Assignment 2

Part 1

Aim: In this assignment, you are required to visualise a data set of Irish covid case numbers per county.

Theory/Working:

The data is in the form of a shape file (.shp) which holds the data for map coordinates, in a multi-polygon format, which is used to make a layout of the map accordingly. The dataset contains the data for COVID-19 confirmed and daily cases from February 2020 to January 2022 day-wise, according to each county present in Ireland.

- ***Question 1:*** *A visualization that allows the reader to accurately compare the cumulative number of cases per 100,000** of population per county on 21 December 2021. County Galway should be highlighted.*

The bar chart is a good choice of visualisation in this question as it allows readers to accurately show the impact of COVID-19 cumulative cases on 21st December 2021. The bar chart is reordered from highest to lowest, and Galway is colored differently from the rest as that county is to be compared to the rest. The bar chart in Figure 1 clearly shows Monagrah as the county with the highest cases, while Wicklow has the lowest. To calculate the Cumulative Cases per 100k population, we manipulated the ConfirmedC column in the data using the `mutate` function from *dplyr* package, like this:

*mutate(ConfirmedC_per_100k = round(100000 * ConfirmedC/Population,1)*

The sketch in Figure 1 shows the results as expected from the programming.

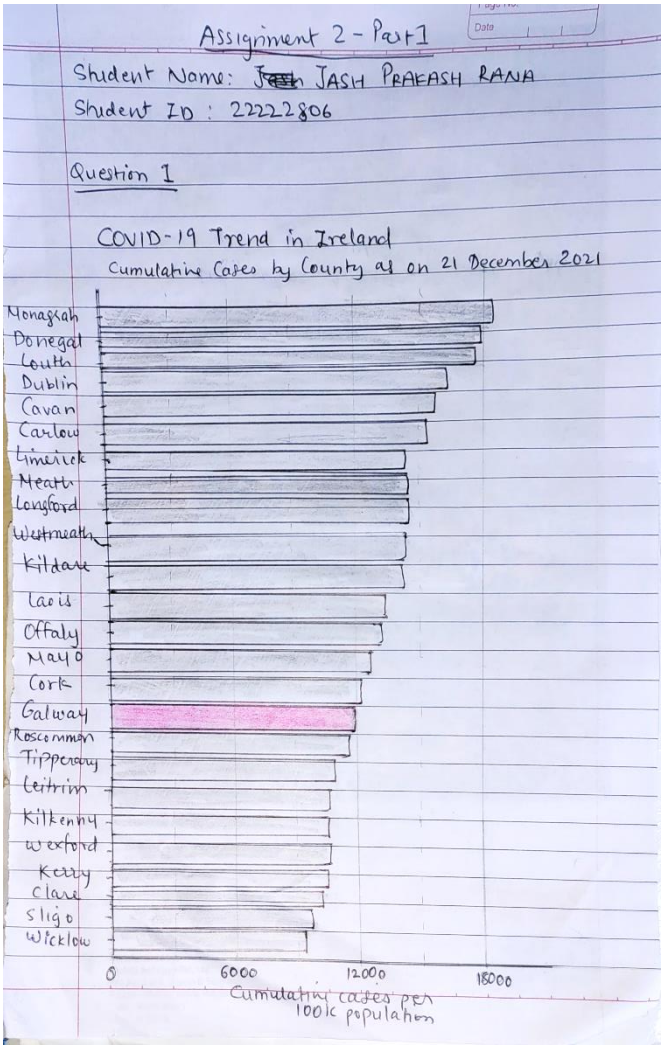


Figure 1: Bar Chart for Covid-19 Cumulative Cases as on 21st Dec 2021.

- ***Question 2:*** *A visualisation that allows the reader to read how each county diverges from the mean cumulative number of cases (per 100,000) in the country as of 21 December 2021. You may also use a daily figure in this section. County Galway should be highlighted.*

The diverging lollipop chart makes it a very interesting fit for this question as the mean line can serve as the point dividing the data into two different groups. Figure 2 shows us the mean line in *blue* and its value to be '13,578.99' for the counties and the points in lollipop fashion reordered from highest to lowest cumulative cases as on 21st December 2021. The points are connected to the mean line showing the distance from the mean, visually appealing to show how much an individual point is far off from the mean data value. Galway is again highlighted similarly to show it in the foreground. An interesting find is that the diverging lollipop chat is different from the bar chart in the sense that this has an X-axis scale starting from 10,000 units, with breaks of 2000 units.

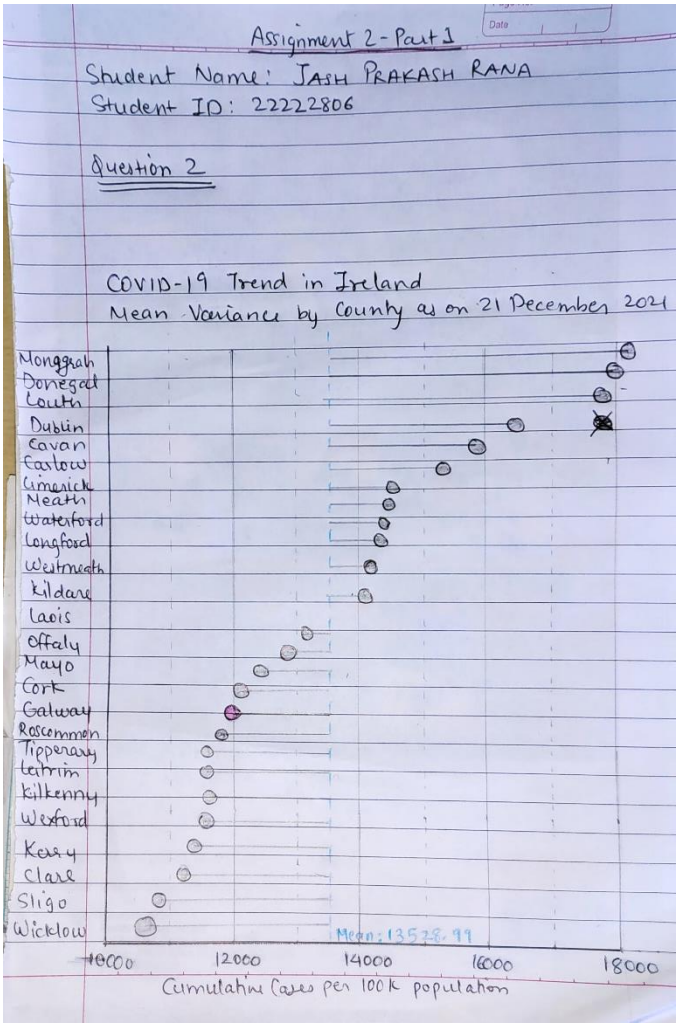


Figure 2: Diverging Lollipop Chart showing Mean Variance for each County on 21st Dec 2021.

- ***Question 3:*** *A visualisation showing the daily number of confirmed covid cases in one county in Ireland for an 18-week period. This visualisation should help the reader to perceive the trend in the data.*

For this solution, we utilise the `line graph with a fill` graph which shows us the start of COVID-19 cases in Galway county i.e. from March to mid-July 2020, completing an 18-week period. Figure 3 clearly depicts the confirmed number of cases for Galway, right from the start of them, and how it rapidly started growing which shows the pandemic's unrestricted rise and failure of the control over it. Galway is again highlighted similarly, and the fill is a lighter color than the line showing the rise. These graphs are used to show how much actual impact something can have over the course of time, and no other graphs have this visual aspect to show an impact. The readers would be amazed to know that this is a very simple graph and doesn't require any calculations as such.

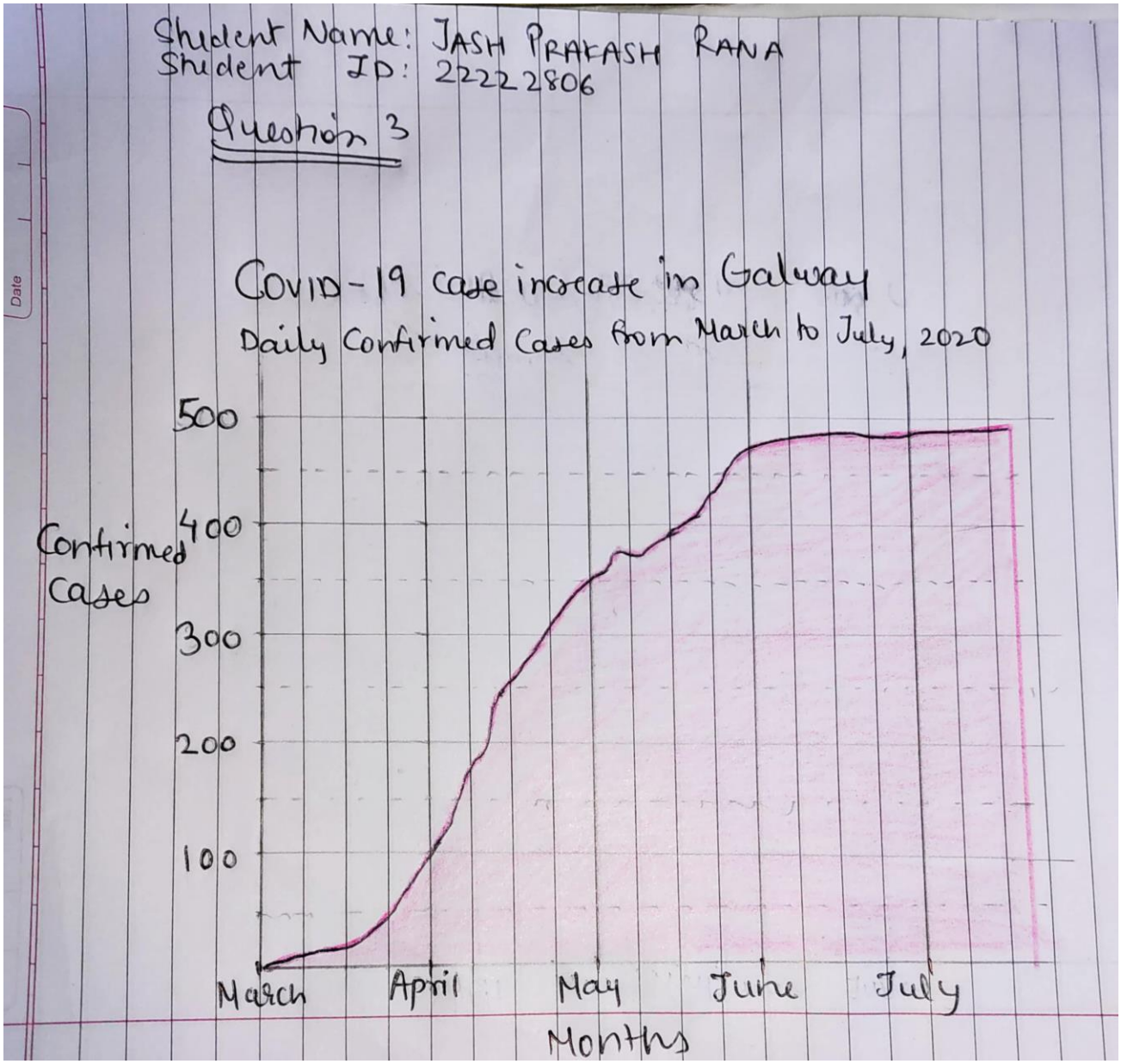


Figure 3: Line with a fill Chart showing the rise of COVID-19 cases from March-July 2020.

- ***Question 4:*** *A visualisation that highlights the cumulative number of cases per 100,000 in Galway and two other counties representing counties that have had the lowest and highest number of cases per 100,000 over the full timeline of the dataset. The visualisation must also show the cumulative case number for all other counties in Ireland in the same plot.*

To highlight three counties showing time-series data over the whole period of the dataset within the data of all the other counties, the line graph is a better fit for showing this. It has the capacity to visually connect the time with the progression of the data. It's a simple graph but can get cluttered, but it is easy to foreground the visual aspects required using coloring and choosing the alpha values for the background to be mild. To show this, we have used Galway, Monagrah, and Wicklow as three counties to compare with, all with a distinct color, while others are kept at grey as seen in Figure 4. The figure shows exactly how many confirmed cases for each county are traced with the change in time.

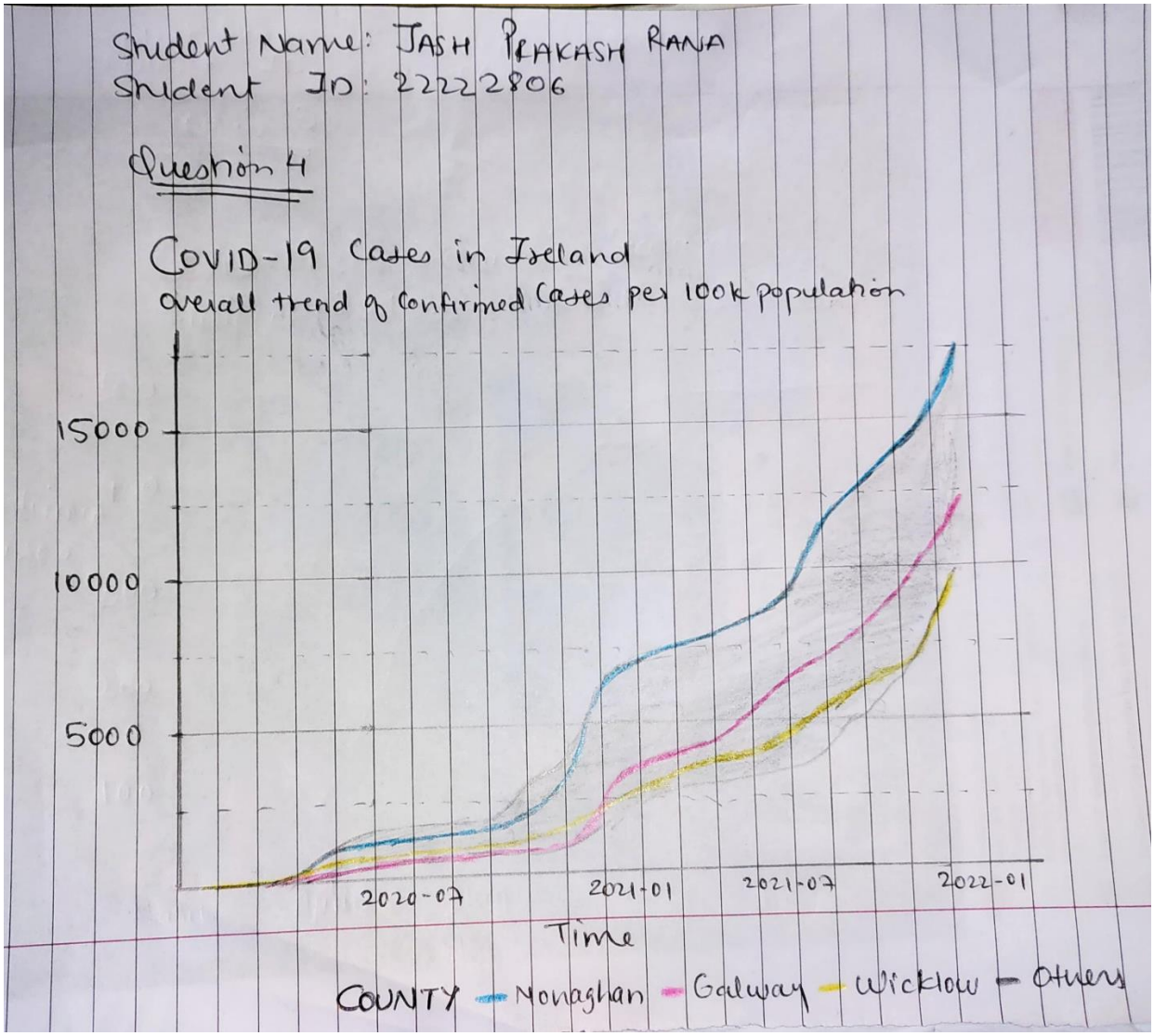


Figure 4: Line Chart showing the rise of COVID-19 cases for each county, over the period of time.

- ***Question 5:*** *A choropleth visualisation of the counties of Ireland showing total new confirmed cases (per 100,000) for a 4-week period (of your choice) for each county. The choropleth should show how each county diverges from the mean number of new confirmed cases (per 100,000) per county for that 4-week period.*

The four weeks intended to show are from September to December 2021. The data is filtered out using the `filter` method of the "dplyr" package. For this visualisation, the "GnBu" palette (Green-Blue) is the preferred palette from my perspective as I feel it shows a better aesthetic visual to the reader. An example of the palette is shown in figure 5.

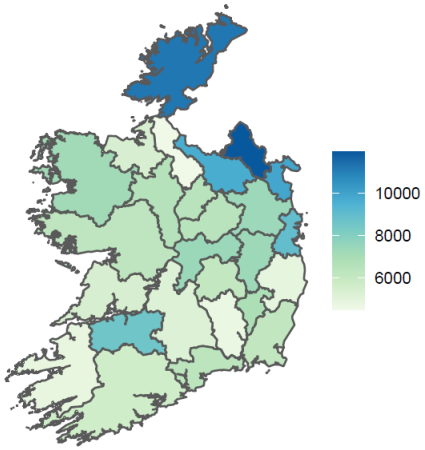


Figure 5: Example of "GnBu" palette on Choropleth visual.

Assignment 2

Name: Jash Prakash Rana, ID: 22222806

2023-03-20

The COVID-19 data set shows us a data of time-series format for each county in Ireland and also the map co-ordinates in multi-polygon data type, starting right from Feb 27 2020 up to January 2022. The two main columns to focus is the Daily Confirmed Cases which tells us about the per day number of cases, and Confirmed Cases which adds the daily numbers to the past data making it incremental (see Question 4 Diagram for reference). This data set is used to compare the cases by the counties and find meaningful insights into the COVID-19 spread in Ireland and what parts were affected the most and least, mean cases, etc.

I have performed all the 5 visualisations although only 3 were asked to show. I was intruiged to make the other two as well and so apologies for that.

```
library(sf)
file <- "./data/CovidCountyStatisticsIreland_v2.shp"
df <- st_read(file, quiet = TRUE)
```

We are adding two new columns here, where one sees the DailyCCases and ConfirmedC columns converted into per 100,000 population numbers. This helps us to scale the population per 100k and also get the numbers to a smaller format, easy to compare and visualise.

```
library(lubridate)
library(dplyr)

df<- df%>%
  mutate(ConfirmedC_per_100k =  round(100000  * ConfirmedC/Population,1))%>%
  mutate(DailyCCase_per_100k =  round(100000  * DailyCCase/Population, 1))%>%
  group_by(TimeStamp)

# df <- subset(df, select = -c(DailyCCase, ConfirmedC))
df$TimeStamp <- ymd(df$TimeStamp)
```

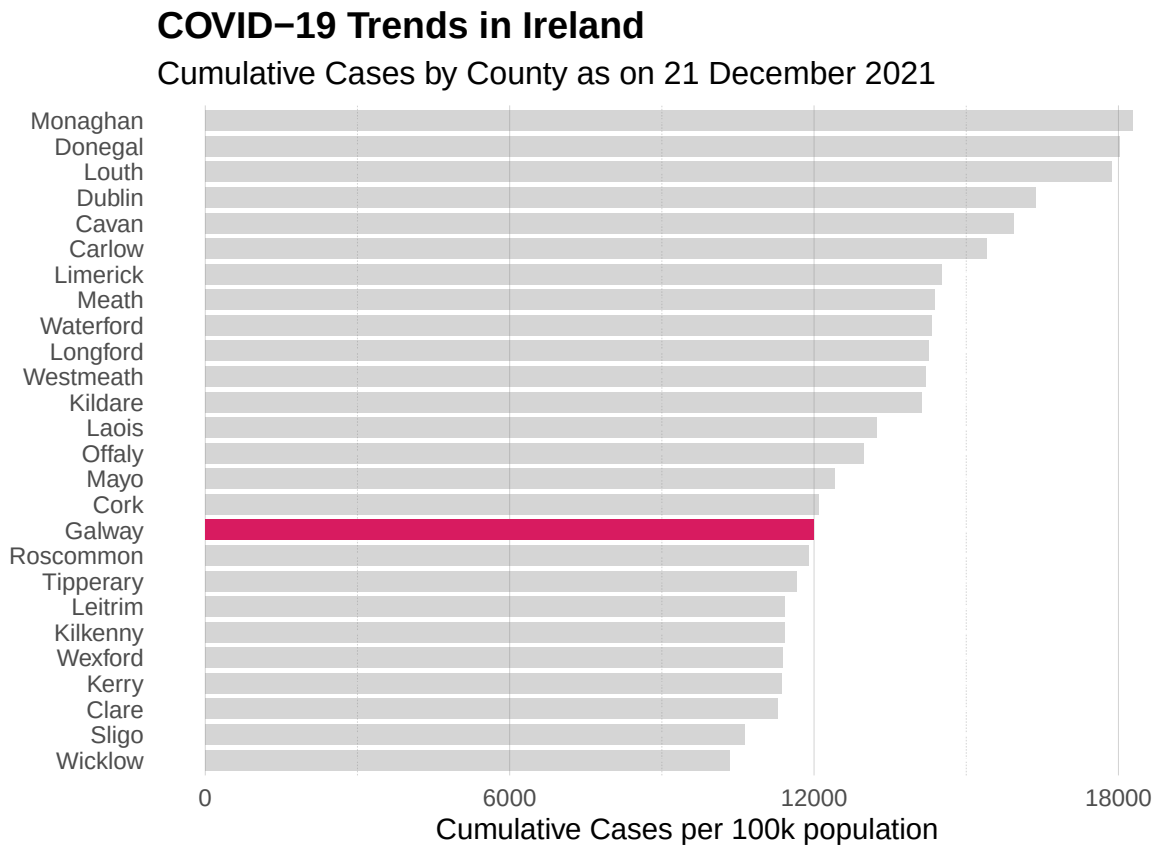
```
library(dplyr)

dec21_data <- df%>%
  dplyr::filter(TimeStamp == ymd("2021-12-21"))%>%
  group_by(TimeStamp)

dec21_data <- subset(dec21_data, select = -c(DailyCCase, ConfirmedC))
```

Part 1: A visualisation that allows the reader to accurately compare the cumulative number of cases per 100,000 of population per county on the 21 December 2021. County Galway should be highlighted.**

The best way I feel to compare the cumulative number of cases per 100,000 of population per country on 21 Dec 2021 is to use a bar chart, descending order of the number of cases while highlighting Galway with a different color while showing other counties with grey and setting alpha lower for them. The Galway highlight clearly shows its position in the number of cases which is 10 above from the lowest case, and is visible due to techniques used to foreground it from the rest. A bar chart accurately shows us the position of each counties in the order of cases, and the length of the bar shows the impact of the cases at which COVID-19 was operating in each county. Only the X-axis here has the grid lines as to show which number is the bar crossing or near to it, highlighting the numbers on the scale at each given point on the graph. The design is way similar to Part 1 as I do not think that any change is required which shows better visuals and non-technical understanding than what you saw in Part 1.

```
#Question 1
library(ggplot2)

ggplot(data = dec21_data, aes(x = reorder(CountyName, ConfirmedC_per_100k),
                              y = ConfirmedC_per_100k, fill = CountyName))+
  geom_bar(stat = "identity", width = 0.8,
           aes(group = ConfirmedC_per_100k,
               fill = CountyName,
               alpha = CountyName == "Galway")) +
  scale_fill_manual(values = c("Galway" = "#D81B60", Others = "lightgrey"),
                                guide = "none",
                    aes(group = ConfirmedC_per_100k)) +
  scale_alpha_manual(values = c(0.32, 1), guide = "none")+
  labs(title = "COVID-19 Trends in Ireland",
      subtitle = "Cumulative Cases by County as on 21 December 2021")+
  xlab(element_blank()) +
  ylab("Cumulative Cases per 100k population") +
  scale_y_continuous(breaks = seq(0,19000, by = 6000), limits = c(0,19000))+
  coord_flip(clip = "off")+

  theme(panel.background = element_rect(fill = "white"),
    panel.grid.major.x = element_line(size = 0.12, linetype = "solid",
    colour = 'lightgrey'),
    panel.grid.minor.x = element_line(size = 0.068, linetype = "dashed",
    colour = 'lightgrey'),
    axis.ticks = element_blank(),
    axis.text = element_text(size = 9),
    axis.title = element_text(size = 11),
    axis.title.y = element_text(size = 11,
    angle = 0,
    vjust=0.5,
    hjust=0,
    margin = margin(r=4)),
    legend.title = element_text(face="italic", size=11),
    legend.text = element_text(face="italic", size = 9),
    legend.box.background = element_rect(fill = 'lightgrey',
                                         colour = 'lightgrey'),
    legend.key = element_rect(fill = "white"),
    legend.position = "bottom",
```

```
    legend.direction = "horizontal",
    legend.justification = "center",
    legend.box.margin = margin(0.5, 0.5, 0.5, 0.5),
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(size = 12),
    panel.border = element_blank(),
    panel.spacing = unit(1, "lines"))
```



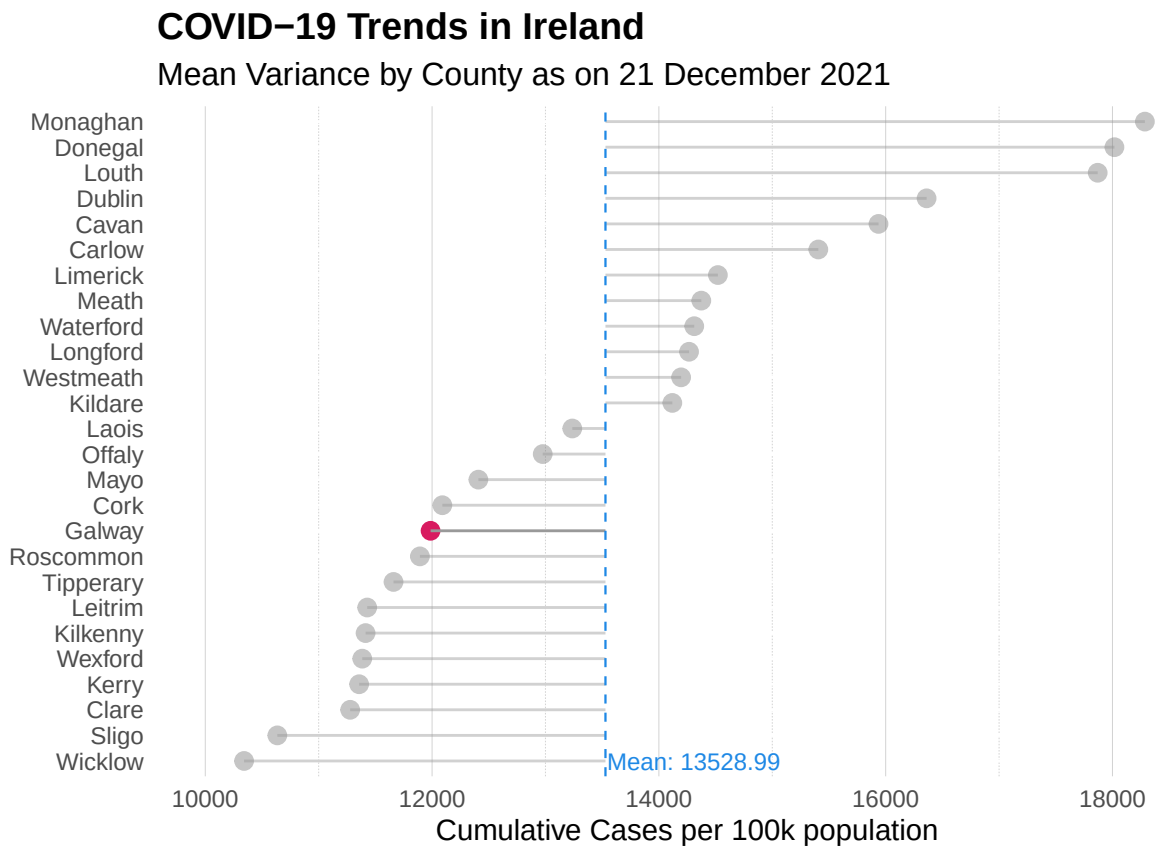
Part 2: A visualisation that allows the reader to read how each county diverges from from the mean cumulative number of cases (per 100,000) in the country as at the 21 December 2021. You may also use a daily figure in this section. County Galway should be highlighted.

As show in Part 1, a diverging lollipop chart can be one way to show a diverging data from the mean in a very less cluttered way. Although the points are not the best depiction of the exact number in the scales, it still gives a better position of the points from the mean, showing us how the distribution in the data is actually helpful and what is the best course of action ahead. The points on the graph are connected to the mean line (mean is 13,528.99) using a line drawn using 'geom-segment' and show us how much divergence is there from the mean. Galway is again highlighted using the same color and foregrounded with higher alpha value while the others are backgrounded with grey color and less alpha value. No change was observed from Part 1 again in this case.

```r
mean <- mean(dec21_data$ConfirmedC_per_100k)

mean_cases <- dec21_data%>%
  mutate(values = ConfirmedC_per_100k - mean)
```

```r
library(ggplot2)

ggplot(mean_cases, aes(x = reorder(CountyName, ConfirmedC_per_100k),
                       y = ConfirmedC_per_100k,
                       colour = CountyName,
                        fill = CountyName,
                        alpha = CountyName == "Galway"))+
  geom_point(stat = "identity", size = 2.8) +
  geom_segment(aes(yend = mean,
                   xend = CountyName),
               color = "#999999")+
  scale_fill_manual(values = c("Galway" = "#D81B60", Others = "lightgrey"),
                               guide = "none",
                    aes(group = ConfirmedC_per_100k)) +
  scale_colour_manual(values = c("Galway" = "#D81B60", Others = "lightgrey"),
                               guide = "none",
                    aes(group = ConfirmedC_per_100k))+
  scale_alpha_manual(values = c(0.45, 1), guide = "none")+
  labs(title = "COVID-19 Trends in Ireland",
      subtitle = "Mean Variance by County as on 21 December 2021")+
  xlab(element_blank()) +
  ylab("Cumulative Cases per 100k population") +
  scale_y_continuous(breaks = seq(10000,18500, by = 2000),
                     limits = c(10000,18500))+
  coord_flip(clip = "off")+
  geom_hline(yintercept = mean, size = 0.4, linetype = "dashed",
             colour = "#1E88E5")+
  annotate("text", x=1, y=mean+780, label=sprintf("Mean: %#.2f", mean),
           size = 3.2, colour = "#1E88E5")+

  theme(panel.background = element_rect(fill = "white"),
    panel.grid.major.x = element_line(size = 0.12, linetype = "solid",
    colour = 'lightgrey'),
    panel.grid.minor.x = element_line(size = 0.068, linetype = "dashed",
    colour = 'lightgrey'),
    axis.ticks = element_blank(),
    axis.text = element_text(size = 9),
    axis.title = element_text(size = 11),
    axis.title.y = element_text(size = 11,
    angle = 0,
    vjust=0.5,
    hjust=0,
    margin = margin(r=4)),
    legend.title = element_text(face="italic", size=11),
    legend.text = element_text(face="italic", size = 9),
    legend.box.background = element_rect(fill = 'lightgrey',
                                         colour = 'lightgrey'),
    legend.key = element_rect(fill = "white"),
```

```
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.justification = "center",
    legend.box.margin = margin(0.5, 0.5, 0.5, 0.5),
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(size = 12),
    panel.border = element_blank(),
    panel.spacing = unit(1, "lines"))
```

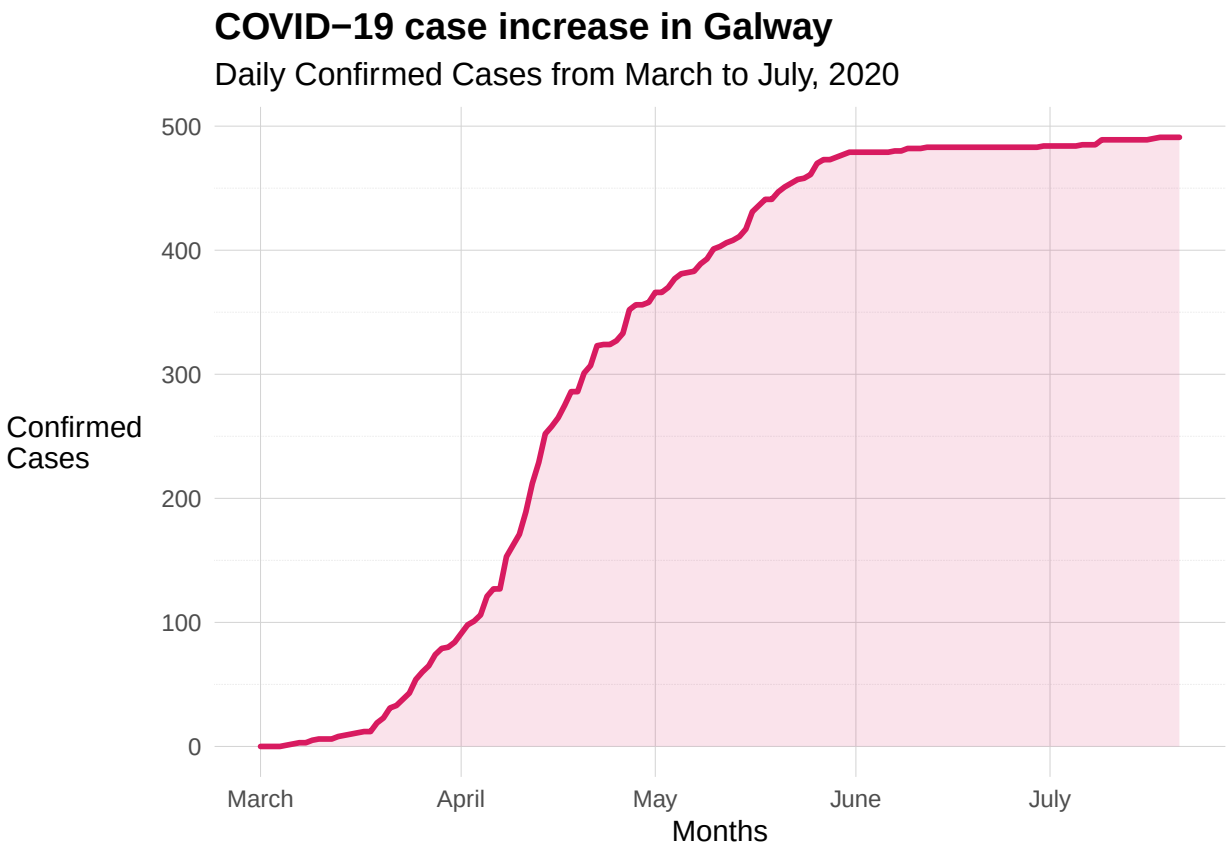


Part 3: A visualisation showing the daily number of confirmed covid cases in one county in Ireland for a 18-week period. This visualisation should help the reader to perceive the trend in the data.

A `ridgeline` or 'line with a fill' plot is the best way to show an impact with an area filled with a lighter shade. We show the first 18 weeks of COVID-19 data in Galway where we can see the rise in Confirmed Cases for the span of the duration. The sudden rise in cases is shown with the lines, and the area impact is show with the fill under the line giving us a feel of the magnitude of the event. We see that within first 4 months itself the numbers rose upto 500 patients which is a very useful insight on how this disease spreads along the airborne way. On the X-axis, we only show the major grid lines as to show months on the scale, while we use all the grid lines on Y-axis to show the numerical data. Its a very simplistic graph showing a very simple data trend over the time.

```r
#Question 3
dates <- seq(ymd("2020-03-1"),ymd("2020-7-21"), by="days")
galway_dailyCCase <- df%>%
  dplyr::filter(CountyName == "Galway")%>%
  dplyr::filter(TimeStamp %in% dates) %>%
  group_by(TimeStamp)
```

```r
library(ggplot2)
library(ggridges)
library(scales)

ggplot(galway_dailyCCase, aes(x = TimeStamp, y = ConfirmedC))+
    geom_ridgeline(aes(height = ConfirmedC, y=0),
                   color = "#D81B60",
                   fill = "#D81B6020",
                   size = 1)+
  scale_x_date(name = "Months", breaks = "1 month", labels=date_format("%B")) +
  labs(title = "COVID-19 case increase in Galway",
       subtitle = "Daily Confirmed Cases from March to July, 2020")+
  ylab("Confirmed \nCases") +


   theme(panel.background = element_rect(fill = "white"),
    panel.grid.major.x = element_line(size = 0.1, linetype = "solid",
                                      colour = 'lightgrey'),
    panel.grid.major.y = element_line(size = 0.12, linetype = "solid",
                                      colour = 'lightgrey'),
    panel.grid.minor.y = element_line(size = 0.068, linetype = "dotted",
                                      colour = 'lightgrey'),
    axis.ticks = element_blank(),
    axis.text = element_text(size = 9),
    axis.title = element_text(size = 11),
    axis.title.y = element_text(size = 11,
    angle = 0,
    vjust=0.5,
    hjust=0,
    margin = margin(r=4)),
    legend.title = element_text(face="italic", size=11),
    legend.text = element_text(face="italic", size = 9),
    legend.box.background = element_rect(fill = 'lightgrey',
                                         colour = 'lightgrey'),
    legend.key = element_rect(fill = "white"),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.justification = "center",
    legend.box.margin = margin(0.5, 0.5, 0.5, 0.5),
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(size = 12),
    panel.border = element_blank(),
    panel.spacing = unit(1, "lines"))
```

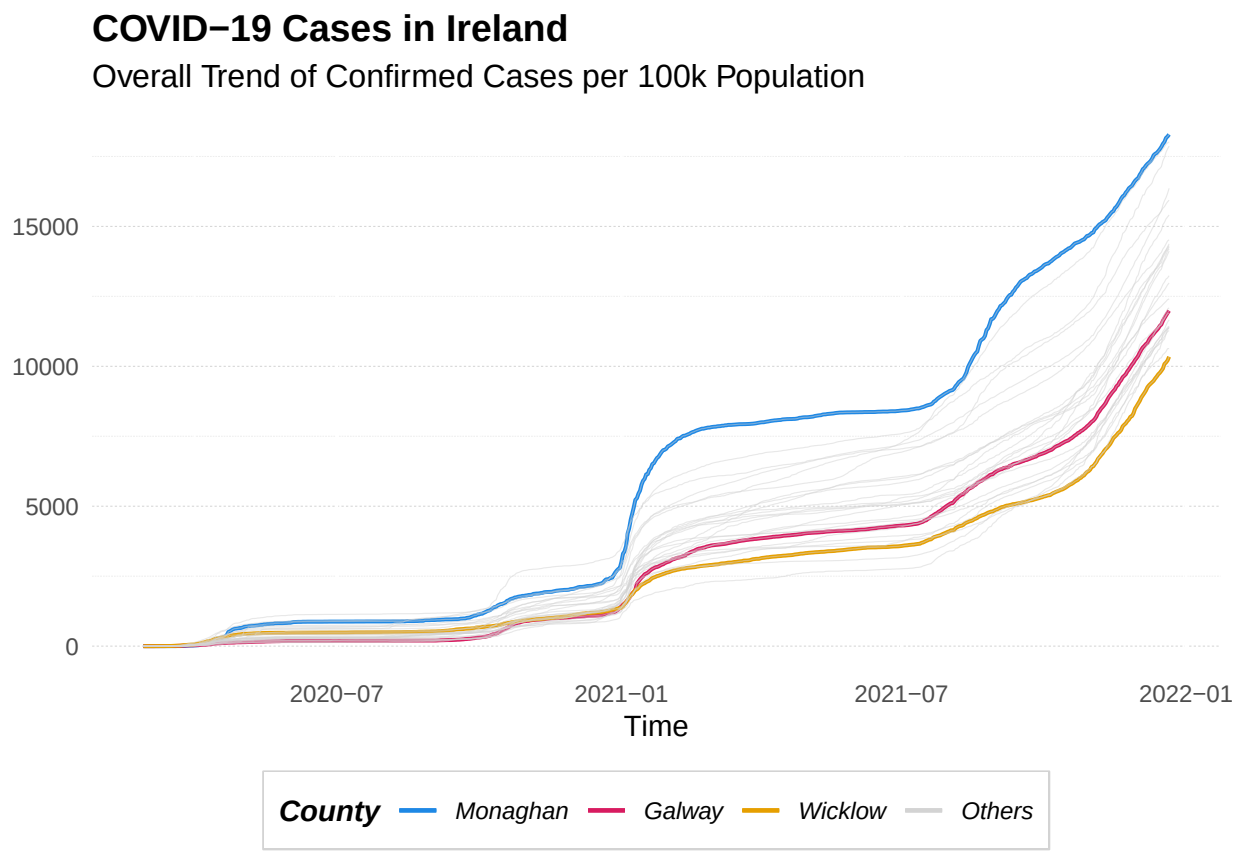


Part 4: A visualisation that highlights the cumulative number of cases per 100,000 in Galway and two other counties representing counties that have had the lowest and highest number of cases per 100,000 over the full timeline of the dataset. The visualisaton must also show the cumulative case number for all other counties in Ireland in the same plot. However, the three selected counties (Galway and two other counties) must be highlighted)

This is a very tricky form of question as a data for so many counties can result in cluttered visual, and the same has happened in the actual representation. Even though we use line chart for this solution, we see that the background data creates many lines in visuals, but its good to note that our highlighted counties are visible clearly, which is good to go for us. We considered Monagrah, Galway & Wicklow to be the highest, our county and lowest cases respectively, all highlighted in different colors as shown in the legend. The trend shows the confirmed cases per 100k population for the whole duration of the time. This is again a very minimal visual in which X-axis has no grid lines to show the flow of time in linear format while Y-axis has both the grid lines to show the cases.

```
#Question 4
library(ggplot2)
Counties <- c("Monaghan", "Galway", "Wicklow", "Others")
colours <- c("#1E88E5", "#D81B60", "#E69F00", "lightgrey")

ggplot(df, aes(x = TimeStamp, y = ConfirmedC_per_100k, group = CountyName))+
  geom_line(data = df%>%filter(CountyName == Counties[1]),
            size = 0.7, alpha = 1, aes(color = "#1E88E5"))+
```

```
  geom_line(data = df%>%filter(CountyName == Counties[2]),
            size = 0.7, alpha = 1, aes(color = "#D81B60"))+
  geom_line(data = df%>%filter(CountyName == Counties[3]),
            size = 0.7, alpha = 1, aes(color = "#E69F00"))+
  geom_line(data = df%>%filter(CountyName != Counties),
            size = 0.2, alpha = 0.52, aes(color = "lightgrey"))+

  labs(title = "COVID-19 Cases in Ireland",
      subtitle = "Overall Trend of Confirmed Cases per 100k Population")+
  xlab("Time") +
  ylab(element_blank()) +
  scale_color_manual(values = colours, name = "County", labels = Counties)+

  theme(panel.background = element_rect(fill = "white"),
    panel.grid.major.y = element_line(size = 0.1, linetype = "dashed",
                                      colour = 'lightgrey'),
    panel.grid.minor.y = element_line(size = 0.068, linetype = "dotted",
                                      colour = 'lightgrey'),
    axis.ticks = element_blank(),
    axis.text = element_text(size = 9),
    axis.title = element_text(size = 11),
    axis.title.y = element_text(size = 11,
    angle = 0,
    vjust=0.5,
    hjust=0,
    margin = margin(r=4)),
    legend.title = element_text(face="bold.italic", size=11),
    legend.text = element_text(face="italic", size = 9),
    legend.box.background = element_rect(fill = 'lightgrey',
                                         colour = 'lightgrey'),
    legend.key = element_rect(fill = "white"),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.justification = "center",
    legend.box.margin = margin(0.5, 0.5, 0.5, 0.5),
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(size = 12),
    panel.border = element_blank(),
    panel.spacing = unit(1, "lines"))
```

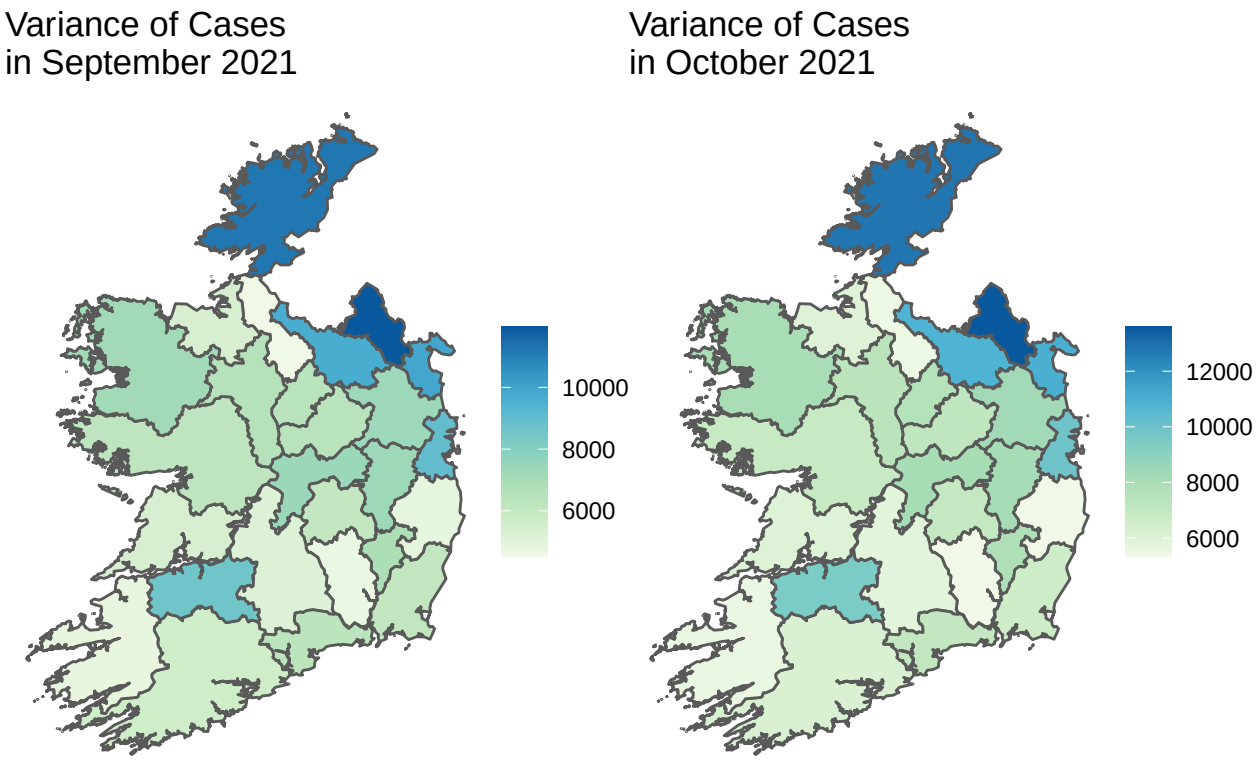


Part 5: A choropleth visualisation of the counties of Ireland showing total new confirmed cases (per 100,000) for a 4-week period (of your choice) for each county. The choropleth should show how each county diverges from the mean number of new confirmed cases (per 100,000) per county for that 4-week period.

A choropleth visualisation requires the data of latitudes and longitudes in a multi-polygon format which then takes shape in a map region. Here we show the map of Ireland and the confirmed cases in each county for the months of September, October, November and December of 2021 to show how varied the numbers were for each month. We use "GnBu" or the `Green Blue` palette where lightest green shows the lowest case while darkest blue shows the highest number in the legend. The visualisation is appealing as it shows exactly how each county was affected when compared to another. We also used theme_void() which doesn't have gridlines to start off with the theme and we do not use much to design the aesthetics of the plot. I also used the grid.arrange to show the plots in 2 columns, and I tried showing it in one plot but it looks way smaller so it was better to split the visuals in two different parts showing 4 maps respectively.

```
library(ggplot2)
library(gridExtra)

p1 <- ggplot(df%>%filter(TimeStamp == ym("2021-09"))) +
  geom_sf(aes(fill = ConfirmedC_per_100k)) +
  ggtitle("Variance of Cases \nin September 2021") +
  scale_fill_distiller(palette = "GnBu", direction =1)+
  theme_void() +
```

```
  theme(legend.title = element_blank())

p2 <- ggplot(df%>%filter(TimeStamp == ym("2021-10"))) +
  geom_sf(aes(fill = ConfirmedC_per_100k)) +
  ggtitle("Variance of Cases \nin October 2021") +
  scale_fill_distiller(palette = "GnBu", direction =1)+

  theme_void() +
  theme(legend.title = element_blank())

grid.arrange(p1,p2, ncol = 2)
```




```
p3 <- ggplot(df%>%filter(TimeStamp == ym("2021-11"))) +
  geom_sf(aes(fill = ConfirmedC_per_100k)) +
  ggtitle("Variance of Cases \nin November 2021") +
  scale_fill_distiller(palette = "GnBu", direction =1)+
  theme_void() +
  theme(legend.title = element_blank())

p4 <- ggplot(df%>%filter(TimeStamp == ym("2021-12"))) +
  geom_sf(aes(fill = ConfirmedC_per_100k)) +
  ggtitle("Variance of Cases \nin December 2021") +
  scale_fill_distiller(palette = "GnBu", direction =1)+
  theme_void() +
  theme(legend.title = element_blank())
```

```
grid.arrange(p3,p4, ncol = 2)
```