COVID

Sam Kluthe DSC 680

Introduction

This project analyzes COVID-19 case-level surveillance data to identify patterns in case volume over time, differences across regions, and outcome relationships (hospitalization and death). The analysis combines exploratory visualizations with classification modeling to estimate the likelihood of death among recorded cases using demographic and epidemiological attributes. The goal is to demonstrate how structured public-health data can be transformed into interpretable insights and a reproducible modeling workflow.

Business Problem

Public-health agencies and healthcare systems need timely insight into how disease burden changes across time and geography, and which factors are most associated with severe outcomes. Even when individual-level predictions are not used operationally, understanding which attributes are most strongly linked to death and hospitalization can support preparedness planning, communication strategies, and resource allocation. This analysis explores whether measurable variables such as age group, region, epidemiological timing, and hospitalization status to show meaningful relationships with death outcomes and whether those variables can produce a usable classification model.

Background and History

COVID-19 created repeated surges that varied by region, season, and variant period. Surveillance systems typically track cases using epidemiological weeks and years to standardize reporting and enable comparisons across time. Over the course of the pandemic, data collection evolved and categories such as “Not Stated” or “Unknown” appear in many public datasets. Because public health decisions often need to be made quickly and transparently, this project emphasizes simple, interpretable visuals paired with baseline models that can be explained to non-technical audiences.

Data Explanation

The dataset used in the notebook is loaded as **covid.csv** and contains categorical surveillance fields commonly used in epidemiological reporting. Key variables include COV_REG (geographic reporting region), COV_EY (epidemiological year), COV_EW (epidemiological week, renamed to “Week of Year” in the analysis), COV_AGR (age group), COV_GDR (gender), COV_HSP (hospitalization status), and COV_DTH (death status). To ensure consistency and accurate analysis, surveillance columns were converted to categorical data types, and outcome fields such as COV_DTH were recoded into human-readable categories (Yes, No, Not Stated). Placeholder or non-informative time values, such as Week 99, were removed from time-based visualizations to prevent artificial spikes or distortions. For predictive modeling, categorical predictors were transformed using one-hot encoding so they could be appropriately utilized within scikit-learn classification algorithms.

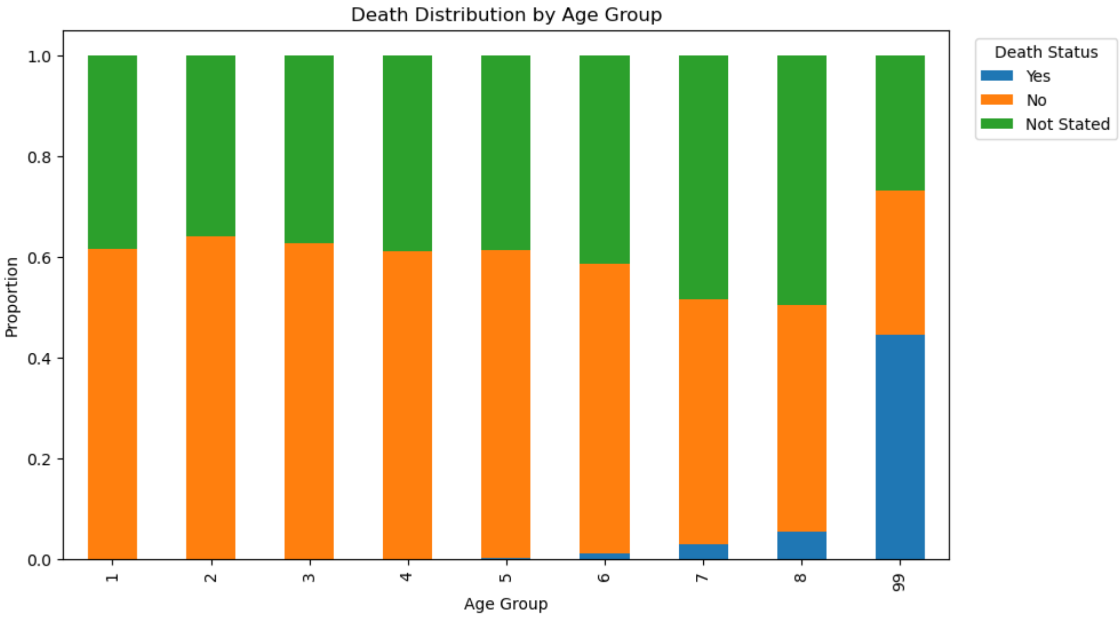
Methods

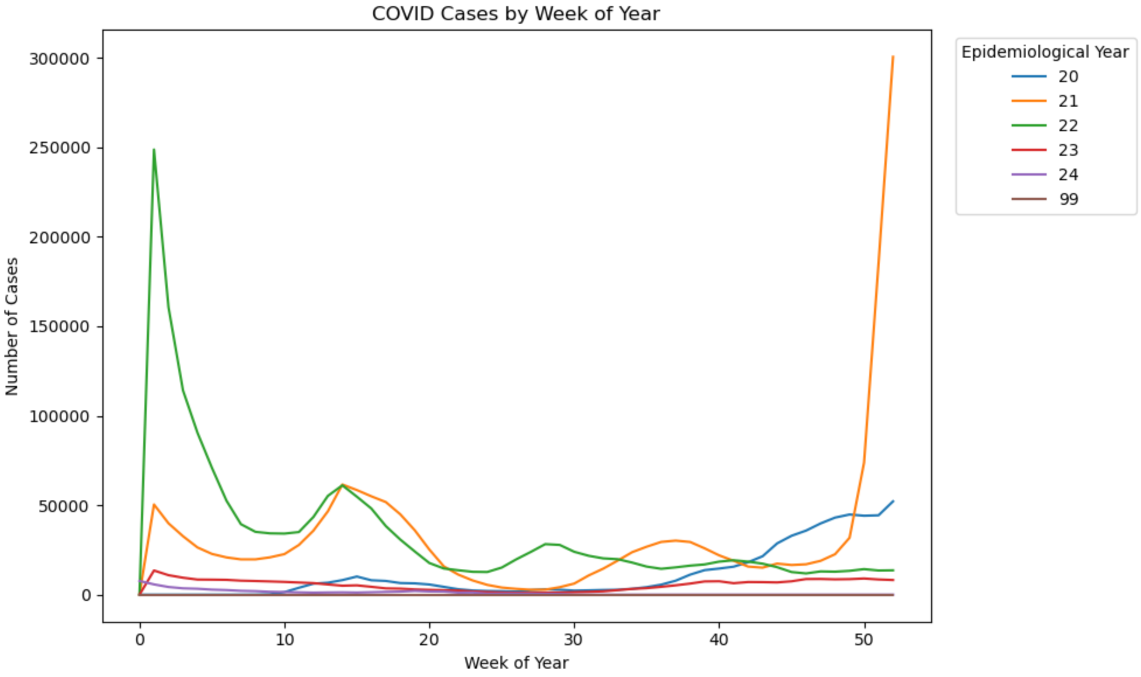
This study used a combination of exploratory data analysis and supervised machine learning techniques to evaluate COVID-19 surveillance patterns and severity outcomes. First, the dataset was cleaned and standardized by converting key variables such as region, epidemiological year, epidemiological week, age group, gender, hospitalization status, and death status into categorical formats to ensure consistent handling. Non-informative values, such as placeholder weeks used for missing data, were excluded from time-based analyses to prevent distortion of trends. Exploratory visualizations were then created to examine case distributions across regions and epidemiological years, weekly surge patterns using heatmaps, and proportional outcome differences by age group and hospitalization status. Following exploratory analysis, a predictive modeling framework was implemented to estimate the likelihood of death among cases with known outcomes. Categorical predictors were transformed using one-hot encoding, and the dataset was split into training and testing subsets with stratification to address class imbalance. Two models were developed: logistic regression, chosen for interpretability and baseline comparison, and a random forest classifier, selected for its ability to capture non-linear relationships and interactions among predictors. Model performance was evaluated using confusion matrices, precision, recall, F1-scores, and ROC-AUC metrics, and feature importance rankings from the random forest were examined to identify the strongest contributing variables.

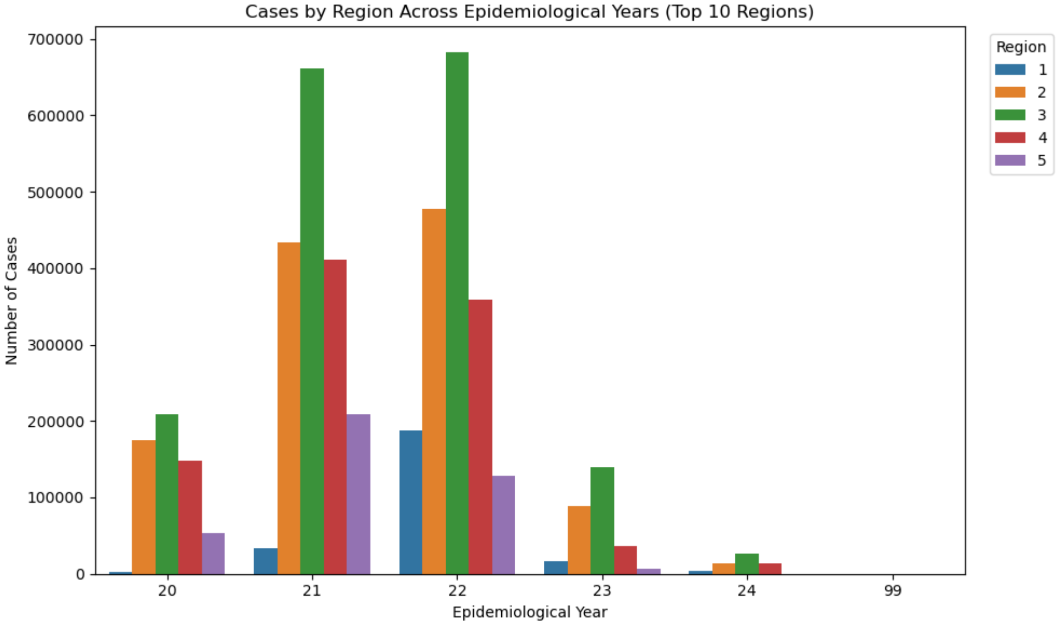
Analysis

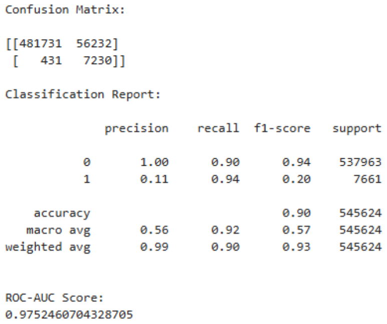
The analysis revealed clear variation in COVID-19 case volume across regions and epidemiological weeks, with distinct surge periods rather than a steady distribution of cases throughout the year. When examining outcomes, proportional visualizations showed that death rates increased substantially across older age groups, reinforcing the well-documented severity gradient associated with age. Graphs comparing hospitalization status further demonstrated that individuals who were hospitalized experienced markedly higher death rates than non-hospitalized cases, highlighting the strong association between clinical severity and mortality outcomes. A stacked bar chart of death distribution by age group clearly illustrated how the composition of “Yes” outcomes shifts toward older categories, while a comparative bar chart of death rate by hospitalization status emphasized the disparity between hospitalized and non-hospitalized populations.

Although the correlation matrix provided a general overview of relationships among encoded variables, predictive modeling results reinforced these visual findings by identifying age group and hospitalization indicators as among the most influential predictors of death. Together, the graphical analysis and model outputs demonstrate that even high-level surveillance variables contain meaningful signal for understanding severity patterns, and that visualizing death rates across demographic and clinical groupings provides clearer insight than case counts alone.

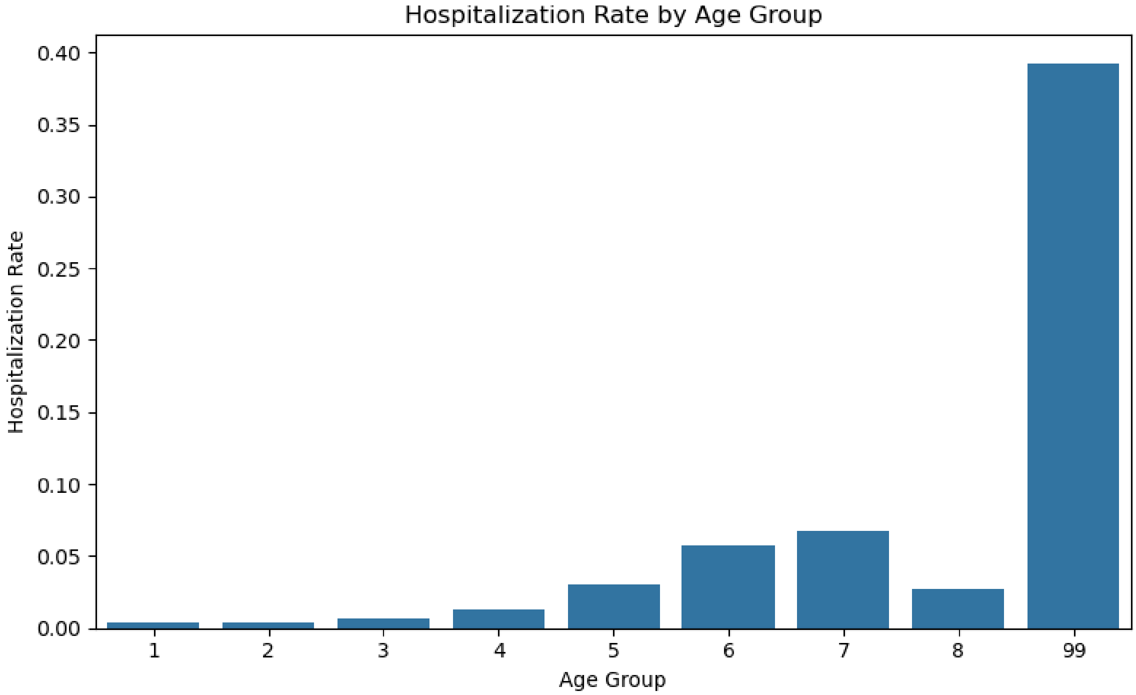



COVID Cases by Week of Year
Number of Cases
300000
250000
200000
150000
100000
50000
0
0
10
20
30
40
50
Week of Year
Epidemiological Year
20
21
22
23
24
99


Cases by Region Across Epidemiological Years (Top 10 Regions)
Number of Cases
700000
600000
500000
400000
300000
200000
100000
0
20
21
22
23
24
99
Epidemiological Year
Region
1
2
3
4
5



```
Confusion Matrix:

[[481731  56232]
 [   431   7230]]

Classification Report:

              precision    recall  f1-score   support

           0       1.00      0.90      0.94    537963
           1       0.11      0.94      0.20      7661

    accuracy                           0.90    545624
   macro avg       0.56      0.92      0.57    545624
weighted avg       0.99      0.90      0.93    545624


ROC-AUC Score:
0.9752460704328705
```



Conclusion

In conclusion, this analysis demonstrates that COVID-19 surveillance data can provide meaningful insight into temporal trends, regional variation, and severity patterns when explored through structured visualization and predictive modeling. Case volumes varied significantly across epidemiological weeks and regions, reflecting distinct surge periods rather than uniform spread. More importantly, outcome-focused analysis showed that death rates were strongly associated with age group and hospitalization status, with older and hospitalized individuals experiencing substantially higher mortality proportions. The modeling results reinforced these visual findings, indicating that even without detailed clinical variables, demographic and surveillance indicators contain measurable predictive signal. Overall, the project highlights the value of combining exploratory data analysis with interpretable machine learning techniques to better understand public health patterns and communicate findings in a clear, evidence-based manner.

Assumptions, Limitations, Recommendations and Challenges

This analysis assumes that the surveillance data was accurately and consistently reported across regions and time periods. Limitations include potential reporting bias, missing or "Not Stated" values, and the absence of detailed clinical variables such as comorbidities or vaccination status. Additionally, correlations based on encoded categorical variables must be interpreted cautiously, and the models identify associations rather than causal effects.

Future improvements could include incorporating population-adjusted rates, adding more health-related predictors, and evaluating model fairness across demographic groups. A key challenge was balancing clear, interpretable visualizations with meaningful predictive modeling while managing large categorical datasets.

Ethical Assessment

This analysis uses aggregated, non-identifiable surveillance data, which minimizes direct privacy risks. However, ethical responsibility remains important when interpreting and presenting results. Findings were communicated as associations rather than causal claims to avoid overstating conclusions, particularly when discussing higher death rates among certain age groups or hospitalized individuals. Care was taken not to frame results in a way that stigmatizes specific regions or demographic groups, recognizing that outcomes are influenced by broader social, healthcare, and structural factors not captured in the dataset.

Additionally, predictive models were used for analytical insight rather than decision-making about individuals. Model outputs should not be treated as deterministic predictions, especially given limitations in reporting consistency and missing clinical variables. Transparent reporting of assumptions, limitations, and uncertainty is essential to ensure responsible use of public health data and to prevent misinterpretation of severity patterns.

10 Questions

What was the primary objective of this COVID-19 analysis?

Why is it important to examine cases by epidemiological week instead of only total case counts?

What patterns were observed when comparing case volume across regions?

How did death rates differ across age groups?

What relationship was identified between hospitalization status and death outcomes?

Why were proportional death rates more informative than raw counts?

What was the purpose of using both logistic regression and random forest models?

Which variables appeared to be the strongest predictors of death?

Why should correlations in categorical data be interpreted cautiously?

What are the main limitations of using surveillance data for predictive modeling?

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