COVID in Canada

Data Analysis at Work DSC680

Introduction

This project analyzes confirmed COVID-19 case data in Canada. The goal of the analysis is to examine patterns in reported cases over time and to explore demographic and regional factors associated with infection trends and health outcomes. By using publicly available national surveillance data from Statistics Canada and the Public Health Agency of Canada, this project applies data analysis techniques to better understand pandemic trends and to demonstrate how structured public health data can support evidence-based decision-making.

Business Problem

Public health agencies rely heavily on data to monitor disease spread, allocate healthcare resources, and guide policy decisions. During the COVID-19 pandemic, timely and accurate analysis of case data was critical for informing interventions such as lockdowns, travel restrictions, vaccination strategies, and hospital capacity planning.

The business problem addressed in this project is identifying measurable demographic and temporal factors associated with COVID-19 case trends and outcomes across Canada. Specifically, this analysis seeks to examine patterns related to age groups, gender, regional distribution, episode week, and hospitalization outcomes. By answering these questions, the analysis illustrates how historical case data can inform preparedness planning,

healthcare system management, and public communication strategies during future public health emergencies.

Datasets

The dataset used for this project was sourced from Statistics Canada, Catalogue number 13-26-0003, in collaboration with the Public Health Agency of Canada. The dataset contains confirmed COVID-19 cases reported across Canada and includes demographic and case-level variables derived from standardized reporting systems. Each row in the dataset represents a confirmed case of COVID-19. Variables include episode year, episode week, region, age group, gender, and hospitalization status. The dataset is structured in tabular format and provides consistent coding for categorical variables such as age ranges and geographic regions. Because the data originates from national surveillance systems, it provides a strong foundation for time-series and demographic analysis. However, the dataset may include reporting lags, changes in testing availability over time, and revisions to case definitions, which can affect interpretation. Additionally, the dataset focuses on confirmed cases and does not capture undiagnosed infections, limiting conclusions about total population spread.

Methods

This project will use a combination of exploratory data analysis and predictive modeling to examine COVID-19 case trends and associated health outcomes.

Exploratory data analysis will first be conducted to evaluate case counts by episode week and episode year, allowing visualization of infection waves over time. Additional analysis will examine demographic distributions, including age groups and gender categories, to identify which populations experienced higher reported case counts. Regional comparisons will also be conducted to explore geographic variation in case frequency.

New variables may be engineered to capture cumulative case counts, rolling averages, or wave indicators to better represent temporal trends. Aggregated hospitalization rates by age group and region will also be examined to assess severity patterns.

Following exploratory analysis, a supervised classification approach may be applied to predict hospitalization status based on demographic and regional variables. Input features may include age group, gender, region, and episode timing. Model performance will be evaluated using accuracy, precision, recall, and confusion matrices to assess how effectively demographic factors are associated with hospitalization outcomes.

The objective of the modeling process is not to predict individual health outcomes with certainty, but to assess whether meaningful and repeatable patterns exist within publicly available case data that can inform future preparedness planning.

Ethics

The use of public health data raises important ethical considerations related to privacy, interpretation, and communication. Although this dataset is anonymized and aggregated, researchers have a responsibility to ensure that findings do not stigmatize specific demographic groups or regions.

It is also ethically important to clearly communicate the limitations of the dataset. Confirmed case data depends heavily on testing availability, reporting practices, and evolving case definitions. Overstating the precision of findings could contribute to misinformation or misinterpretation of public health trends.

Additionally, predictive modeling in healthcare contexts must be approached cautiously. While patterns may emerge in aggregate data, individual health outcomes are influenced by many factors not captured in the dataset, such as comorbidities, vaccination status, and healthcare access. The analysis presented here is intended for educational and analytical purposes rather than clinical decision-making.

Challenges

The major challenge involves distinguishing between true changes in infection rates and changes caused by policy shifts, vaccination campaigns, or behavioral responses. Without additional contextual variables, causal interpretation is limited.

Finally, the dataset does not include detailed clinical variables such as underlying health conditions or vaccination status, which restricts the depth of outcome modeling. As a result, findings should be interpreted as descriptive patterns rather than definitive explanations of disease severity.

Reference Support

To support the results of this project, findings will be compared with official reports published by Statistics Canada and the Public Health Agency of Canada, as well as peer-reviewed public health research examining COVID-19 trends in Canada.

Appendix

Statistics Canada. (2020). Dataset on confirmed cases of COVID-19, Public Health Agency of Canada (Catalogue no. 13-26-0003). Government of Canada. Retrieved from https://www150.statcan.gc.ca/n1/pub/13-26-0003/2020001/COVID19-eng.zip