**Household spending by household quintile in Canada**

**Introduction**

One of the key contributors to a nations GDP is the household spending, it is not only the indicator of economic well-being but also sheds light on the consumer behaviour, which shows how Canadian families allocate their resources across the basket of goods and services. Numerous factors influence the spending such as income levels, family size, economic status and demographic features. Delving deeper in the intricacies of household spending can unveil essential insights about the priorities and financial difficulties faced by Canadian population.

Analysis of spending patterns across the income levels and regions can help businesses and decision makers to understand consumer demand and behaviour in a better way which in return leads to informed decisions on resource allocation, economics planning and social programs. It is equivalently important to track changes in household spending over time as it helps in predicting economic trends and evaluate the effect of factors like unemployment, inflation and changes in government policy on Canadian households.

Spending by household differs greatly by region and province. For instance, the provinces that have higher costs of living such as British Columbia and Ontario, housing is the main expense of households. Comparatively, in regions with lesser living costs such as Atlantic Canada, people tend to spend more on transportation and recreation. Understanding these distinctions can provide important information about economic challenges faced by Canadian in diverse parts of the country and how such difficulties impact the spending impact the spending habits of the population.

This dataset ranks households from lowest to highest according to the value of their before tax income. The ranked population in then divided into five groups of equal number of units, called Quintiles. Where each quintile represents 20% of the total number households. According to the census 2016, incomes quintiles are divided as follows:

* The lowest quintile- incomes from $1 to $35,808.
* The second quintile-incomes from $35809 to $59,521.
* The third quintile- incomes from $59,521 to $88,658.
* The fourth quintile- incomes from $ 88,659 to $133,468.
* The highest quintile- incomes above $133,469.

The dataset provides an immersive overview of household spending patterns across different regions and provinces of Canada. It also provides details about the expenditures across important categories like housing, transportation, food, healthcare, and recreation divided by income levels from the lowest quintiles to the highest quintile.The project will further investigate that what are the main drivers for each spending category within the different quintiles. And also how regional economic factors correlate with the spending patterns.

The dataset used in this research was downloaded from Statistics Canada and it includes 242268 instances and 17 variables. And the aim of the project is to analyse the distribution and trend of highest and lowest expenditure category, forecast the total expenditure of the 10 provinces of Canada and to see the trend of total expenditure of all provinces over the years.

**Literature Review**

There were not many research papers about determining the distribution of household spending among household quintiles in Canada. The report on spending patterns in Canada 1998 by Statistics Canada portrays the results of 1998 survey of household spending conducted in January till March 1999. It illustrates the spending habits of the population, trends of expenditure across the provinces and more.

Spending pattern by income level is also presented using a clustered bar graph. For this analysis four primary expenditure category are selected i.e. shelter, food, transportation and personal taxes which is a great way of categorizing expenditure streams into a broader stream and to include the categories that are of utmost importance. With the lowest quintile, 1st quintile, 2nd quintile, 3rd quintile, 4th quintile and highest quintile on the x axis and the percentage of household budget spent on the y axis.

According to this report, population in the lowest quintile spends the highest on food and those in highest quintile spends on personal taxes the most.

Also, Shelter was the second largest expenditure category across all the income quintiles in 1998.

Through this report, an understanding of the segregation of expenditure category into broad categories and different approaches to depict and visualize the distribution of household spending based on income quintile is gained to develop an analysis that is accurate and diverse.

Secondly, the 2021 Survey of Household Spending (SHS) conducted in 2021, gathered a sample of 12,575 households across all the provinces of Canada. It was observed in this survey that the household spending averaged $67,126 on goods and services in 2021 which was 2.7% less than that in 2019 accounting to the pandemic and the consumer inflation of +4.1%. According to the previously reviewed report of 1998, shelter was the largest expenditure category, and it remained the same in the year 2021. The household spending on transportation fell amid the pandemic, this decrease was seen in all the provinces. “The reduction in average expenditure for air travel was consistent with a [large decline in the number of airplane passengers from 2019 to 2021](https://www150.statcan.gc.ca/n1/daily-quotidien/230120/dq230120c-eng.htm).”

The one-person households allocated a large share of their spending to shelter. The households spending in different provinces were compared to each other. Whitehorse and Yellowknife were compared with average spending on goods and services of Yellowknife higher than that of Whitehorse.

**Methodology**

**Objective**

The purpose of the study is to review the distribution of expenditures, define correlation between the variables, perform trend analysis by expenditure category for each province and to forecast the future expenditure on each category. To achieve the objective, Exploratory data analysis, trend analysis and forecasting method were used to analyse the mentioned area.

**Data Preprocessing**

It is the process of transforming the raw, uncleaned data into clean data which is good for analysis. It is the most essential part of data analysis process as it prepares the data by improving its quality, reducing noise and making it easier to work with which is an important criterion for achieving accurate insights from the data analysis and machine learning process. Therefore, it is essential to spend a good amount of time to this step of the data analysis process.

* Removing unnecessary columns-

Firstly, the columns that were not relevant for the analysis were removed from the dataset. There were in total 10 columns excluded, namely- uom, uom\_id, scalar\_factor, scalar\_id, vector, coordinate, status, symbol, terminated and decimals. It helps in reducing the dimensionality of the dataset keeping focus on the important variables resulting in increased efficiency and reduction of noise in the data. After that the cleaned dataset contains 7 columns that are considered important for the analysis.

* Standardizing columns’ names-

To ensure that the code is readable and consistent throughout, the column names were converted to snake case. It is a good practice during analysis and helps in easy integration with the functions and packages.

* Checking for missing values and duplicate rows-

The dataset’s inspection it was found that there were no duplicates in the dataset and

The missing values present in the dataset were searched. The count of total number of distinct expenditure category was found, after that expenditure category that has missing values in any year was identified and any expenditure category that contained missing values were extracted and thereby removed from the dataset. The dataset was inspected for remaining missing values on column-by-column basis. Also, the “value” column consisted of 79,751 values that were missing (NA) most of which were values from the year 2018 and 2021 because it is mentioned that after 2017 the data was collected after every 2 years so during 2018 and 2021, data was not collected which can be a reason for missing instances in the dataset.

**Feature Engineering**

The income quintile in the dataset is filtered to only include the rows where the income quintile is “All quintiles” to ensure that the mean values from only this category is included.

Two new features that are Average\_expenditure and Total\_expenditure are created by aggregating the expenditure data cross various categories and by using these the highest and the lowest expenditures per year were identified.

**Exploratory Data Analysis**

Exploratory data analysis is a technique that helps in summarizing and visualizing the data to derive insights into the patterns, trends and relationship between various variables. R was used to perform EDA on the dataset to have a deeper understanding of its structure and distribution.

The average and the total expenditures by expense category helps to find the distribution of expenditures across categories. A line graph is used to depict the trend of the highest expenditure category i.e. shelter over the years. Also, line plot is used to portray the trend of total expenditure across the provinces which allows for an effective comparison of trends between provinces.

**Machine learning method**

ARIMA stands for Auto-Regressive Integrated Moving Average that is a statistical model used for time series forecasting. It is divided in three components that are as follows:

1. AR (Auto- Regressive) -This part shows the relationship between the observation and a specified number of previous observations.
2. I (Integrated) – It helps in differencing of the raw observations to remove any trend or seasonal variation from the data to make it easy to compare.
3. MA (Moving Average) – This part models the relationship between the observation and a moving average of past residual errors from the model or it shows how random errors in the past data affect the forecast.

Based on the given data, the ARIMA model is used to forecast the future expenditures in the provinces across Canada.

For every forecasted value, the model also calculates 80% confidence interval that indicates the range in which the future predicted values might lie. The forecasted value and fitted values are plotted alongside the past data making it possible to see how the forecasted values line up with the historical data.

**Result**

After loading and cleaning the data, the structure of the dataset is analyzed because it is extremely important to understand the structure and the type of data. The str() function was used for this. It also provides numerous statistical values like mean, max and min for different variables.

A white text with black text

Description automatically generated with medium confidence

**Fig. 1**: Summary of the dataset.

It was reviewed that Shelter is the expenditure category on which the population spends the most, so the trend of annual expenditure on shelter through a line plot is visualized. With year from 2010 to 2021 on the x axis and expenditure amount on the y axis. It can be observed that the annual expenditure on shelter have gradually increased over the years peaked in the year 2021 i.e. around $240000.

A graph showing the growth of a number of people

Description automatically generated

**Fig. 2**: Annual expenditure on shelter

The ARIMA model is used to forecast total expenditure for each province across Canada. The data is first filtered based on “Total expenditure” for each province. Using the model, forecast for 3 years i.e. 2022, 2023 and 2024 is produced. The prediction also depicts the confidence interval taken as 80% which shows the range of values within which the future values will fall.

The graphs present the actual and the forecasted expenditure for the provinces with years on the x-axis and expenditure on the y-axis. The black line in the graph represents the actual data values from the dataset. It captures the historical view of expenditure values and showcases a trend to analyse the patterns if any. The red dashed line symbolizes the fitted data from the ARIMA model which shows the efficiency of the model to adapt to the historical expenditure trend. On the other hand, the forecasted expenditure values are depicted by blue dashed line. The shaded blue band depicts the 80% confidence interval. It provides the range of occurrence of the future values. A wider interval denotes to more uncertainty whereas a narrow interval represents a highly confident forecast.

By analyzing the graphs of different provinces, a varying trend across the provinces can be observed i.e., some provinces exhibit a steady rise in expenditure on the contrary some have unstable and stable trends. The forecast of Ontario can be seen as the most effective as it has a narrow confidence interval whereas the future expenditure of Alberta can vary greatly because the broader blue region represents higher uncertainty.

A graph of forecasting for canada

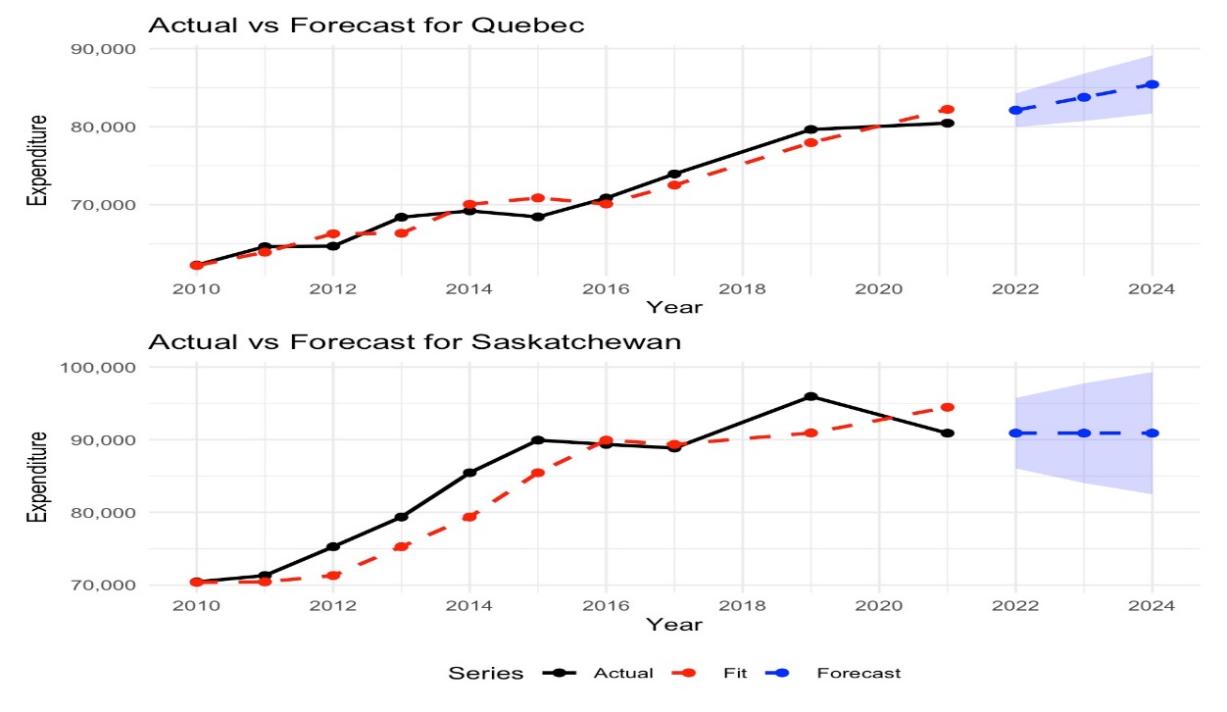
Description automatically generated with medium confidence

**Fig. 3**: Actual and forecasted expenditure

A graph of different types of forecasts

Description automatically generated with medium confidence

**Fig. 4**: Actual and forecasted expenditure



**Fig. 5**: Actual and forecasted expenditure

This visualization shows the trend of total expenditure of each province over the years, with years on the x-axis and total expenditures on the y-axis. Each province has been assigned a different scheme of color for easier comparison. It can be noticed that Alberta’s population has the highest expenditure over the years experienced a sharp fall in the year 2019. On the contrary, the total expenditure was the lowest in Prince Edward Island in the beginning but it has gradually increased after the year 2016.

A graph showing the growth of the country

Description automatically generated

**Fig. 6**: Trend of total expenditure across provinces

**Discussion**

The result of the analysis presents shelter as the dominant expenditure category, this trend is like the economic realities such as soaring property prices and rental costs. Also, the variation in spending pattern and habits is quite prominent across the provinces. On the other hand, spending pattern also differ with income quintiles as it is seen that individuals in lowest income quintile spend a huge proportion of their income on necessities like food and shelter whereas individuals in higher income quintile splurge on recreation and personal taxes. The rural provinces have shown a steady growth in spending over the years whereas urban areas experience more fluctuations due to increased unpredictability of market and policy changes.

Some of the external factors could have an impact on household spending such as the Covid-19 pandemic which resulted in fall in airfare and spending on recreation. Secondly, inflation could be the reason behind the increase in spending on shelter.

On the analysis of past studies and reports, shelter consistently ranked as the highest expenditure category. The result from this analysis affirms the long-term trend highlighting the fact that housing toped the economic problems for Canadian families. The analysis introduced machine learning (ARIMA) to foresee the trends and aimed to present a predictive approach to the topic.

This area of study is central to the Canadian society and deeper research on this could prove to be fruitful. The impact of demographic factors such as age, sex and education to gain insights into the spending pattern of an individual in each province and quintile and the change in the patterns post COVID-19 pandemic could be explored.

To summarize, this analysis provides the current trends and forecasts the future expenditure in each province providing a mechanism for the government to formulate policies to cater to the current and upcoming needs of the population ensuring better resource allocation and overall welfare of the population.

**References**

<https://www150.statcan.gc.ca/n1/daily-quotidien/231018/dq231018a-eng.htm>

<https://publications.gc.ca/Collection-R/Statcan/62-202-XIE/0009862-202-XIE.pdf>