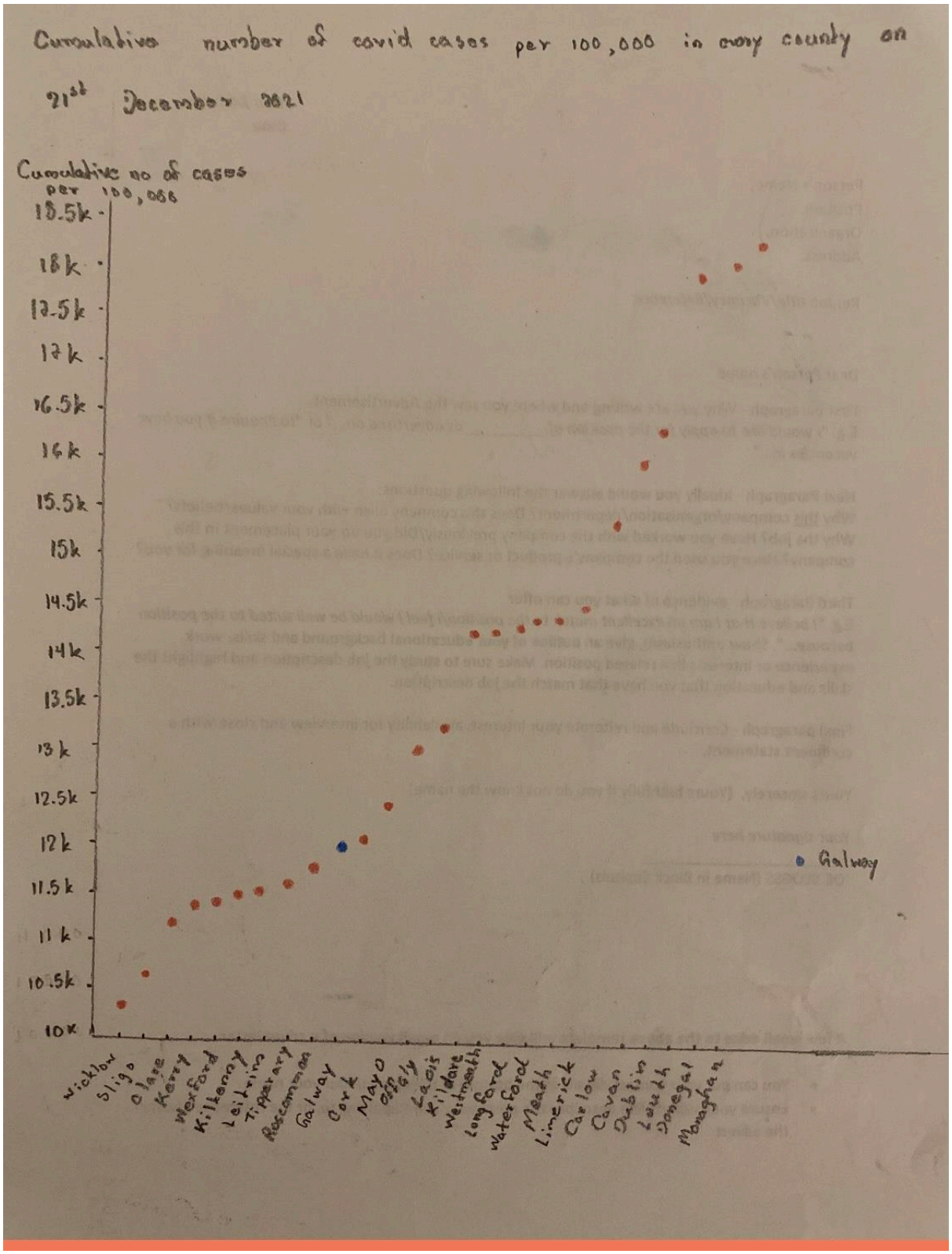
Data Visualization
Assignment 2 Part 1

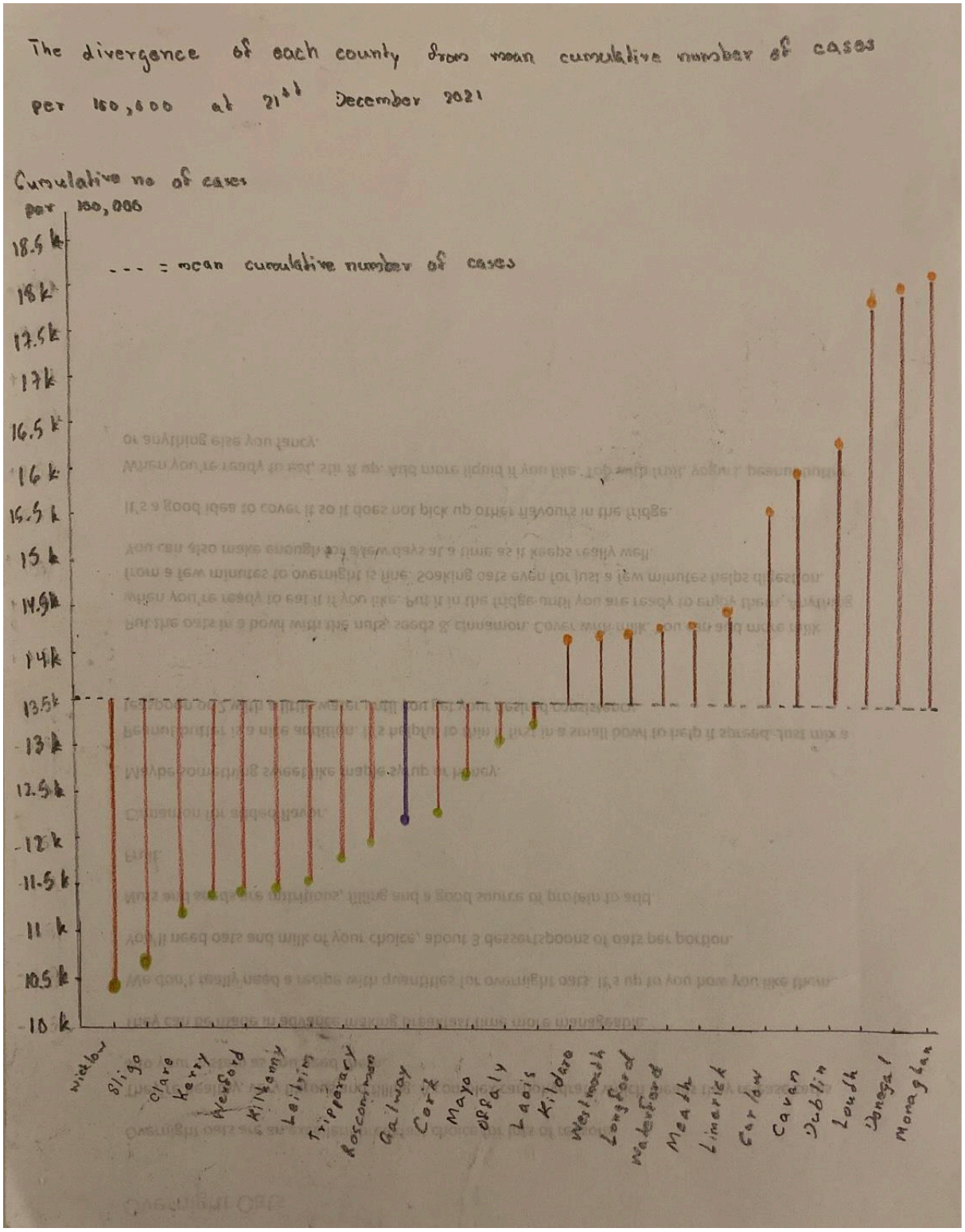
1.



For accurate comparison, dot plot visualization is chosen. The data is collected from the column of the cumulative number of cases per 100,000(calculated in R). On the X-axis, the

counties are placed in ascending order. On the Y-axis, the values of cumulative number of cases per 100,000 are put as 10 K instead of 100,00. To highlight Galway data, Galway is different colour (blue) while the other counties share the same colour (red).

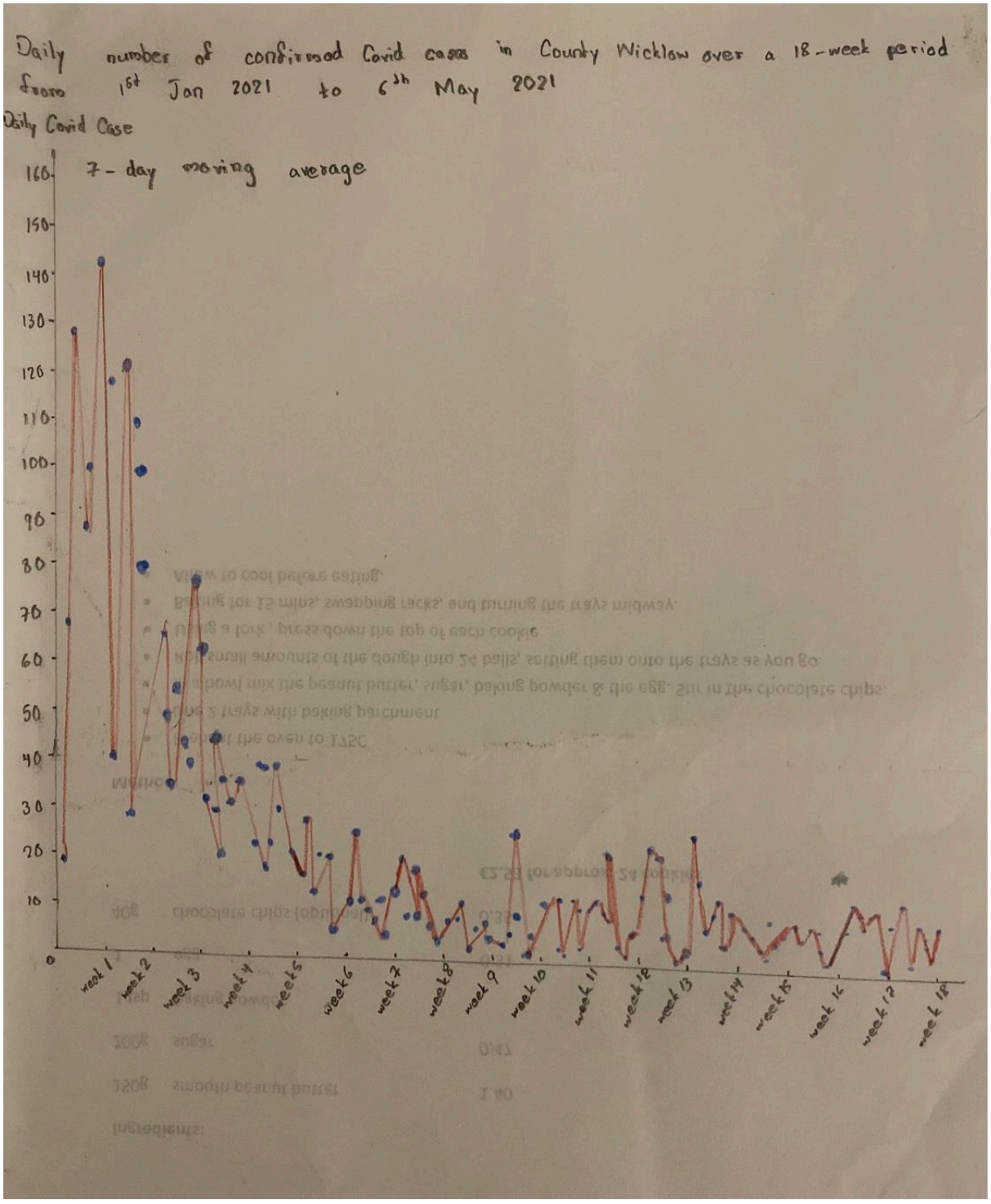
2.



Firstly, we calculate the mean value of the data from the column of cumulative cases per 100,000. The mean value is 13528. In the visualization, dotted horizontal line serves as

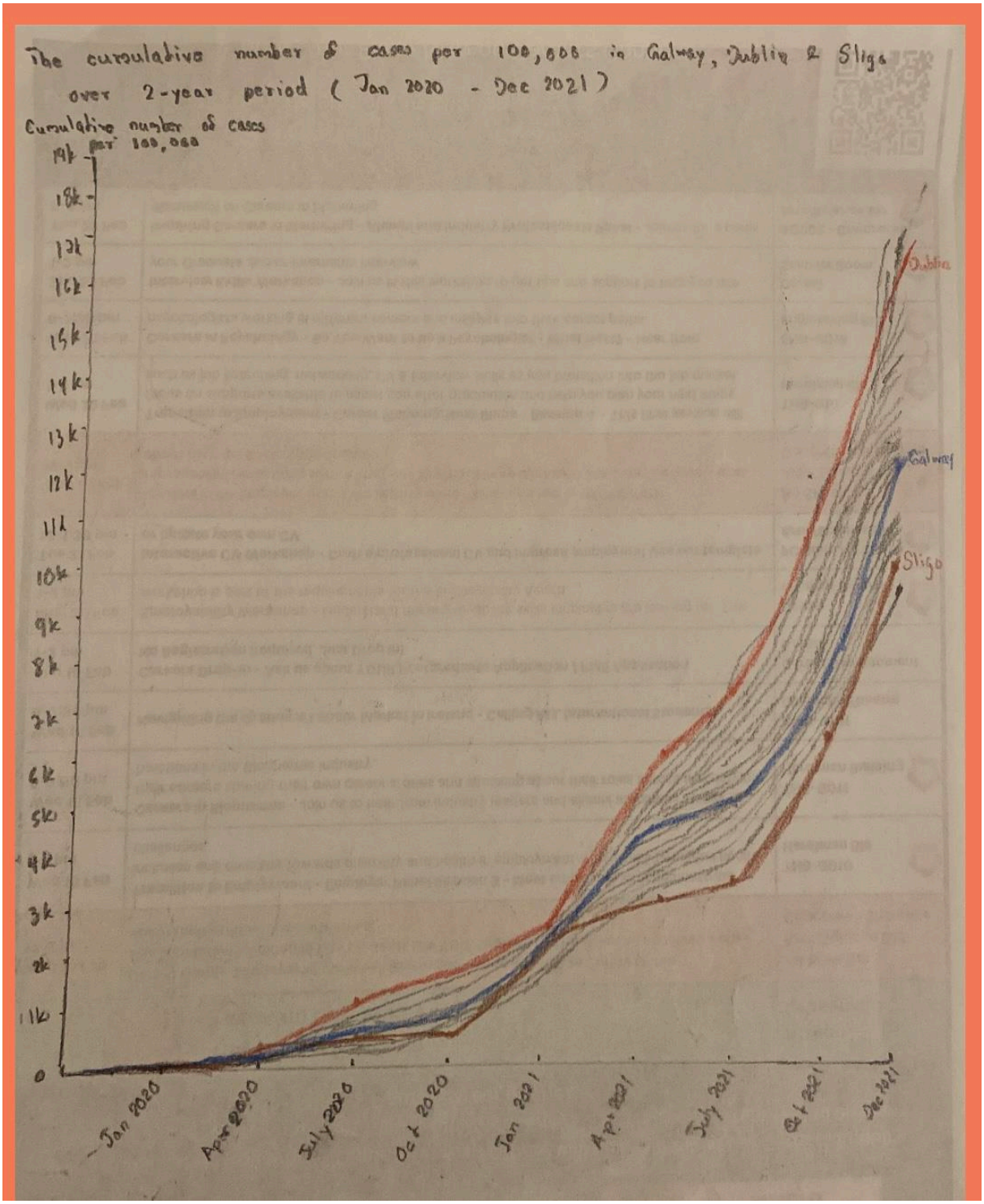
mean. On the X- axis & Y-axis, they are the same as the previous plot. In this lollipop plot, each dot shows individual value of each county and the associated stick shows how this data is divergence from the mean (i.e the longer stick means the greater divergence from mean). For the data below mean value, red and green colour are used for stick and dot respectively. For those above mean value, brown and orange colour are used. For the emphasis of Galway data, we will use another different colour (violet).

3.



To show the trend, line graph is appropriate. I choose the County Wicklow dataset from 1st January 2021 to 6th May 2021. There are a lot of data in this plot. I think smoothing will be needed and use 7-day moving average.

4.



First of all, I calculate mean number of cumulative number of cases per 100,000 of each county over the full time. I selected Sligo and Dublin as lowest and highest number of cases respectively. For the remaining 23 counties data, it will be put at background as grey colour and low alpha value. For highlighting 3 counties, red, blue and brown colour are used for Dublin, Galway and Sligo respectively.

5. For a choropleth visualization, I choose 4-week period from 1st November 2021 to 28th November 2021. Firstly, I collected and added the values of new confirmed cases per 100,000 on each day (total 28 days) of each county. I get 26 values of cumulative number of cases per 100,000 during that time frame (for example , cumulative value and new confirmed case value of the particular county at 1st November is the same). And then, I will find the mean value of these 26 cumulative ones.

I will use diverging scales palette which shows the divergence of data from a neutral mid point (in this visualization, mean value). In R colorBrewer Diverging palette, RdYlGn will be used. The above data of mean value will be shown as red and the below data is described as green. The data around the mean value is yellow. The more divergence from the mean value will give more darker red or green colour.

DataVisualisationAssignment2

aungmyatphyo22221184

2023-03-14

```
library(sf)
```

```
## Linking to GEOS 3.10.2, GDAL 3.4.2, PROJ 8.2.1; sf_use_s2() is TRUE
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

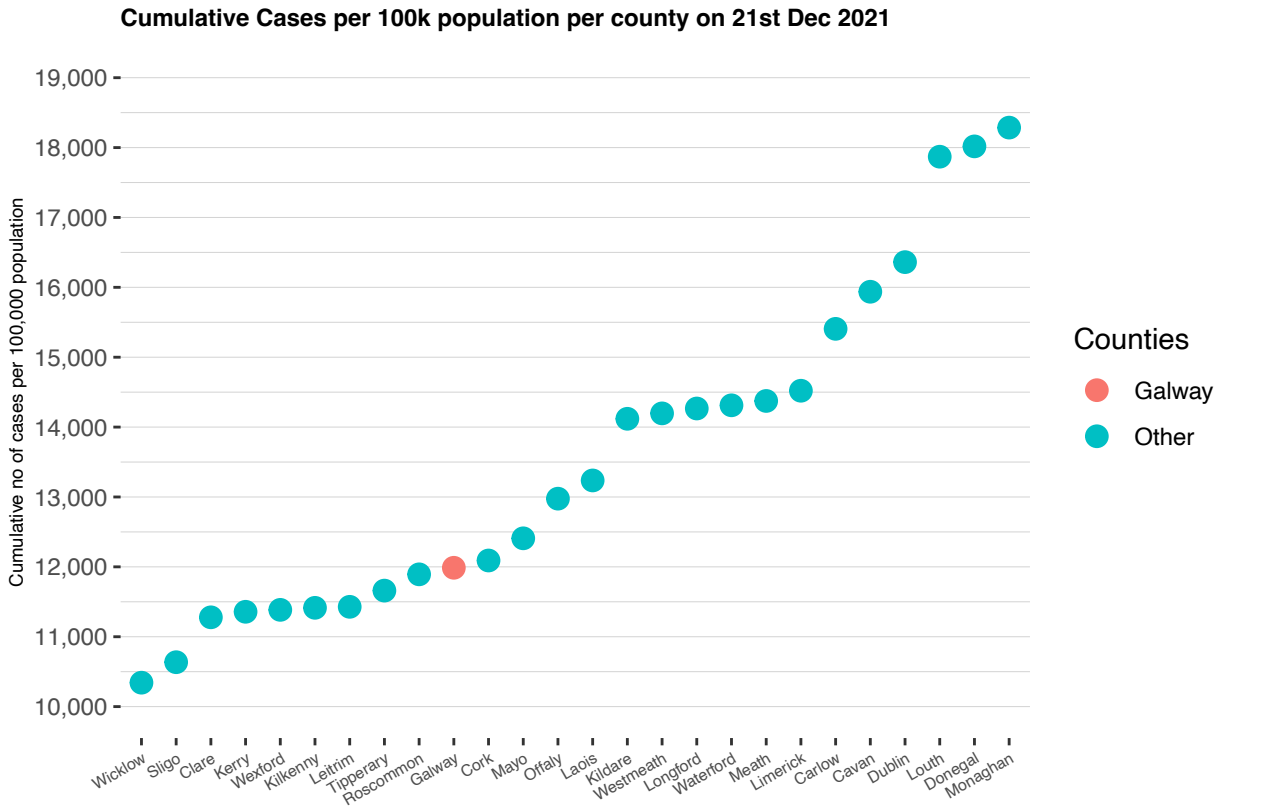
```
library(scales)
library(ggplot2)
library(tidyr)
library(colorspace)
library(colorblindr)

setwd("~/Downloads")
file1<-"/Users/aungphyo/Downloads/DV2/CovidCountyStatisticsIreland_v2.shp"
IRL_counties_Covid19 <- st_read(file1, quiet = TRUE)

#filter the data of all counties on 2021,12,21
#normalize DailyCCase and Confirmed C columns per 100,000 population
q1<- 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))

#colour variable
cb<-desaturate(palette_OkabeIto,0.25)
cb1<-cb[c(1,2)]
```

```r
#new variable for difference colour between galway and others
Counties=c('Other','Other','Other','Other','Other',
           'Other','Galway','Other','Other','Other',
           'Other','Other','Other','Other','Other','Other',
           'Other','Other','Other','Other','Other','Other',
           'Other','Other','Other','Other')
#bind a new variable Counties column to q1 dataframe
q1<-cbind(q1,Counties)

q11<-ggplot(q1,
            #order by ascending value of confirmedC per 100k value
            aes(x=reorder(CountyName,ConfirmedC_per_100k),
                y=ConfirmedC_per_100k,color=Counties))+
  geom_point(size=3.5)+
  ggtitle('Cumulative Cases per 100k population per county on 21st Dec 2021 ')

q11<-q11+theme(panel.spacing = unit(1,'lines'))

#change the y label name and discard x label
q11<-q11+xlab('')+
  ylab('Cumulative no of cases per 100,000 population')

#change the scales of y axis and add comma after 3 digits
q11<-q11+scale_y_continuous(breaks = seq(10000,19000,by=1000),
                            limits = c(10000,19000),
                            labels = comma)

q11<-q11+scale_fill_manual(values=cb1)

#theme
q11<-q11+theme_bw()+theme(
  panel.background = element_rect(fill = "white"),
  panel.grid.major = element_line(size = 0.1,
                                  linetype = 'solid',
                                  colour = "lightgrey"),
  #remove x axis major grid line
  panel.grid.major.x = element_blank(),
  panel.grid.minor = element_line(size = 0.075,
                                  linetype = 'solid',
                                  colour = "lightgrey"),
  #reduce the text size of x and angle into 30
  axis.text.x = element_text(angle=30,vjust=1,hjust=1,size=6),
  #bold the title name
  plot.title = element_text(face='bold',size=9),
  axis.title.y = element_text(size=7),
  axis.line.y = element_blank(),
  axis.line.x = element_blank(),
  panel.border = element_blank()
 )
```

```
## Warning: The `size` argument of `element_line()` is deprecated as of ggplot2 3.4.0.
## i Please use the `linewidth` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
```

```
## generated.
```

```
q11
```



For question 1, my hand drawing graph & ggplot graph are similar. Firstly, there is a difference in y-axis scale upper limits (19000) to cover highest data points and text style from 10k to 10,000. In ggplot graph, there are major and minor y axis grid lines which means easier to read the data. For colour, colour blind package is adapted for considering about the people with colour problem. For legend, Counties is described as legend title and 2 categories such as Galway and other which will help the reader for better understanding this graph. Finally, the plot title is modified.

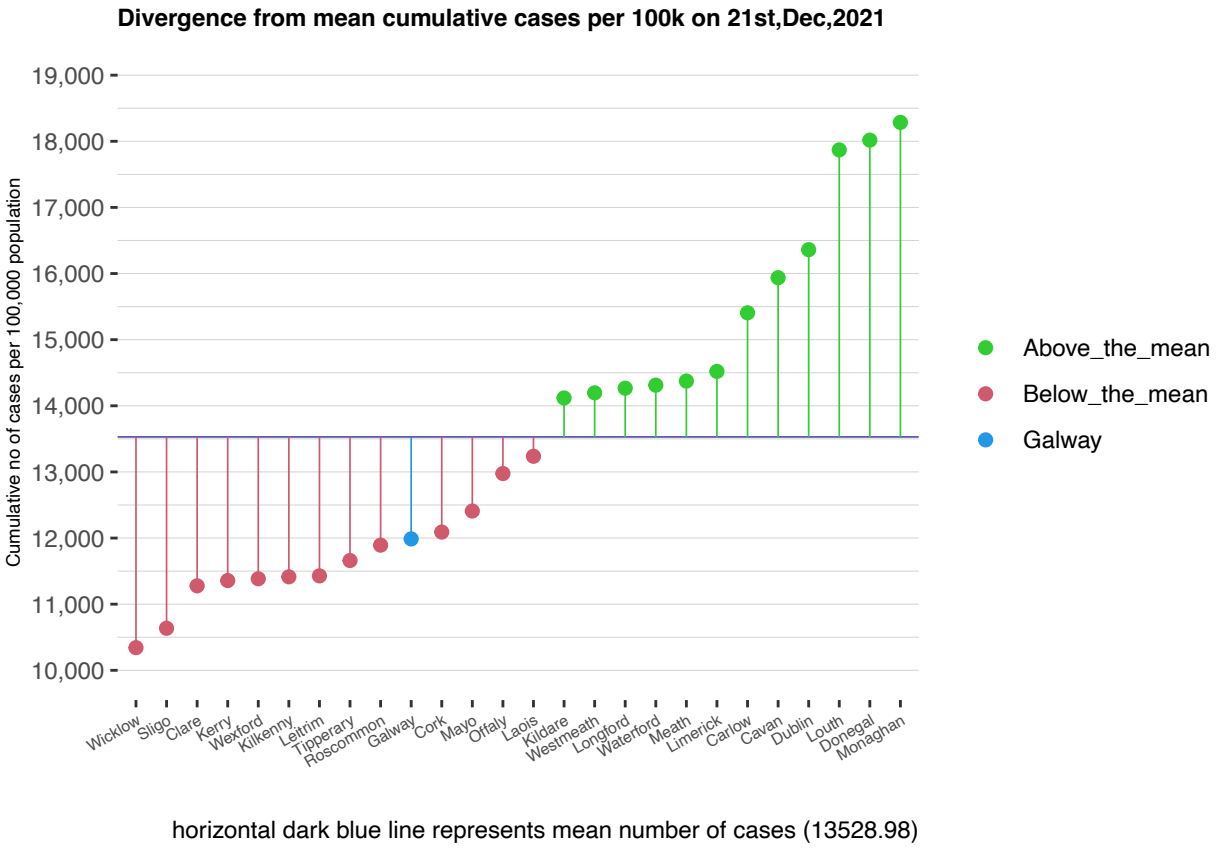
```
library(sf)
library(dplyr)
library(lubridate)
library(scales)
library(ggplot2)
library(tidyr)
library(colorspace)
library(colorblindr)

setwd("~/Downloads")
file1<-"/Users/aungphyo/Downloads/DV2/CovidCountyStatisticsIreland_v2.shp"
IRL_counties_Covid19 <- st_read(file1, quiet = TRUE)

#filter the data of all counties on 2021,12,21
#normalize DailyCCase and Confirmed C columns per 100,000 population
q2<- 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))

#colour variable
cb<-desaturate(palette_OkabeIto,0.25)
cb1<-cb[c(1,2)]

#mean value of ConfirmedC per 100k column
m<-mean(q2$ConfirmedC_per_100k)
#mean 13528.98

#to differ the colours among below the mean counties,
#above the mean counties and galway (3 colours)
mm<-c('a','a','b','b','a','a','g','b','a',
      'b','b','b','a','a','a','b','a','a',
      'b','b','b','b','a','a','b','b')
#bind new variable mm column to dataframe q2
q2<-cbind(q2,mm)


q22<-ggplot(q2,
            #order by ascending values of ConfirmedC per 100k value
            aes(x=reorder(CountyName,ConfirmedC_per_100k),
               y=ConfirmedC_per_100k,colour=mm))+
  #draw a mean horizontal line by colour dark blue
  geom_hline(yintercept = m,
             linetype="solid",
             size=0.2,
             colour='#00008B')+
  #to show divergence from mean horizontal line
  geom_segment(aes(x=reorder(CountyName,ConfirmedC_per_100k),
                   y=m,
                   xend=CountyName,
                   yend=ConfirmedC_per_100k),
               size=0.3,
               show.legend = FALSE)+
  geom_point(size=2)+
  #title name
  ggtitle("Divergence from mean cumulative cases per 100k on 21st,Dec,2021")+
  #caption explanation
  labs(caption = "horizontal dark blue line represents mean number of cases (13528.98)")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
#adding manual color and labels
q22<-q22+scale_colour_manual(values= c("#32cd32","#ce5a6c","212f85"),
                             name = "",
                         labels = c("Above_the_mean","Below_the_mean","Galway"))
#theme
q22<-q22+theme(panel.spacing = unit(1,'lines'))+
  #remove x label
```

```
  xlab('')+
  #change y label name
  ylab('Cumulative no of cases per 100,000 population')+
  #change the scales of y axis and add comma after 3 digits
  scale_y_continuous(breaks = seq(10000,19000,
                                  by=1000),
                     limits = c(10000,19000),
                     labels = comma)+
  theme_bw()+theme(
    panel.background = element_rect(fill = "white"),
    panel.grid.major = element_line(size = 0.1,
                                    linetype = 'solid',
                                    colour = "lightgrey"),
    #remove x axis major grid line
    panel.grid.major.x = element_blank(),
    panel.grid.minor = element_line(size = 0.075,
                                    linetype = 'solid',
                                    colour = "lightgrey"),
    #reduce the text size of x and angle into 30
    axis.text.x = element_text(angle=30,
                               vjust=1,
                               hjust=1,
                               size=6),
    #bold the title name
    plot.title = element_text(face='bold',
                              size=9),
    axis.title.y = element_text(size=7),
    axis.line.y = element_blank(),
    axis.line.x = element_blank(),
    panel.border = element_blank()
  )

q22
```

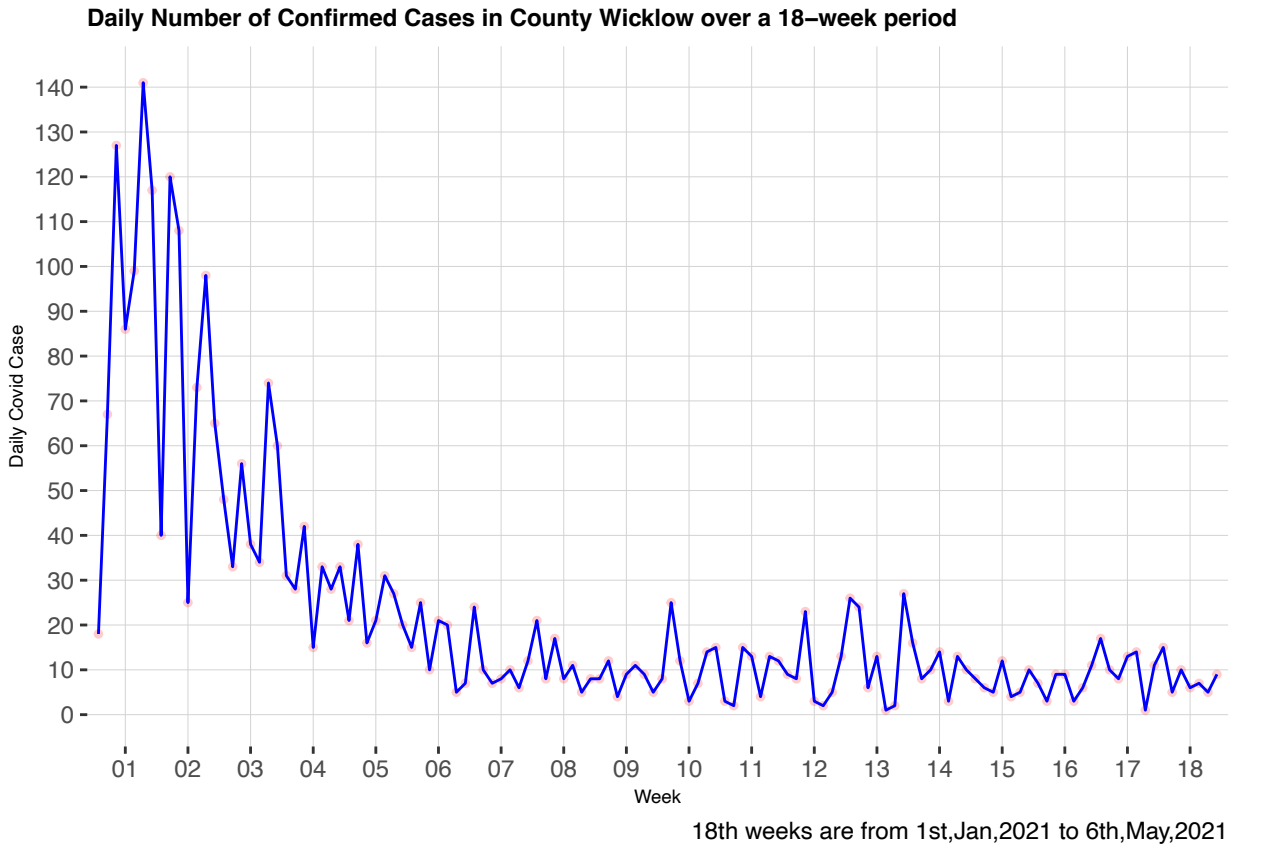


For question2, both graphs are similar. Firstly, there is a difference in y-axis scale upper limits (19000) to cover highest data points and text style from 10k to 10,000. In ggplot graph, there are major and minor y axis grid lines which means easier to read the data. The solid dark blue mean line is drawn rather than dotted ones and green colour for above the mean counties, red for below the mean counties and blue for galway is applied. Each dot means individual values and vertical sticks in the lollipop plot describes how much to differ from the mean value. Moreover, a legend is added and there is a caption about mean line below the graph. These features will provide better visualisation. Finally, the plot title is modified.

```
library(sf)
library(dplyr)
library(lubridate)
library('ggplot2')
library('tidyr')
library(scales)

setwd("~/Downloads")
file1<-"/Users/aungphyo/Downloads/DV2/CovidCountyStatisticsIreland_v2.shp"
IRL_counties_Covid19 <- st_read(file1, quiet = TRUE)

#filter the County Wicklow data
county<-subset(IRL_counties_Covid19,CountyName=="Wicklow")
#filter the 18th week duration between 1st Jan 2021 to 6th May 2021
Wicklow<- county%>%
  filter(TimeStamp >=as.Date('2021-01-01')& TimeStamp<=as.Date('2021-05-06'))


#plot with x axis timestamp and y axis DailyCCase
```

```
q3<-ggplot(data =Wicklow, aes(x =TimeStamp, y = DailyCCase)) +
  #daily data during 18th week
  geom_point(alpha = 0.2, colour='red',
             size = 1,
             na.rm = TRUE) +
  #title
  ggtitle('Daily Number of Confirmed Cases in County Wicklow over a 18-week period')+
  #explain 18th week duration
  labs(caption = "18th weeks are from 1st,Jan,2021 to 6th,May,2021")

#change y label and discard X label
q3<-q3+labs(x = " ",
            y = "Daily Covid Case")

#change x label into week and breaks as 1 week
q3<-q3+scale_x_date(
  name = 'Week',
  breaks = "1 week",
  labels = date_format("%W"),
  expand = c(0.01, 0))

#to show trends, draw the blue line
q3<-q3+geom_line(color = "blue",
                 size = .5)

q3<-q3+theme(panel.spacing = unit(1,'lines'))

#change the y scales
q3<-q3+scale_y_continuous(breaks = seq(0,140,
                                       by=10),
                          limits = c(0,142))

#theme
q3<-q3+theme_bw()+theme(
  panel.background = element_rect(fill = "white"),
  panel.grid.major = element_line(size = 0.1,
                                  linetype = 'solid',
                                  colour = "lightgrey"),
  panel.grid.minor = element_line(size = 0.075,
                                  linetype = 'solid',
                                  colour = "lightgrey"),
  #remove minor x and y grid lines
  panel.grid.minor.y = element_blank(),
  panel.grid.minor.x = element_blank(),
  #bold the title
  plot.title = element_text(face='bold',
                            size=9),
  axis.title.x = element_text(size=7),
  axis.title.y = element_text(size=7),
  axis.line.y = element_blank(),
  axis.line.x = element_blank(),
  panel.border = element_blank())
```

```
q3
```



For question 3, both graphs are similar. Firstly, y-axis upper limit scales is changed into 140 and x-axis scales are described as 1,2,3 and x-axis label is as Week. In ggplot graph, smoothing process is not applied because the data is not too big or too complexed for smoothing. Major gridlines of x-axis and y-axis are applied for better visualisation. For colour, line is drawn as blue and dots on each data is as red. In these difference colour, blue line shows the trend and each red dots serve as the individual data value. Finally, the plot title is modified and a caption is added to mention about the 18th week period.