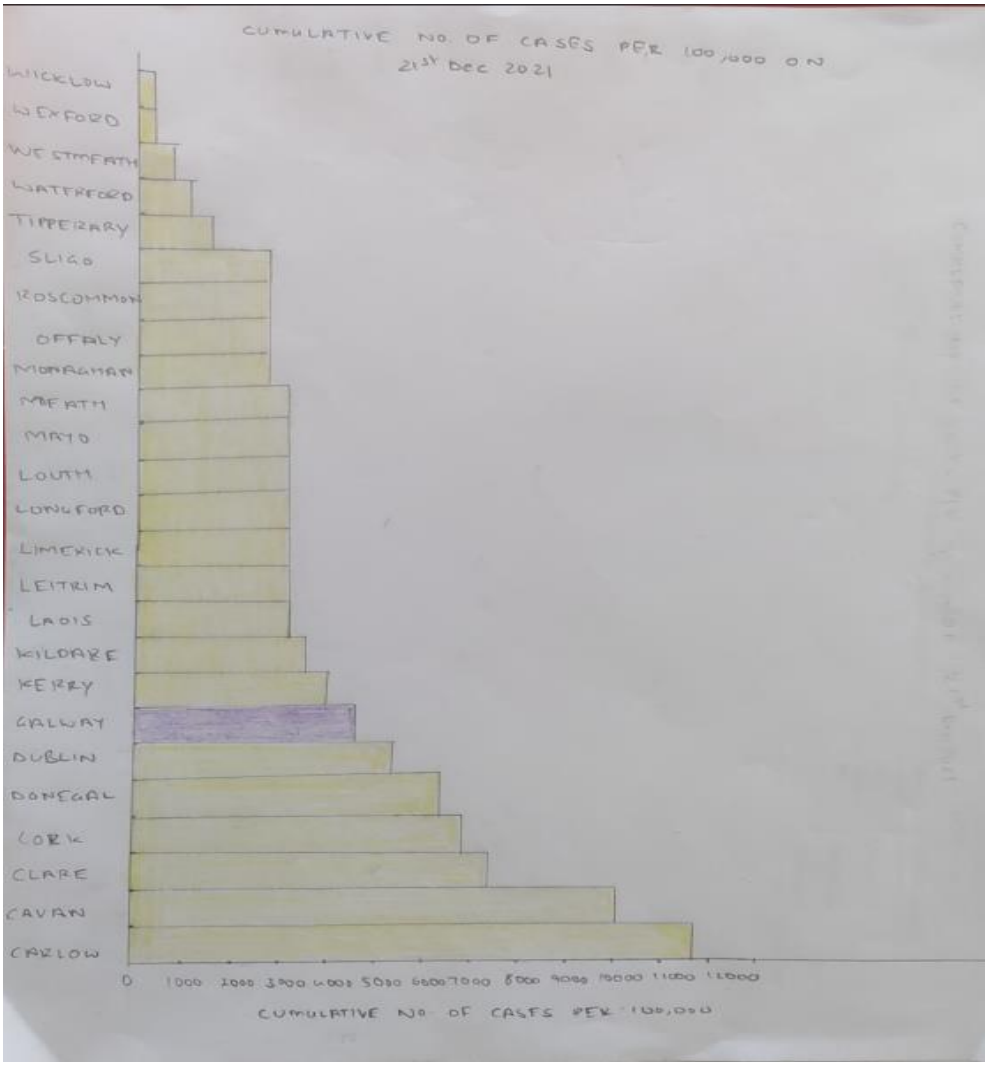
Assignment 2

1. A visualisation 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.

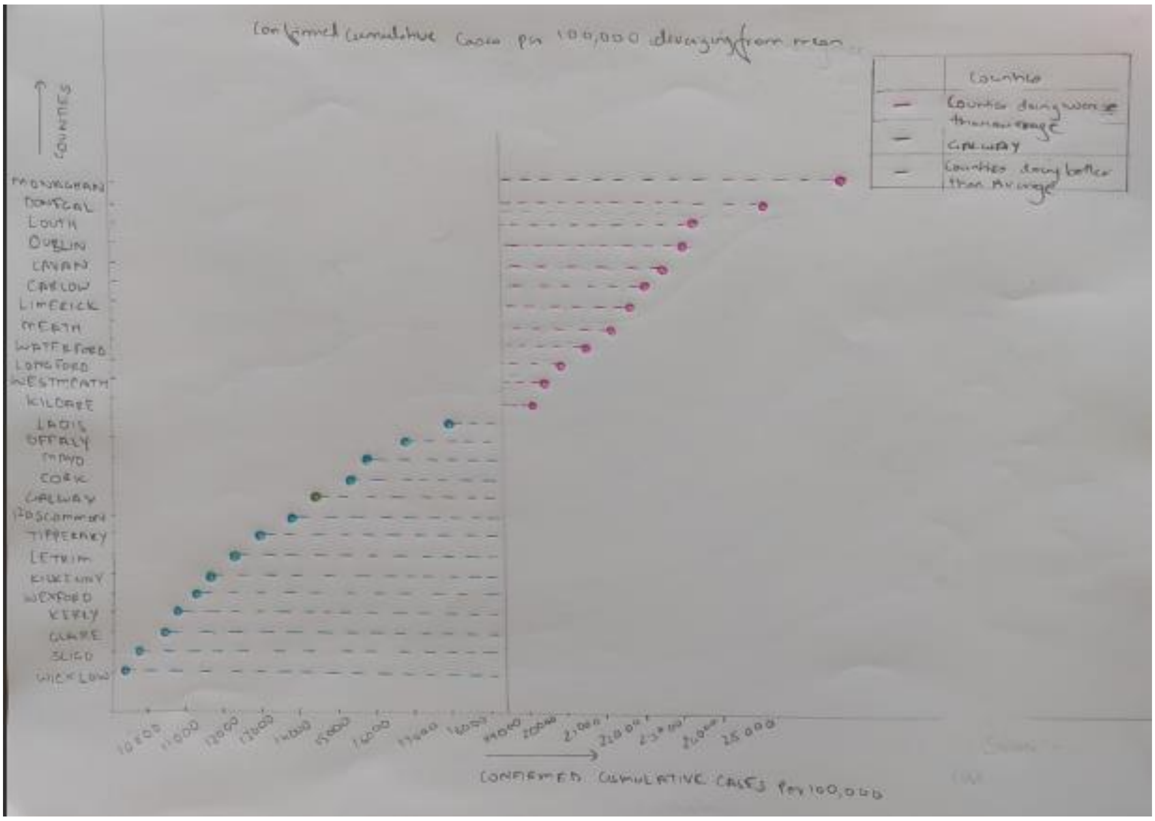
Answer: A simple bar plot with the bar for county Galway highlighted should work for this problem. As the data points to be plotted are for a single day, we can plot a bar plot sorted in order of the cumulative number of cases per 100,000 so that it is easy for the eye to follow the order with Galway highlighted.



Calculations Required: Data for date ’21-12-2021’ should be filtered and sorted by the value in column ‘ConfirmedC_per_100k’.

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.

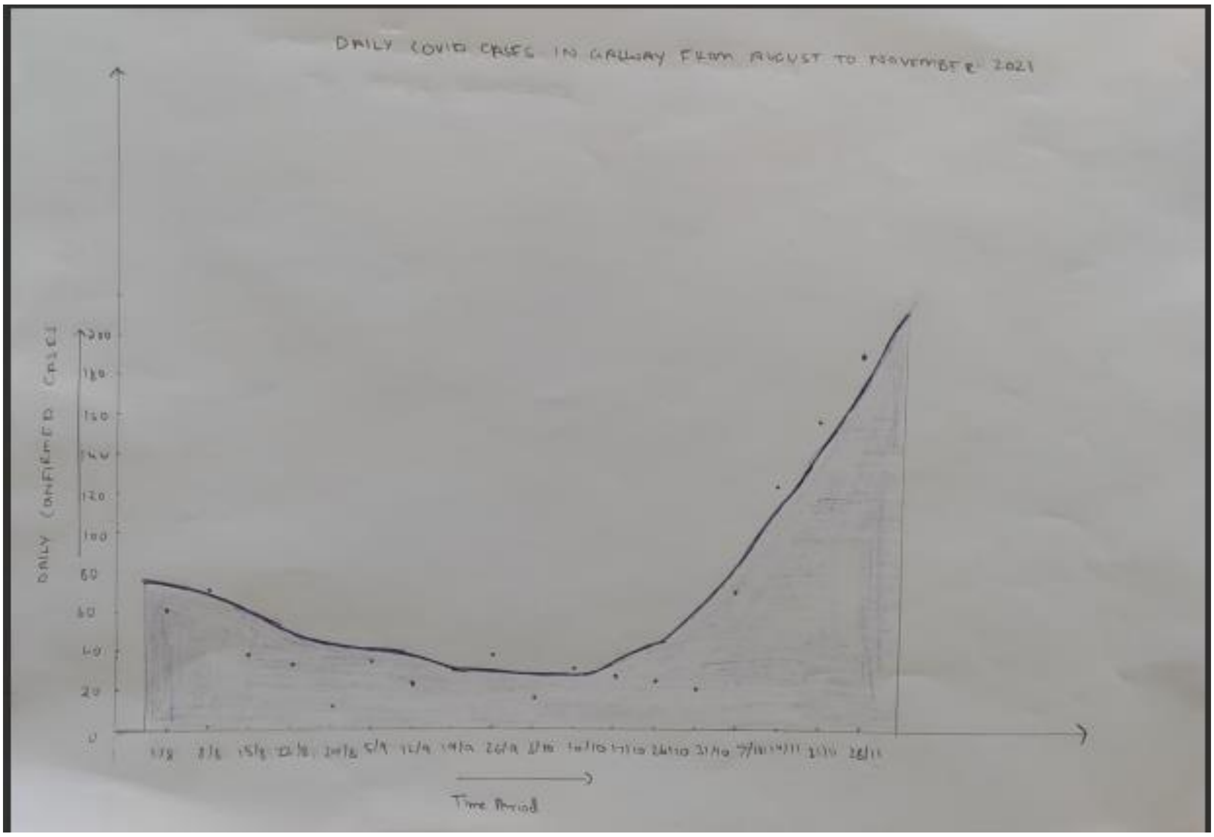
Answer: As we have to show how each county diverges from the average no. of cases in Ireland, we can use a diverging bar chart to showcase how much each county deviates from the mean value. A diverging chat usually has a color encoding for points above the mean and points below the mean. Hence, different colors would be used for counties for their divergence from the mean. Also, the County Galway would be highlighted in a different color.



Calculations Required: Same as previous code.

3. A visualisation showing the daily number of confirmed covid cases in one county in Ireland for 18 weeks. This visualisation should help the reader to perceive the trend in the data.

Answer: A trend line can be created using the geom_smooth() function with a filled area under the line. For time-series data, a trend line with the dots can show the trajectory of information changes with respect to time effectively. As the data showcased is cumulative i.e summed over time highlighting the area would give the idea of change in the volume of cases with respect to time.

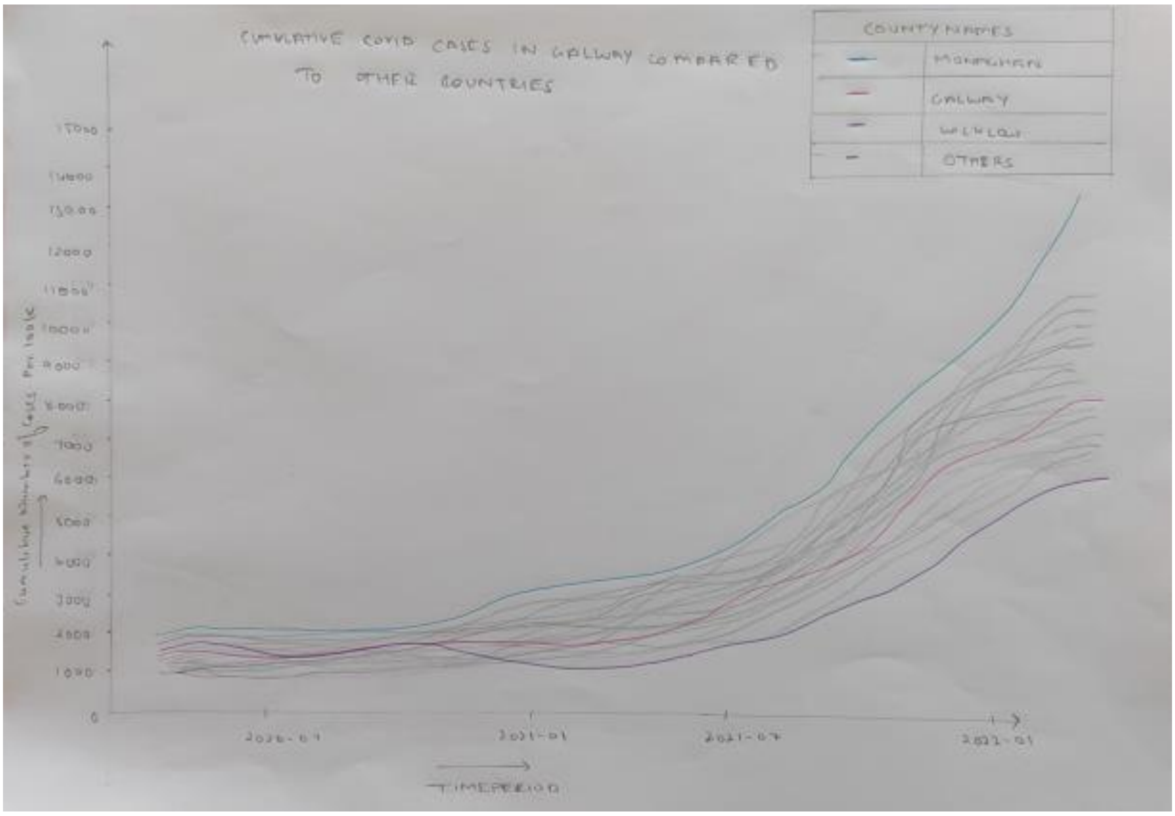


Calculation Required: One of the Counties data (Galway in this case) should bee filtered along with a time period of 18 weeks.

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)

Answer:

As per the problem statement, we have to plot tred off cumulative number of cases per 100,000 with three counties highlighted. We can plot trend lines for all the counties with respect to confirmed cases . However it would look messy and data would be indistinguishable for all the counties because there are a lot of them. We can highlight the three counties and faded the remaining ones for better visualisaton.



Calculation Required: We need to find the counties with highest and lowest cumulative cases, we can do that by finding the county with minimum and maximum values in cumulative column.

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.

Answer: To show how the number of new confirmed cases diverge over a period of 4 weeks we need to find the difference between cases occurring between that period, and then plot in on a cholorpeth. As the goal here is to show the number of cases of a disease, a warm sequential palette like ‘inferno’ pallete would be appropriate as the dark colors helps in conveying the spread of disease.

Calculations required: We need to calculate number of new cases along a period of 4 weeks and then plot the deviation from the mean of the values for all counties. The 4 week period for plotting would be from 2021-01-01 to 2021-02-01

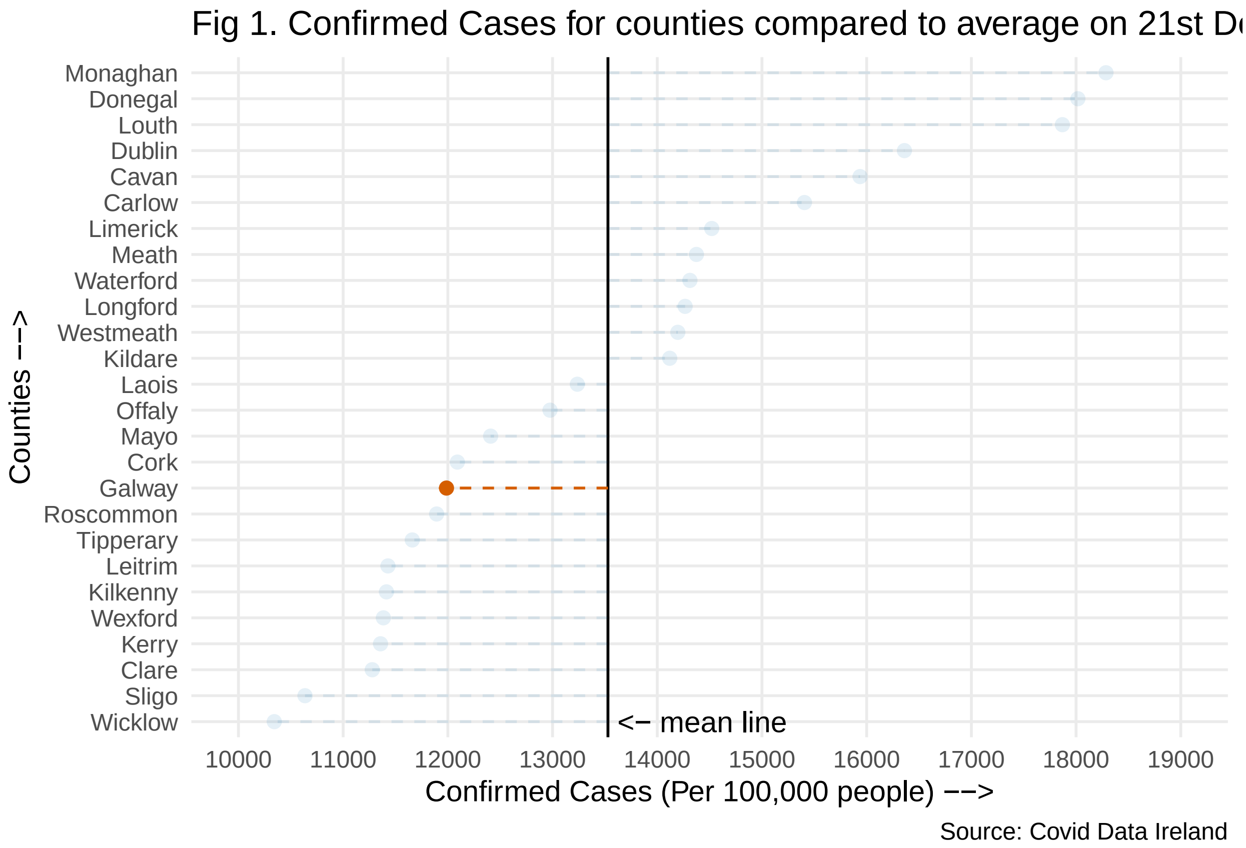
Assingment 2 - Part 2

Smitesh Patil

2023-03-06

Q2. 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

```
#loading colorblind colors
palette <- colorblindr::palette_OkabeIto

#normalising dailyccases and confirmdccases
IRL_Covid19_2021_12_21<- IRL_counties_Covid19%>%
  filter(TimeStamp == ymd("2021-12-21"))%>%
  mutate(ConfirmedC_per_100k =  round(100000  * ConfirmedC/Population,1))%>%
  mutate(DailyCCase_per_100k =  round(100000  * DailyCCase/Population, 1))

#getting the mean of confirmed cases per 100k for plotting
mean_daily_cases <- IRL_Covid19_2021_12_21 %>%
  select(ConfirmedC_per_100k) %>%
  st_drop_geometry() %>%
  unlist() %>%
  mean()

#for plotting the graph
IRL_Covid19_2021_12_21 %>%
  # creating color column for plotting color different for galay, less than mean and
  #more than mean
  mutate(color = ifelse(CountyName == "Galway", "1", "2")) %>%
  #loading aesthetics for the graphy reordering for sorting
  ggplot(aes(x = ConfirmedC_per_100k, y = reorder(CountyName,ConfirmedC_per_100k)))+
  #geom_point for dot plot
  geom_point(size = 2, aes(color = color, alpha = ifelse(color == "1", 1, 0.9)))+
  #mean line
  geom_vline(aes(xintercept = mean_daily_cases))+
  #support line
  geom_linerange(aes(xmin = mean_daily_cases, xmax = ConfirmedC_per_100k, color = color,
                     alpha = ifelse(color == "1", 1, 0.9)),
                 linetype = "dashed")+
  #setting colors and lables
  scale_color_manual(values = c(palette[6], palette[5]),
                    guide='none')+
  #dropping alpha legend
  scale_alpha(guide = 'none')+
  #changeing x ticks for graph
  scale_x_continuous(limits = c(10000, 19000),
                     breaks = seq(10000, 19000, by = 1000),
                     name = "Confirmed Cases (Per 100,000 people) -->")+
  scale_y_discrete(name = "Counties -->")+
  #adding the mean line mark
  annotate(x=mean_daily_cases+900, y=1, label="<- mean line", color="black",
           geom = "text", lineheight = .6)+
  #theme set
  theme_minimal()+
  #title
  ggtitle("Fig 1. Confirmed Cases for counties compared to average on 21st Dec 21")+
  #caption
  labs(caption = "Source: Covid Data Ireland")+
  #for legend position
  theme(legend.position = "top",
        legend.title = element_blank(),
         panel.grid.minor.x = element_blank())
```

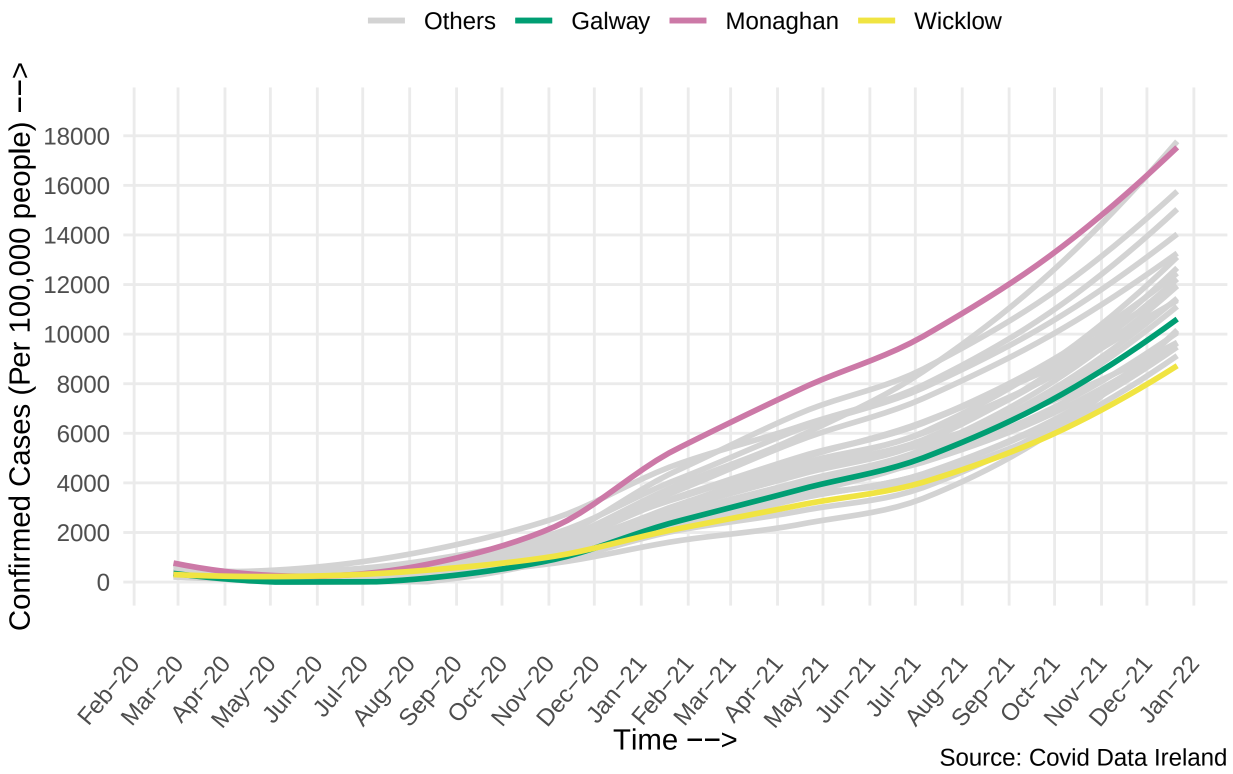


Analysis and Changes made: I have gone with the same diverging dot plot (mistakenly mentioned as bar chart in part 1) mentioned in the part 1, but there are a couple of changes that I have made, First, the color for counties that diverge above the value of mean and below are the same (except for county galway which is highlighted) as they lie on the opposite sides of the mean line they appear distinct and hence, they don't need to be colored differently. Second, the alpha values of counties other than galway is decreased so that County Galway is highlighted compared to others.

Q4. 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)

```r
#normalising for plotting
IRL_Covid19_plot2 <- IRL_counties_Covid19%>%
  mutate(ConfirmedC_per_100k =  round(100000  * ConfirmedC/Population,1))%>%
  mutate(DailyCCase_per_100k =  round(100000  * DailyCCase/Population, 1))

# getting a list of means by counties and taking the first and last form highest
#and lowest values of cumulative data
mean <- IRL_Covid19_plot2 %>%
  st_drop_geometry() %>%
  filter(TimeStamp == "2021-12-21") %>%
  arrange(ConfirmedC_per_100k) %>%
  select(CountyName, ConfirmedC_per_100k)

# get the data of galway and county with lowest and highest cumulative scores
select_county_data<- IRL_Covid19_plot2%>%
  filter(CountyName %in% c("Galway", head(mean$CountyName, 1), tail(mean$CountyName, 1)))

# for rest of the counties
other_counties<- IRL_Covid19_plot2%>%
  filter(!CountyName %in% c("Galway", head(mean$CountyName, 1), tail(mean$CountyName, 1)))

#for plotting plot2
IRL_Covid19_plot2 %>%
  #loading aesthetics for graphy
  ggplot(aes(x = TimeStamp, y=ConfirmedC_per_100k, color = color))+
  # for other counties a faded grey shade
  geom_smooth(data = other_counties,aes(group = CountyName, colour = "#d3d3d3" ),
              size = 1, alpha = 0.9, na.rm = TRUE, method = "loess", se = FALSE)+
  # individual colors for galway and highest, lowest counties
  geom_smooth(data = select_county_data, aes(group = CountyName, color = CountyName),
              size = 1, alpha = 0.8, na.rm = TRUE, method = "loess", se = FALSE)+
  # settings colors
  scale_color_manual(values = c("#d3d3d3", palette[3],palette[7] , palette[4]),
                     labels = c("Others", "Galway", "Monaghan", "Wicklow"))+
  # setting labels on a sequence of values for y axis
  scale_y_continuous(limits = c(0, 19000),
                     breaks = seq(0, 19000, by = 2000),
                     name = "Confirmed Cases (Per 100,000 people) -->")+
  # setting dates on the x axis
  scale_x_date(date_breaks = "months", date_labels = "%b-%y",
               name = "Time -->")+
  #title
  ggtitle("Fig 2. Cumulative Covid cases in Galway compared to Other counties")+
  #labels
  labs(caption = "Source: Covid Data Ireland")+
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 50, vjust = 0.5, hjust=1),
        axis.title.x = element_text(vjust = -2.5),
        legend.position = "top",
        legend.title = element_blank(),
         panel.grid.minor = element_blank())
```

Fig 2. Cumulative Covid cases in Galway compared to Other counties



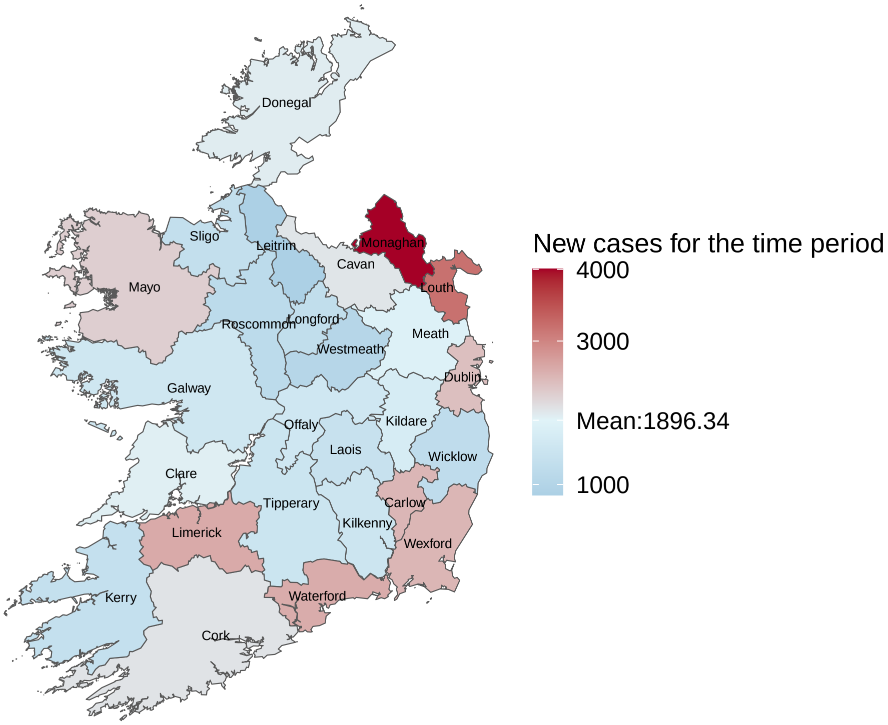
Source: Covid Data Ireland

Analysis and changes made : while plotting the trend lines, the alpha values of counties other thaa galway and other 2 were set to be less for them to be in the background. for getting the counties with highest and lowest number of cumulative cases the number of cumulative cases on last day in the dataframe 21-12-2021 was consider and the highest and lowest values were filtered.

Q5. 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.

```
library(tidyr)
library(RColorBrewer)

#normalise on per 100k population
plot_3 <- IRL_counties_Covid19 %>%
  mutate(ConfirmedC_per_100k =  round(100000  * ConfirmedC/Population,1))%>%
  mutate(DailyCCase_per_100k =  round(100000  * DailyCCase/Population, 1))

#get the cases for all counties on a 4 week period
plot_3 <- plot_3[plot_3$TimeStamp == ymd("2021-01-01") | plot_3$TimeStamp ==ymd("2021-01-29"),]%>%
  select(CountyName, ConfirmedC_per_100k, TimeStamp) %>%
  #subtract the previous value from current as in the dataframe we have two rows for one county
  # with the data between the range of 4 weeks
  mutate(new_cases_4weeks = ConfirmedC_per_100k - lag(ConfirmedC_per_100k,
                                          default =  first(ConfirmedC_per_100k)))

# as we have two entries for each county bases on difference we only keep the difference
# value between the same county and discard the the other
plot_3 <- plot_3[seq(2, nrow(plot_3), 2), ]
#get the mean value of new cases found during the 4 week period
mean = mean(plot_3$new_cases_4weeks)

# selecting columns necessary for plotting
plot_3<- plot_3 %>%
  select(CountyName, geometry, new_cases_4weeks)

ggplot(plot_3) +
  #sf geom for map based on geometry filling the case data
  geom_sf(aes(fill = new_cases_4weeks))+
  geom_sf_text(aes(label = CountyName), size = 1.8)+
  #hex values for RdYlBu pallete
  scale_fill_gradient2(#colors = c("#d73027", "white", "#4575b4"),
                       #values = rescale(c(4000, mean, 0)),
                       # for positioning the colorbar
                       low = "#74add1",
                       mid = "#e0f3f8",
                       high = "#a50026",
                       #setting the gradient midpoint on mean the value
                       midpoint = mean,
                       breaks = c(4000,3000, mean, 1000),
                       labels = c("4000", "3000", paste0("Mean:", round(mean,2)), "1000"),
                       #for setting the color bar
                       guide = guide_colorbar(
                       label.position = "right",
                       title = "New cases for the time period",
                       barwidth = grid::unit(0.4, "cm"),
                       barheight = grid::unit(3, "cm"))
                       )+

  ggtitle("Fig 3. Comparing new covid cases in Ireland from 1st Jan 21 to 29th Jan around the mean") +
  labs(caption = "Source: Covid Data Ireland")+
  theme_void()+
  #formatting the graph
  theme(plot.title = element_text(size = 10),
        legend.title = element_text(size = 10))
```

Fig 3. Comparing new covid cases in Ireland from 1st Jan 21 to 29th Jan around the mean



Source: Covid Data Ireland

Analysis and changes made : New confirmed cases were plotted for time range 1st of Jan 2021 to 29th of Jan 2021, The calculations were calculated from the similar way as described in part1 with a minor difference. Inilially, I was normalising the new confirmed data by subtracting the value by the mean, so the values would be diverging around 0, hence, the counties with positive values would have seen a increase in cases compared to mean new cases in Ireland and vice versa for negative values. But the problem with that approach was that a important statistic, the mean increase in new case would be missing as we normalised around the mean. Therefore, I have not normalised the data and have set the cholorpeth midpoint at mean for a divering color bar.

Additionally, The RdYlBu pallete was not visually coherent on the choropleth as the mean value lies slighly below the median of data, hence, I have went with the RdBu (Red Blue) diverging pallette for the choropleth.