Documentation - DS203 Final Project
CICCC

Toa Umehara
Mario Sandoval

Census Profile, 2021 Census of Population
Link to dataset:
https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/details/download-telecharger.cfm?Lang=E

1. Introduction:

The goal of our project was to explore a large, multi-file dataset using PySpark, perform data preprocessing, execute analytical operations, and generate insights using Spark SQL and DataFrame transformations. For our dataset, we selected the 2021 Census Profile Data from the Government of Canada. This contains demographic, population, income, housing, education, and language information for different geographical regions across Canada.

This dataset is ideal for PySpark because:

- It contains multiple CSV files.
- The full dataset is approximately 1GB in size, meeting project requirements.
- It includes rich, structured data suitable for filtering, aggregations, joins, and window functions.

Our objective was to load the dataset into Spark, examine its structure, perform analytical queries, and document our findings. Besides general demographic exploration, this project focuses on analyzing one-person household patterns in British Columbia. By comparing health standards by regions in the different cities and micro-level census subdivisions, we aim to demonstrate how large geographic averages can mask important local realities.

2. **Dataset Description**

The dataset used in this project was downloaded from Statistics Canada's official database.

Dataset Characteristics

- Source: Statistics Canada – 2021 Census Profile
- Format: Multiple CSV files

- Total Size: Approx. 1GB (varies depending on selected regions)
- Rows: Millions of observations depending on selected geographies
- Columns:
  Typical columns included (these vary by file):
- *Geographic code (GEO_CODE)*
- *Geographic name (GEO_NAME)*
- *Total population*
- *Population by age groups*
- *Household income*
- *Languages spoken*
- *Housing characteristics*
- *Education levels*

(The exact columns depend on which CSVs you load; include screenshots of your schema.)

This dataset allows analysis at multiple geographic levels, which supports both regional and local comparisons.

Why use this dataset

- National-level, government-quality data
- Supports analytical questions about Canada's demographics
- Large enough to demonstrate Spark's distributed processing
- Multi-file structure works well with Spark's wildcard CSV reader

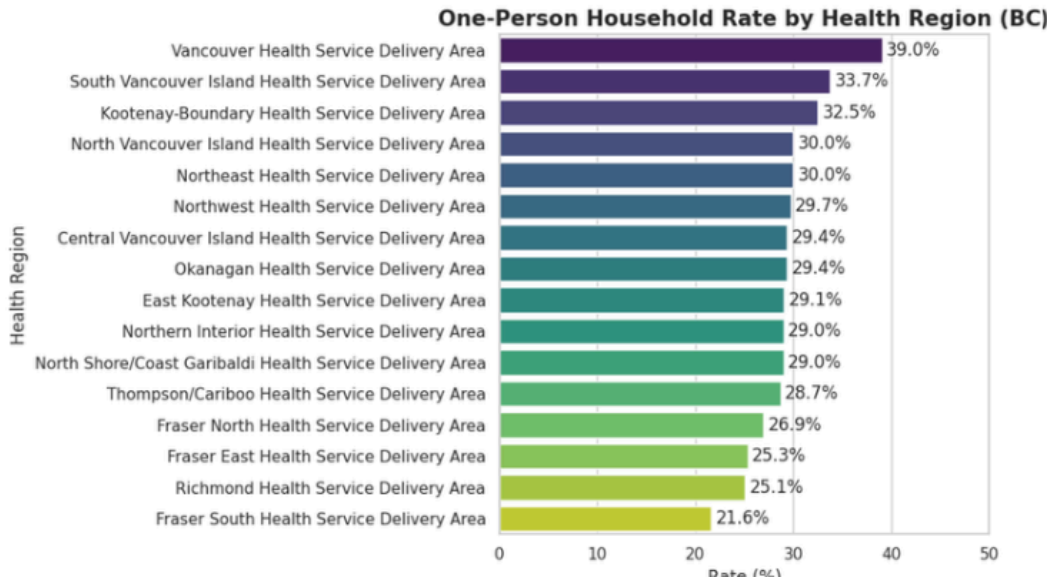
3. Data Import and Initial Analysis

The dataset was imported into a PySpark environment using Spark's distributed CSV reader. Multiple CSV files were loaded simultaneously into a single DataFrame using a wildcard path.

Spark automatically inferred the schema for each column, allowing numeric and categorical fields to be identified. The schema was verified using Spark's schema inspection tools to ensure consistency across files.

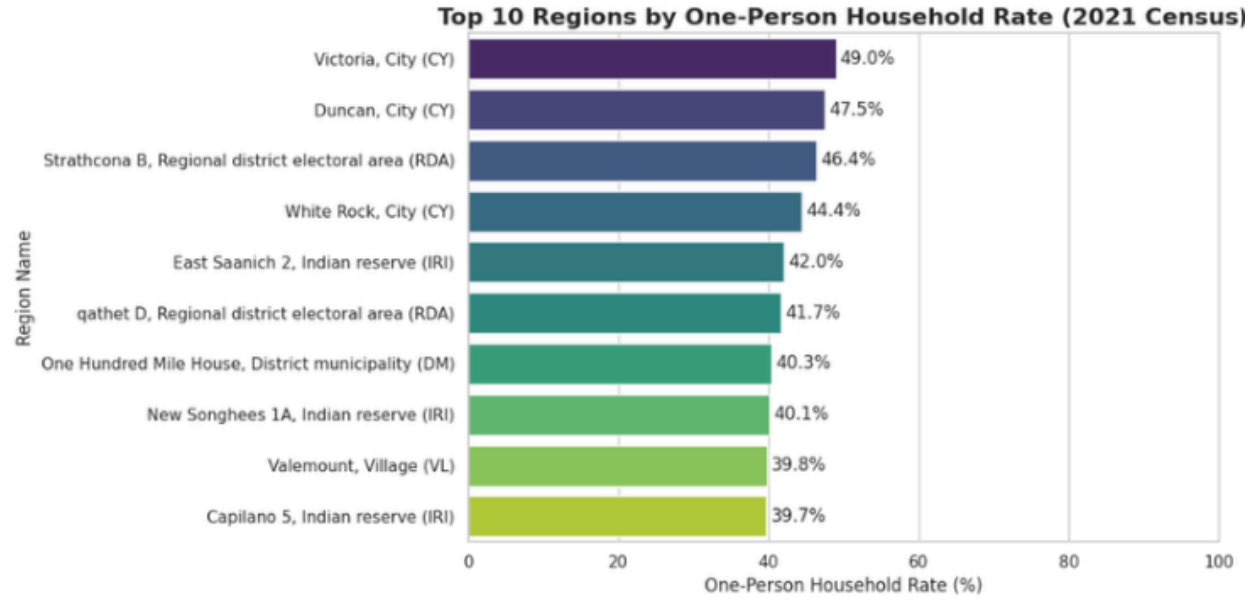
We also calculated the total number of rows and columns to confirm successful data ingestion. Descriptive statistics were generated to understand the distribution of key variables such as population counts and income-related fields.

graph1

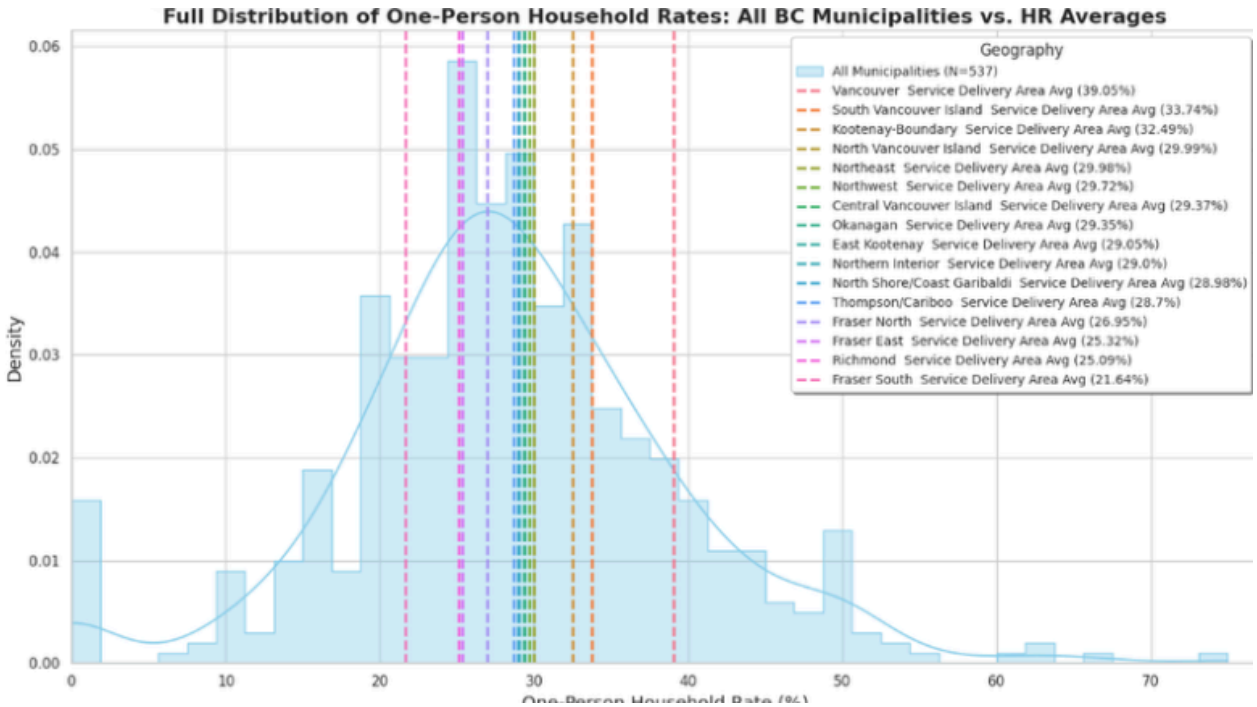


graph2

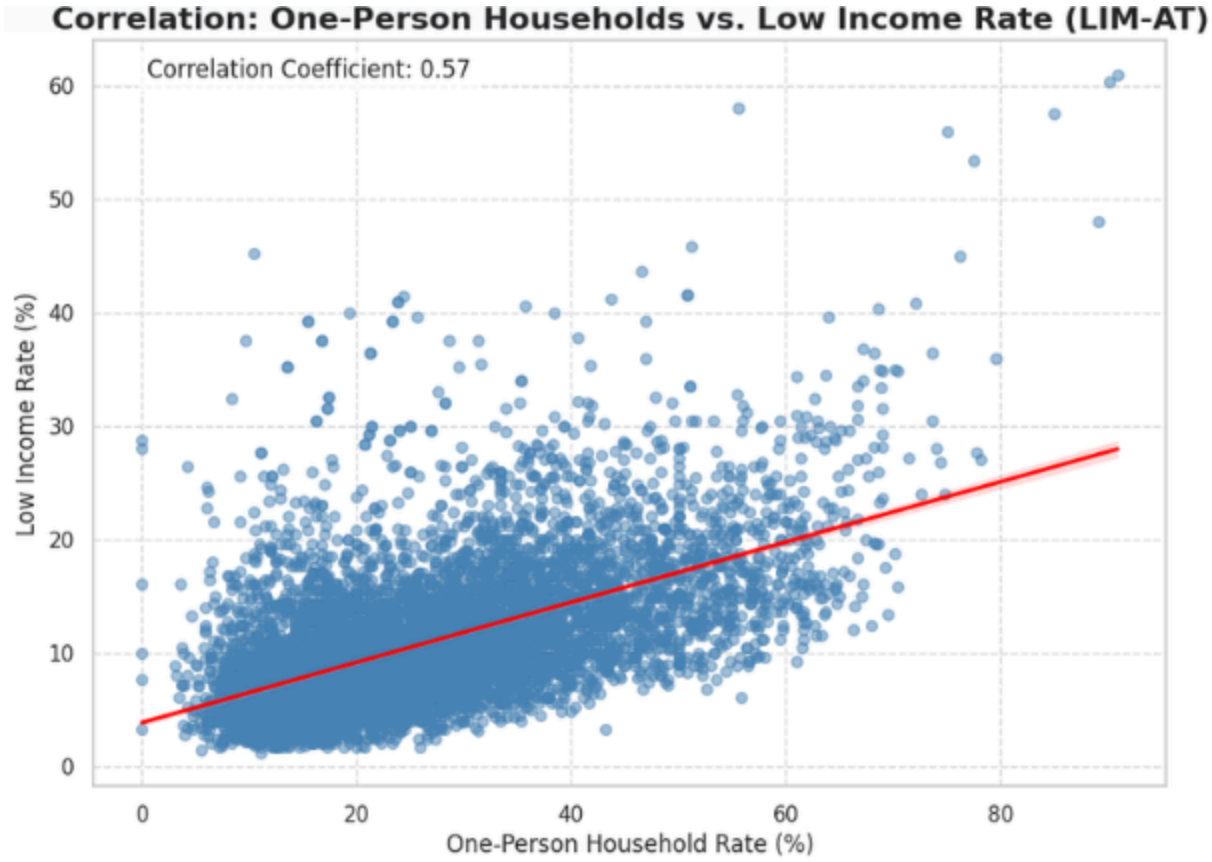
--- Generating Bar Chart ---



graph3



graph 4



4. **Findings and Observations**

Based on our analysis:

- Some provinces showed significantly higher population density than others.
- Income per capita varied widely across major cities.
- Language diversity was strongly tied to specific geographic regions.
- Window functions revealed population ranks and cumulative distribution across provinces.
- Aggregations provided useful insight into demographic and economic differences.

These observations demonstrate how Big Data tools can extract meaningful insights from large datasets of government information.

Our analysis revealed significant variation in one-person household rates across British Columbia.

At the Health Region (HR) level, average one-person household rates appeared relatively stable, generally around 30%. However, when analyzing Census Subdivisions (CSDs), we observed much higher variability, with some municipalities approaching or exceeding 45%.

Ranking and window function analysis showed that smaller communities often exhibit higher one-person household rates than larger regional averages suggest.

Additionally, correlation analysis indicated a moderate positive relationship between one-person household rates and low-income rates, suggesting potential socioeconomic implications.

5. Challenges

During the project, we experienced several challenges:

1. Dataset Size

The nearly 1 GB dataset required cloud-based processing to avoid RAM limitations.

2. Schema Consistency

Different CSV files sometimes use slightly different column names or data formats.

3. Data Cleaning

Some numeric columns contained commas or missing values that required attention.

4. Processing Time

Some Spark operations took longer depending on cluster size.

5. Understanding Window Functions

It took time to correctly partition and order the data for ranking and cumulative calculations.

6. Conclusion

Using PySpark and the 2021 Statistics Canada Census dataset, we successfully performed data ingestion, filtering, aggregation, joins, sorting, and window function operations.

This project demonstrated the value of Spark for handling large datasets and extracting insights efficiently.

The experience improved our understanding of distributed computing, SQL-like operations in PySpark, and the challenges of real-world data analysis.