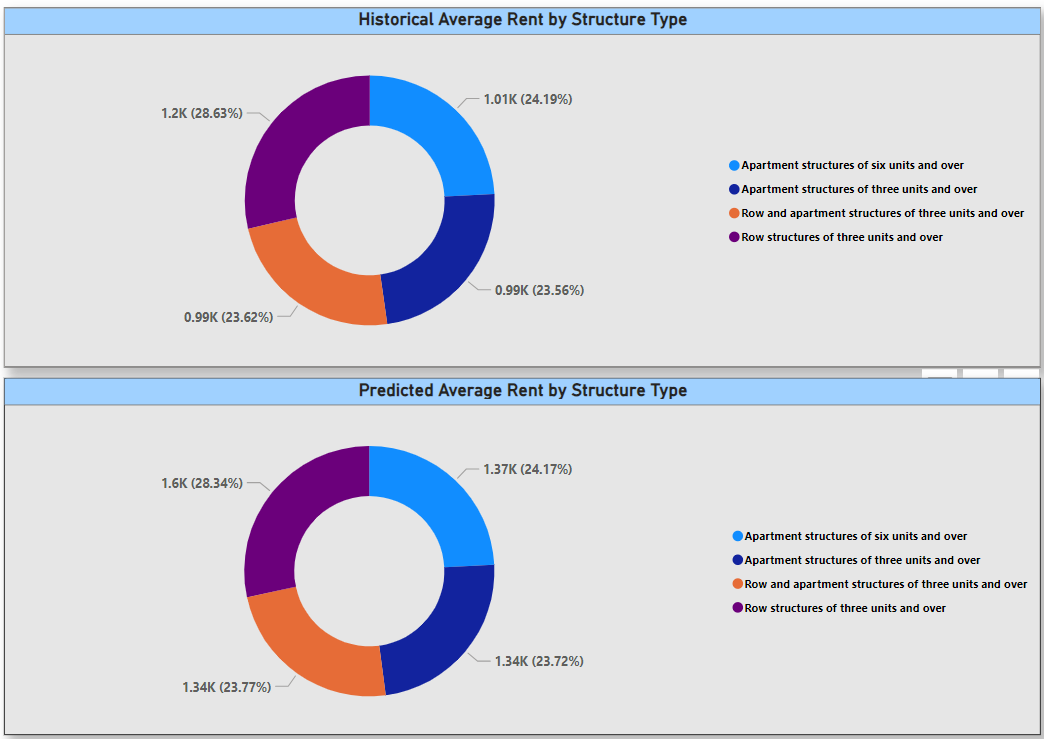
According to Fig. 1.3, from the year 2020 to 2023 the average rent has a continuously rising trend. According to the projected data for 2025-2027, such a tendency will persist, and soon rents will rise sharply. Based on the analysis of the given history and the projected trend, considerations can be made that there may exist an issue of rising rent prices with the increasing population in the future. Generalization of the charts: Such charts, when taken together, are coherent in presenting the average rent prices' ability to increase in history and future projections.

**Comparing Historical and Predicted Average Rent price values based on Province**

**Fig 1.3 – Average rent calculated by Province, Historical dataset vs New Predicted dataset**

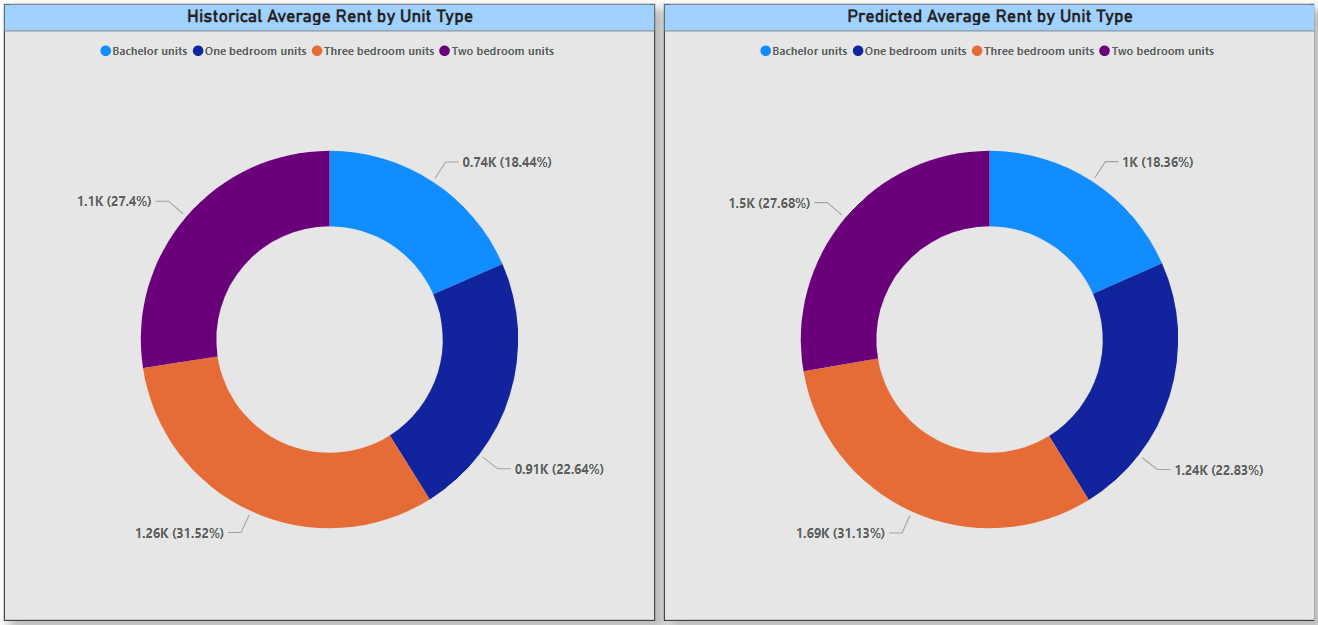
As you can see from above fig 1.3, the use of the charts presents the variation and trend of rentals across the provinces and the expected future rentals. The forecasted average rent is a bit higher than the historical average rent for all the provinces considered. As for the historical and predicted data analysis Northwest Territories remains to be the region with the highest average rents. British Columbia, Ontario, and Alberta are still among the provinces with the higher rent values in both the historical and the predicted data.

**Comparing Historical and Predicted Average Rent price values based on Structure type**

**Fig 1.4– Historical vs Predicted average rent by Structure type**

The predicted average rents for each kind of structure are higher than the historical values, which indicates the increase in rent prices. The analysis of both historical and predicted data reveal that the row structures of three units and over have the highest average rents that confirm the tendency, according to which the specified structure type is substantially more expensive. Thus, the charts illustrate the historical and predicted average rent prices change for the different structure types, pointing to the greatest rate of the rent prices raise for the row structures of three and more units.

**Comparing Historical and Predicted Average Rent price values based on Unit**

**Fig 1.5 – Historical vs Predicted average rent by Unit type**

The predicted data above is in Fig 1. 5 illustrates the current average rent on the units to be higher than the average rent based on the historic records, on all categories of units. Units with three bedrooms are the most expensive in average rent in historical and projected data. The average rent by unit types in historical and predicted datasets are similar in proportion.

**The fit of the new solution into the existing IT architecture**

**Integration into Existing IT Architecture**

The approach of solving the problem of analysing and predicting rent prices in Canada suggests the further smooth integration of the solution into the existing IT environment, using the solid data structure of Statistics Canada as the basis. The new IT ecosystem will also comprise of various kinds of data collection thus, all these components will be integrated into the new solution to improve its functionality and efficiency.

1. Data Collection and Sources: With access to data that is already collected by government agencies and non-governmental business, the proposed solution will be provided with a large amount of information that can help analyse sociological and economic trends.

2. Data Integration: The business models that have been developed are to implement and function strictly and solely in our own systems. This model has also been positively built in this organization. It is worth recalling that now, we have not incorporated our system with back-ends of any APIs from the authorities of Canada, including those offered by Statistics Canada. This is to mean that our model does not independently access real-time data feeds from the official Statistics Canada API services.

3. Software Tools and Platforms:

* Today’s software that are available for statistical analysis and data management will remain useful in the new offering solution. Software tools such as Anaconda Navigator, Jupyter Notebook and power BI were used into making this business model.
* Jupyter Notebook allows for interactive data analysis, and we could run Python codes individually, view results immediately and make iterative changes based on the output.
* Power BI excels in creating interactive and visually appealing dashboards and reports. We could present our analysis results in an engaging and accessible format by analyzing data from multiple sources.
* Due to Power BI’s advanced data visualization capabilities, we could uncover trends and anomalies in rent prices. The interactive nature of Power BI dashboards will allow future users to explore the data dynamically.

**Benefits of the New Solution**

1. Enhanced Data Utilization: This new solution will enhance the benefits of the existing IT architecture, which is having effective data collection and integration functionalities. This is because more information will be available and thus better analysis on the prices of rent and market trends will be given.

2. Improved Decision-Making: By using the developed solution, the identified stakeholders will receive the data dashboard that will help in decision making. Here, the stakeholders can really paint the picture of affordability indices, relative to region on district, or year, they could even paint the picture of vacancy rates.

3. Predictive Analytics: Incorporation of time series predictive model will enable the rents to be forecasted hence giving the stakeholders adequate time to act on the changing market. This predictive capability will be vital for directing and for the design of guidelines and other strategies.

4. Scalability and Flexibility: This means that the new solution is to incorporate the balance of growth in both data volumes and users to the social media. The architecture will allow for future gains, for instance, the inclusion of other databases and the development of the analytical features.

5. User Engagement and Accessibility: The solution will be easy to use through incorporating the use of dashboards as well as APIs for the end-users. This is a future predictive model which will help the users to predict the possible outcome.

To sum up, the new solution will complement in the IT environment and enhance ways of using, analysing, and analytical predictiveness of data. Therefore, listing out the main findings, we prerecord the applicability and the relevance of the presented solution.

**Impact Analysis**

* **Risk analysis of new model/features:** The new model faces several risks, including data quality issues such as redundancy, incorrect entries, and inconsistent formats, system design issues like inadequate scalability, process design challenges in integrating diverse data sources, technical issues like network configuration, server performance, and security vulnerabilities, and data loss risks. To mitigate these, regular data cleaning, preprocessing, and standardization practices are recommended. Regular audits, performance testing, and strong security protocols are also necessary.
* **Different application modules:** Data is collected from various sources, including surveys, administrative data, and web scraping, and integrates it using APIs. API Integration facilitates data retrieval and integration through Web Data Service (WDS) and other APIs.4
* **Perform Test scenarios:** Test scenarios cover various aspects of the new solution's integration and functionality, ensuring it meets the requirements and enhances the existing IT architecture for rent price analysis in Canada. Power BI is used for Data Cleaning and pre-processing in which Power Query tool is used for cleaning, reshaping, removing duplicates, splitting columns, and transformations.
* **Model Testing**: The Random Forest Regressor, K-Nearest Neighbor, Catboost Regressor, and LightGBM Regressor that are used to build models have been thoroughly tested. With the lowest root mean square error (RMSE) of all the studied models, the Random Forest Regressor can identify data patterns most effectively and generate the most accurate predictions. Among the models examined, the Random Forest Regressor results in accurate predictions with the lowest RMSE value of 89.258, as previously mentioned. This model testing provided us with insights to help us opt for the most effective model.

**Suggestions to further enhance the business model**

1. **Enhanced Data Governance:** Increased surveillance over data needs to be enforced to ensure that the data collected is credible, reliable and coherent in all places and from all units that compile it. This includes activities such as formatting of data, decision on who is responsible for a specific data or information and enhancing the normally accepted norms of data stewardship.
2. **Improve Scalability and Flexibility**: Design IT systems which can easily accommodate additional pressures from users or volumes of data in the future. It should be extensible and modular so that new services or data can be incorporated later if available.
3. **Enhance User Experience and Accessibility:** Based on Statistics Canada’s IT context and the mentioned findings, it can be concluded that a consistent improvement of the accessibility and utility of currently and potentially available dashboard or API will indeed be systematic and will need continuous work. These include updates and changes in accordance with the feedback and testing from the users to ensure easy reach, comprehensible graphics, and flexible operations to fulfill as many users’ needs as possible.
4. **Enhance Database Management:** This includes deploying automated tools and workflows within the IT infrastructure to streamline these processes. The model can be further enhanced to have live data feeding done from the Canadian Housing website into the model to make predictions for every year going ahead. The dashboard can be scheduled to run daily through ETL pipelines from the backend that updates it.
5. **Improved User Experience:** Focus on making data more accessible and user-friendly through dashboards and APIs. To guarantee usefulness across all stakeholder groups and to iterate on design changes, conduct user input sessions.
6. **Advance Security measures:** Enhance cybersecurity by implementing advanced threat detection and response tools, such as proactive threat hunting, ongoing monitoring, and the newest security technologies like endpoint protection and next-generation firewalls.
7. **External factors affecting rent prices:** In the future, the rental prices can be influenced by the following changing factors: Government regulations that wish to control the rates that landlords charge, weather conditions that can lead to fluctuations in rental prices depending on the season, political & economic conditions that influence the economy of the country where the property is located, and natural calamities that may cause changes in rental prices. Some of these factors can complement the business model to provide a more agile and accurate model to enable more realistic and exact prices.

**Challenges**

1. **Data Quality Issues:**

* Standardization: There were inconsistencies in format and structure of data, these needed to be standardized to maintain data consistency and integrity throughout the project.
* Data Entry Errors - There were many cases of incorrect data entry, such as numbers not within the permissible range or not formatted correctly. They need to be fixed to guarantee data consistency and correctness throughout the project.

1. **System design issues:** Since the system must manage growing amounts of data and user demand without experiencing performance degradation, scalability is essential. Since the data sources and services develop in the future, this entails creating an architecture that can grow with it.
2. **Process design issues**: Combining data from different sources, including open data, administrative data, and linked microdata, is referred to as data integration, and in this case, it is virtually always challenging to ensure that data are integrated with no clashes. Firstly, so that large chunks of data can be managed reliably and in a short span of time, efficient data processing is implemented. This entails data conversion where data collected is transformed into workable forms, checks, and ensuring that updates are received on time. Furthermore, the paradigm shift towards user inclusion requires development of user-friendly Interfaces such as dynamic dashboards and APIs alongside enhanced training and support services to ensure optimality in utilization of all the data compiled.
3. **Technical issues**: Technical issues at this level are understanding how best to configure the network to properly integrate ports, drivers, firewall and other system components while also ensuring scalability of the architecture to accommodate more data and additional users and assuring that the servers that provide data can effectively handle the loads being generated by new data processing and retrieval functions. These problems need to be solved to ensure that the system is strong, safe, and adaptable enough to stand and develop amidst the existing and emerging competitions.
4. **Loss of Data**: Loss of data happens during the transmission, storage, or processing, and this must be avoided at all costs. This risk encompasses errors in data processing procedures, including poor data transfer between systems, corruption or loss of data because of storage hardware failure, or inherent weaknesses in data processing algorithms that inadvertently overwrite storage media or sections thereof. To minimise this risk, adequate provisions ought to be made for data backup and recovery, ensure that data is secured during transit by encoding among other methods and regularly conduct checkups and investigations to discover processes and situations that lead to data loss.
5. **Prediction Model error:** When trying to develop the predictive model for the correlation coefficient, even though the code was executed repeatedly, the values and the mean for each year were identical. The expanded features proved to improve the predictive model.

|  |  |  |
| --- | --- | --- |
| **Group contribution** | | |
| **S no.** | **Content** | **Stakeholder** |
| 1. | Process data diagram | Gurleen, Shrutika, Yshika |
| 2. | Existing IT Architecture | Apeksha, Tanya |
| 3. | Solution Design and Choice of Methodology | Apeksha, Tanya |
| 4. | The fit of the new solution in the existing IT architecture | Yshika |
| 5. | Impact Analysis | Gurleen |
| 6. | Challenges and future enhancement opportunities | Shrutika |

**References**

* Statistics Canada. (2017, February). Review of Information Technology Infrastructure - Access and Continuity Management (Project Number: 8059097). Retrieved from <https://www.statcan.gc.ca/en/about/audit/pn-80590-97>
* Statistics Canada. (n.d.). *Statistical methods*. Retrieved July 18, 2024, from <https://www.statcan.gc.ca/en/subjects-start/statistical_methods>
* Statistics Canada. (n.d.). *Where do our data come from?* Retrieved July 18, 2024, from <https://www.statcan.gc.ca/en/our-data/where>
* Statistics Canada. (n.d.). *Application Programming Interface (API)*. Retrieved July 18, 2024, from <https://www.statcan.gc.ca/en/microdata/api>
* Statistics Canada. (n.d.). *Web data services*. Retrieved July 18, 2024, from <https://www.statcan.gc.ca/en/developers/wds>
* Shared Services Canada. (n.d.). *Network security strategy*. Retrieved July 18, 2024, from <https://www.canada.ca/en/shared-services/corporate/publications/network-security-strategy.html>
* Statistics Canada. (n.d.). *Table 34-10-0133-01: Mortgage lenders, estimated number of new residential mortgage loans by mortgage insurance status, type of lender and amortization period*. Retrieved July 18, 2024, from <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3410013301>