# Ireland Covid Statistics (2020-22)

## 1. Introduction

Covid-19, also known as Corona Virus, is caused by SARS-CoV-2 virus. It was declared a pandemic by the World Health Organisation on 11th March 2020. The virus is considered to be very contagious as it has affected more than 600 million people across the globe and caused more than 6 million deaths. In Ireland, is has affected more than 1.68 million people and caused more than 8270 deaths. Analysing the key statistics related to the spread of disease can aid in planning effectively and help to decrease the spread of this virus and reduce the load on healthcare infrastructure.

## 2. Technologies Used

I have used below technologies to create the visualisations:

- **R**
  R was used to pre-process and manipulate the data provided by the Government of Ireland. I have used libraries like DPLYR and PLYR to perform various operations on the data. R Studio was used to write all the code for this assignment.
- **GGPLOT2 and Plotly**
  GGPLOT and Plotly libraries were used to create the visualisations for this assignment. SF library was used to visualise the spatial data.
- **Shiny**
  Shiny library was used to create the dashboard for the visualisations. All the panels, dropdowns present in the dashboard were also created using this library. SHINYTHEMES library was used to provide a suitable theme to the dashboard.

## 3. The Dataset

The dataset used in this visualisation is provided by the Government of Ireland. Following are the details of this dataset:

**Dataset:** COVID-19 HPSC County Statistics Historic Data of Ireland

**URL:** https://data.gov.ie/dataset/covid-19-hpsc-county-statistics-historic-data?package_type=dataset

**Type:** Shapefile

**Dataset Description:** The dataset contains the county-wise statistics for Total Confirmed cases for each day starting from 27th February 2020 to 12th December 2022. It also contains to total population and proportion of covid cases based on population of each county. Furthermore, the dataset also has spatial information like Shape_Area, Shape_Length, Latitude and Longitude for each county.

**Dataset Attributes:**

- Attributes present in the dataset are: OBJECTID, ORIGID, CountyName, PopulationCensus16, TimeStamp, IGEasting, IGNorthing, Lat, Long, UGI, ConfirmedCovidCases, PopulationProportionCovidCases, ConfirmedCovidDeaths, ConfirmedCovidRecovered, SHAPE_Length, SHAPE_Area
- There is categorical data present in the dataset like CountyName
- There is Continuous Quantitative data type in dataset like PopulationCensus16, TimeStamp, IGEasting, IGNorthing, Lat, Long, UGI, ConfirmedCovidCases, PopulationProportionCovidCases, ConfirmedCovidDeaths, ConfirmedCovidRecovered, SHAPE_Length, SHAPE_Area

**Data Pre-processing**

- I have checked for null rows in ConfirmedC column which represents the Total confirmed cases in each county. If any null row was found, it was removed from the dataset.
- For all the visualisations, I have first normalised covid cases by dividing the Total Confirmed Cases for each county by its Population and then multiplying with 100,000. This gave us normalised Caseload per 100K population for each county The primary reason for normalising the data is to ensure that data for each county is standardised as population of each county is different and case load will depend on the population. Therefore, normalising this would make us to better analyse the data.

- For all the visualisation, I divided the dataset into 3 parts based on the year i.e., Cases in each county as of 31st December 2020, 31st December 2021 and 12th December 2022. The primary reason for this to make 3 visualisations for each statistical plot as this will better help us to understand the spread of the virus in each county on yearly basis.
- For Choropleth map, normalised caseloads in each county were categorised into different interval ranges. A colour palette was also generated to for these ranges.
- For Time Series Visualisation, I have first identified counties with Maximum Confirmed Cases and minimum confirmed cases. Then, I created subset of the data for these counties along with data for Dublin. This was done in order to highlight these counties on the time series graph.
- For Difference from Mean Cases Visualization, I have calculated difference from mean for total confirmed cases for each county.

## 4. The Visualizations

- **Choropleth Visualisation for the total number of cases per 100K population**
  **Idiom Used:** Spatialized Data Arrangement in form of Choropleth Map
  **Visual Encoding Channels:**
  - **Position**: Positional encoding is used to highlight different counties on the choropleth map.
  - **Brightness:** Brightness has also been used as an encoding channel in Choropleths with darker shade of colour representing higher values and lighter shade of colour representing lower value.

  **Task**
  Some of the key tasks associated with the visualisation are categorize, compare, and distinguish. The Choropleth map can be used to highlight the number of total number of cases in each county in different categories of case load, which makes it very useful to analyse the data quickly. It can be used for exploratory data analysis as one can easily identify the counties that have higher caseload than others. The pop-up that comes on hovering the map makes it easier for the user to understand the values represented in the choropleth map as it highlights the county name and caseload. Generating year wise choropleth maps can help to see the spread of virus in each county on yearly basis.
  **Justification:**
  A choropleth map is a coloured map that is used to display the divided geographical areas or regions with colour of the region related to a numeric variable. I have used choropleth map as positional encoding because it makes it easier for the user to identify the county they want to analyse. Also, I have used brightness as it is easily distinguishable and it makes it easier to highlight counties with higher caseload than others. To achieve this, I created a custom palette to represent the caseload in each and I chose “Plasma” palette as the colours of this palette are easily distinguishable and I also reversed the order of colours as I wanted light colours first to represent smaller caseload range and dark colours to represent higher caseload range

- **Time Series analysis graph for all counties in Ireland**
  **Idiom Used:** Line graph in form of Time Series
  **Visual Encoding Channels:**
  - **Position:** Positional encoding is used to indicate the month and the caseload in each county or Ireland.
  - **Colour:** Colour encoding is used to highlight counties that had maximum and minimum caseload at the end of each year and one colour is used to highlight Dublin, which is capital of Ireland

  **Task**
  Some of the key tasks associated with the visualisation are associate, compare, distinguish, and identify. The main analysis that can be done from the visualisation is that one can easily identify the county that has minimum and maximum caseload as these are specifically highlighted in the visualisation. One can analyse which counties have fared well in containing Covid cases over the period of time. It can be used for both exploratory and explanatory data analysis to find the tipping points in the visualisation that can help to understand the important factors related to increase or decrease in the caseload in each county, for instance effectiveness of change in covid related policies, vaccination drives, etc. The pop-up that comes on hovering the line graph makes it easier for the user to understand the values represented in the time series graph as it highlights the county name and caseload for that particular month.

**Justification:**
I have used Positional Encoding because time series is the study of the evolution of one or several variables through time and a time series line graph is used to provide a visualization of the evolution of one or several numeric variables over time by connecting the data points by the line segments. In our case, we wanted to understand the evolution of the Total confirmed cases month by month for each county. Therefore, it is an apt channel for this type of data. Also, I have used colour as encoding channel to highlight counties with highest and lowest caseload and Dublin County. This makes it easier for user to identify these counties.

- **Difference in Total Confirmed cases from mean cases for each county**
  **Idiom Used:** Horizontal Bar graph
  **Visual Encoding Channels:**
  - **Position:** Positional encoding is used to indicate the month and the caseload in each county or Ireland.
  - **Colour:** Different colours are used to highlight counties that had more cases and less cases than mean confirmed cases.

  **Task**
  Some of the key tasks associated with the visualisation are associate, categorise, compare, distinguish, and identify. It can be used for both exploratory and explanatory visualisation as one can easily identify the counties that have more cases than mean cases and make effective contingency plan for these counties. By comparing these visualisations for period of 3 years one can identify the effectiveness of covid containment policies in each county. The pop-up that comes on hovering the bar graph highlights the county name and difference from mean.
  **Justification:**
  I have used Positional Encoding in form of bar graph because it makes it very easy to understand and is best suited to show comparisons from a base value like mean. Also, I have used colour as encoding channel with red colour to highlight counties that have more cases than mean case load and blue colour to highlight countries that have less cases than mean caseload in the country.

## 5. The Novelty and Complexity Analysis

From novelty and complexity point of view, Covid-19 data has been visualised in different ways by multiple platforms. I have chosen these visualisations because these can be used for both explanatory and exploratory analysis. Quick comparisons between data from each county of Ireland can be made and it can be used for effective management of this virus. As per my understanding, I have not seen this data being visualised in terms of Time Series and Difference from mean plots for Ireland. The dashboard allows user to interactively choose the visualisation and the year user wants to analyse. At the same time, user can also see the remaining plots that can also aid in critical analysis. Furthermore, each of these visualisations combine more than one encoding channels which added to overall complexity. Also, manipulating and pre-processing such a large dataset was also very complex as I have to create new columns from existing columns and subset of the dataset so that it can be used in visualisations.

## 6. Strengths and Weakness Analysis

**Strengths**
The main strength of these visualisations is that these are fairly easy to understand and one can perform critical analysis from it. It supports both exploratory and explanatory analysis. I have also used the screen real-estate judiciously by dividing the dashboard into 2 panels – the main panel (70% of screen) and a bottom panel (30% of screen). The main panel contains a sidebar from which user can choose year and. type of visualisation. The dashboard is interactive and allows user to modify the main view by choosing the main visualisation and year which is presented in main panel of dashboard. At the same time, it shows other 2 visualisation in side panels without overcrowding the overall screen. The main visualisation supports task like panning, zooming and auto scaling.

**Weakness**
The key weakness is that more filters could be provided like month-wise filter and county-wise filters. The data also does not contain essential parameters like vaccination and deaths related data. Daily covid cases analysis could not be performed using this dashboard.

## References:

### Video, Code and Dataset

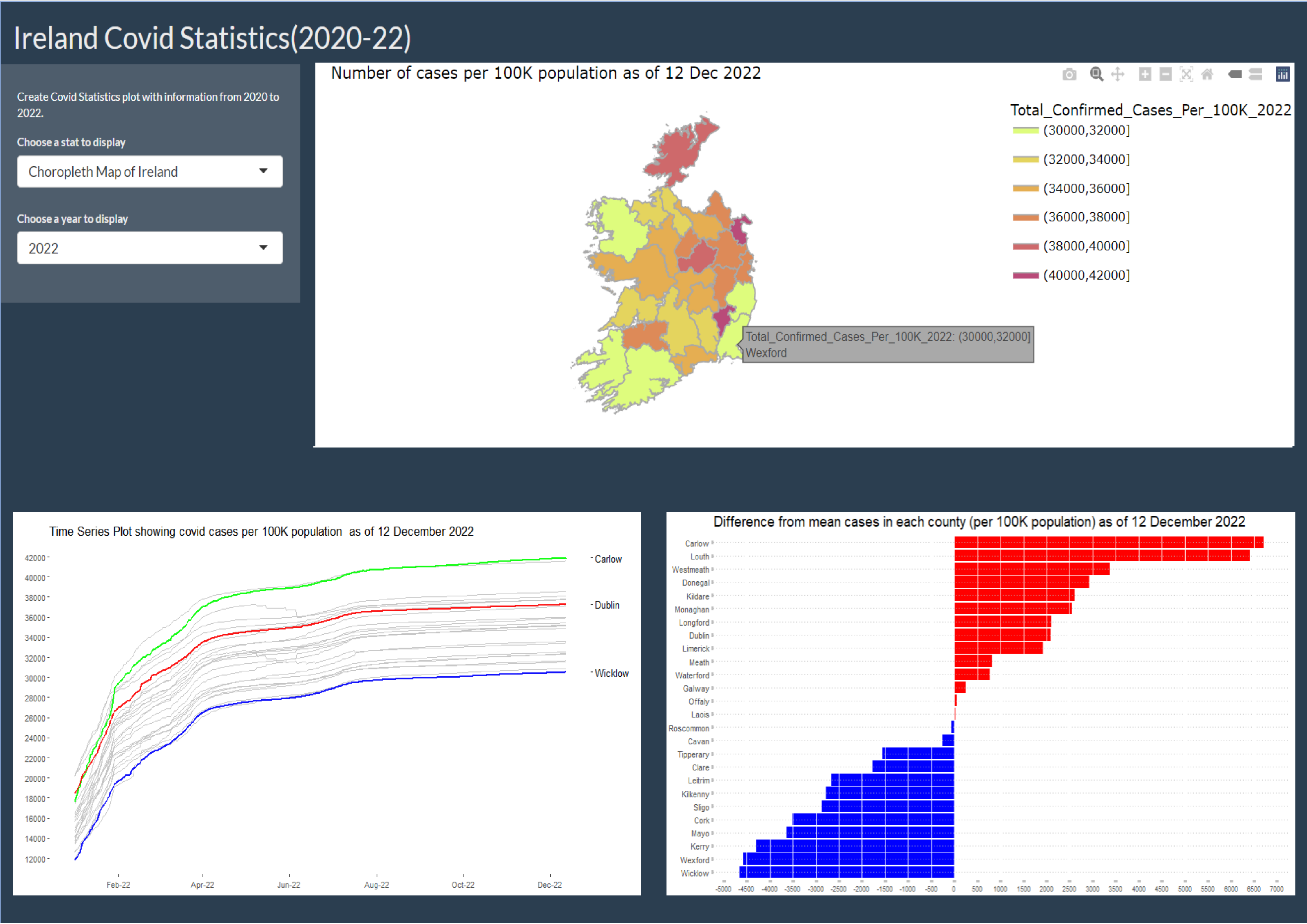
https://drive.google.com/drive/folders/1Dtmt-W9WhDBGbiHQtgy9StpoM4SczpzA?usp=sharing

### Running Instructions:

1. Download Code and Dataset from Google drive.
2. Place Code and dataset in same folder
3. Open R Studio
4. Open file Assignment3_DuaKaran_21331391_Code.R
5. Install the required libraries, if required. By default, these lines are commented as part of submitted code.
6. To set current working directory, right click on the file name in R studio and choose 'Set Working Directory' option.
7. Execute code.
8. Once Shiny window appears, click on 'Open in Browser'

Also, I have attached code along with this report where I have added comments on each line and tried to explain the purpose of each line of code.

Please refer next page for screenshot of the dashboard.



# Ireland Covid Statistics(2020-22)

Create Covid Statistics plot with information from 2020 to 2022.

Choose a stat to display: Choropleth Map of Ireland

Choose a year to display: 2022