

**CAPSTONE PROJECT**

Assignment 2

**Analyzing and predicting rent prices in Canada**

*Professor: Samer Al-Obaidi*

*Submitted on 7th July 2024*

**Group: Ottawa**

Team Members

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

**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 data set 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. Values in the rent column were imputed in the form of zeroes and a new status column was created to highlight values as either ‘known’ or ‘unknown.’ All tables were consistently named, and three more-dimension tables were developed to improve the possibilities for the future when categorising the data and making analysis.

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, and these are: determining genuine data source, handling with ineffective and repetition data, handling with unavailable data and constructing right kind of data presentation. 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.

**Data sources**

The Canada Mortgage and Housing Corporation (CMHC) is the source of the average rent statistics for locations with a population of 10,000 or more. The national statistics office, Statistics Canada is the vehicle through which the Government of Canada gathers and manages the data. This dataset focuses on Canadian cities with high population densities, provides comprehensive data on average rental costs in several Canadian regions.

One of the main sources of information about the Canadian housing market, the Canada Mortgage and Housing Corporation, is where the data comes from. To provide information on rental trends, affordability, and housing conditions across the nation, CMHC regularly polls the public and compiles statistics about the housing and rental markets.

The official website of the Government of Canada offers users the ability to see and download the dataset. The data is accessible to the public and may be employed for diverse analytical objectives, guaranteeing openness and evidence-based decision-making about housing-associated issues.

Statistics Canada. [Table 34-10-0133-01 Canada Mortgage and Housing Corporation, average rents for areas with a population of 10,000 and over](https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3410013301) .

**Key data entities**

* Rent Price
* Housing Types
* Regions
* Economic Indicators
* Population Data

**Data flow**

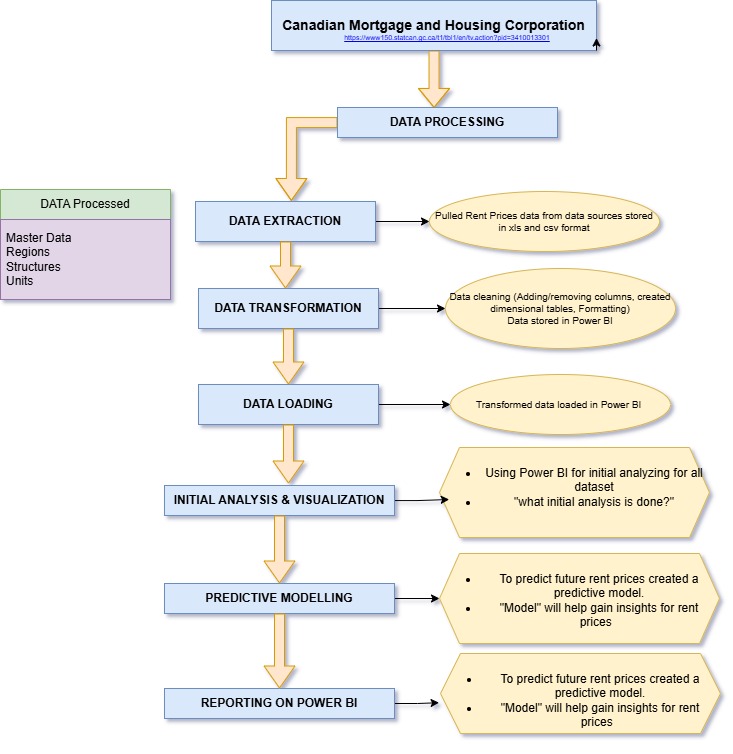


Fig (1.1)

**2.2 Data Dictionary**

**MASTER TABLE:**

|  |  |  |  |
| --- | --- | --- | --- |
| **MASTER TABLE** | | | |
| **Column Name** | **Data Type** | **Format** | **Description** |
| RentID | Text | Alpha-Numeric | Primary key for the data uniquely defining each record |
| Year | Whole numbers | YYYY | The year for a particular rent record |
| Region | Text | Alpha | The region where the property is located |
| Province | Text | Alpha | The province where the property is located |
| Structure Type | Text | Alpha | Type of structure (e.g., apartment, house) |
| UnitType | Text | Alpha | Type of unit (e.g., Bachelor, 1-Bedroom, 2-Bedroom) |
| RentValue | Whole numbers | Currency | The value of rent in currency (CAD $) |
| Status | Text | Alpha | The value of rent is known or unknown |

**DIMENSION TABLES:**

|  |  |  |  |
| --- | --- | --- | --- |
| **REGIONS** | | | |
| **Column Name** | **Data Type** | **Format** | **Description** |
| Region ID | Whole numbers | Numeric | A unique identifier for each region to distinguish between different regions. |
| Region | Text | Alpha | Name of the region or designation given to the region |
| Region Type | Text | Alpha | Classification of region for e.g. - County, Municipality |
| Province | Text | Alpha | Name of the province in which region is located |
| AreaSqKm | Decimal number | Decimal numbers | The geographical area of the region is square kilometers. This indicates the size of the region in terms of its land area. |

|  |  |  |  |
| --- | --- | --- | --- |
| **STRUCTURE** | | | |
| **Column Name** | **Data Type** | **Format** | **Description** |
| StructureType ID | Whole numbers | Alpha | A unique key identifier to describe each Structure Type |
| StructureType Name | Text | Alpha | Type of structure (e.g., apartment, house) |
| MinUnits | Text | Numeric | Minimum number of units within the Structure type |
| MaxUnits | Text | Text | Maximum number of units within the Structure type |
| TypicalFloors | Whole numbers | Whole numbers | The typical number of floors in the structure Type |
| HasElevator | Text | Alpha | indicates whether the structure type typically has an elevator |
| ParkingAvailable | Text | Alpha | Indicates whether the structure type typically has parking available |

|  |  |  |  |
| --- | --- | --- | --- |
| **UNITS** | | | |
| **Column Name** | **Data Type** | **Format** | **Description** |
| UnitType ID | Whole numbers | Roman Numbers | A unique key identifier to describe each Unit Type |
| UnitType Name | Text | Alpha | Type of unit (e.g., Bachelor, 1-Bedroom, 2-Bedroom) |
| NumberOfBedrooms | Text | Numeric | Indicates number of Bedrooms in each unit |
| MaxOccupancy | Text | Numeric | Maximum number of occupants allowed in the unit |
| SqFoot | Whole numbers | Numeric | Indicates the size of the unit in terms of floor space |
| IsStudio | Text | Alpha | Indicates whether the unit is a studio apartment |
| EstimatedUtilities | Text | Numeric | An estimate of monthly utility expenses associated with the unit |

**2.3 Data Cleaning**

**1. Dataset Scope Refinement**

The original timeline for dataset scope refinement was 1987-2023, with a focus on recent, relevant market data, and the refined timeline is 2019-2023.

**2. Column Optimization**

* Removed irrelevant columns
* Symbols
* Decimal
* UOM\_ID
* Units
* Scalar\_ID
* UOM
* Vector
* DGUID
* Coordinates

**3. Value Standardization**

The 'Value' column was scanned for 480 blank rows out of 17,360 total, resulting in a 0.028% error. The action was to replace these values with 0.

**4. Status Column Implementation**

The 'Status' column has been implemented, with T (True) representing known values in the 'Value' column and F (False) representing 0 values in the 'Value' column.

**5. Column Renaming**

The application of standardized naming conventions was implemented to maintain consistency across the dataset.

**6. Dimension Tables Creation**

Three new dimension tables were created, enhancing categorization, data organization, relationships, and enabling more detailed analysis capabilities.

**7. Outcome**

The outcome was a refined, standardized dataset with enhanced data structure and relationships, providing a solid foundation for real estate market analysis.

**Data storage**

Data in flight: Currently, Excel files and Power BI are used to handle data. Data is momentarily stored in Excel files during intermediate processing steps and in RAM during processing

Data at Rest: Data at rest is kept in Excel files or Power BI datasets while not in use. While data within the Power BI service is maintained as Power BI datasets, Excel files are preserved either on secure file systems or in the cloud.

Storage till date: At this stage, the overall storage capacity for Power BI files is 312 KB and for Excel files is 1113 KB.

Although there aren't any security guidelines for accessing the source and output data, Power BI's behavior when files are opened on various computers is a crucial factor to consider. Errors might occur when a Power BI file is accessed from a different device because the data source route / pathway changes. This is because the first configuration of the network location or file path is required. Make sure the data sources are reachable from all available devices or utilize relative routes and centralized storage locations where you can, to prevent problems like this.

Below is an example of the error:

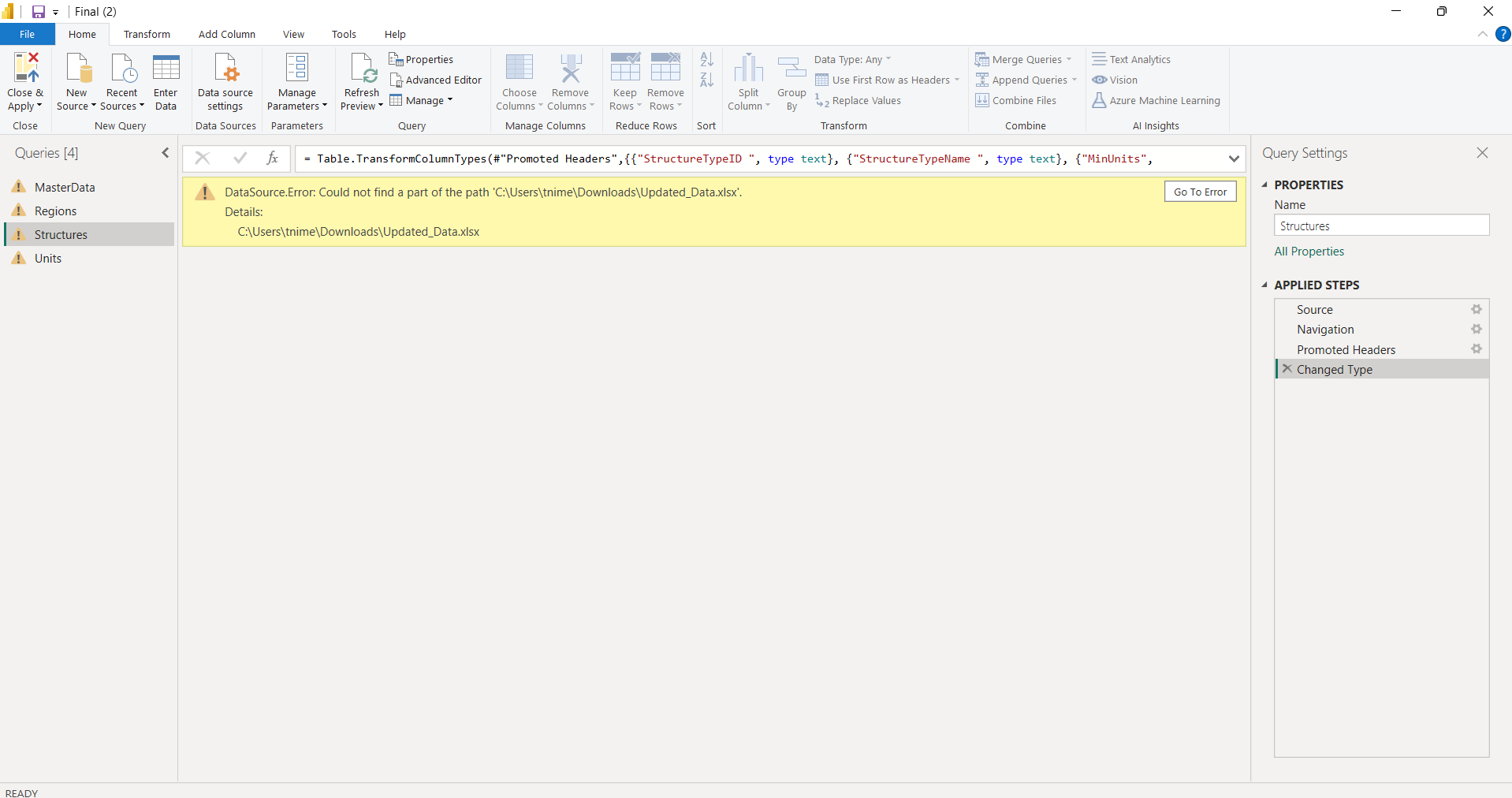
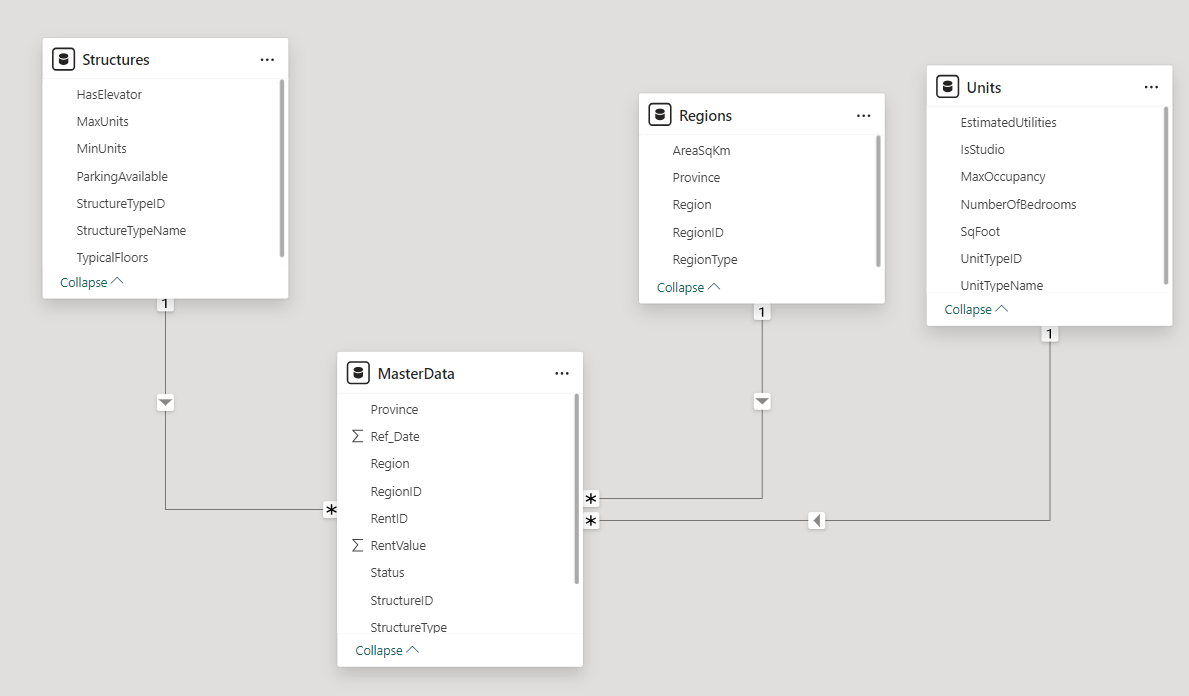


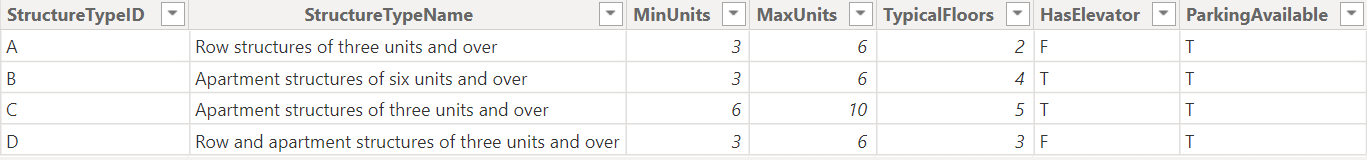
Fig (1.2)

**Data output**

In fig 1. 1, you can view model aspects of the dataset peculiarities. Master data contains a one-to many relationship; it has foreign key to relate the data with tables.

Fig (1.3)

First the ‘Structures’ Table field can be described as information about structures – different constructions such as buildings, facilities, etc. It has field names for instance ‘StructureID,’ ‘MaxUnits,’ ‘ParkingAvailable,’ and other related fields.

Fig (1.4)

Subsequently, the ‘Regions’ Table is used to store information in relation to various geographical regions and includes fields such as ‘RegionID,’ ‘Province,’ and ‘RegionType’.

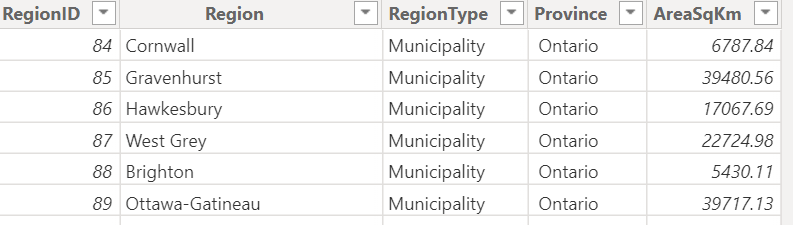
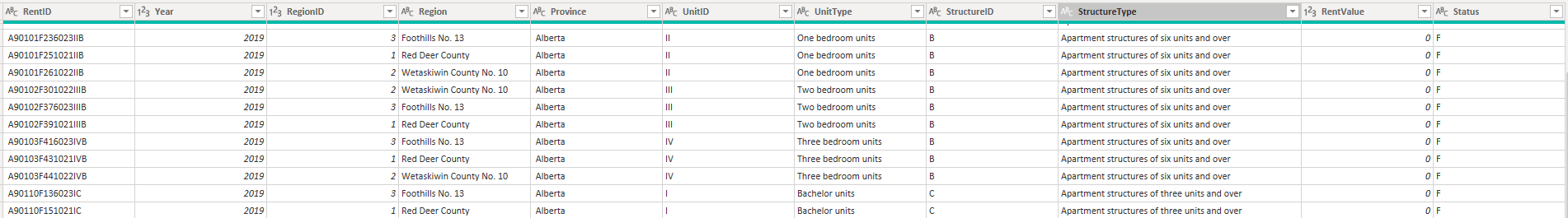
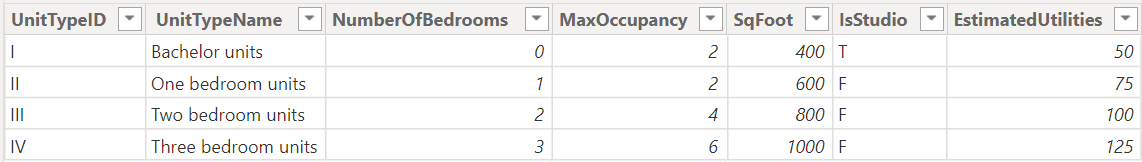


Fig (1.5)

The ‘MasterData’ Table contains information of rental property and the fields of ‘RentID,’ ‘Rentvalue,’ ‘Status,’ and other relevant columns containing information from rest of the tables to strengthen each record.

  
  
Fig (1.6)

Lastly, the ‘Units’ Table contains information on various units such as apartments, condos among others. They are labels including UnitTypeID, IsStudio, and NumberOfBedrooms.

Fig (1.7)

**Challenges**

1. **Data Source Identification**

Finding trustworthy data sources was really challenging. We felt like treasure hunters in a vast desert. We slogged through piles of useless info to find gems. Cross-checking took hours to make sure we had solid relevant facts. The whole thing dragged on forever - talk about a needle in a haystack! It left us worn out, but we kept at it. No quick wins here just good old-fashioned detective work to dig up the goods. Man, what a slog - but worth it in the end to get the real deal.

1. **Data Cleaning and Preprocessing**

Dealing with inaccurate data, duplicate records, and missing information was one of the main obstacles. Several datasets have gaps that needed to be filled up or removed to avoid interfering with the research. We worked hard to find and eliminate duplicate records to keep the data accurate and clean. We also had to carefully identify and fix 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.

1. **Missing Data in dimension table**

The fact table has several missing portions, making it difficult to link records with the dimension tables appropriately. Each piece of data must be manually matched and identified one at a time. The lack of different data necessitated our tedious procedure of painstakingly cross-referencing every element in the fact table with the relevant area in the dimension table. It took a great deal of time.

1. **Data Transformation**

We must normalize the data consistently to preserve consistency across all datasets. Throughout this procedure, the numbers measured on different scales need to be modified to come inside a given range. We had to be very careful while normalizing the data to make sure we didn't lose any crucial information or distort the data. Standardizing the data while maintaining its connections and integrity was essential, necessitating careful planning and accurate calculations.

1. **Handling Outliers**

Managing outliers posed a significant obstacle during the normalization process. These large numbers of complicated things might drastically change the scale and lead to distorted results. It might be difficult to find significant patterns if, for example, the range of the remaining values is compressed while normalizing the data without considering the few extremely high or low values in the dataset.

1. **Data Source error in Power Bi**

When a Power BI file is accessed from a different device, errors may arise because the data source path is altered. This is because setting up the file path or network location first is necessary. Prevent such issues by ensuring that the data sources are accessible from all available devices or by using relative routes and centralized storage sites where possible.

1. **Data Visualization Challenge**

Powerful visualizations that accurately portray complex data insights proved to be a difficult task, especially considering the amount and complexity of the data. To create thorough reports that elegantly communicate these results, tedious preparation and precise clarity were required.

|  |  |  |  |
| --- | --- | --- | --- |
| **Group contribution** | | | |
| **Page no** | **Points** | **Content** | **Stakeholder** |
| 1. | 10 | Data sources and key data entities and flows | Gurleen |
| 2. | 40 | Data dictionary | Apeksha  Tanya |
| 3. | 10 | Data cleaning | Gurleen |
| 4. | 10 | Data storage | Shrutika |
| 5. | 20 | Data output | Yshika |
| 6. | 10 | Challenges | Shrutika |

**References**

Data source: [**https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3410013301**](https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3410013301)