CT5100 Data Visualisation: Assignment 2

Jiarong Li 20230033, 1MAI1

The School of Computer Science
University of Galway
j.li11@nuigalway.ie

1 Introduction

In this assignment you are required to one visualisation for each question from this data.

- 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.
- 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.
- 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 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.
- 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.

2 Part 1

2.1 For question 1 to 4, briefly justify the visualisation approach you will use. If calculations are required, clearly say what they are. Using coloured pencils* make a sketch of the approach. This will be your design for part 2. Do not include ggplot visualisations in part 1.

- Question 1.

  The goal of question 1 is to allow the readers to do the comparison accurately. In addition, the difference in values is small. Hence, I use a dot plot here and the position of each dot shows the cumulative number of cases per 100, 000 of the population of the corresponding county on 21 December 2021. I subset the sf object with time $stamp == 2021 - 12 - 21$. I create a new column named "ConfirmedC100000" representing the cumulative number of cases per 100, 000 of population per county on 21 December 2021 by the calculation of $ConfirmedC/Population * 1000000$ and reorder the data set based on the new column in descending order. I highlight Galway with black colour and other counties with grey colour.

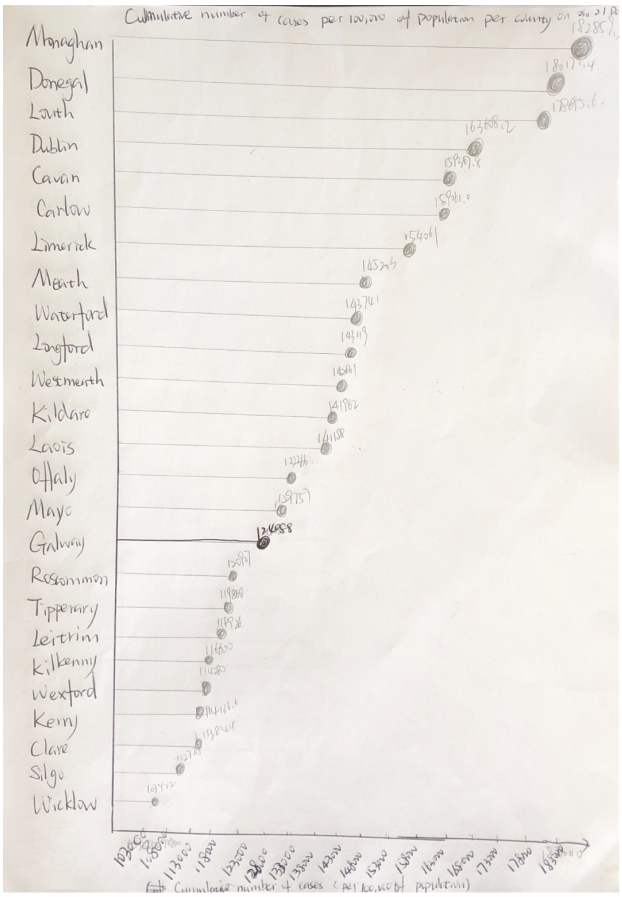

Figure 1: Cumulative Number of Covid Cases per County of Ireland.

- Question 2.

  In order to allow the reader to read how each county diverges from from the mean cumulative number of cases. I calculate the average value of the values of the created column – "ConfirmedC100000"– by coding "$summarize(AvgTMean = mean(ConfirmedC1000000, na.rm = TRUE))$" based on the dataframe I made in Question 1.

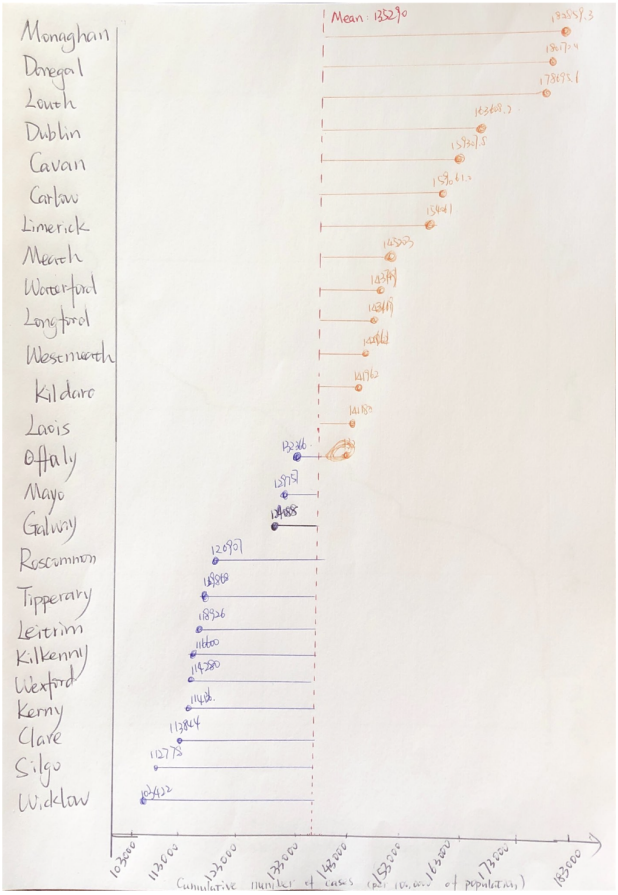

Figure 2: Divergence from the mean cumulative number of covid cases per county of Ireland.

- Question 3.

  I choose Carlow to visualise the daily number of confirmed covid cases for an 18-week period. I create a subset named "carlow_daily" by selecting Carlow and the corresponding confirmed daily cases from the dataframe. Then calculate the number of indexes that need to be selected for a period of 18-week by $18 * 7$. Then I select the corresponding rows by coding "$carlow_daily[0:126,]$".

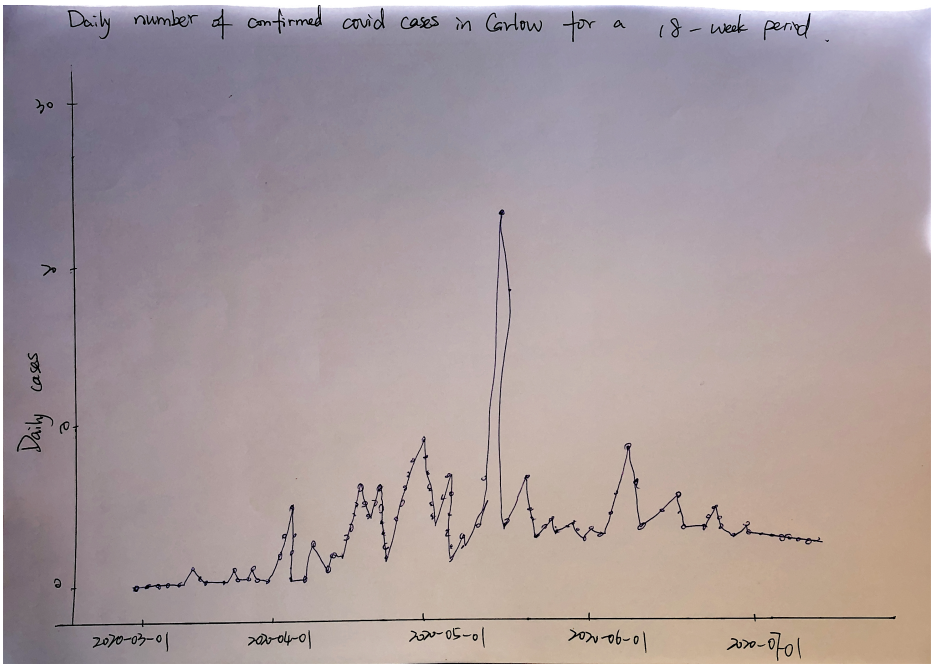

Figure 3: The daily number of confirmed covid cases in Carlow for a 18-week period.

- Question 4.

  In order to visualise the cumulative number of covid cases per county over the full timeline, I am using a time series line plot. Based on the data frame created in question 1, I found the county (Monaghan) representing the highest number of cases and the county (Wickklow) representing the lowest number of cases. Then I highlight these three counties (Galway, Monaghan and Wickklow) with blue, red and black colours, and colour the other counties with grey colour.

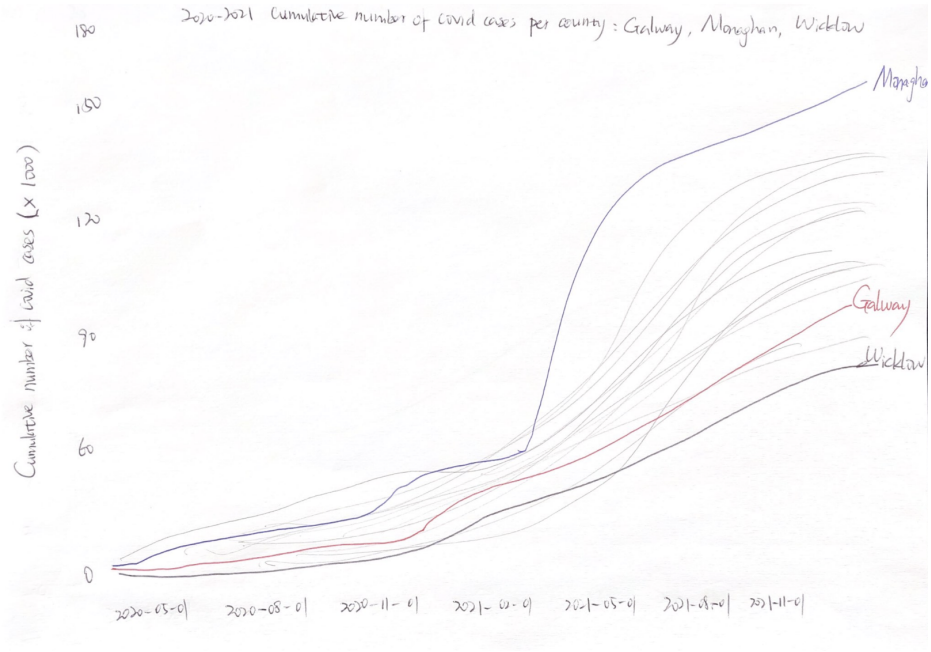

Figure 4: 2020-2021 Cumulative number of covid cases per county: Galway, Monaghan and Wickklow.

2.2 For question 5, indicate the 4-week period you intend to show. If calculations are required, clearly say what they are. Explain the type of palette you will use and give an example of such a palette.

I choose the 4-week period from 2020.02.27 to 2020.03.30. I create a subset of this period and select the converted cumulative confirmed cases (per 100, 000) on 2020.03.30. I calculate the mean of the values of the column by summing up the values of this column and call the mean function by coding "$mean(date_4weeks_Confirmed_C100000\$Confirmed_C100000)$". In order to show how each county diverges from the mean number of new confirmed cases for the 4-week period, a diverging palette which puts equal emphasis on mid-range values with light colours and extremes at both ends of the data range with contrasting dark colours is suitable in this case. I use a diverging palette from the RColorBrewer library and to void using red and green as the dark colour at the end I use '$scale_fill_continuous_diverging(palette = "Purple - Green")$'.

3 Part 2

3.1 Using the ggplot2 visualisation library, produce a visualisation of three of the five of the visualisation questions. Show your code including any code required for calculations.

- Question 3.

  I choose Carlow to visualise the daily number of confirmed covid cases for an 18-week period. I create a subset named "carlow_daily" by selecting Carlow and the corresponding

confirmed daily cases from the dataframe. Then calculate the number of indexes that need to be selected for a period of 18-week by $18 * 7$. Then I select the corresponding rows by coding "$carlow_daily[0:126,]$".

```
# read the shape file
library(sf)
library(dplyr)
library(readr)
library(lubridate)
library(ggplot2)
library(scales)


file1 <- "CovidCountyStatisticsIreland_v2.shp"
IRL_counties_SF <- st_read(file1, quiet = TRUE, as_tibble = TRUE)

# Subset the sf object with time stamp == 2021-12-21
IRL_counties_sub_time <- IRL_counties_SF[IRL_counties_SF$TimeStamp ==
↪  "2021-12-21", ]


# create a column which stores the ConfirmedC per 100,000 per county
IRL_counties_sub_per = IRL_counties_sub_time %>%
  dplyr::mutate(ConfirmedC100000 = ConfirmedC/Population * 100000) %>%
  arrange(-ConfirmedC100000)

# Subset the sf object with time stamp == 2021-12-21
IRL_counties_sub_carlow <- IRL_counties_SF[IRL_counties_SF$CountyName
↪  == "Carlow", ]

carlow_daily <- IRL_counties_sub_carlow %>%
 select(TimeStamp, DailyCCase)

# 18*7
carlow_daily_18weeks <- carlow_daily[0:126, ]

# plot
ggplot(carlow_daily_18weeks, aes(TimeStamp, DailyCCase)) +
  geom_line(color = "#0072B2") +
 geom_point(color = "white",fill = "#0072B2",
            shape = 21, size = 2) +
 scale_y_continuous(limits = c(0, 30),
                    name = "Daily cases") +
 scale_x_date(name = "month", breaks = "1 month",
              #labels=date_format("%Y")
```

```
                ) +
  ggtitle("Daily number of confirmed covid cases in Carlow for a
   ↪  18-week period.")+
  theme_classic() +
  theme(plot.margin = margin(7,7,3,1.5),
        axis.title.x=element_blank())
```

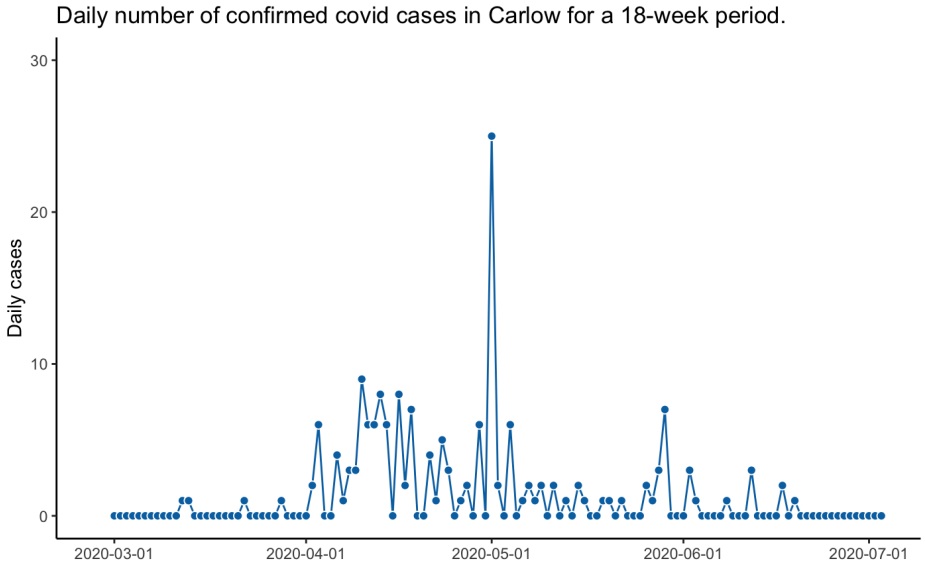

Figure 5: The daily number of confirmed covid cases in Carlow for a 18-week period.

- Question 4.

  In order to visualise the cumulative number of covid cases per county over the full timeline, I am using a time series line plot. Based on the data frame created in question 1, I found the county (Monaghan) representing the highest number of cases and the county (Wickklow) representing the lowest number of cases. Then I highlight these three counties (Galway, Monaghan and Wickklow) with blue, red and black colours, and colour the other counties with grey colour.

```
# read the shape file
library(sf)
library(ggplot2)
library(ggridges)
library(lubridate)
library(ggrepel)
library(knitr)

file1 <- "CovidCountyStatisticsIreland_v2.shp"
IRL_counties_SF_2 <- st_read(file1, quiet = TRUE, as_tibble = TRUE)
 ↪  %>%
  dplyr::mutate(ConfirmedC100000 = ConfirmedC/Population * 100000) %>%
```

```
  arrange(-ConfirmedC100000)

# Plot background layer
p_counties <- ggplot(IRL_counties_SF_2,
                     aes(x =TimeStamp, y=ConfirmedC100000,
                         group = CountyName)) +
  geom_line(aes(group = CountyName),size= 0.35, na.rm = TRUE,
   ↪  color="grey90", alpha =0.7, show.legend = FALSE ) +
  scale_y_continuous(breaks=seq(0,186000, by = 30000),
                     name = "Cumulative number of cases (x 1000)",
                     labels = seq(0,186, by = 30)) +

  scale_x_date(name = "year", breaks = "3 month", expand=c(0,0) ) +
  theme_classic() +
  theme(panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank(),
        axis.title.x=element_blank(),
        axis.text.x = element_text(  vjust = .5),
        legend.key = element_rect(fill = NA, colour = NA, size = 0.25)
,
        plot.margin = margin(14, 14, 8, 14))

#foreground countries
target_counties <- c("Galway","Monaghan", "Wicklow")
#foreground data
counties_data_targets<- subset(IRL_counties_SF_2, CountyName %in%
 ↪  target_counties)

x_pos_label<- as.Date("2022-01-01")

# gives the final count value that you want to show on the right
 ↪  axis
counties_data_targets_final <- filter(counties_data_targets, TimeStamp
 ↪  == ymd("2021-12-21"))

# foreground layer
p_counties2 <- p_counties +

 geom_line(data=counties_data_targets, size =1, alpha=0.85,
  ↪  show.legend = TRUE, (aes(x =TimeStamp, y=ConfirmedC100000,
  ↪  colour= CountyName, group = CountyName))) +

```

```
48    geom_label(data=counties_data_targets_final,
      ↪  aes(x=x_pos_label,y=ConfirmedC100000, label=CountyName,
      ↪  color=CountyName), size=2.75, hjust=0, fill="white",label.size
      ↪  = NA) +
49
50    # note how the limits attribute allows us to specify the order of
      ↪  items in the legend
51    scale_colour_manual(values = c("green4","#D55E00", "#0072b2"),name
      ↪  = NULL, limits = c("Galway","Monaghan", "Wicklow")) +
52
53    ggtitle("2020-2021 cumulative number of covid cases per county (x
      ↪  1000) : Galway, Monaghan, United Wicklow") +
54
55    coord_cartesian(xlim=c(as.Date("2020-02-27"),as.Date("2021-12-21"
      ↪  )), clip = 'off') +
56
57    theme(
58       legend.position = "none",
59       axis.ticks.y.right = element_blank(),
60       axis.ticks.y = element_blank(),
61       axis.ticks.x = element_blank(),
62       axis.title.y= element_blank(),
63       axis.text.y.right = element_text(colour="black", size =8),
64       plot.margin=margin(r=70)
65    )
66
67    p_counties2
```

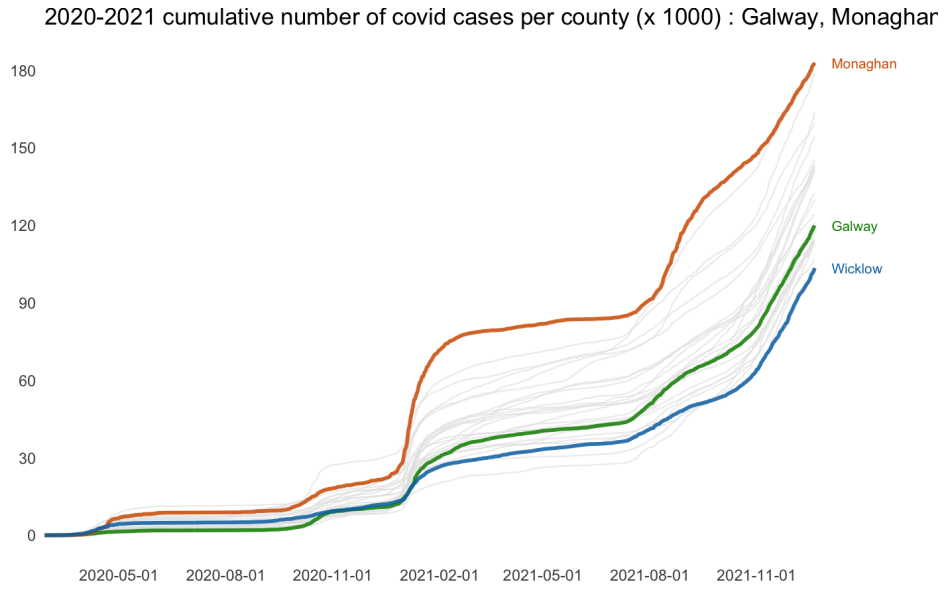

Figure 6: 2020-2021 Cumulative number of covid cases per county: Galway, Monaghan and Wickklow.

- Question 5.

  I choose the 4-week period from 2020.02.27 to 2020.03.30. I create a subset of this period and select the converted cumulative confirmed cases (per 100, 000) on 2020.03.30. I calculate the mean of the values of the column by summing up the values of this column and call the mean function by coding "$mean(date_4weeks_ConfirmedC100000\$ConfirmedC100000)$". In order to show how each county diverges from the mean number of new confirmed cases for the 4-week period, a diverging palette which puts equal emphasis on mid-range values with light colours and extremes at both ends of the data range with contrasting dark colours is suitable in this case. I use a diverging palette from the RColorBrewer library and to void using red and green as the dark colour at the end I use '$scale_fill_continuous_diverging(palette = "Purple - Green")$'.

```
# read the shape file
library(sf)
library(dplyr)
library(ggplot2)
library(RColorBrewer)
library(scales)
library(colorblindr)

file1 <- "CovidCountyStatisticsIreland_v2.shp"
IRL_counties_SF_2 <- st_read(file1, quiet = TRUE, as_tibble = TRUE) %>%
  dplyr::mutate(ConfirmedC100000 = ConfirmedC/Population * 100000) %>%
  arrange(-ConfirmedC100000)

IRL_counties_SF_daily <- IRL_counties_SF_2 %>%
  dplyr::mutate(daily100000 = DailyCCase/Population * 100000)

date_4weeks <- IRL_counties_SF_daily %>%
  filter(TimeStamp >= '2020-02-27' & TimeStamp <= '2020-03-30')

date_4weeks_ConfirmedC100000 <- date_4weeks %>%
  filter(TimeStamp == '2020-03-30')

date_4weeks_ConfirmedC100000_mean <-
  mean(date_4weeks_ConfirmedC100000$ConfirmedC100000)

date_4weeks_ConfirmedC100000$pos <-
  date_4weeks_ConfirmedC1000000$ConfirmedC1000000 -
  date_4weeks_ConfirmedC100000_mean


ggplot(date_4weeks_ConfirmedC100000) +
  geom_sf(aes(fill = pos),
```

```
            color = "darkgrey",
            linetype = 1,
            lwd = 0.4) +

  scale_fill_continuous_diverging(palette = "Purple-Green", l1 = 10,
  ↪  l2 = 100, p1 = .9, p2 = 1.2) +

  ggtitle("Total new confirmed cases (per 100,000) for a 4-week period
  ↪  for each county of Ireland.") +

  theme_void() +
  theme(legend.title = element_blank(),
     legend.text.align = 0.5,
     legend.justification = c(0, 0),
     legend.position = c(-0.2, 0.58))
```

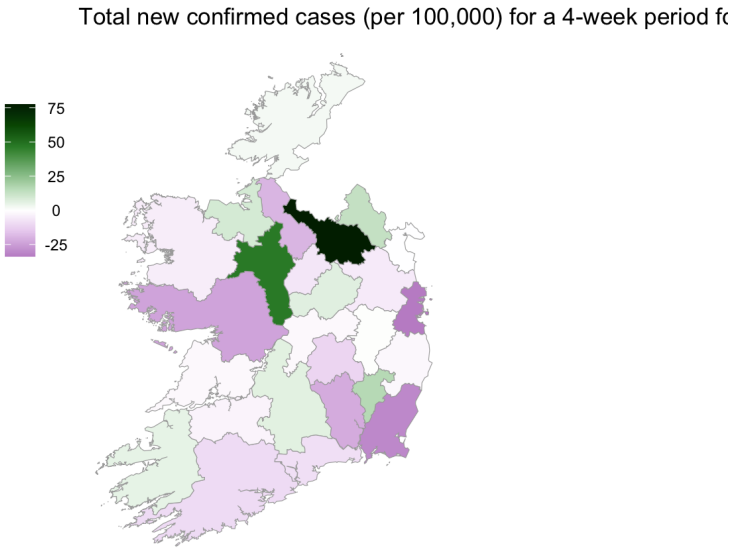

Figure 7: Total new confirmed cases (per 100,000) based on the average number of new cases for a 4-week period for each county of Ireland.