McGill University

INSY 695: Advanced Topics in Information Systems

24Seven Store Prediction & Forecasting

Presented to

Prof. Vivek Saahil

Shamas, Mahrukh (261149453)

Sohail, Syed Ammad (261162304)

Suri, Jatin (261152263)

Tang, Yusen (260890520)

Tian, Xiaorong (261150556)

Introduction to the Problem

In the evolving retail landscape, the strategic expansion of convenience store chains like "24Seven" is paramount to maintaining competitive advantage and fulfilling customer demands efficiently. The transition from a local shop to a chain with 10 stores across Canada underscores the brand's commitment to accessibility and customer service. As 24Seven prepares to open its 11th store, the challenge lies in leveraging data-driven insights to pinpoint the optimal location for expansion. This involves an intricate analysis of various factors including customer demographics, historical sales data, product preferences and geographic accessibility.

Proposed Solution

As "24Seven" embarks on the expansion of its successful convenience store chain with the opening of its 11th store in Canada, the decision-making process for selecting a new store location will be deeply rooted in data-driven insights. Our proposed solution is to develop an advanced, interactive dashboard that integrates various datasets, providing comprehensive analytics and visualizations to facilitate strategic decision-making.

Our approach harnesses the synergy of multiple datasets, encompassing sales trends, customer demographics, product popularity, geographical distribution, regional climate influences and broader census demographics. By integrating this wealth of data, we aim to construct a multi-faceted analytical model that not only highlights current performance metrics but also forecasts future trends.

To determine the most viable location for a new store, we have devised a multifaceted strategy that hinges on the calculation of an "Expansion Score". This score is an aggregate metric derived from a weighted average of several key indicators: the Total Store Revenue by City, Revenue Forecasting and a Saturation Score. The Saturation Score inversely correlates with the number of existing stores, ensuring that areas with fewer stores are considered more favorably to prevent market oversaturation. By integrating these elements, the Expansion Score provides a balanced view, considering various dimensions of market potential and competitive landscape, to guide our strategic store placement decisions.

Methodology

Datasets Used:

1. Historical Sales Data: This dataset includes transaction details such as Sales Date, Product Category Key, Customer Key, Store Key and Total Amount, enabling us to analyze sales trends, popular products and store performance.
2. Customer Data: This dataset provides insights into customer demographics including age, gender and location, which are crucial for understanding the target market and customer preferences.
3. Product Categories Data: Lists the primary categories of products sold, helping us identify which products are the most popular and potentially guiding inventory decisions for the new store.

4. Stores Data: Contains the unique store key, the store name and the postal code for each location. This dataset would be used to map the current stores' distribution across Canada.

5. Regional Climate Data: To address the impact of regional climate on sales and development of 24Seven, we collected and organized the regional climate data of target cities. The dataset was downloaded from the official sources and included temperature and rainfall information. (Source Link – https://climate.weather.gc.ca/climate_normals/index_e.html)

6. Census Demographic Data: To enhance our understanding of customer demographics and market potential, we integrated census demographic data from Statistics Canada. This dataset includes detailed information on age, gender, income level after tax and household type/size, enabling targeted marketing and strategic store placement. (Source Link – https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=E)

7. Economic indicators: We provide a comprehensive examination of important economic metrics and the state of the Canadian economy, with particular attention to information on GDP, population, unemployment and the economic output of the retail sector in Montreal, Calgary and Toronto. (Source Link - https://www150.statcan.gc.ca/n1/en/type/data?MM=1#tables)

8. Traffic Flow Data: Our comprehensive analysis provides an in-depth look at the dynamics of traffic flow across key urban areas, leveraging detailed Annual Average Daily Traffic (AADT) data. This report focuses on quantifying and understanding traffic patterns, including the movement of bicycles, pedestrians, and motor vehicles. By examining traffic counts meticulously gathered over various periods, we offer valuable insights into the volume and trends of urban mobility in Montreal, Calgary, and Toronto.

   (Source links: https://open.toronto.ca/dataset/traffic-volumes-at-intersections-for-all-modes/
   https://data.calgary.ca/Transportation-Transit/Bike-and-Pedestrian-Counts/pede-tz7g/about_data,
   https://www150.statcan.gc.ca/n1/pub/71-607-x/71-607-x2022018-eng.htm)

The exploration, pre-processing and the usage for the external dataset is provided in the Appendix-1.

Power Query steps:

In our comprehensive data preparation process, we carefully refined and transformed the datasets to ensure optimal analysis and visualization outcomes. Initially, we standardized the structure across all tables by designating the first row as headers and validating that each column adhered to its appropriate data type. For the demographics-related data, which included variables such as Age, Annual Income, Household Size, and Gender, common preparatory steps involved the removal of extraneous rows from both the top and bottom of the datasets. Given that we possessed individual-level demographic data for three distinct cities, we employed the Append Query function to consolidate this information into a single, unified dataset.

Regarding the traffic-based data, our approach entailed additional steps to facilitate in-depth analysis. Specifically, we generated separate columns to aggregate the data, followed by the application of the Unpivot Columns feature. This critical step enabled us to create dynamic charts that categorically distinguish among bikes, pedestrians, and cars. Furthermore, we implemented a filter across all tables to exclusively include data from 2021 onwards, ensuring the relevance and timeliness of our analysis.

For the Population table, which was initially not formatted to display population figures by city, we executed a series of transformations. After discarding irrelevant bottom rows, we used the Unpivot Columns functionality to reshape the data, thereby enabling population comparisons across different cities.

In addressing the missing values within the Calgary traffic flow dataset, we utilized Power Query's Statistics functionality to determine the median, which we then employed as a basis for imputing missing values through the Replace Values feature. The integration of bike/pedestrian and car traffic data for Calgary required a Merge Queries operation, based on common date and location fields, to produce an aggregated traffic flow table.

To enhance data accuracy and consistency, we corrected discrepancies in city naming conventions, such as adjusting “Montréal” to "Montreal," using the Replace Values tool. Additionally, we introduced a "Data Refreshed" table using the DateTime.LocalNow() formula, enabling users to easily identify the date and time of the last dashboard update, thus ensuring data synchronization.

Moreover, to improve clarity for end-users and facilitate a smoother dashboard development process, we strategically renamed several columns. This step not only aids users in navigating and comprehending the dashboard but also streamlines our analytical workflow.

Relationship Mapping:

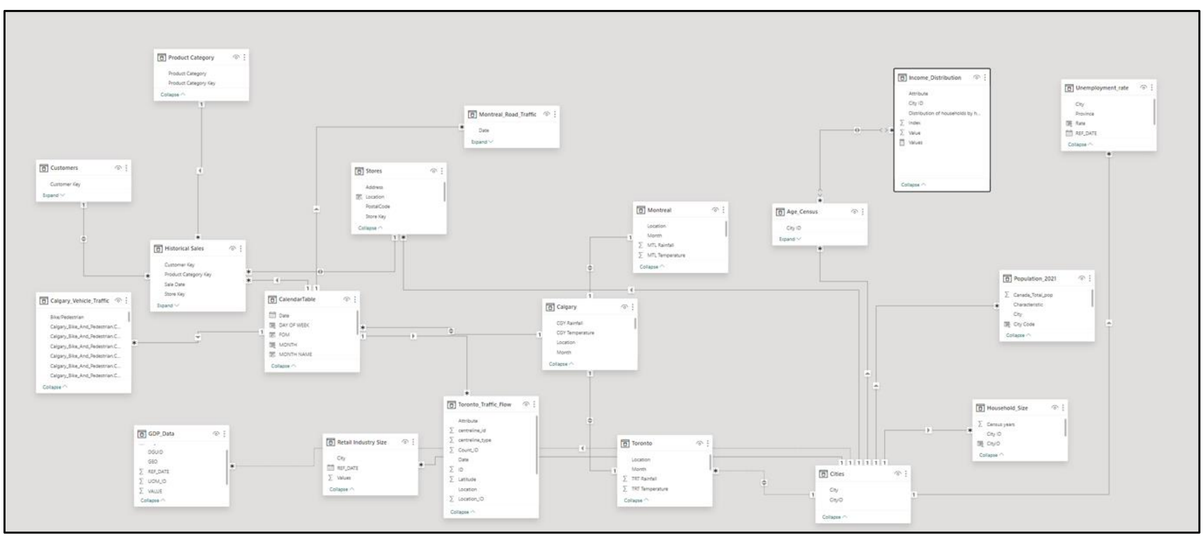


Figure 1: Relationship mapping for project

Our analytical model is constructed around the "Historical Sales" table, which serves as the central fact table, intricately linked to several dimension tables such as "Customers," "Stores," "Product Category," and "CalendarTable." This configuration enables a comprehensive, multidimensional analysis, allowing for the dissection of sales data across various vectors including customer demographics, store attributes, product types and temporal trends.

The "Customers" table is adeptly linked to the "Historical Sales" through the "Customer Key," facilitating an in-depth examination of how sales trends correlate with customer demographics. Similarly, the "Stores" and "Product Category" tables are connected via their respective keys, "Store Key" and "Product Category Key," enabling an analysis of sales performance by store locations and product categories.

To enrich our model with external factors that may influence sales, we integrate additional tables such as "Calgary_Vehicle_Traffic," "Montreal_Road_Traffic," "Toronto_Traffic_Flow," and "GDP_Data." These datasets, while indirectly connected to the sales data, provide a nuanced view of the external environment, including economic conditions and traffic patterns, which could have a significant impact on store performance.

The "CalendarTable" is pivotal for temporal analysis, acting as the anchor for all time-related data across various datasets. This ensures a harmonized approach to date-based reporting and analysis.

Moreover, demographic and socio-economic tables like "Age_Census," "Income_Distribution," "Unemployment_rate," "Population_2021," and "Household_Size" are tied to the sales data through geographic identifiers such as "City ID" or "City Name." This structure allows us to weave demographic factors such as age, income, unemployment rates, population and household sizes into our sales analysis, offering a richer context for understanding sales performance within demographic and economic landscapes.

In essence, our data model is a tapestry of interconnected datasets that together provide a 360-degree view of the sales ecosystem. The model is designed with the objective of identifying the optimal new store location by leveraging insights from various data points that collectively drive strategic business decisions.

Calculation Steps:

Some calculations were done to create some new columns and measures:

Columns:

We created the Calendar Table to be used for visualizations and to connect with other tables. Calculations included:

- For CalendarData:

```
CalendarTable =
VAR MinDate = MIN('Historical Sales'[Sale Date])
VAR MaxDate = DATE(2024, 12, 31)
RETURN
```

```
CALENDAR(MinDate, MaxDate)

DAY OF WEEK = WEEKDAY('CalendarTable'[Date])
FOM = STARTOFMONTH('CalendarTable'[Date])
MONTH = MONTH(CalendarTable[Date])
MONTH NAME = FORMAT('CalendarTable'[Date], "MMMM")
MONTH SHORT = UPPER(left('CalendarTable'[MONTH NAME],3))
QUARTER = QUARTER('CalendarTable'[Date])
START OF YEAR = STARTOFYEAR('CalendarTable'[Date])
WEEKDAY/WEEKEND = IF(WEEKDAY('CalendarTable'[Date]) = 1 || WEEKDAY('CalendarTable'[Date]) =
7 , "Weekend","Weekday")
YEAR = YEAR('CalendarTable'[Date])
```

- For Calgary traffic flow:

```
Latitude =
VAR WKTString = 'Calgary_Vehicle_Traffic'[WKT]
VAR StartParenthesis = SEARCH("(", WKTString, 1, LEN(WKTString)) + 1
VAR Space = SEARCH(" ", WKTString, StartParenthesis, LEN(WKTString))
VAR EndParenthesis = SEARCH(")", WKTString, Space, LEN(WKTString))
VAR LatitudeString = MID(WKTString, Space + 1, EndParenthesis - Space - 1)
RETURN
VALUE(LatitudeString)

Longitude =
VAR WKTString = 'Calgary_Vehicle_Traffic'[WKT]
VAR StartParenthesis = SEARCH("(", WKTString, 1, LEN(WKTString)) + 1
VAR Space = SEARCH(" ", WKTString, StartParenthesis, LEN(WKTString))
VAR LongitudeString = MID(WKTString, StartParenthesis, Space - StartParenthesis)
RETURN
VALUE(LongitudeString)

Total Count = Calgary_Vehicle_Traffic[Vehicles_count] +
Calgary_Vehicle_Traffic[Calgary_Bike_And_Pedestrian.Total]
```

We also created a City table to connect to different economic indicators and demographics

- For Historical sales:

```
Month = MONTH('Historical Sales'[Sale Date])
```

- For Montreal traffic flow:

```
TotalCountByDirection =
    Montreal_Road_Traffic[Count_East_Approach] +
```

```
Montreal_Road_Traffic[Count_Eastbound_Left_Turn] +
Montreal_Road_Traffic[Count_Eastbound_Right_Turn] +
Montreal_Road_Traffic[Count_Eastbound_Through] +
Montreal_Road_Traffic[Count_North_Approach] +
Montreal_Road_Traffic[Count_Northbound_Left_Turn] +
Montreal_Road_Traffic[Count_Northbound_Right_Turn] +
Montreal_Road_Traffic[Count_Northbound_Through] +
Montreal_Road_Traffic[Count_South_Approach] +
Montreal_Road_Traffic[Count_Southbound_Left_Turn] +
Montreal_Road_Traffic[Count_Southbound_Right_Turn] +
Montreal_Road_Traffic[Count_Southbound_Through] +
Montreal_Road_Traffic[Count_West_Approach] +
Montreal_Road_Traffic[Count_Westbound_Left_Turn] +
Montreal_Road_Traffic[Count_Westbound_Right_Turn] +
Montreal_Road_Traffic[Count_Westbound_Through]
```

- For Population:

```
YEAR =
IF(
    AND('Population_2021'[Index] >= 3, 'Population_2021'[Index] <= 5),
    2016,
    2021
)
```

- For Unemployment Rate:

```
Rate = Unemployment_rate[Sum of VALUE]*0.01
```

Measures:

- For Calgary Weather ('Calgary'):

```
AVERAGE RAINFALL = AVERAGE(Calgary[CGY Rainfall])
AVERAGE TEMPERATURE = AVERAGE(Calgary[CGY Temperature])
Max Rainfall = MAX('Calgary'[CGY Rainfall])
Max Temperature = MAX(Calgary[CGY Temperature])
Min Rainfall = MIN('Calgary'[CGY Rainfall])
MIN TEMPERATURE = MIN('Calgary'[CGY Temperature])

NOTE: In similar way, we did for Montreal['Montreal'] and Toronto['Toronto'] Weather data
```

- For Calgary Traffic flow:

(For forecasting 2023 traffic we used)

```
2023_CUMULATIVE_COUNT =
CALCULATE([2023_FORECAST],FILTER(ALLSELECTED(CalendarTable),CalendarTable[Date] <=
MAX(CalendarTable[Date])))
2023_FORECAST = CALCULATE([Count_LY],FILTER(CalendarTable,CalendarTable[YEAR] = 2023))
Count_LY = CALCULATE(
    'Calgary_Vehicle_Traffic'[TOTALTRAFFICCOUNT],
    SAMEPERIODLASTYEAR('CalendarTable'[Date])
)
FORECAST VS COUNTS = IF( ISBLANK([TOTALTRAFFICCOUNT]),
BLANK(),
[2023_CUMULATIVE_COUNT]-[CUMULATIVE_COUNT])
CUMULATIVE_COUNT = IF(ISBLANK([TOTALTRAFFICCOUNT]),
BLANK(),
CALCULATE([TOTALTRAFFICCOUNT],FILTER(ALLSELECTED(CalendarTable),CalendarTable[Date] <=
MAX(CalendarTable[Date]))))
```

Then we used this to forecast Sales Conversion Rate for 2024 and calculate for current data
NOTE: Same type of measures were created for Montreal traffic flow and Toronto Traffic Flow

```
Sales Conversion Rate 2023 = DIVIDE(
    CALCULATE(
        'Historical Sales'[REVENUE],
        'CalendarTable'[YEAR] = 2023
    ),
    [2023_FORECAST]
)
SCR_LY = CALCULATE(
    'Calgary_Vehicle_Traffic'[Sales Conversion Rate 2023],
    SAMEPERIODLASTYEAR('CalendarTable'[Date])
)
2024_SCR_FORECAST = CALCULATE([SCR_LY],FILTER(CalendarTable,CalendarTable[YEAR] = 2024))
2024_CUMULATIVE_SCR =
CALCULATE([2024_SCR_FORECAST],FILTER(ALLSELECTED(CalendarTable),CalendarTable[Date] <=
MAX(CalendarTable[Date])))
CUMULATIVE_SCR = IF(ISBLANK([Sales Conversion Rate 2023]),
BLANK(),
CALCULATE([Sales Conversion Rate 2023],FILTER(ALLSELECTED(CalendarTable),CalendarTable[Date]
<= MAX(CalendarTable[Date]))))
```

- For Customers:

```
Average Purchase Value = AVERAGE('Historical Sales'[Total Amount])
Purchase Frequency = DIVIDE(COUNT('Historical Sales'[Total Amount]),
DISTINCTCOUNT(Customers[Customer Key]))
Customer Lifetime Value = [Average Purchase Value]*[Purchase Frequency]
```

- For GDP:

```
AVERAGE GDP = AVERAGE(GDP_Data[VALUE])
```

- For Historical Sales:

(We did forecasting once again for sales)

```
2024_CUMULATIVE_SALES =
CALCULATE([2024_FORECAST],FILTER(ALLSELECTED(CalendarTable),CalendarTable[Date] <=
MAX(CalendarTable[Date])))

2024_FORECAST = CALCULATE([SALES_LY],FILTER(CalendarTable,CalendarTable[YEAR] = 2024))

2024NormalizedREVENUEFORECAST =
DIVIDE(
    [2024_FORECAST] - [GlobalMin2024FORECASTREVENUE],
    [GlobalMax2024fORECASTREVENUE] - [GlobalMin2024FORECASTREVENUE],
    0
)

CUMULATIVE_SALES = IF(ISBLANK([REVENUE]),
BLANK(),
CALCULATE([REVENUE],FILTER(ALLSELECTED(CalendarTable),CalendarTable[Date] <=
MAX(CalendarTable[Date]))))

FORECAST VS SALES = IF( ISBLANK([REVENUE]),
BLANK(),
[2024_CUMULATIVE_SALES]-[CUMULATIVE_SALES])

MoM Growth =
VAR CurrentMonthSales = [REVENUE]
VAR PreviousMonthSales = CALCULATE([REVENUE], DATEADD(CalendarTable[Date], -1, MONTH))
RETURN DIVIDE(CurrentMonthSales - PreviousMonthSales, PreviousMonthSales)

MoM_ Triangle Indicator =
IF (
    [MoM Growth] > 0,
    UNICHAR(9650),  // Unicode for upward-pointing triangle
    IF (
        [MoM Growth] < 0,
        UNICHAR(9660),  // Unicode for downward-pointing triangle
        ""  // You can choose what to display when [Your Measure] is 0
    )
)

REVENUE = SUM('Historical Sales'[Total Amount])
```

```
SALES_LY = CALCULATE(
    'Historical Sales'[REVENUE],
    SAMEPERIODLASTYEAR('CalendarTable'[Date])
)
```

```
Triangle Indicator =
IF (
    [YoY Growth] > 0,
    UNICHAR(9650),  // Unicode for upward-pointing triangle
    IF (
        [YoY Growth] < 0,
        UNICHAR(9660),  // Unicode for downward-pointing triangle
        ""  // You can choose what to display when [Your Measure] is 0
    )
)
```

```
YoY Growth =
VAR CurrentYearSales = [REVENUE]
VAR PreviousYearSales = CALCULATE([REVENUE], SAMEPERIODLASTYEAR(CalendarTable[Date]))
RETURN (CurrentYearSales - PreviousYearSales) / PreviousYearSales
```

- For Population:

```
Population_2021_Sum =
SUMX(
    FILTER(
        Population_2021,
        CONTAINSSTRING('Population_2021'[Characteristic], "Population, 2021")
    ),
    'Population_2021'[City_population]
)
```

- For Retail Industry Size:

```
AVERAGE RETAIL INDUSTRY SIZE = AVERAGE('Retail Industry Size'[Values])
```

- For Stores:

```
SATURATIONSCORE = 1 / COUNT(Stores[Store Key])
ExpansionScore = ((
    ([NormalizedREVENUE] * 0.65 +
    [2024NormalizedREVENUEFORECAST] * 0.15 +
    [SaturationScore]
    )) + (1/Customers[Total Customers])*10 + Population_2021[Population_2021_Sum]*0.0000001)
```

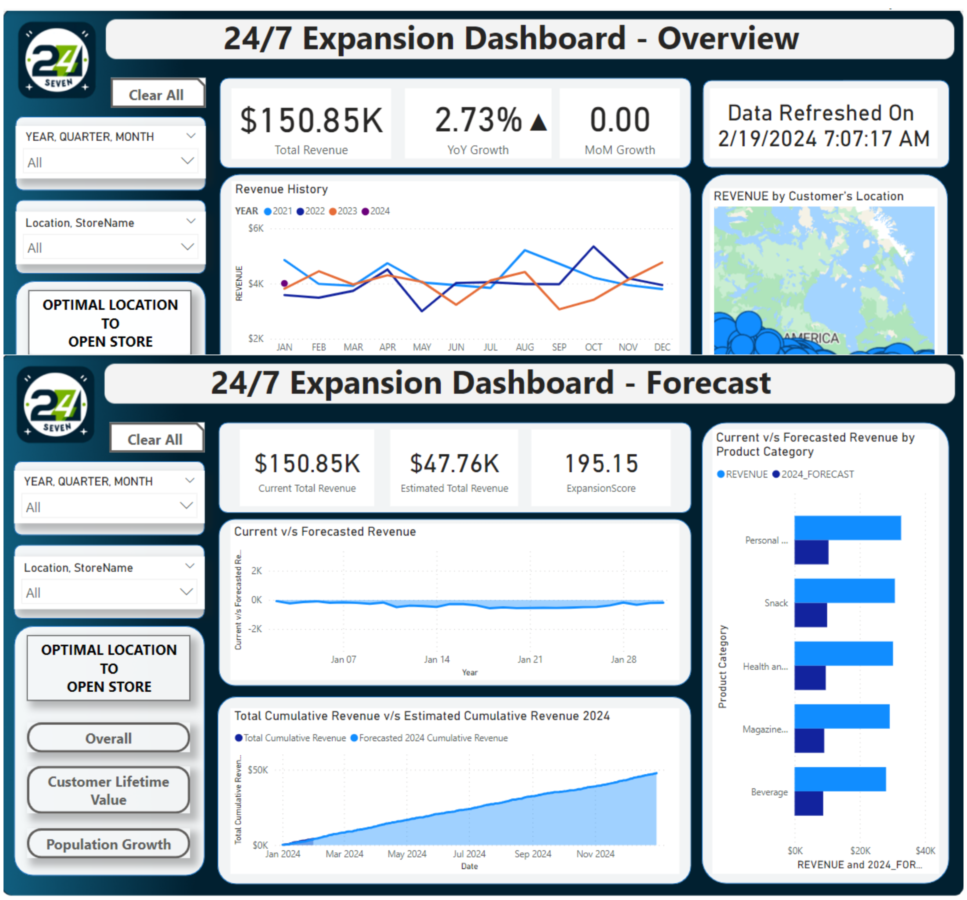
- For Unemployment Rate:

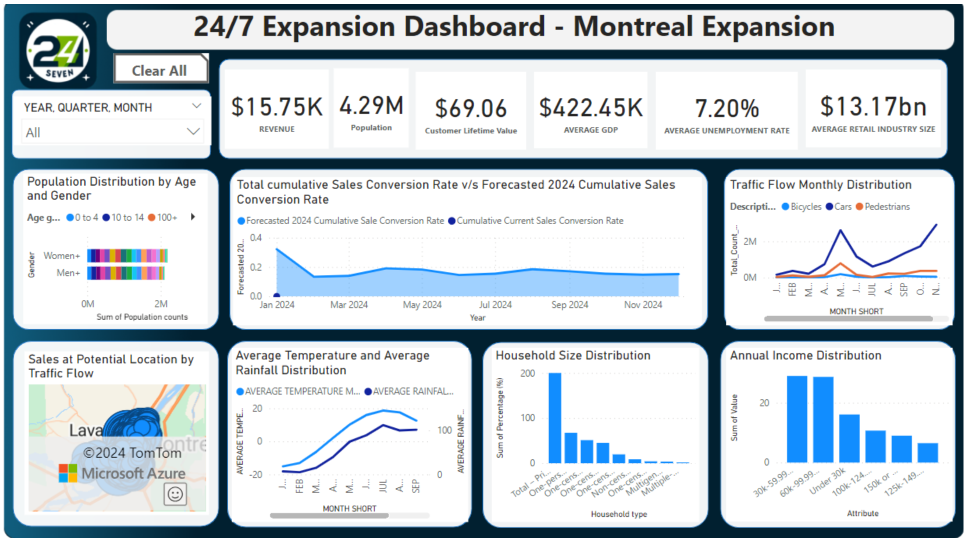
```
AVERAGE UNEMPLOYMENT RATE = AVERAGE(Unemployment_rate[Rate])
```

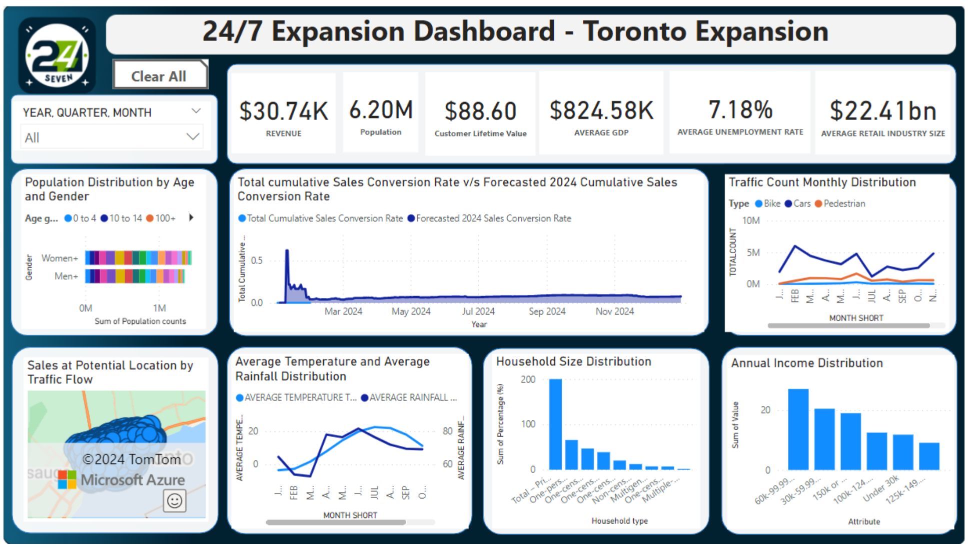
Final Visualizations & Results

Our dashboard provides a holistic overview of 24Seven's business performance and expansion strategy, focusing on key metrics such as revenue, customer lifetime value and average GDP in different locations. It also highlights the impact of regional climate on sales, showing temperature and precipitation information in Toronto, Calgary and Montreal, which are critical to understanding customer behavior and product demand. The dashboard also includes a detailed breakdown of sales by product category, allowing trends and preferences to be identified in different regions. This information is crucial for adapting marketing strategies and product offerings to local demand. In addition, our dashboard provides a forecast for future sales and expansion opportunities, with metrics such as saturation value and expansion value providing insight into market potential and competitive position. The inclusion of cumulative sales and forecast data enables a comparison of actual performance against forecasts, aiding strategic decision making and resource allocation.

As per our analysis, the new store should be opened in Toronto, since it has the highest value of Expansion Score.







Value for Stakeholders

The Power Bi report provides stakeholders with an invaluable visual and analytical tool, empowering them with real-time insights to drive strategic decisions for expansion and growth. The dashboard is a dynamic synthesis of crucial business metrics, from sales data to demographic trends, which brings the following values to stakeholders:

Informed Decision-Making: With the "Optimal Location to Open Store" feature, stakeholders are equipped with data-driven recommendations for store expansion, considering overall market potential, customer lifetime value, and population growth. This holistic approach ensures that new store locations are positioned to capture maximum market share and align with demographic trends.

Comprehensive Revenue Analysis: The dashboard presents a detailed breakdown of total revenue, year-over-year growth, and month-over-month growth, supplemented by a revenue history that tracks performance over time. Stakeholders can evaluate financial health at a glance and make timely decisions to capitalize on growth trends or address emerging challenges.

Geo-Analytic Customer Insights: Through the integration of a map showing revenue by customer's location, stakeholders can visualize market penetration and identify regions with high revenue potential or areas that may be underserved. This geo-analytic approach enables targeted marketing and resource allocation.

Segmentation and Personalization: By segmenting revenue by customer gender and product category, the dashboard allows stakeholders to tailor product mixes and marketing strategies to the preferences of specific customer segments, enhancing personalization and customer satisfaction.

Appendix-1

Economic Indicators Exploration:

In our research on the impact of macroeconomic conditions on the retail industry, we incorporate a thorough analysis of key economic indicators and Canadian domestic economic conditions, focusing on datasets related to GDP, unemployment rates and the economic value generated by the retail industry in Montreal, Calgary and Toronto. GDP figures are crucial for understanding economic activity and its correlation with consumer spending and retail performance. The unemployment rate indicates labor market health and consumer purchasing power, directly affecting retail sales. Lastly, analyzing the economic value created by the retail sector provides insights into its performance, revenue trends and its role in the local economy. Together, these indicators offer a comprehensive view of how macroeconomic variables influence the retail landscape, enabling us to identify trends, challenges and opportunities for growth and resilience within the industry amidst economic changes.

Collecting data from the Government Portal, we started the data cleaning and organization process using power queries. For the dataset about GDP, we kept the columns reference date, city names, provinces and GDP values. To obtain the values final GDP values, we used the power query `GDP = FORMAT('GDP_City Level'[VALUE] * 1000000, "Currency")` to add a new column.

Also, exploring datasets for GDP values at the province level, specifically for Alberta, Quebec and Ontario, offers critical insights into the economic performance and health of these regions. This analysis not only highlights the diverse economic landscapes and sectorial contributions within each province but also aids in understanding regional economic strengths and vulnerabilities.

For the dataset about unemployment rate, we also kept the columns reference date, city names, provinces and unemployment rate values. And we calculated the average employment rate by using measure `AVERAGE UNEMPLOYMENT RATE = AVERAGE('unemployment rate'[VALUE])` and the calculating the average employment rate by city for comparation by using: Average Unemployment Rate by City = `CALCULATE([Average Unemployment Rate],ALLEXCEPT('Cities','Cities'[City]))`

For the dataset about retail sales amounts, we also kept the columns reference data, city names, provinces and sales amount. To obtain the values final retailing values, we used the power query retailing = FORMAT('GDP_City Level'[VALUE] * 1000000, "Currency") to add a new column.

Climate Conditions Exploration

Our research reveals the great impact of climate conditions on purchasing behavior, affecting both the amount and types of products bought. Therefore, it is important to examine the diverse climate conditions across regions before determining the most suitable locations for expansion. Among various climate indicators, temperature and rainfall stand out as the most influential. Based on a thorough analysis of climate data from various cities, we have selected Toronto, Montreal and Calgary for a more detailed examination.

The relationship between temperature and sales volume demonstrates significant variation across different cities. In Calgary, for example, sales reach the peak in September, whereas this period marks a low point in sales for both Toronto and Montreal. Also, the sales trends for five distinct categories—beverages, health and wellness products, magazines and newspapers, personal care items and snacks—show different correlations with weather patterns, underscoring the complexity of the interaction between climate and consumer purchasing behavior.

Census Data Exploration

In our strategic exploration of market potential and customer segmentation, we have integrated an in-depth analysis of census demographic data, focusing on the intricacies of age, gender, income level after tax and household type/size across key Canadian locales, including Montreal, Calgary and Toronto. This data is instrumental in delineating the socio-economic landscape, offering insights into the diverse fabric of potential customer bases. Age and gender demographics allow for a nuanced understanding of market segments, tailoring product offerings to meet specific needs. Income levels post-tax provide a lens through which purchasing power and consumer spending habits can be discerned, crucial for pricing strategies and product positioning. Household types and sizes offer a glimpse into living arrangements and lifestyle choices, impacting buying behaviors and preferences. This comprehensive demographic overview aids in the customization of marketing strategies, store location planning and inventory management, ensuring alignment with the demographic profile of each region. By leveraging data from Statistics Canada, we ensure our analysis is rooted in reliable and up-to-date information, enabling informed decisions that drive growth and customer engagement in the competitive retail landscape.