Data Wrangling and Preliminary Analysis

Sometimes, it is important to view the raw data as it helps in identifying the problems with the data. The data that I worked on is related to the confirmed cases of COVID-19 in Canada which is published by Statistics Canada. The data contains different variables such as the Case Identifier Number (which is a unique value for each data record), Region, Gender, Age group etc. The data is downloaded from the data source in the Comma Separated Value (CSV) format. It is difficult to perform computations on the data that is stored in CSV format. Therefore, I used pgAdmin 4, which is a PostgreSQL Database management tool to store the data. When compared to the data in CSV files, accessing and querying the data is easier if the data is stored in a database. Firstly, the database schema is created in pgAdmin 4 and then a database table is created under the schema with the data columns of integer type because the data contains only numeric values. Subsequently, the data that is in the CSV file is loaded into the database table. PostgreSQL queries are used to interact with the data. For better visualization, it is important to clean the data and perform some preliminary analysis and preprocessing. Before cleaning up the data and performing preliminary analysis, the metadata is reviewed which summarizes the data and describes what each data column is and how the data specific to a column is represented. This data is crucial as it provides vital information regarding the data columns that play an important role in visualizing the data. The data that I worked on contains the null values and sentinel values. All the data records with null values are eliminated and are not included as a part of data visualization. The data that contains the sentinel values are retained because these values are used in visualizing the data and the data with sentinel values are visible in the visualization. Few variables that I used in visualization are Onset Week of Symptoms, Onset Year of Symptoms, and the Age group. Instead of the week and the year number, I want to represent the data in the date format. Thus, Onset Week of Symptoms and Onset Year of Symptoms variables are pre-processed so that these variables are represented by the start date and end date of the week. The start date and end date of the week are stored in two different columns in the data table. The age group variable, which is a categorical data is stored in numeric format. These values are decoded, and the categorical values of the age group variable are stored in a separate column. Apart from the confirmed cases of COVID-19 data, the population of different regions and GeoJSON data that encodes the geography of Canada are also used to visualize the data.