Conestoga College
School of Applied Computer Science & Information Technology

SENG8080 - Case Studies Big Data

Topic - Data Analysis on Covid-19 positive cases, fatal rate, and Vaccine

By
Shraddha jitendrabhai Kapadia (8757108)
Nikhlesh Deepakbhai Mulrajani (8840531)
Christina Arokiaswamy (8923489)
Dhwani savani (8897043)

October 20, 2023

Abstract

In this report, we will present a comprehensive analysis of COVID-19 data, focusing on positive cases, the fatal rate, and vaccine distribution. This analysis aims to provide insights into the trends, patterns, and impact of the COVID-19 pandemic.

Introduction

The COVID-19 pandemic has swept the globe, demanding immediate attention and data-driven responses. This report undertakes a focused analysis of COVID-19 data, specifically concerning positive cases, fatal rates, and vaccine distribution. Our goal is to uncover critical insights to guide decision-making in these challenging times.

We have three primary objectives: to identify trends in the progression of COVID-19 cases, assess the severity of the pandemic through fatal rates, and evaluate the impact of vaccination efforts. This analysis is not just data exploration; it's an indispensable resource for healthcare professionals, policymakers, and the broader community as we navigate the complexities of the COVID-19 crisis. In the subsequent sections, we detail our data sources, collection methods, quality assurance measures, and visualization techniques, with the aim of providing valuable perspectives to aid in our collective response to the pandemic.

Data Research and Integration

Our data analysis primarily relies on publicly accessible data sources provided by the Ontario government know as Ontario public site. The official Ontario COVID-19 website serves as our primary data hub, offering up-to-date information on positive cases, testing, vaccinations, and public health guidelines. This transparency enables us to conduct a comprehensive analysis of COVID-19 trends within the province.

To ensure data accuracy, we follow stringent integration and preprocessing procedures. We harmonize data from different sources, address naming variations, and perform data cleaning. This meticulous process results in a high-quality dataset for reliable analysis. Our commitment to utilizing publicly available data aligns with open access principles, promoting transparency and facilitating broader validation of our findings.

Data Collection

API for Data Collection:

1. https://covid-193.p.rapidapi.com/statistics
2. https://data.ontario.ca/dataset/covid-19-vaccine-data-in-ontario/resource/c08620e0-a055-4d35-8cec-875a459642c3

We gather COVID-19 statistics continuously from the COVID-19 statistics API on RapidAPI . This API provides up-to-date information on cases, testing, fatalities, and vaccinations. Our ongoing data collection process ensures we have the most current data for analysis.

Data Cleaning

The purpose of this data cleaning script is to prepare a dataset for analysis by performing various cleaning and preprocessing tasks. The Python code is used to clean and prepare vaccine data for analysis. The cleaning process involves the following key steps:

Removing Empty Columns:
Empty columns in the dataset are eliminated. Columns with all missing values are dropped to reduce data redundancy.

Handling Missing Values:
Missing data is handled by replacing empty values with zeros. This ensures uniform treatment of missing data throughout the dataset.

Eliminating Duplicate Rows:
Duplicate rows are removed to eliminate redundant information and maintain dataset integrity.

Cleaning 'agegroup' Column:
The 'agegroup' column is cleaned to extract the highest age values. Symbols ('+', '-', 'yrs') are removed, and '-' is replaced with 'to'. Rows with 'ALL' values are excluded. The highest age is determined from the age ranges, and the cleaned data is stored in a new 'highest_age' column.

Cleaning 'date' Column:
The 'date' column is transformed into a datetime data type, enabling date-related operations. The date part is extracted and retained within the 'date' column.

Handling Outliers with IQR:
Outliers are identified and managed using the Interquartile Range (IQR) method. Lower and upper bounds are calculated based on the IQR, and rows with values falling outside these bounds are removed.

Saving Cleaned Data:
The final cleaned dataset is saved as 'cleaned_dataset.csv' in CSV format. The 'index' parameter is set to 'False' to exclude the index column when saving.

Purpose:

The purpose of this data cleaning process is to ensure that the vaccine data is ready for further analysis. It involves removing irrelevant columns, addressing missing values, eliminating duplicates, extracting relevant information, and handling outliers. The cleaned dataset provides a reliable and consistent basis for meaningful analysis and insights.

Data Storage and Maintenance

To keep all the COVID-19 data safe and organized, we need a strong storage system. This data includes information about cases, tests, and vaccines, and it's a lot of information. We have to follow rules to make sure the data is private and secure, like a secret code.

We use two tools for this job. First, we use Python to prepare and organize the data. It's like sorting and cleaning the data to make it ready for analysis. Then, we use MsSQL, which is like a safe and big storage room for the data. It helps us keep all the information in one place, and it's good at protecting the data. This way, we can work with the data effectively and keep it safe.

Data Quality

Maintaining the quality, accuracy, and reliability of our COVID-19 data is paramount. Data validation procedures are employed to double-check data accuracy by cross-referencing multiple sources, detecting inconsistencies, and addressing errors. For data cleaning, we utilize Python, a versatile tool. Python scripts are crafted to automatically identify and rectify issues such as missing or inconsistent data. This cleaning process involves filling in gaps, removing duplicates, and standardizing data formats. These measures ensure that our dataset remains accurate, consistent, and reliable, forming the bedrock for trustworthy and meaningful analysis of COVID-19 data.

Data Analysis and Visualization

We've looked at COVID-19 data, focusing on cases, how many people got really sick, and vaccines. We found important things. For cases, we saw when more people got sick and when it got better. Fatal rates helped us see how serious the virus is for different groups. And we checked how vaccines are being given and how they help. We used Tableau to make pictures and charts to show these things clearly. This makes it easier for doctors, leaders, and everyone to understand and make good choices in managing the virus.

Extension

We have the opportunity to make our project even more informative by including data from more sources and a wider geographic area. This means we can look at COVID-19 in more places and get a better overall picture.

To do this, we need to plan for having enough space to store all the extra data. We also have to make sure the new data fits with the old data and is accurate. This can be a complex job that might need more people and better computers. We'll create a clear plan with dates for each step

of this expansion. Our goal is to offer a deeper understanding of COVID-19's impact on a larger scale.

Project Timeline

Date	**Deliverable**	**Responsible**
Oct 20	Data Collected and planned	Christina, Shraddha
Nov 4	1st Draft Circulated to Team	Christina, Shraddha, Dhawani, Nikilesh
Nov 11	1st Draft of Presentation Circulated	Christina, Shraddha, Dhawani, Nikilesh
Nov 11	User testing by the team and errors/refinements identified.	
Nov 17	Final Adjustments made and checked	
Nov 18	Process and Report Due at 10pm	