Assignment 3 Visualizing Irish Covid figure

Vishakha Prakash Ratnakar (21238738)

09/03/2022

Introduction

In this report, we look at how many COVID cases there are in Ireland, broken down by county. We use various visualizations to show how COVID cases are increasing day by day and also the collective confirmed cases of counties ranging from 1st Jan 2020 to 21st Dec 2021.

Load all the required libraries

```
#=============================================================================
#Loading All required libraries
#=============================================================================
library(sf)
library(plyr)
library(dplyr)
library(ggplot2)
library(colorspace)
library(scales)
library(lubridate)
library(e1071)
library(knitr)
library(kableExtra)
```

```
#=============================================================================
#read the file
#=============================================================================
f<-"C:\\Users\\vishakha\\Desktop\\CovidCountyStatisticsIreland\\CovidCountyStatisticsIreland_v2.shp"

covid_data <- st_read(f, quiet = TRUE)
```

The given dataset is a .shp file that contains data on Irish COVID case numbers by county from January 1, 2020, to December 21, 2021. This dataset consists of 5 fields: CountyName, Population, TimeStamp, DailyCCase, ConfirmedC, and geometry.

CountyName: 26 Irish county names. eg: Cork, Sligo, Dublin etc.

Population: Population of that county

TimeStamp: Dates ranging from “2020-01-01” to “2021-12-21”. Dates at which data is recorded

DailyCCase: Number of cases per day in particular timestamp

ConfirmedC: Total number of confirmed cases till date.

geometry: co-ordinates to plot the Ireland map with county distribution

There are total 17212 number of rows.

Part 1

Part 1: A visualization that allows the reader to accurately read and compare the number of cases per 100,000 of population per county on the 21 December 2021.

The data is initially normalized, which entails transforming raw data into something more useful. As a result, we normalize the ConfirmedC data by dividing each counties total population by 100,000 (ConfirmedC/population * 100000). This normalized data is then filtered with TimeStamp = "2021-12-21" because just this TimeStamp data is required for the question's visualization. CountyName, Population, TimeStamp, DailyCCase, ConfirmedC, geometry, and number of cases per 100,000 population are among the 26 features with 6 fields returned.

```
#normalization of data

normalized_data <- covid_data %>%
  select(CountyName,Population,TimeStamp,DailyCCase,ConfirmedC) %>%
  mutate(num_of_cases_per_100000_population = (ConfirmedC/Population)*100000)


#extract with 21st dec 2021 timestamp


question_1_data <- normalized_data %>%
  group_by(CountyName) %>%
  filter(TimeStamp == "2021-12-21")

question_1_data
```

```
## Simple feature collection with 26 features and 6 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -1182398 ymin: 6695905 xmax: -667637.5 ymax: 7449865
## Projected CRS: WGS 84 / Pseudo-Mercator
## # A tibble: 26 x 7
## # Groups:   CountyName [26]
##    CountyName Population TimeStamp  DailyCCase ConfirmedC
##  * <chr>           <int> <date>          <int>      <int>
##  1 Carlow          56932 2021-12-21         39       8771
##  2 Cavan           76176 2021-12-21         44      12140
##  3 Clare          118817 2021-12-21         48      13400
##  4 Cork           542868 2021-12-21        264      65637
##  5 Donegal        159192 2021-12-21         56      28682
##  6 Dublin        1347359 2021-12-21       2460     220439
##  7 Galway         258058 2021-12-21        208      30933
##  8 Kerry          147707 2021-12-21         25      16775
##  9 Kildare        222504 2021-12-21        275      31415
## 10 Kilkenny        99232 2021-12-21        112      11326
## # ... with 16 more rows, and 2 more variables: geometry <MULTIPOLYGON [m]>,
## #   num_of_cases_per_100000_population <dbl>
```

Choice of Visualization

The figure below gives us insights into the number of COVID cases per 100,000 population per county on December 21st, 2021. The plot uses a horizontal bar chart whose primary aim is to compare the values for

the set of categorical variables. Categorical variables, in our case, are the county names. We swap the x and y-axis as there are more than 6 categories to be placed on the x-axis. After flipping the axes, we get a more compact figure with all elements horizontally orientated, including the text. As a result, the figure is far more readable. Each bar is the same length as the value associated with the data category, and all bars run from left to right. The bars are colored with "#5069be". Therefore, the chart consists of ordered bars, labels on the y axis indicating different counties in Ireland, and the number of COVID cases on the x-axis. A ranking of counties and an approximation of COVID cases per 100,000 population were generated.

Here we have used coord_flip() function to swap the x and y axis which helps to increase the readability of the bar chart. To order the bars according to the number of cases in ascending order we have used reorder() function.

```
theme_set(theme_classic())

ggplot(question_1_data,
       aes(x= reorder(CountyName,num_of_cases_per_100000_population),
           y=num_of_cases_per_100000_population)) +

  #alpha to reduce the intensity of the color
  #fill is used customize the color fill of bar
  geom_col(alpha=0.8, fill="#5069be", width=0.8) +


  scale_y_continuous(limits = c(0, 19e3),
                     expand = c(0, 0),
                     breaks = seq(2e3,18e3, by =2e3), #axis breaks
                     labels = seq(20,180,by=20)) + # label the values

  ggtitle("Covid Cases per 100,000 population (x 100) - Dec 21 2021 ") +

  # flip the x and y axis
  coord_flip(clip = "off") +

  theme(
    axis.title = element_blank(), #remove axis title
    axis.line.y = element_blank(), #remove y line
    axis.ticks.y = element_blank(), #remove y ticks
    axis.line.x = element_blank(), #remove x line
    axis.ticks.x = element_blank(), # remove X axis line
    axis.title.y = element_blank(), #remove y axis line
    plot.title = element_text(hjust = 0.5, face = "italic", size = 11),
    panel.grid.major.x = element_line(size = 0.2,
                                      linetype = 'solid',
                                      colour = "Darkgrey"))
```

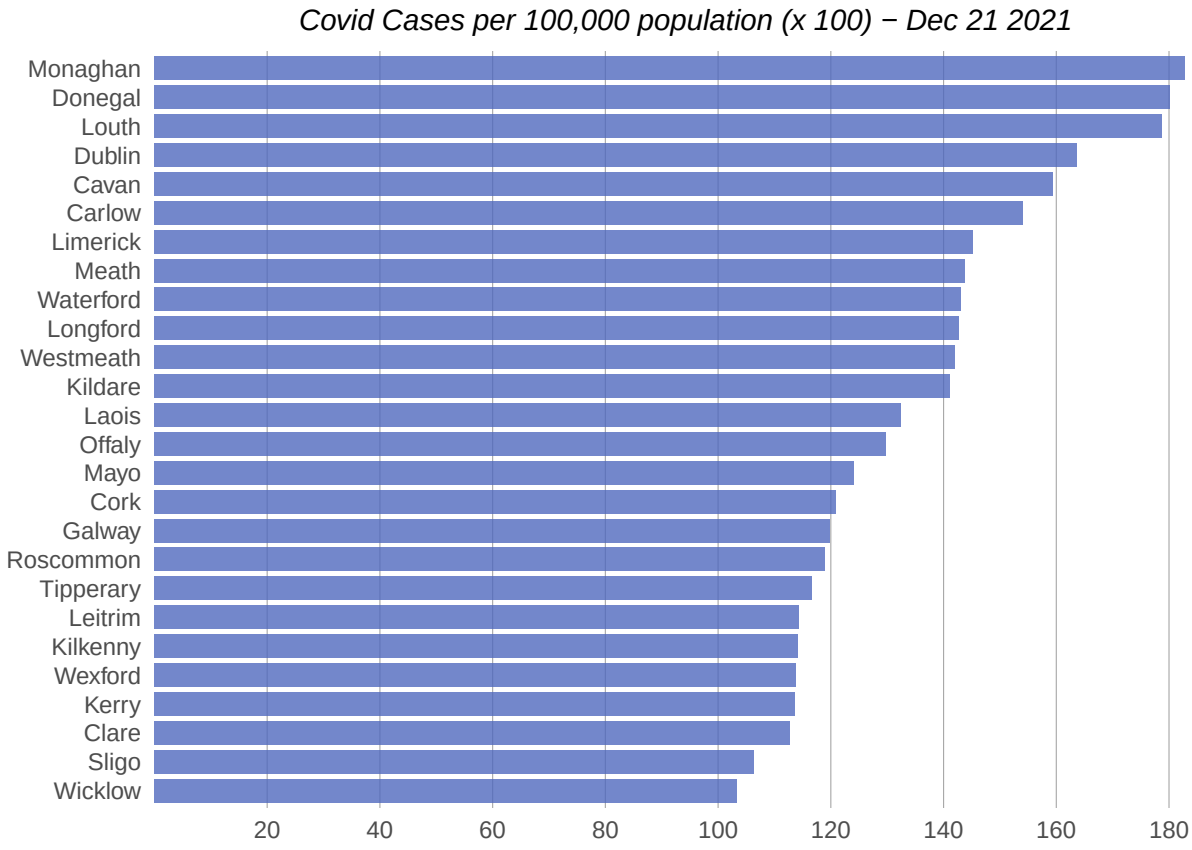

Figure 1: Number of covid cases per 100,000 of population of county on the 21 December 2021

Observation:

According to the December 21st, 2021 record, all counties in Ireland have more than 100k COVID cases, with Monaghan ranking top with 182k cases and Wicklow ranking lowest with 103k cases. Wicklow has the fewest COVID cases, with about 80K lesser than the highest-ranking county. Limerick, Meath, Waterford, Longford, Westmeath, and Kildare, Ireland's six counties, exhibit little variance in the number of COVID cases. Similarly, Leitrim, Kilkenny, Wexford, Kerry, and Sligo are all very consistent, yet the former group ranks higher than the latter. Cavan ranks 5th in the country with 15,900 COVID cases. Dublin has 4000 more cases than Cavan, putting it in fourth place, and Cairo has 5000 fewer cases than Cavan, putting it in sixth place. Monaghan, Donegal, and Louth are the counties with the largest number of COVID cases, whereas Wicklow, Sligo, and Clare have the lowest number of COVID cases.

Part 2

Part 2: A visualization that allows the reader to read how each county differs from the mean number of cases (per 100,000) in the country as at the 21 December 2021.

Choice of plot

In this section, we'll compare how each county differs from the country's average number of cases. The Divergence bar chart is employed for this purpose. Divergence bar charts are just two aligned bar charts that can handle both negative and positive numbers. Positive and negative divergences are determined in relation to a reference value, which can be zero or the mean. In our case, it's the mean value.

We began by calculating the mean number of cases per 100,000 up to December 21, 2021, obtaining 13529 as the mean. Following that, we only select the columns that are required for the plot. Positive and negative numbers are usually color-coded in a diverging bar chart. We established a new column called "pos" for this purpose. Values greater than or equal to the mean value will be saved as "TRUE," while values less than the mean value will be stored as "FALSE." This allows us to map the colors as an aesthetic to the value of the pos field. In order to visualize the values as positive and negative, we calculate the difference from the mean value and store it in the "difference" variable.

Now we'll make a diverging bar chart with the difference values on the x-axis and the county names on the y axis. We have used coord_flip() to flip the x and y-axis because there are 26 categorical values and it would be difficult to visualize these on the x-axis. We color the bars according to the pos values. We utilize two colors to color the pos values: red and blue, with red indicating increase and blue indicating decrease ("#0000FF," "#FF0000").

```
#calculate mean
mean_value <- (mean(question_1_data$num_of_cases_per_100000_population))

guestion_2_data <- question_1_data %>%
  select(CountyName,num_of_cases_per_100000_population) %>%
  mutate(pos = num_of_cases_per_100000_population >= mean_value) %>%
  mutate(difference = num_of_cases_per_100000_population - mean_value)

#colors for positive and negative bars
color_for_pos <- c("#0000FF","#FF0000")

ggplot(guestion_2_data,
       aes(x = reorder(CountyName,difference),
       y = difference, fill = pos,na.rm = TRUE)) +

  geom_col(alpha=0.6, width = 0.7) +

  ggtitle("Divergences of covid cases from mean number of cases in Ireland (x 1000)") +

  scale_y_continuous(limits = c(-3.5e3, 5.5e3),
                     breaks = seq(-3e3, 5e3, by= 500) ,
                     labels = seq(-3, 5, by =0.5),
                     expand=c(0,0)) +

  coord_flip(clip = "off") +

  scale_fill_manual(values=color_for_pos) +
```

```
theme(
  axis.title.y = element_blank(),
  axis.line.y = element_blank(),
  axis.ticks.y = element_blank(),
  axis.line.x = element_blank(),
  axis.ticks.x = element_blank(),
  axis.title.x = element_blank(),
  plot.title = element_text(hjust = 0.5, face = "italic", size = 11),
  panel.grid.major.y = element_line(size = 0.2, linetype = 'solid',colour = "grey88"),
  legend.position = "none"

)
```

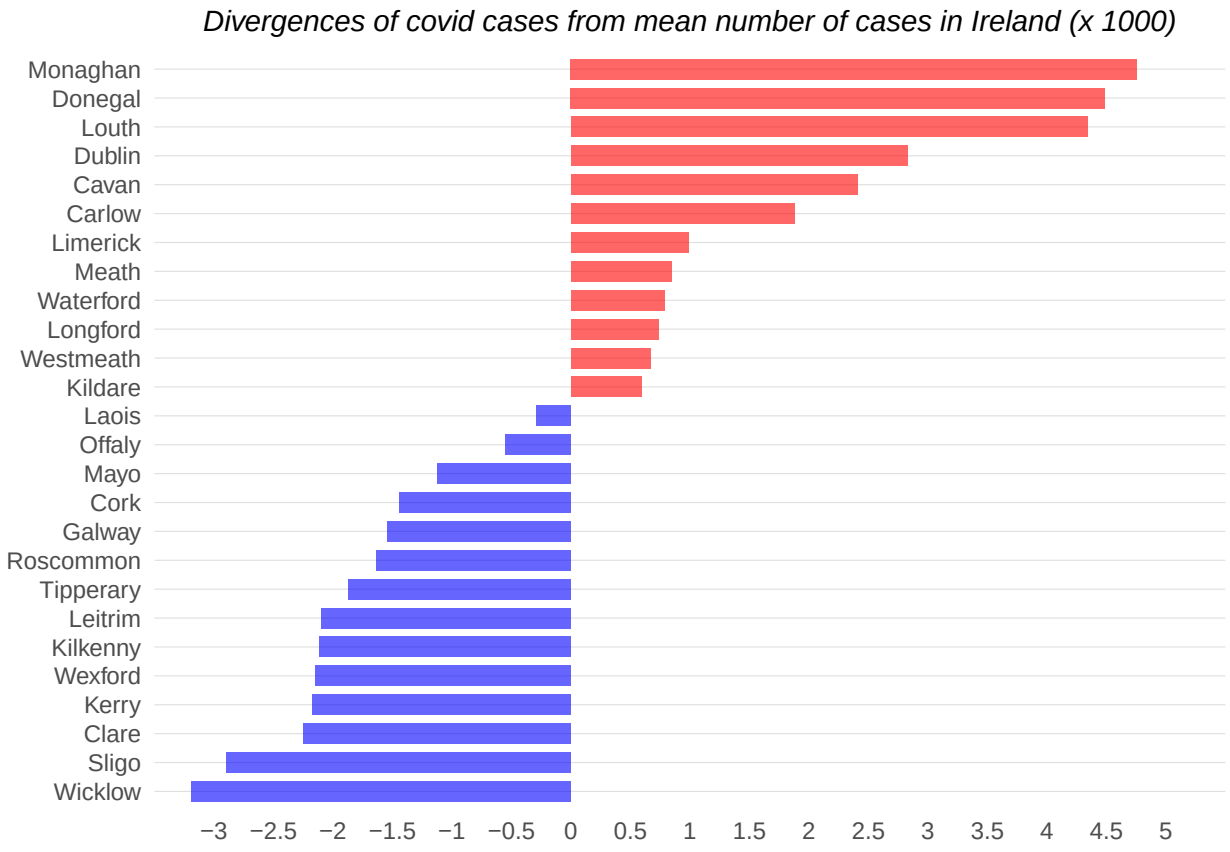

Figure 2: Divergences of covid cases from mean number of cases in Ireland on 21st Dec 2021

Observation

We can see that 12 of the 26 counties have seen an increase in the number of cases compared to the mean value, while the remaining 14 counties have seen a drop. As the mean value is approximately 13k, the lowest decrease is in county Laois and the lowest increase is in county Kildare by 0.3k and 0.6k. Wicklow (3.1k) has the largest decrease, whereas Monaghan has the largest increase (4.7k). Tipperary and Carlow show the same amount of decrease and increase in the number of COVID cases from the mean value, which is 1.8k

respectively. Counties such as Leitrim, Kilkenny, Wexford, and Kerry have similar amount decreases from the mean value.

Part 3

Part 3: A choreopleth visualisation of the cumulative number of cases per 100,000 on the 21 December 2021 and on 21 December 2020. These should be placed side by side on the page and must use the same scale so that they can be directly compared.

Part 1 has already normalized the data. We extract only the rows with the TimeStamps of December 21st, 2020, and December 21st, 2021 from the same data. The min and max values from the normalized data are then used to produce the intervals for the labels. Now we'll make 1000 distance discretised value breaks that start at the minimum and end at the maximum. After that, the continuous data was scaled down to the same size.

We've chosen a color palette (inferno) to display low-to-medium numeric numbers.

Facet wrap() divides a single plot into two plots, one for each timestamp. Because the data only has two timestamps, for December 21st, 2020, and December 21st, 2021, this results in two graphs, one for each timestamp. Also we have used labeller function to customize the predefined Timestamp in the plot.

```
#=============================================================================
#Question 3
#=============================================================================


question_3_data <- normalized_data %>%
  group_by(CountyName) %>%
  filter(TimeStamp == "2021-12-21" | TimeStamp == "2020-12-21")


#minimum value
scale_min <- round_any(min(question_3_data$num_of_cases_per_100000_population),
                       500, f = floor)

#maximum value
scale_max <- round_any(max(question_3_data$num_of_cases_per_100000_population),
                       1000,f = ceiling)

breaks<-seq(scale_min,scale_max, by =1000)


question_3_data$cases_D<-cut(question_3_data$num_of_cases_per_100000_population,
                         breaks = breaks,
                         dig.lab = 5)




#number of colors needed
no_of_colors<- nlevels(question_3_data$cases_D)

# Discretise the Palette
palette_data <- hcl.colors(no_of_colors, "Inferno", rev = TRUE)

# change the intensity of the colors
desaturated_palette <-desaturate(palette_data,amount = 0.2)
```

```r
#Create custom labels for the legend - e.g. (0k-10k]
labs_data_general <- breaks/1000
labs_plot <- paste0("(", labs_data_general[1:no_of_colors], "k-",
                                labs_data_general[1:no_of_colors+1], "k]")



labels_for_TimeStamp <-  c("2020-12-21" = "21st December 2020",
                           "2021-12-21" = "21st December 2021")

ggplot(question_3_data) +
  geom_sf(aes(fill = cases_D),
          color = "darkgrey",
          linetype = 1,
          lwd = 0.4) +

  facet_wrap(~TimeStamp,labeller = labeller(TimeStamp = labels_for_TimeStamp)) +


  ggtitle("Covid Cases per 100,000 population in Ireland") +

  # Custom palette
  scale_fill_manual(values = desaturated_palette,
                    drop = FALSE,
                    na.value = "grey80",
                    label = labs_plot,
                    # Legend
                    guide = guide_legend(direction = "vertical",
                                         label.position = "right")) +
  # Theme
  theme_void() +
  theme(
    strip.text = element_text(size = 10),
    legend.position  = "right",
    legend.title = element_blank(),
    legend.text = element_text(size=8),
    legend.key.height = grid::unit(0.4, "cm"),
    plot.title = element_text(hjust = 0.5, face = "italic", size = 15))
```

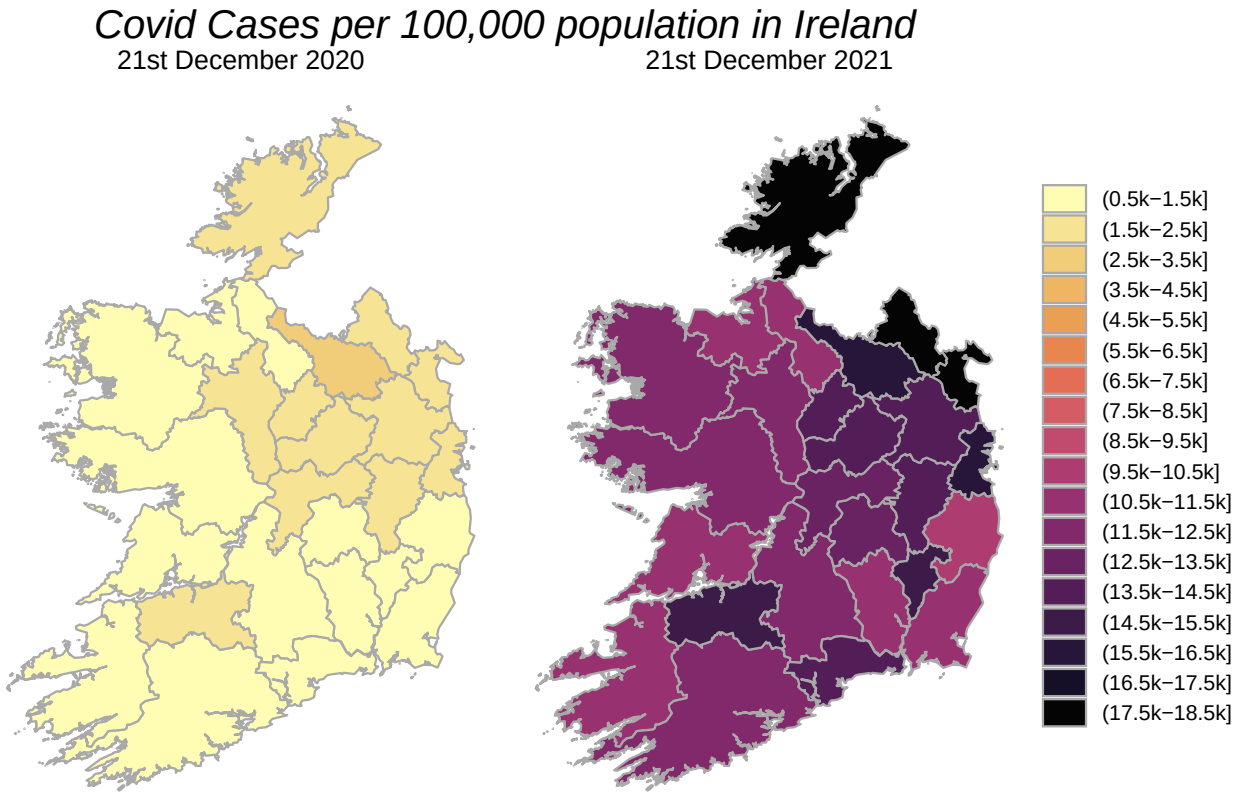

Figure 3: Choreopleth visualisation of the cumulative number of cases per 100,000 on the 21 December 2021 and 21 December 2020

Observation

On December 21st, 2020, the number of covid cases ranges from 0.5k to 3.5k, whereas on December 21st, 2021, it ranges from 10.5k to 18.5k. A difference of almost 7k can be seen.

In 2020, the majority of counties have COVID cases ranging from 0.5k to 1.5k. As of December 21st, 2020, Cavan has the highest number of COVID cases. 11 counties out of 26 have cases in the range of 1.5k to 2.5k.

In comparison to 2020, we can see a considerable increase in covid cases on December 21st, 2021. County Monaghan, Donegal, and Louth have the highest number of COVID cases, ranging from 17.5k to 18.5k, while county Wicklow has the lowest number of COVID cases, ranging from 10.5k to 11.5k. 5 counties range between 13.5k and 14.5k, of which 4 lie in the midlands east and 1 in the southeast. Galway, Mayo, and Roscommon counties on the west side had covid cases ranging from 11.5k to 12.5k.

The counties with COVID cases ranging from 1.5 to 2.5 thousand in 2020 has climb to 13.5 to 18.5 thousand in 2021. When compared to December 21st, 2020, the number of cases in Monaghan has risen by about 15,000. Similarly, there has been a 10k to 14k growth on the west and south sides. Cavan, which had the most cases in 2020, has between 14.5k and 15.5k in 2021.

Part 4

Part 4: A time series bar graph of the daily number of confirmed covid cases in one county in Ireland for one of the following periods: 3-month, 6 months or 1 year (1 year ~ January 1st 2021 to December 21 2021). This bar graph should also have a line representing the 7-day average for this period.

We've drawn a bar chart displaying the DailyCase of County Dublin for three months, from January 1st to March 31st, 2020, in this section. We also show a 7-day average line beside the bar chart.

A moving average is produced to capture some key points in the data by removing irrelevant minor details.

We created a 7-day moving average for this task. We take a time window of 7. The first window, according to our data, will run from January 1st to January 7th. The average of these 7 days is then calculated. Following that, we advance the window by one day and repeat the process. Now the window will be from January 2nd to January 8th.

We first extract the Covid data from 2021-01-01 to 2021-03-31.Moving average function (from Lecture notes). Calculate the 7-day average using a 7-day window size and mutate the data. geom_bar() and geom_line() are used to create the bar chart and the moving average line, respectively . The date scale on the x axis has a 10-day interval, while the y axis scale ranges from 0 to 3800 with a 200 gap with labels from 0 to 380 by 20. The bars are colored "steelblue" and the 7-day average line is colored "#E4D00A."

```
#================================================================================
#question 4
#================================================================================

start_date = "2021-01-01"
end_date = "2021-03-31"

question_4_data <- normalized_data %>%
  select(CountyName,TimeStamp,DailyCCase) %>%
  filter(TimeStamp >= start_date & TimeStamp <= end_date &
           CountyName == "Dublin")

question_4_data <- as_tibble(question_4_data)


#moving_ave function (Used the function from lecture worksheet)
moving_ave <- function(date, value, range, center = TRUE) {

  # This code was developed by Claus Wilke
  if (isTRUE(center)) {
    offset <- ceiling(range/2)
  } else {
    offset <- range
  }
  vapply(
    1:length(value),
    function(i, date, value) {
      focal_day <- date[i]
      first_day <- focal_day - offset + 1
      last_day <- focal_day - offset + range
      idx <- date >= first_day & date <= last_day
      if (head(date, 1L) > first_day || tail(date, 1L) < last_day) {
        NA_real_
      } else {
```

win	kurtosis
0	8.605936
7	8.605936

```
        mean(value[idx])
      }
    },
    double(1),
    date,
    value
  )
}


move_average<- question_4_data %>%
  mutate(
    close_7days_ave = moving_ave(TimeStamp, DailyCCase, 7, center = TRUE),
  )

#kurtosis value of data without smoothing and with smoothing of 7 days average

k0<-kurtosis(question_4_data$DailyCCase)
k7<-kurtosis(question_4_data$DailyCCase, na.rm=TRUE)

kurtosis_values<- data.frame("win" = c(0,7), "kurtosis" = c(k0,k7))


kable(kurtosis_values) %>%
  kable_styling( full_width = F)
```

```
ggplot(move_average, aes(TimeStamp, DailyCCase)) +

  geom_bar(stat = "identity",fill = "steelblue") +

  geom_line(aes(TimeStamp, close_7days_ave, color = "7d"), size = 1, na.rm = TRUE) +

  ggtitle("Daily Covid cases reported (x 10)") +

  scale_color_manual(values = c(`7d` = "#E4D00A"),
                breaks = c("7d"),labels = c("7-days average"),name = NULL) +

  scale_x_date(limits = c(dmy(start_date), dmy(end_date)),
               expand = c(0, 0),
               date_breaks = "15 days",
               date_labels="%d %b %Y") +

  scale_y_continuous(breaks = seq(0,3.7e3, by =300), #axis breaks
                     labels = seq(0,370,by=30)) +

  xlab(NULL) + ylab("Daily cases reported") +

  theme_minimal()+
```

```
theme(
  legend.position = c(0.9,0.9),
  legend.title = element_text(size = 15),
  plot.title = element_text(hjust = 0.5, face = "italic", size = 15))
```

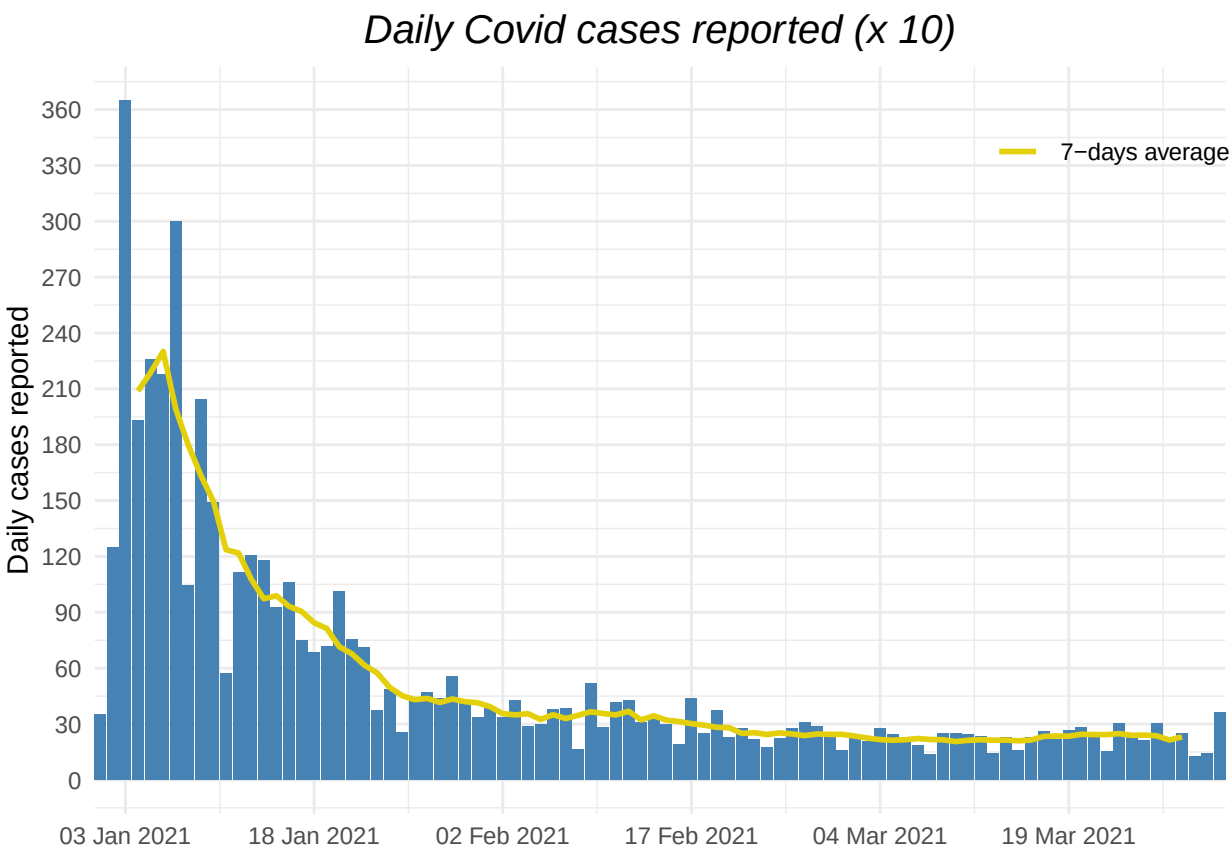

Figure 4: Time series bar chart of daily number of covid cases in Dublin with 7 days average of 3 months

Observation:

We can see that the highest number of cases, around 3650, was recorded in Dublin on January 3, 2021. After this day, the number of COVID cases dropped sharply from 3650 to 1000 approx in just 18 days. Every day since January 23rd, 2021, a consistent number of cases has been reported. The COVID case data appears to fit well with the 7-day moving average.

We also check Kurtosis which is a statistical measure that refers to the degree of presence of outlier in the distribution. We compare the Kurtosis measure of non-smooth data with the 7-day average data. There is no change in both values. This represents that the 7-day average line fits perfectly for the data.

Part 5

Part 5: A time series line graph that shows 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. This time series line graph must also show the time series of all other counties in Ireland. However, the three selected counties (Galway and two other must be highlighted)

We give a time series line graph in this part that displays the cumulative a number of cases per 100,000 in the counties of Galway, Monaghan, and Wicklow, while keeping data from other counties as a background.

For each county, we first generate a line time-series graph. As the background plot, this graph contains all lines in grey with a low alpha value. We next use the data from question 1 to discover the counties with the lowest and highest number of cases per 100,000, as they already have the data for the 2021-12-21 timestamp. The min and max functions are used for this. From this, we can infer that Monaghan has the highest number of cases, while Wicklow has the lowest. Along with this, we also pull statistics from Galway county. We then create a subset of these 3 counties, which is our foreground layer. We next apply this foreground layer to the preceding line graph and use the colors "#32CD32,"#FF6347," and "#1E90FF" to emphasize the three counties of Galway, Monaghan, and Wicklows, respectively. Also, we have removed the legends and inserted a duplicate y-axis by using function sec.axis = dup_axis() and labeled the three highlighted counties. We generated a second data frame named sec label data that solely contains the values of these three counties for labeling on the duplicate y-axis.

```
#================================================================================
#question 5
#================================================================================

#plot for background having grey lines for all counties
all_county_lines<- ggplot(normalized_data,
                          aes(TimeStamp,num_of_cases_per_100000_population)) +

  geom_line( aes(group = CountyName),size= 0.35, na.rm = TRUE, color="grey90",
             alpha =0.7, show.legend = FALSE ) +

  scale_x_date(name = "TimeStamp", breaks = "3 months",
               date_labels=("%b %y"), expand=c(0,0) ) +

  scale_y_continuous(labels = seq(from = 0, to =180, by=20),
                     breaks =seq(from = 0, to =18000, by=2000),
                     expand=c(0,0),
                     name= "Covid"
  ) +

  # this theme clears away grid lines, makes background white
  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))


#County with lowest cases
lowest_cases_county <- question_1_data %>%
  filter(num_of_cases_per_100000_population==
           min(question_1_data$num_of_cases_per_100000_population))
```

```r
lc <- lowest_cases_county$CountyName
print(paste0("county with lowest number of cases per 100,000: ",lc))
```

```
## [1] "county with lowest number of cases per 100,000: Wicklow"
```

```r
#county with highest cases
highest_cases_county <- question_1_data %>%
  filter(num_of_cases_per_100000_population==
           max(question_1_data$num_of_cases_per_100000_population))

hc <- highest_cases_county$CountyName
print(paste0("county with highest number of cases per 100,000: ",hc))
```

```
## [1] "county with highest number of cases per 100,000: Monaghan"
```

```r
#foreground counties
requried_counties <- c("Galway",hc, lc)


#foreground data
foreground_data<- subset(normalized_data, CountyName %in% requried_counties)
sec_label_data <- foreground_data %>%
  filter(TimeStamp == "2021-12-21")


#final plot with foreground layers

foreground_layer <- all_county_lines +
  geom_line(data=foreground_data, size =1, alpha=0.85,
            (aes(TimeStamp,num_of_cases_per_100000_population,
                 colour= CountyName, group = CountyName))) +

  ggtitle("2020-2021 Cumulative number of covid cases (x 100):
          Galway, Monaghan, Wicklows") +

  scale_colour_manual(values = c("#32CD32","#FF6347", "#1E90FF"),name = NULL,
                      limits = requried_counties) +

  scale_y_continuous(labels = seq(from = 0, to =180, by=20),
                     breaks =seq(from = 0, to =18000, by=2000),
                     expand=c(0,0),
                     name= "Covid Cases",
                     sec.axis = dup_axis(
                       breaks = sec_label_data$num_of_cases_per_100000_population,
                       labels = sec_label_data$CountyName,
                       name = NULL,
                     )) +

  theme(legend.position = "none",
        plot.title = element_text(hjust = 0.5, face = "italic", size = 15))

foreground_layer
```

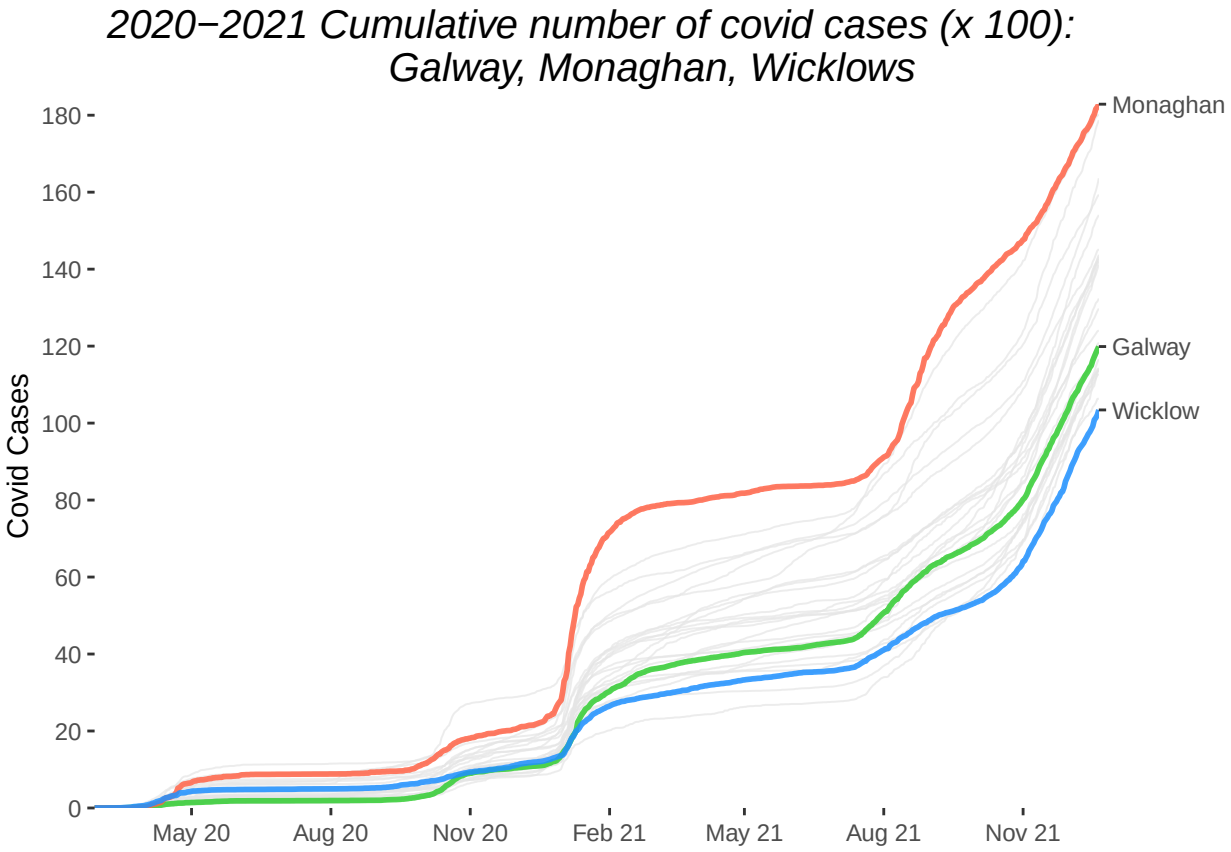

Figure 5: The cumulative number of cases per 100,000 from 2020 to 2021

Observation

From January 1, 2020, to December 21, 2021, the cumulative number of COVID cases per 100,000 population increases. There is no significant increase in the number of COVID cases from January 2020 to December 2020, however, there is a drastic increase in the number of COVID cases from January 2021 onwards. From 26 counties, we have highlighted 3 counties: Galway, Monaghan, and Wicklow which is the foreground data that gives a sense of where these 3 counties lie among other counties. Because the target counties are colored differently, the difference in the cumulative number of COVID cases is clearly visible. The red line represents Monaghan, which has the highest cases, while the green line represents Galway, which is closer to Wicklow and further away from Monaghan in terms of COVID cases. Wicklow, which is symbolized by the color blue, has the lowest number.

Conclusion

We used a variety of visualizations to display the Irish covid cases data, including a bar chart, a time series line chart, and a choropleth. On December 21st, 2021, the largest number of cases per 100,000 population was discovered in Monaghan, while the lowest was discovered in Wicklow. The number of covid cases grown massively after December 2020.

Reference

Lecture WorkSheets