209041841_GY7702_CW1

Gowtham Nallathambi

18/11/2020

#GY7702-R-for-Data-Science

##*The University of Leicester* **CourseWork 1** The link to the GitHub Repository

#Question 1

```
rm(list=ls())       # To clear environment
library(tidyverse)  # Load library in relevant script
```

##Question 1.1:

Vector list of 25.

```
survey_ans<- c(NA, 3, 4, 4, 5, 2, 4, NA, 6, 3, 5, 4, 0, 5, 7, 5, NA, 5, 2, 4,
               NA, 3, 3, 5, NA)
```

To check whether all participants to the survey either completely disagree or completely agree once the missing values are excluded.

```
missing_value<-is.na(survey_ans)          # Missing values are excluded
value_vector<-
  survey_ans[!missing_value]              # Creating complete value vector using logical operator

for(iterator in value_vector) {
  if (iterator == 3) {
    cat(iterator, "= somehow disagree\n")
  } else if (iterator == 4) {
    cat(iterator, "= neither agree nor disagree\n")
  } else if (iterator == 5) {
    cat(iterator, "= somehow agree\n")
  } else if (iterator == 2) {
    cat(iterator, "= disagree\n")
  } else if (iterator == 6) {
    cat(iterator, "= agree\n")
  } else if (iterator == 7) {
    cat(iterator, "= completely agree\n")
  } else {
    cat("0 = No Response\n")
  }
}
```

```
## 3 = somehow disagree
## 4 = neither agree nor disagree
## 4 = neither agree nor disagree
## 5 = somehow agree
## 2 = disagree
## 4 = neither agree nor disagree
## 6 = agree
## 3 = somehow disagree
## 5 = somehow agree
## 4 = neither agree nor disagree
## 0 = No Response
## 5 = somehow agree
## 7 = completely agree
## 5 = somehow agree
## 5 = somehow agree
## 2 = disagree
## 4 = neither agree nor disagree
## 3 = somehow disagree
## 3 = somehow disagree
## 5 = somehow agree
```

```
rm(iterator)                                  # After the loop delete the iterator
```

##Question 1.2:

The code necessary to extract the indexes related to the participants in the survey who at least somehow agree or more.

```
survey_index_extract<- c(5,6,7)                    # Create a sub vector to extract index
cat("The participants in the survey who
at least somehow agree or more : ",
which(value_vector %in% survey_index_extract))     # Using which condition to extract
```

```
## The participants in the survey who
## at least somehow agree or more :  4 7 9 12 13 14 15 20
```

#Question 2

```
rm(list=ls())                                      # To clear environment
#install.packages('palmerpenguins')                # Install palmerpenguins packages
library(tidyverse)                                 # Load tidyverse library in script
library(palmerpenguins)                            # Load palmerpenguins library in script
```

##Question 2.1:

Installed palmerpenuins in above chunk.

##Question 2.2:

To create a table showing species, island, bill length and body mass of the 10 Gentoo penguins in the penguins table with the highest body mass.

```
palmer_penguins <-
  palmerpenguins::penguins                    # Import the palmerpenguins data
gentoo_penguins <- palmer_penguins[           # Subset the Gentoo penguins
  palmer_penguins$species == "Gentoo",
  c("species","island", "bill_length_mm",
    "body_mass_g")
]
body_mass<-arrange(                           # Arranging body mass using arrange function
  gentoo_penguins,
  -body_mass_g)
head(body_mass, n=10)                         # Displaying 10 rows from table
```

```
## # A tibble: 10 x 4
##    species island bill_length_mm body_mass_g
##    <fct>   <fct>           <dbl>       <int>
##  1 Gentoo  Biscoe           49.2        6300
##  2 Gentoo  Biscoe           59.6        6050
##  3 Gentoo  Biscoe           51.1        6000
##  4 Gentoo  Biscoe           48.8        6000
##  5 Gentoo  Biscoe           45.2        5950
##  6 Gentoo  Biscoe           49.8        5950
##  7 Gentoo  Biscoe           48.4        5850
##  8 Gentoo  Biscoe           49.3        5850
##  9 Gentoo  Biscoe           55.1        5850
## 10 Gentoo  Biscoe           49.5        5800
```

##Question 2.3:

Calculate average bill length for each Island.

```
average_bill_length<- palmer_penguins %>%   # Assigning variables
  group_by(island)%>%                       # Using group_by functions with pipe operator
  summarize(
    average_bill_length =
      mean(bill_length_mm,
           na.rm = TRUE
           ),
            .groups = "drop")%>%            # Summarize the mean
  print(average_bill_length)
```

```
## # A tibble: 3 x 2
##   island    average_bill_length
##   <fct>                   <dbl>
## 1 Biscoe                   45.3
## 2 Dream                    44.2
## 3 Torgersen                39.0
```

##Question 2.4:

Calculating Minimum, median and maximum proportion between bill length and bill depth by species.

```r
species_analyse<-palmer_penguins %>%            # Assigning variables
  group_by(species)%>%                          # Grouping by species using pipe operator
  summarise(                                    # Summarize the values
    Median_bill_length = median(bill_length_mm,
                                na.rm = TRUE), # To calculate median
    Median_bill_depth = median(bill_depth_mm,
                               na.rm = TRUE),
    Min_bill_length = min(bill_length_mm,
                          na.rm = TRUE),        # To calculate minimum
    Min_bill_depth = min(bill_depth_mm,
                         na.rm = TRUE),
    Max_bill_length = max(bill_length_mm,
                          na.rm = TRUE),        # To calculate maximum
    Max_bill_depth = max(bill_depth_mm,
                         na.rm = TRUE),
    .groups = "drop"
  )
print(species_analyse)                          # Print the results
```

```
## # A tibble: 3 x 7
##   species Median_bill_len~ Median_bill_dep~ Min_bill_length Min_bill_depth
##   <fct>              <dbl>            <dbl>           <dbl>          <dbl>
## 1 Adelie              38.8             18.4            32.1           15.5
## 2 Chinst~             49.6             18.4            40.9           16.4
## 3 Gentoo              47.3             15              40.9           13.1
## # ... with 2 more variables: Max_bill_length <dbl>, Max_bill_depth <dbl>
```

#Question 3

```r
rm(list=ls())                                   # To clear environment
#install.packages('lubridate')                  # install packages
library(tidyverse)                              # Load tidyverse library in script
library(lubridate)                              # Load lubridate library in script
```

##Question 3.1

Import covid data.

```r
covid_data <-
  readr::read_csv("covid19_cases_20200301_20201017.csv")          # using readr
```

##Question 3.2:

Creating a new tibble assigned to my student ID.

```r
covid_data %>%
  dplyr::arrange(specimen_date) %>%       # Arrange tibble using dplyr by pipe operator
  tidyr::fill(newCasesBySpecimenDate ,
              cumCasesBySpecimenDate) %>% # Replace NA values with the value available
  tidyr::replace_na(list(                 # Replace the remaining NA values with 0
    newCasesBySpecimenDate = 0,
    cumCasesBySpecimenDate = 0)
    )
```

```
## # A tibble: 82,969 x 4
##    specimen_date area_name               newCasesBySpecimen~ cumCasesBySpecimen~
##    <date>        <chr>                                 <dbl>               <dbl>
##  1 2020-03-01    Aberdeen City                             0                   0
##  2 2020-03-01    Aberdeenshire                             0                   0
##  3 2020-03-01    Angus                                     0                   1
##  4 2020-03-01    Antrim and Newtownabbey                   0                   0
##  5 2020-03-01    Ards and North Down                       0                   0
##  6 2020-03-01    Argyll and Bute                           0                   0
##  7 2020-03-01    Armagh City, Banbridge~                   0                   0
##  8 2020-03-01    Barking and Dagenham                      1                   1
##  9 2020-03-01    Barnet                                    0                   1
## 10 2020-03-01    Belfast                                   0                   0
## # ... with 82,959 more rows
```

```
print(covid_data, n=5)
```

```
## # A tibble: 82,969 x 4
##   specimen_date area_name             newCasesBySpecimenDa~ cumCasesBySpecimenD~
##   <date>        <chr>                                 <dbl>               <dbl>
## 1 2020-03-01    Aberdeen City                             0                   0
## 2 2020-03-01    Aberdeenshire                             0                   0
## 3 2020-03-01    Angus                                     0                   1
## 4 2020-03-01    Antrim and Newtownab~                     0                   0
## 5 2020-03-01    Ards and North Down                       0                   0
## # ... with 82,964 more rows
```

```
#Subset area assigned to my student ID

trafford_complete_covid_data <-           # Storing the data in a new tibble
  covid_data[
  covid_data$area_name == "Trafford",     # Dropping the area_name
  c("specimen_date",
    "newCasesBySpecimenDate",
    "cumCasesBySpecimenDate"
    )
]

print(trafford_complete_covid_data, n=5)
```

```
## # A tibble: 223 x 3
##   specimen_date newCasesBySpecimenDate cumCasesBySpecimenDate
##   <date>                         <dbl>                  <dbl>
## 1 2020-03-02                         3                      3
## 2 2020-03-03                         0                      3
## 3 2020-03-04                         1                      4
## 4 2020-03-05                         0                      4
## 5 2020-03-06                         0                      4
## # ... with 218 more rows
```

##Question 3.3:

Combine the trafford data with new cases and last day cases.

```
trafford_day_before <-
  trafford_complete_covid_data                              # load data

trafford_day_before <-                                      # mutate data with lubridate
  mutate(trafford_day_before,
         day_to_match =
           trafford_day_before$specimen_date +
           days(1)
         )

trafford_day_before <-
  dplyr::select(trafford_day_before,
                newCasesBySpecimenDate,
                day_to_match) %>%
  rename(
    newCases_day_before = newCasesBySpecimenDate            # subset table
    )

final_data = trafford_complete_covid_data %>%
  full_join (trafford_day_before, by = character())         # Join the table

final_data <- mutate(final_data,                            # Calculate percentage
                     percentage = (
                       (
                       final_data$newCasesBySpecimenDate/
                       final_data$newCases_day_before
                                    )*100
                                  )
                     )

print(final_data, n=5)                                      # print data
```

```
## # A tibble: 49,729 x 6
##   specimen_date newCasesBySpeci~ cumCasesBySpeci~ newCases_day_be~ day_to_match
##   <date>                   <dbl>            <dbl>            <dbl> <date>
## 1 2020-03-02                   3                3                3 2020-03-03
## 2 2020-03-02                   3                3                0 2020-03-04
## 3 2020-03-02                   3                3                1 2020-03-05
## 4 2020-03-02                   3                3                0 2020-03-06
## 5 2020-03-02                   3                3                0 2020-03-07
## # ... with 49,724 more rows, and 1 more variable: percentage <dbl>
```

##Question 3.4:

###The Covid - 19 daily cases trends on Trafford

All over the UK, the cases has been increasing,but in Trafford we haven't seen any high rise in the positive case. eventhough, in some areas as per the daily cases are register in surbans. As per the analyzed data, In recent times,there huge pike in the new cases in the trafford comparatively in March'2020 as the first wave. There are some chances to spike in getting expose of the covid. Nevertheless, trafford is small town with low density.

#Question 4

##Covid Data analysis with UK datasets

```
rm(list=ls())            # To clear environment
library(tidyverse)       # Load tidyverse library
library(plotly)          # Library to produce the data plot in graphical representation
library(knitr)           # knitr Library
```

To load the population data into a variable and join this information with new table

```
# Load .csv file using readr

covid_data <-
  readr::read_csv("covid19_cases_20200301_20201017.csv")

lad19_population <-                                  #Rename the column to merge two tables
  readr::read_csv("lad19_population.csv")%>%
  dplyr::rename(area_name = lad19_area_name)

covid_population <-                                  #Join the table using full_join
  full_join(covid_data,
            lad19_population,
            by = "area_name"
            )

trafford_covid <-                                    #Extract the trafford covid data
  covid_population[
  covid_population$area_name == "Trafford",
  c("area_name", "specimen_date",
    "newCasesBySpecimenDate",
    "cumCasesBySpecimenDate",
    "area_population"
    )
]

print(trafford_covid, n=5)
```

```
## # A tibble: 223 x 5
##   area_name specimen_date newCasesBySpecimen~ cumCasesBySpecime~ area_population
##   <chr>     <date>                      <dbl>              <dbl>           <dbl>
## 1 Trafford  2020-03-02                      3                  3          237377
## 2 Trafford  2020-03-03                      0                  3          237377
## 3 Trafford  2020-03-04                      1                  4          237377
## 4 Trafford  2020-03-05                      0                  4          237377
## 5 Trafford  2020-03-06                      0                  4          237377
## # ... with 218 more rows
```

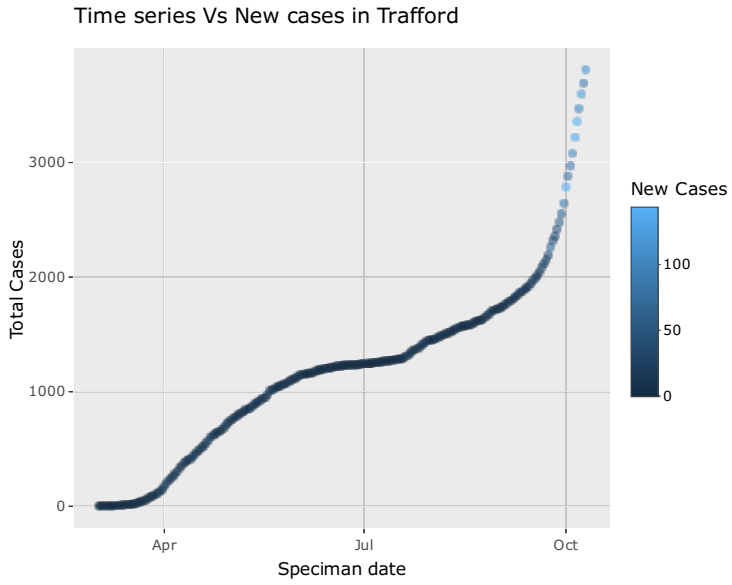
###Visualize the graphical representation of analyzed data

```
#To analyze the new cases with cumulative case with time frame in trafford region

flower_plot <- ggplot(data = trafford_covid) +     # Plotting the data using flower_plot
  geom_point(mapping =                             # Mapping the data
               aes(x = specimen_date,
                   y = cumCasesBySpecimenDate,
```

```
                  color = newCasesBySpecimenDate,
                ),
              alpha = .6
              ) +
  labs(                                          # Adding title and axis table values
    title = "Time series Vs New cases in Trafford",
    x = "Speciman date",
    y = "Total Cases ",
    color = "New Cases"
  )

ggplotly(flower_plot)                            # Calling the flower_plot function
```



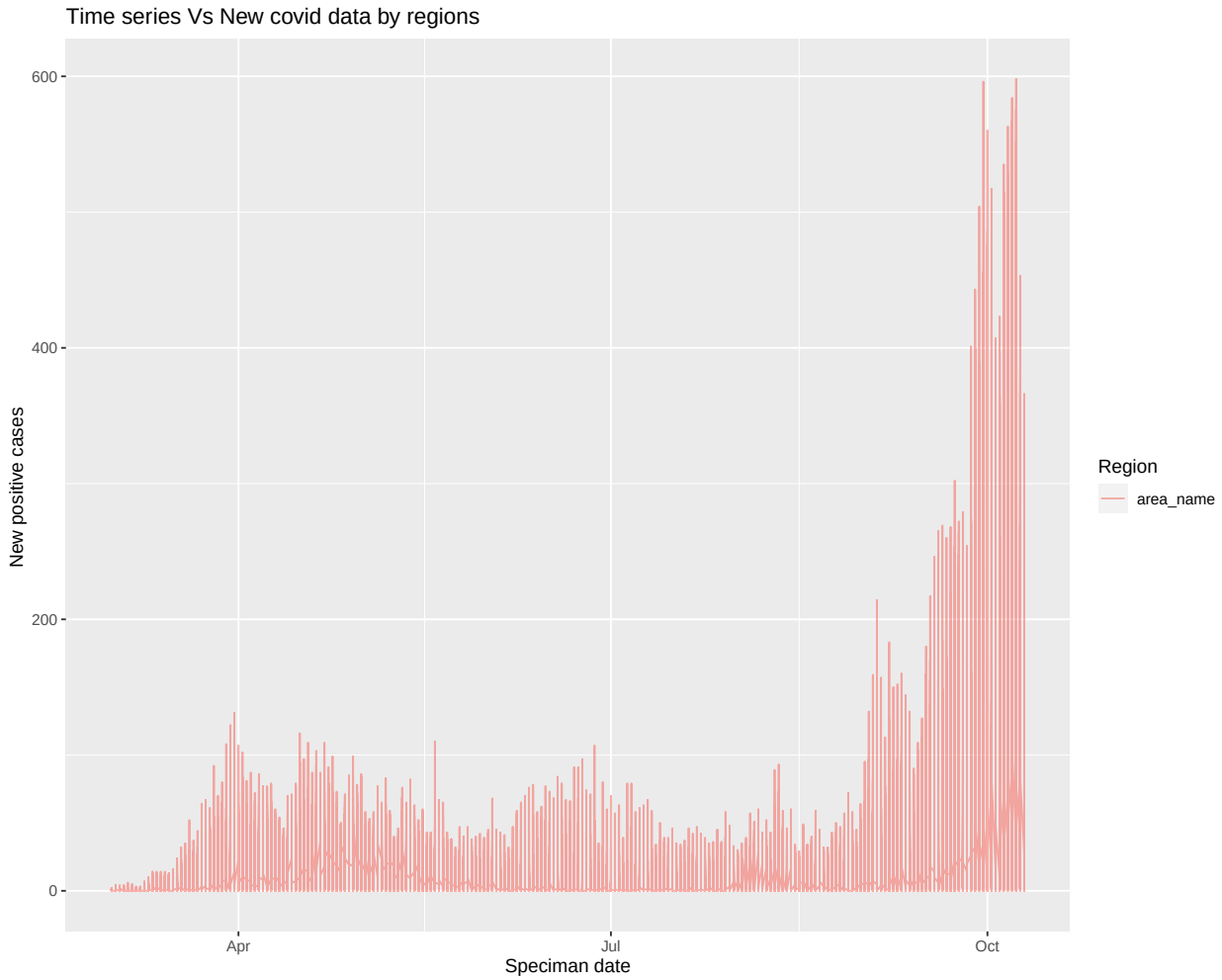

```
# To analyze the new cases with total cases for The united Kingdom region

ggplot(data = covid_population) +              # Plotting the data using ggplot_line plot
  geom_line(
    mapping = aes(x = specimen_date,
                  y = newCasesBySpecimenDate,
                  color = "area_name"
```

```
                ),
  alpha = .6
)+

# Adding title and axis table values

labs(
  title = "Time series Vs New covid data by regions",
  x = "Speciman date",
  y = "New positive cases ",
  color = "Region"
)
```



###The Case study - Covid-19 trends in UK

The UK is well into the second flood of COVID-19,according to temporary information from the Office for National Statistics. The official's of UK Government information show that cases have been rising dramatically since late August, 2020, with increments over all areas in England as of late. Official UK Government information show that cases have been rising dramatically since late August, 2020, with increments over all locales in England as of late. As per the analyzed data, In recent times,there huge pike in the new cases in the United Kingdom comparatively in March'2020 as the first wave. All the regions in the country getting registered more positives cases everyday, specifically, East midlands cities like Nottingham, Leeds, Derby,

Birmingham and so on. We can clearly observe from the above graphical representation, from September the cases has been doubled compared to the past couple of months and keeps climbing on table. Larger cities like Birmingham with most population we have to more worried as new case cases are increasing rapidly, cities like less populous city are registering very low cases, but here too people getting affected.

To conclude, the people should have aware themselves about the Covid-19 impacts. They should follow the government guidelines and stay at indoor, wear face mask and use hand sanitizer, social distancing and so on.