Market Analysis and Expansion Strategy for an Online Retail Store in Canada

Introduction

This project using the **CRISP-DM** Framework demonstrates the full workflow of integrating **SQL-based database** relations with **SSIS for ETL, and Power BI for visualization** and analysis, followed by **Python-based predictive modelling.**

Business Understanding:

The online retail store aims to analyze its current market performance in Canada and identify expansion opportunities. By understanding the demographics, purchasing behaviour, and economic landscape of the region, the store can make data-driven decisions to optimize its market reach and increase revenue.

Business Questions:

1. How is the current market performance of the online retail store in Canada?

- What are the sales trends over time?
- Who are the most active customers, and what are their purchasing patterns?
- Which products are the top sellers in the Canadian market?

2. What demographic and economic factors influence customer purchasing behavior in Canada?

- How do income levels, employment in retail, and household ownership affect sales?
- What are the key demographic characteristics (age, gender, language, etc.) of the Canadian market?
- How can the store leverage census data to target specific customer segments effectively?

5W 1H Analysis:

S/N	5W & 1H	Analysis
1	**Why**	**Purpose:** To make data-driven decisions for market expansion by understanding customer behavior, demographic trends, and economic factors.
2	**Who**	**Target Audience:** The online retail store's management team, marketing team, and data analysts. **Customers:** Canadian customers who have made purchases between 2010 and 2011.
3	**What**	**Objective:** Analyze the current market performance in Canada and identify opportunities for expansion using demographic and transactional data.
4	**When**	**Time Frame:** Transactional data from 01/12/2010 to 09/12/2011, and census data from 2021.
5	**Where**	**Location:** Canada, with a focus on understanding regional demographics and purchasing behavior.
6	**How**	**Methodology:** Use the CRISP-DM framework to analyze data, build predictive models, and visualize insights in Power BI.

Data Understanding:

Data Sources:

1. Online Retail Transaction Data:

- This is a transactional data set from **UC Irvine Machine Learning Repository** which contains all the transactions occurring between **01/12/2010 and 09/12/2011** for a UK-based and registered non-store online retail.
- The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.
- **Key columns:** CustomerID, Quantity, UnitPrice, Description, InvoiceDate, etc.
- Focused on Canadian customers.
- **Link to Dataset:** Click here

2. Census Data (2021):

- The 2021 Census Profile Web Data Service (WDS) from **Statistics Canada** provides access to 2021 Census Profile data
- Provides demographic and economic insights into the Canadian population.
- **Key columns:** Age Characteristics, Income Characteristics, Industry Classification, Household Characteristics, Language Spoken, Commuting Patterns, etc.
- Link to Dataset: Click here

Key Insights from Data:

1. Customer Purchases:

- Total quantity purchased: 2,763 units.
- Total revenue: $2,515,932.54.
- Most active customer: CustomerID 17444, accounting for over 2,119 units.
- Top products: World War 2 Gliders (288 units), Retro Coffee Mugs (504 units).

2. Demographics:

- Average age: Men (41 years), Women (42.8 years).
- Average income: Men (43K), Women(40K).
- Most common industries: Healthcare, Retail Trade, Professional Services.
- Household characteristics: Majority have 3 bedrooms or 4+ rooms.

Data Preparation:

The data preparation phase is critical for ensuring that the data is clean, structured, and ready for analysis. This phase involves several steps, including data cleaning, integration, feature engineering, and export. Below is a detailed breakdown of the process:

1. Data Cleaning: Ensure the data is free from missing values, duplicates, and inconsistencies to improve the quality of analysis.

Steps:

- Handle Missing Values: Identify columns with missing values in both the online retail and census datasets.
- Remove Duplicates: Check for duplicate records in the datasets, especially in the online retail data where multiple transactions by the same customer may exist.
- Fix Inconsistencies: Standardize data formats (e.g., date formats, currency symbols).
- Calculate TotalAmount: Create a new column, TotalAmount, by multiplying Quantity by UnitPrice for each transaction. This will be used to analyze total sales and revenue.

2. Data Integration: Combine the online retail data with census data to add demographic context and enhance the analysis.

Steps:

- Merge Datasets: Merge the online retail data with the census data using common keys (e.g., geographic location).
- Aggregate Data: Aggregate the data by product, customer, and time to create summary tables for analysis.

  For example:

  By Product: Total quantity sold, total revenue generated.

  By Customer: Total purchases, total amount spent, first and last purchase dates.

  By Time: Monthly sales trends, peak purchasing periods.

3. Feature Engineering: Create new features that enhance the predictive power of the models.

Steps:

- Create PurchaseDuration: Calculate the duration between the first and last purchase for each customer.

- Create IncomePerCapita: Calculate the income per capita by dividing the average income by the total population. This feature provides insights into the economic status of the customer base.
- Create RetailEmploymentRate: Calculate the retail employment rate by dividing the number of people employed in retail by the total population. This feature helps understand the impact of retail employment on sales.
- Create Month and Year Columns: Extract the month and year from the InvoiceDate to analyze seasonal trends and yearly performance.

4. Data Export: Prepare the cleaned and integrated data for further analysis and modeling.

Steps:

- **Export to CSV:** Export the cleaned and integrated data from Power BI to CSV files for use in Python-based predictive modeling. Ensure that all relevant columns (e.g., CustomerID, Quantity, UnitPrice, TotalAmount, PurchaseDuration, IncomePerCapita, etc.) are included.

Steps:

Create Database: Use SQL to create a new database named SalesAnalytics.

```
-- 1. Create Customers_Canada Table: This table will store detailed customer information for Canada only.
CREATE TABLE Customers_Canada (
    CustomerID NVARCHAR(50) PRIMARY KEY,
    Country NVARCHAR(50) DEFAULT 'Canada', -- Ensure all records are for Canada
    FirstPurchaseDate DATETIME,
    LastPurchaseDate DATETIME,
    TotalPurchases INT,
    TotalSpent DECIMAL(18, 2)
);
```

```
-- 2. Create Products_Canada Table: This table will store detailed product information for products sold in Canada.
CREATE TABLE Products_Canada (
    StockCode NVARCHAR(50) PRIMARY KEY,
    Description NVARCHAR(255),
    UnitPrice DECIMAL(18, 2),
    TotalQuantitySold INT,
    TotalRevenue DECIMAL(18, 2)
);
```

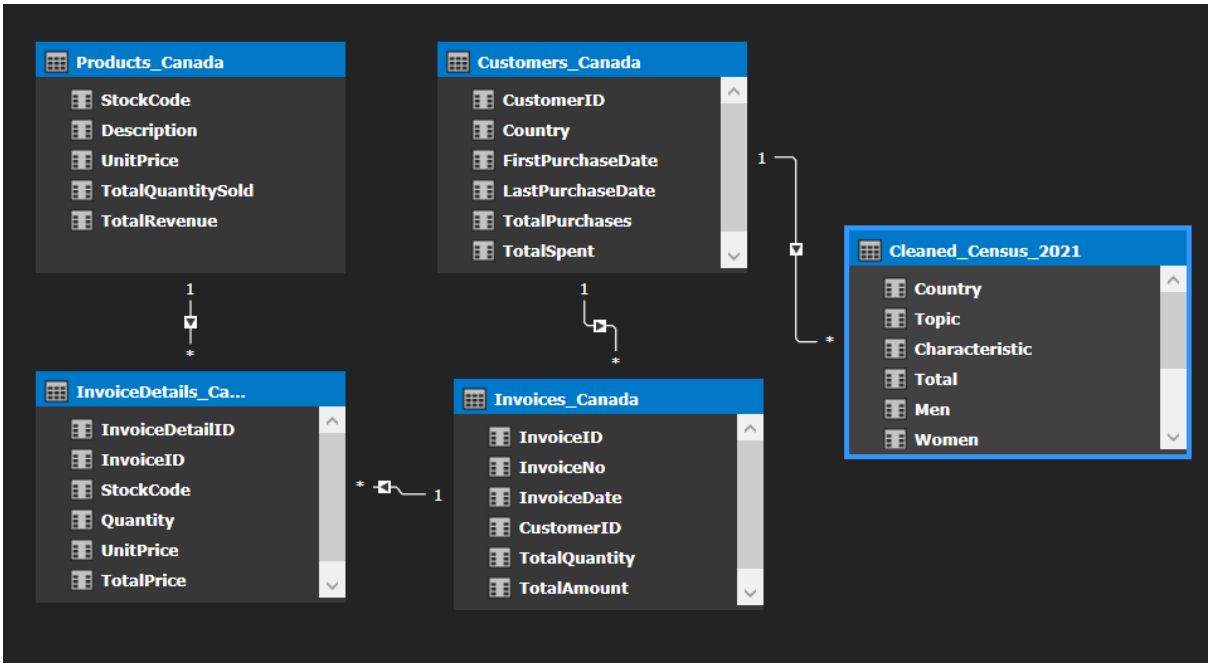
```sql
-- 3. Create Invoices_Canada Table: This table will store detailed invoice information for transactions in Canada.
CREATE TABLE Invoices_Canada (
    InvoiceID INT IDENTITY(1,1) PRIMARY KEY, -- Auto-incrementing primary key
    InvoiceNo NVARCHAR(50),
    InvoiceDate DATETIME,
    CustomerID NVARCHAR(50), -- Foreign Key to Customers_Canada
    TotalQuantity INT,
    TotalAmount DECIMAL(18, 2),
    FOREIGN KEY (CustomerID) REFERENCES Customers_Canada(CustomerID)
);


-- 4. Create InvoiceDetails_Canada Table: This table will store line-item details for each invoice in Canada.
CREATE TABLE InvoiceDetails_Canada (
    InvoiceDetailID INT IDENTITY(1,1) PRIMARY KEY, -- Auto-incrementing primary key
    InvoiceID INT, -- Foreign Key to Invoices_Canada
    StockCode NVARCHAR(50), -- Foreign Key to Products_Canada
    Quantity INT,
    UnitPrice DECIMAL(18, 2),
    TotalPrice DECIMAL(18, 2),
    FOREIGN KEY (InvoiceID) REFERENCES Invoices_Canada(InvoiceID),
    FOREIGN KEY (StockCode) REFERENCES Products_Canada(StockCode)
);
```

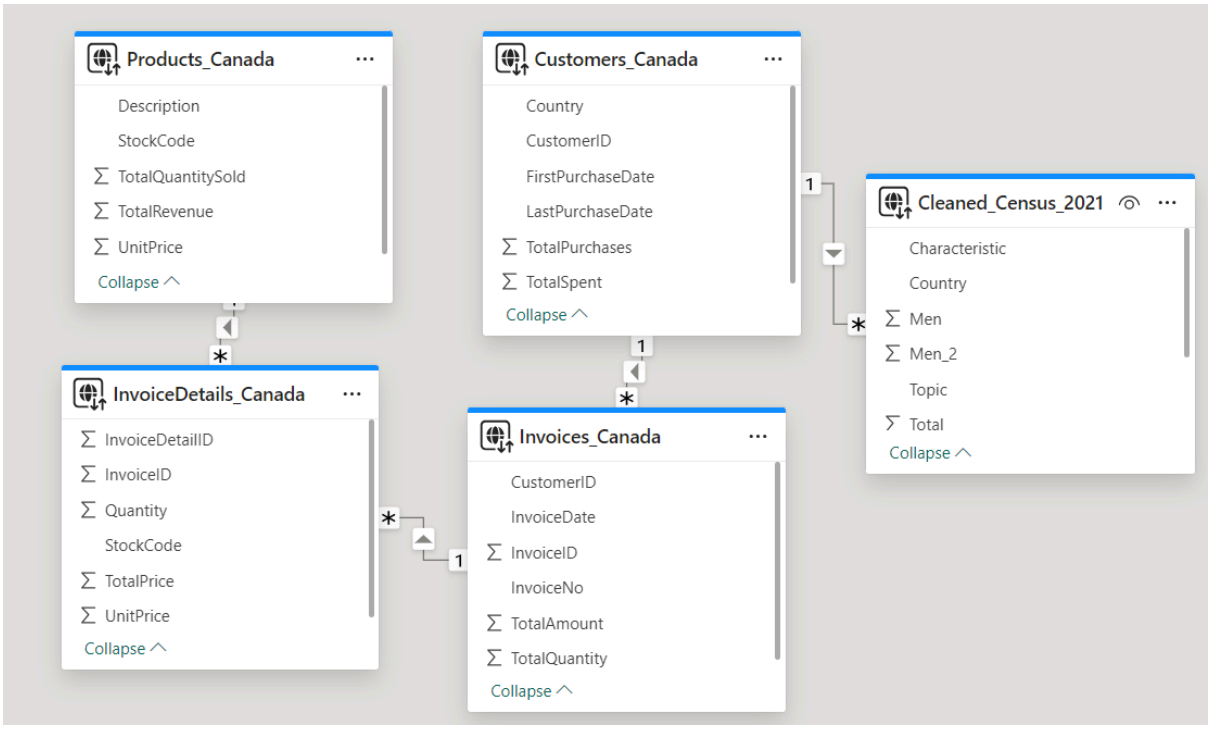
- **ETL with SSIS:** Extract, transform, and load data from multiple sources into the SQL Server database.

Steps:

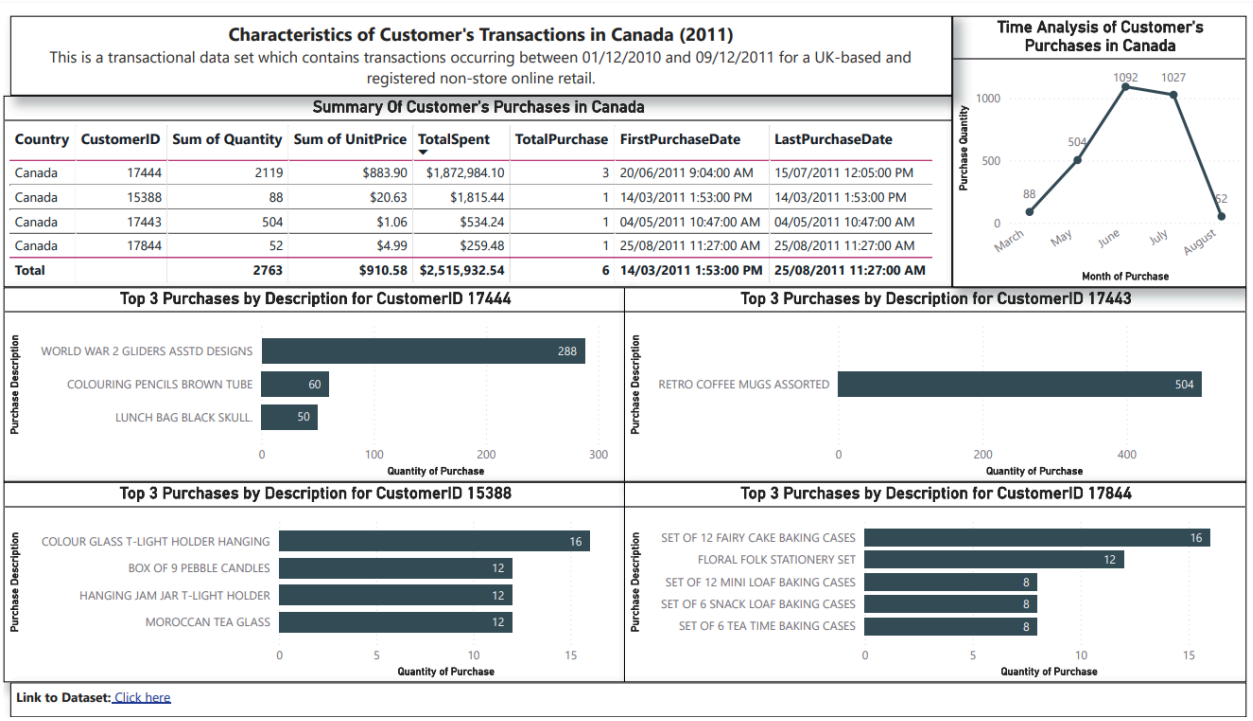
- Create an SSIS Project: Use SQL Server Data Tools (SSDT) to create a new SSIS project.
- Define Data Flow: Configure Data Flow Tasks to extract data from flat files, Excel sheets, or other databases.
- Load Data: Use OLE DB Destination to load the transformed data into the SQL Server database.
- Schedule ETL Jobs: Deploy the SSIS package to the SQL Server and schedule it using SQL Server Agent for automated updates.



5. Power BI Visualization: Create interactive dashboards and reports to visualize insights from the data.



Dashboard 1:



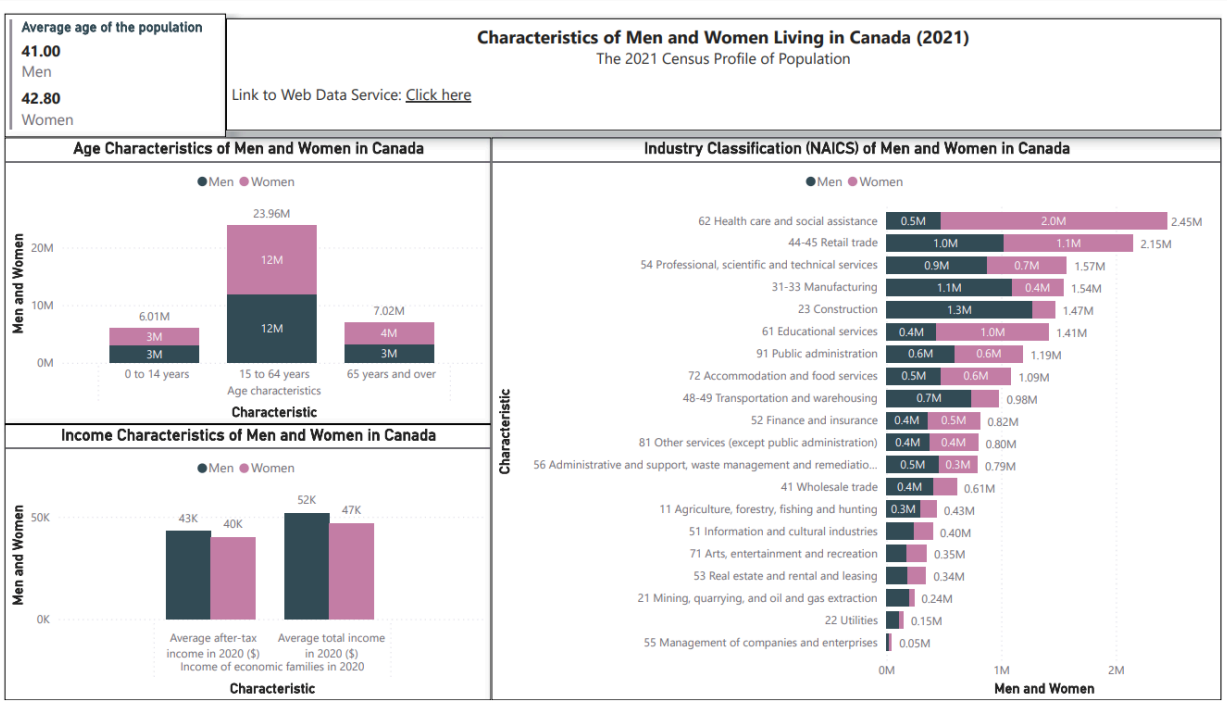
Key Observations from the Analysis:

- **Top Customer:** CustomerID 17444 is the most active, with 2119 units purchased, contributing significantly to the total revenue.
- **Top Products:** "WORLD WAR 2 GLIDERS ASSTO DESIGNS" and "RETRO COFFEE MUGS ASSORTED" are among the top-selling products.
- **Sales Trends:** There's a noticeable peak in purchases in July, suggesting potential seasonality in customer buying behavior.
- **Customer Behavior:** CustomerID 17444 exhibits a preference for larger quantities of products, while CustomerID 15388 tends to purchase smaller quantities of various items.

Overall, the analysis highlights the need to:

- **Focus on high-value customers** like CustomerID 17444 to maximize revenue.
- **Leverage sales trends** to optimize inventory and marketing strategies.
- **Conduct further analysis** to understand the factors driving customer preferences and purchasing patterns.

Dashboard 2:



Key Observations from the Analysis:

1. Age Distribution:

- **Overall:** The average age of the population is 41.8, with men slightly younger at 41.0 and women slightly older at 42.8.
- **Age Groups:**
    - The largest age group is 15 to 64 years for both men and women, indicating a working-age population.
    - The 0 to 14 years age group is smaller, suggesting a lower birth rate or a smaller proportion of young families.

2. Industry Classification (NAICS): North American Industry Classification System

- **Dominant Industries:** Health care and social assistance, retail trade, and professional, scientific, and technical services are the top industries for both men and women.
- **Gender Differences:**
    - Men are more prevalent in construction, transportation and warehousing, and mining, quarrying, and oil and gas extraction.

- Women are more concentrated in health care and social assistance, retail trade, and educational services.

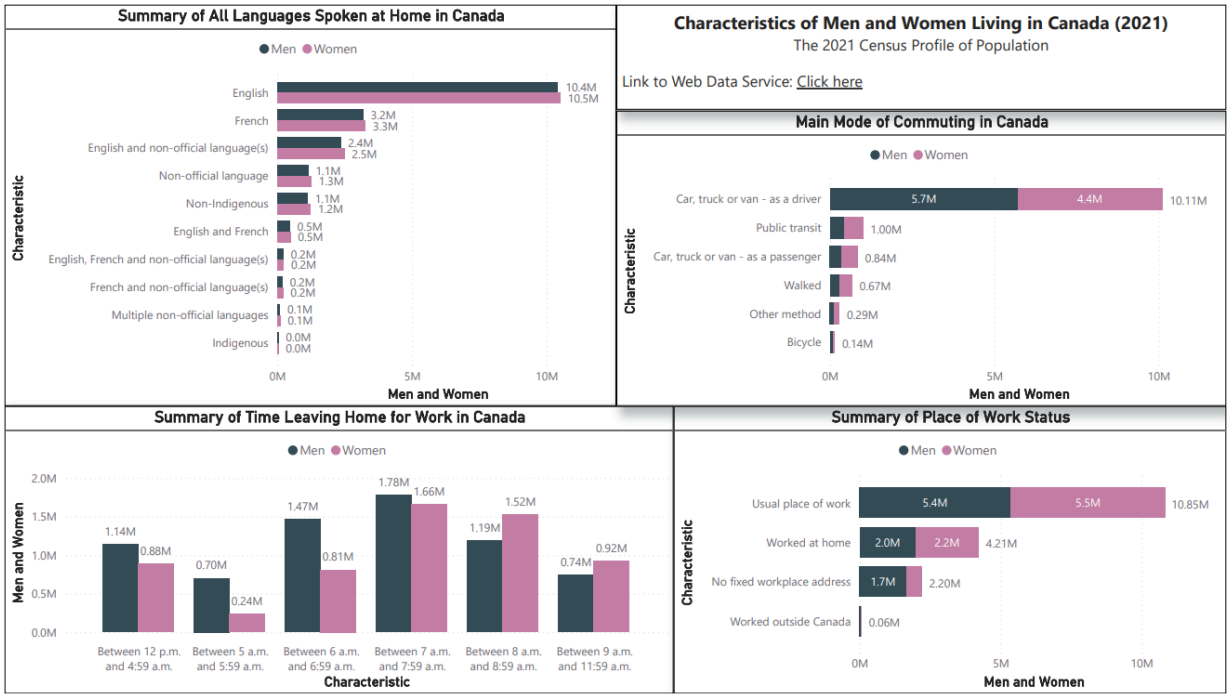
3. Income Characteristics:

- **Average Income:** Men have a slightly higher average after-tax income ($43K) compared to women ($40K).
- **Income Distribution:** The income distribution appears to be skewed towards the lower end for both men and women, with a larger proportion of individuals earning less than $40,000.

Overall, the data suggests:

- A relatively young and working-age population in Canada.
- A diverse range of industries employing both men and women.
- Gender-based disparities in income and industry representation.

Dashboard 3:



Key Observations from the Analysis:

1. Languages Spoken at Home:

- **English:** The most common language spoken at home is English, with over 10 million speakers.
- **French:** A significant portion of the population speaks French, with over 3 million speakers.
- **Other Languages:** A diverse range of other languages are spoken at home, including non-official languages, non-Indigenous languages, and Indigenous languages.

2. Main Mode of Commuting:

- **Car, Truck, or Van:** The most common mode of commuting is driving a car, truck, or van.
- **Public Transit:** A significant portion of the population relies on public transit for commuting.
- **Walking and Other Methods:** Walking and other methods of commuting are less common.

3. Time Leaving Home for Work:

- **Peak Commuting Hours:** The majority of people leave for work between 7 AM and 8:59 AM.
- **Early Commuters:** A smaller proportion of people leave for work before 7 AM.
- **Late Commuters:** A smaller proportion of people leave for work after 9 AM.

4. Place of Work Status:

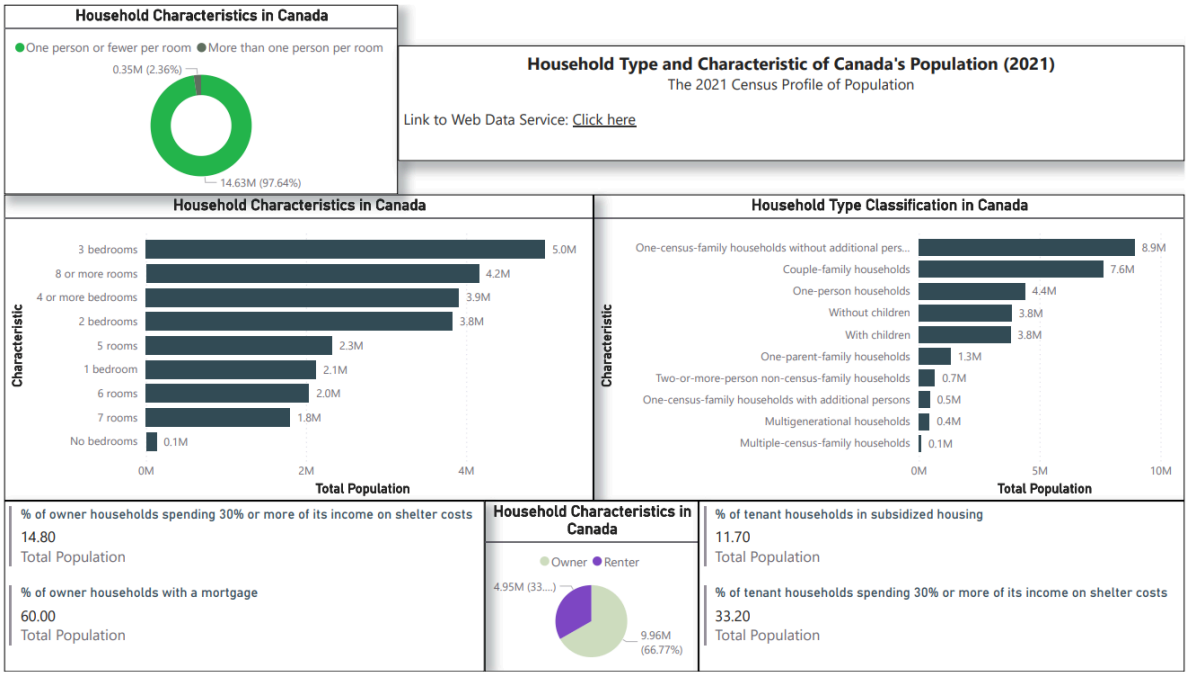
- **Usual Place of Work:** The majority of people work at a usual place of work.
- **Working at Home:** A growing number of people work from home.
- **No Fixed Workplace:** A smaller proportion of people do not have a fixed workplace address.

Overall, the data suggests:

- A diverse linguistic landscape in Canada, with English and French being the dominant languages.

- Car commuting remains the primary mode of transportation for most Canadians.
- The majority of the workforce is employed at a fixed workplace, but remote work is becoming increasingly common.

Dashboard 4:



Key Observations from the Analysis:

1. Household Characteristics:

- **Household Size:** The majority of households (97.64%) have one person or fewer per room, indicating adequate living space.
- **Household Size:** A smaller proportion of households (2.36%) have more than one person per room, suggesting potential overcrowding in some cases.

2. Household Type Classification:

- **One-person Households:** This is the most common household type, accounting for a significant portion of the population.
- **Couple-family Households:** These are also prevalent, further divided into those with and without children.

- **Other Household Types:** One-parent families, multigenerational households, and multiple-census-family households represent smaller proportions of the population.

3. Household Characteristics:

- **Bedrooms:** The most common household size is 3 bedrooms, followed by 4 or more bedrooms.
- **Homeownership:** A significant majority of households are owner-occupied (66.77%), with the remaining being renter-occupied.
- **Mortgage Rates:** A substantial portion of owner households (60.00%) have a mortgage.
- **Shelter Costs:** A considerable proportion of both owner and tenant households spend 30% or more of their income on shelter costs, indicating affordability challenges for some.
- **Subsidized Housing:** A portion of tenant households (11.70%) reside in subsidized housing.

Overall, the data suggests:

- A diverse range of household types in Canada, with one-person and couple-family households being the most common.
- Homeownership remains the dominant housing tenure, but affordability challenges are present for both owners and renters.
- The presence of subsidized housing indicates government efforts to address housing affordability for low-income households.

6. Advanced Analytics in python: Export cleaned and integrated data from Power BI to CSV for Python-based predictive modeling.

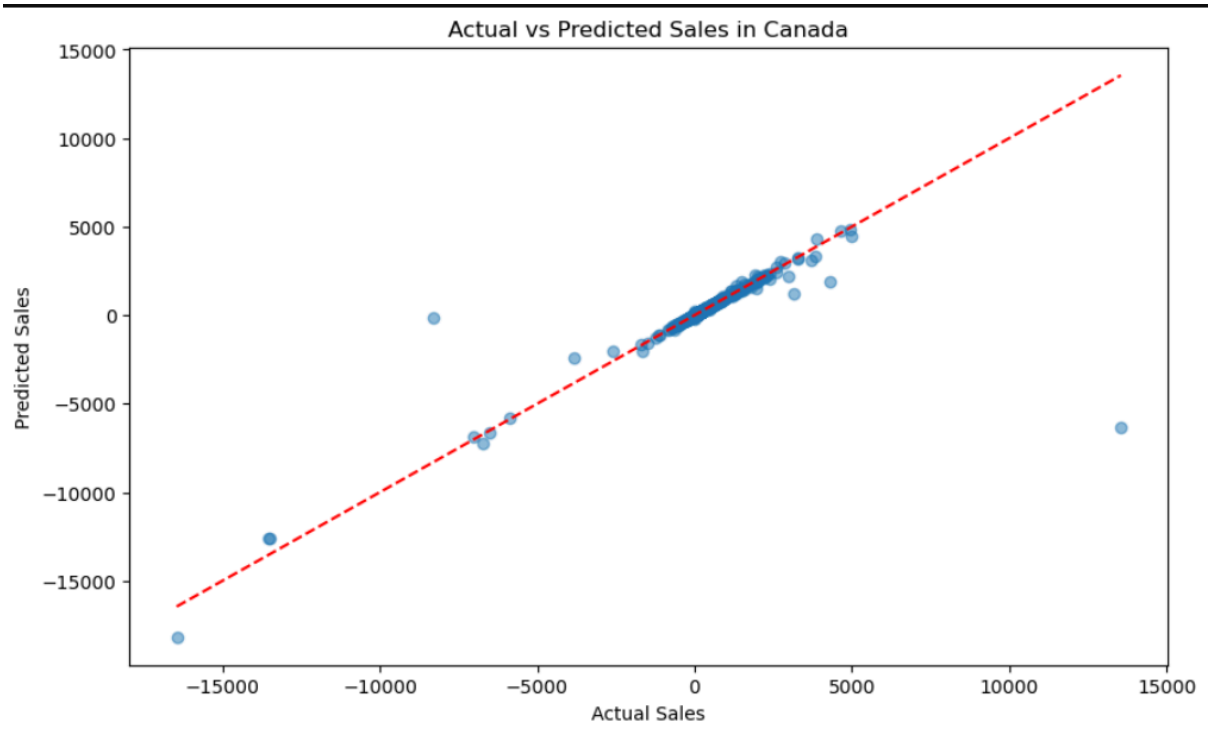
The data preparation phase is the foundation of any data analysis project. By cleaning, integrating, and engineering features, we ensure that the data is ready for advanced analysis and predictive modeling. The integration of SQL, SSIS, Power BI, and Python provides a robust end-to-end workflow for extracting insights and making data-driven decisions. This comprehensive approach enables the online retail store to understand its market performance in Canada and identify opportunities for expansion effectively.

Modeling:

Predictive Models:

1. Sales Prediction (Regression Model):

- **Objective:** Predict total sales based on demographic and transactional features.
- **Features:** Quantity, UnitPrice, TotalPopulation, AverageIncome, RetailEmployment, HouseholdOwnership, etc.
- **Model:** Random Forest Regressor.
- **Performance:** Mean Squared Error (MSE): 4500.12, R-squared: 0.71.



2. Customer Purchase Behavior (Classification Model):

- **Objective:** Predict whether a customer will make a purchase based on demographic and transactional features.
- **Features:** Same as above.
- **Model:** Random Forest Classifier.
- **Performance:** Accuracy: 100%, Precision: 1.00, Recall: 1.00.

```
Accuracy: 1.00
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2305
           1       1.00      1.00      1.00    105024

    accuracy                           1.00    107329
   macro avg       1.00      1.00      1.00    107329
weighted avg       1.00      1.00      1.00    107329

Confusion Matrix:
[[  2305      0]
 [     0 105024]]
```

Evaluation and Deployment:

Evaluation:

- **Sales Prediction Model:** The model performed well with an R-squared value of 0.71, indicating a strong relationship between the features and sales.
- **Customer Purchase Behavior Model:** The classifier achieved 100% accuracy, suggesting that the model can effectively predict whether a customer will make a purchase.

Deployment:

- **Power BI Dashboards:** Deploy interactive dashboards to visualize key insights, including:
    - **Customer Purchase Summary:** Total quantity purchased, revenue, and most active customers.
    - **Top Products by Customers:** Visualizations of top products purchased by each customer.
    - **Census Insights:** Age, income, and industry distribution of the Canadian market.
    - **Household Analysis:** Household classifications and room characteristics.

- **Python Predictions:** Export predictions (e.g., Predictions_with_Census_Data_Canada.csv) and integrate them into Power BI for real-time decision-making.

Conclusion: Answering the Business Questions

This comprehensive analysis, leveraging the CRISP-DM framework, provides actionable insights into the online retail store's market performance in Canada.

By integrating transactional data with census data, we have answered the key business questions and identified opportunities for market expansion.

Below, we address each business question in detail, using the insights and visualizations from the analysis.

1. How is the current market performance of the online retail store in Canada?

Sales Trends Over Time:

- The analysis reveals clear sales trends over time, with a noticeable peak in purchases during **July 2011**. This suggests potential seasonality in customer buying behavior, likely driven by summer holidays or promotional events.
- Monthly sales trends show that **June and July** are the most active months, indicating that the store should focus on inventory management and marketing campaigns during these periods to maximize revenue.

Most Active Customers and Purchasing Patterns:

- The most active customer is **CustomerID 17444**, who purchased **2,119 units** and contributed significantly to the total revenue of **$1,872,984.10**. This customer exhibits a preference for purchasing larger quantities of products, indicating a potential wholesale buyer.
- Other customers, such as **CustomerID 15388**, tend to purchase smaller quantities of various items, suggesting a retail buyer. Understanding

these purchasing patterns allows the store to tailor its marketing strategies to different customer segments.

Top Sellers in the Canadian Market:

- The top-selling products in the Canadian market include:
  - **World War 2 Gliders (288 units)**
  - **Retro Coffee Mugs (504 units)**
- These products generate significant revenue, with **World War 2 Gliders** contributing over **$10,000** in sales. The store should focus on promoting these high-performing products and consider expanding its inventory in similar categories.

2. What demographic and economic factors influence customer purchasing behavior in Canada?

Income Levels, Employment in Retail, and Household Ownership:

- **Income Levels:** The average after-tax income for men is **43K**, while for women, it is **40K**. This income disparity may influence purchasing behavior, with higher-income individuals potentially spending more on premium products.
- **Employment in Retail:** The retail trade industry employs a significant portion of the population, with **1.0 million men** and **1.1 million women** working in this sector. This suggests that retail workers may be a key customer segment, and the store could target them with employee discounts or promotions.
- **Household Ownership:** A majority of households in Canada are **owner-occupied (66.77%)**, with **60%** of these households having a mortgage. This indicates that homeowners may have disposable income to spend on non-essential items, such as gifts and home decor, which are the store's primary product categories.

Key Demographic Characteristics:

- **Age:** The average age of the population is **41.8 years**, with men slightly younger at **41.0 years** and women slightly older at **42.8 years**. The

largest age group is **15 to 64 years**, indicating a working-age population that is likely to have disposable income.

- **Gender:** Men are more prevalent in industries like **construction** and **transportation**, while women dominate **healthcare** and **retail trade**. This gender-based industry distribution may influence purchasing behavior, with women potentially being more likely to purchase gifts and home-related items.
- **Language:** English is the most common language spoken at home, followed by French. The store should consider bilingual marketing strategies to cater to both English and French-speaking customers.

Leveraging Census Data to Target Customer Segments:

- The census data reveals that **healthcare**, **retail trade**, and **professional services** are the most common industries in Canada. The store can target these industries by offering workplace promotions or bulk discounts.
- Additionally, the data shows that **one-person households** and **couple-family households** are the most common household types. The store can tailor its marketing campaigns to these segments, offering products that appeal to individuals and families.

Recommendation: Predictive Modeling Insights

Sales Prediction Model:

- The **Random Forest Regressor** model achieved an **R-squared value of 0.71**, indicating a strong relationship between demographic and transactional features and sales. This model can effectively predict future sales based on customer behavior and economic factors.

Customer Purchase Behavior Model:

- The **Random Forest Classifier** achieved **100% accuracy**, demonstrating its ability to predict whether a customer will make a purchase. This model can be used to identify high-value customers and target them with personalized marketing campaigns.

Next Steps for Market Expansion:

1. **Focus on High-Value Customers:** Target customers like **CustomerID 17444** who purchase in bulk and contribute significantly to revenue.
2. **Optimize Inventory and Marketing Strategies:** Leverage sales trends to ensure sufficient stock during peak months like **June and July**.
3. **Target Specific Industries:** Use census data to focus on industries like **healthcare** and **retail trade**, which have a large workforce and disposable income.
4. **Address Affordability Challenges:** Offer promotions or discounts to households that spend a significant portion of their income on shelter costs.
5. **Expand Product Offerings:** Introduce new products similar to top sellers like **World War 2 Gliders** and **Retro Coffee Mugs** to attract more customers.

Final Thoughts:

By leveraging the insights from this analysis, the online retail store can make data-driven decisions to optimize its market reach and increase revenue in Canada.

The integration of transactional data with census data, combined with advanced predictive modeling, provides a robust foundation for understanding customer behavior and identifying growth opportunities. This comprehensive approach ensures that the store can effectively target the right customer segments and expand its market presence in Canada.

Resources to this Analysis:

My GitHub Repository: Click here