OLLSCOIL TEICNEOLAÍOCHTA
BHAILE ÁTHA CLIATH

TU DUBLIN

TECHNOLOGICAL
UNIVERSITY DUBLIN

School of Computer Science

Data Visualization

SPEC9995

Maksymilian Drzezdzon

C15311966

Degree: TU060/2

Module Coordinator: Emma Murphy

Declaration of Ownership: I declare that the attached work is entirely my own and that all sources have been acknowledged: ☑

Date: 2021/12/05

Data Visualisation Assignment 2

Name: Maks Drzezdzon

ID: C15311966

Class Code: TU060/2

Mode of study: Part-time

Date: 12/08/2021

Dataset: https://www.kaggle.com/kimjihoo/coronavirusdataset

Notebook: https://github.com/Maks-Drzezdzon/Masters-Classes-L-O/blob/master/Data%20Visualisation/assignments/c15311966_assignment_2_notebook.Rmd

Introduction

The goal here is to demonstrate skills developed with R's visualisation libraries supported by theory covered throughout the semester. Having said that I've have more books to read, parched in my library at home.

Problem – Audience – data

Audience/User Story/Problem:

How is covid spreading in South Korea?

The target audience are policy makers in said country. The goal is to provide an overview with the data available assuming this is what I would have been provided. This is to aid the slow down covid infections via visualising the mode that the disease is spreading by and which age groups are spreading them the most.

Given more time I would have wanted to use data from a few countries and compare them.

After reviewing the visualisations given the high elderly population in Gyeongsangbuk-do and those living alone in conjunction with the fact that the highest area of infection is a church this is likely to be the cause of the elderly feeling isolated making them even more inclined to go to church to spend time with people. Safety measure should be taken to tackle this.

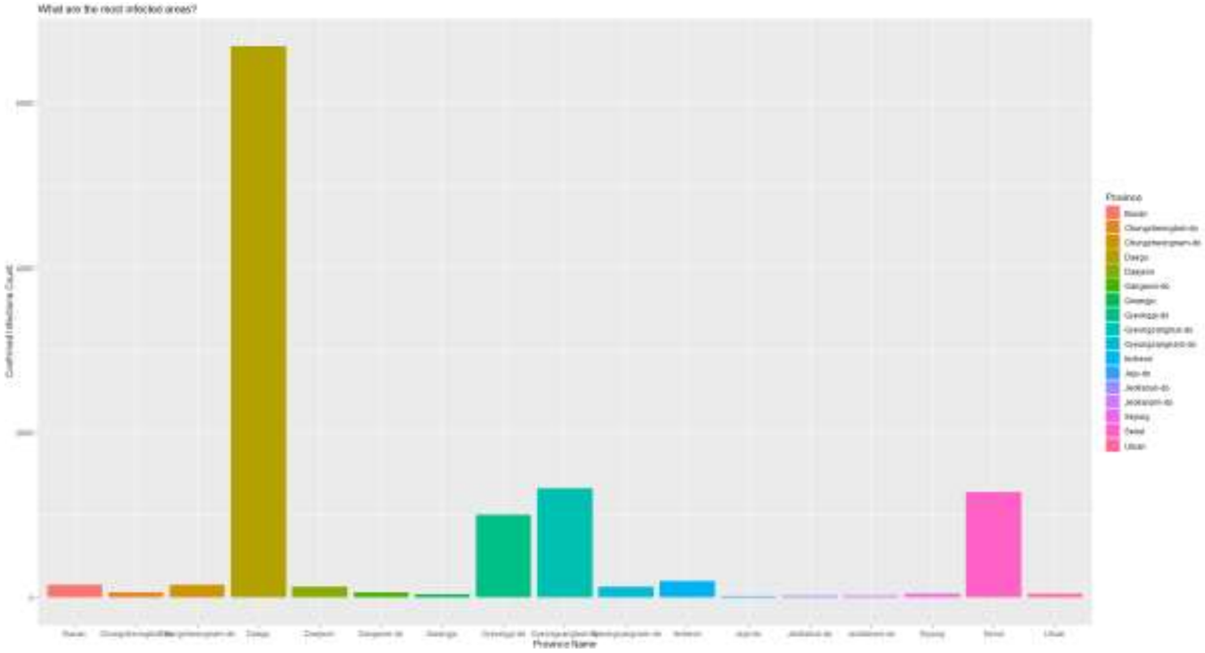
Data:

The dataset consists of 11 files – with most of these files its not possible to interlink them or pull multiple columns into one visualisation because there is no way to map data back to other features. After the 2nd visualisation was created that exhausted any other columns that could be cross referenced.

Visualisations & Previous iterations

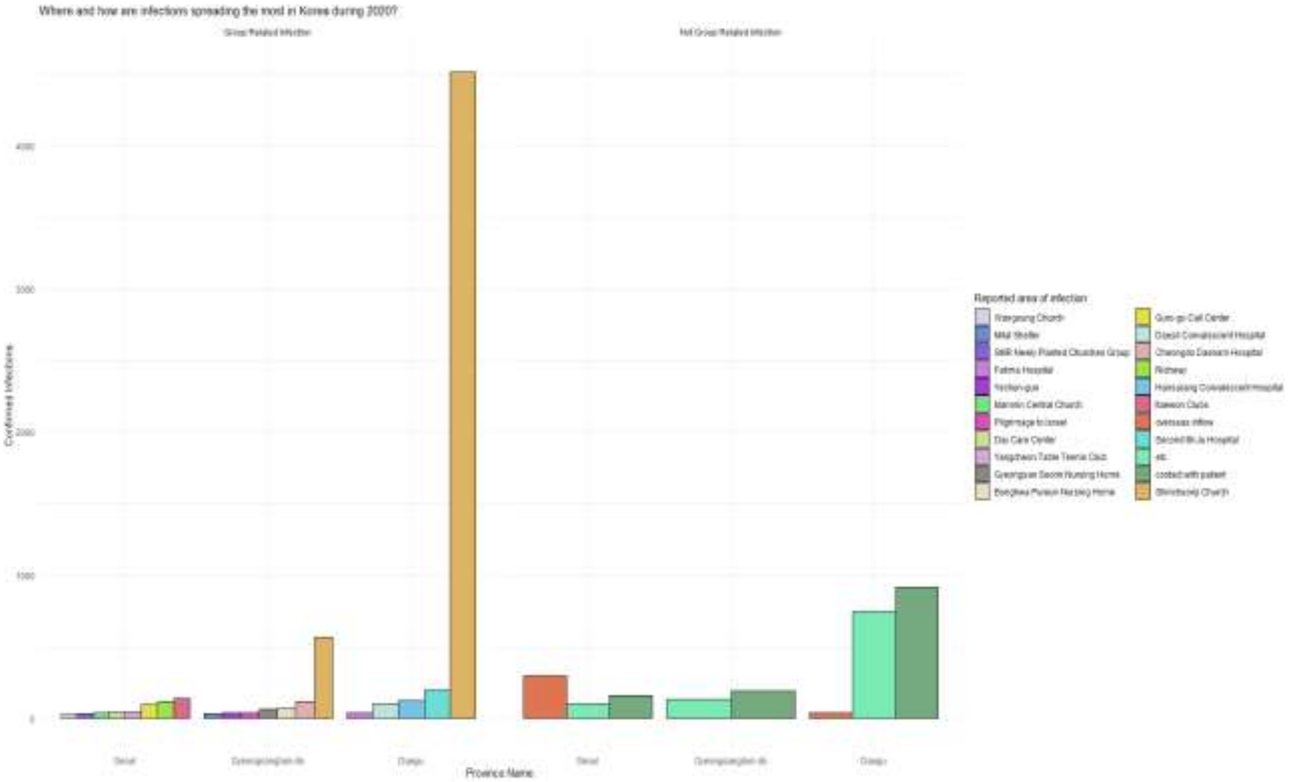
Goals

1. Identify most affected regions (Daegu, Seoul, Gyeonggi-di, Gyeongsangbuk-do)



- This was then used as a first iteration to create the first visualisation
- Other areas have little data or too many missing values to work with, hence the top 4 were chosen
- This is also an issue that the dataset doesn't allow much cross checking/linking of columns/fields to other datasets

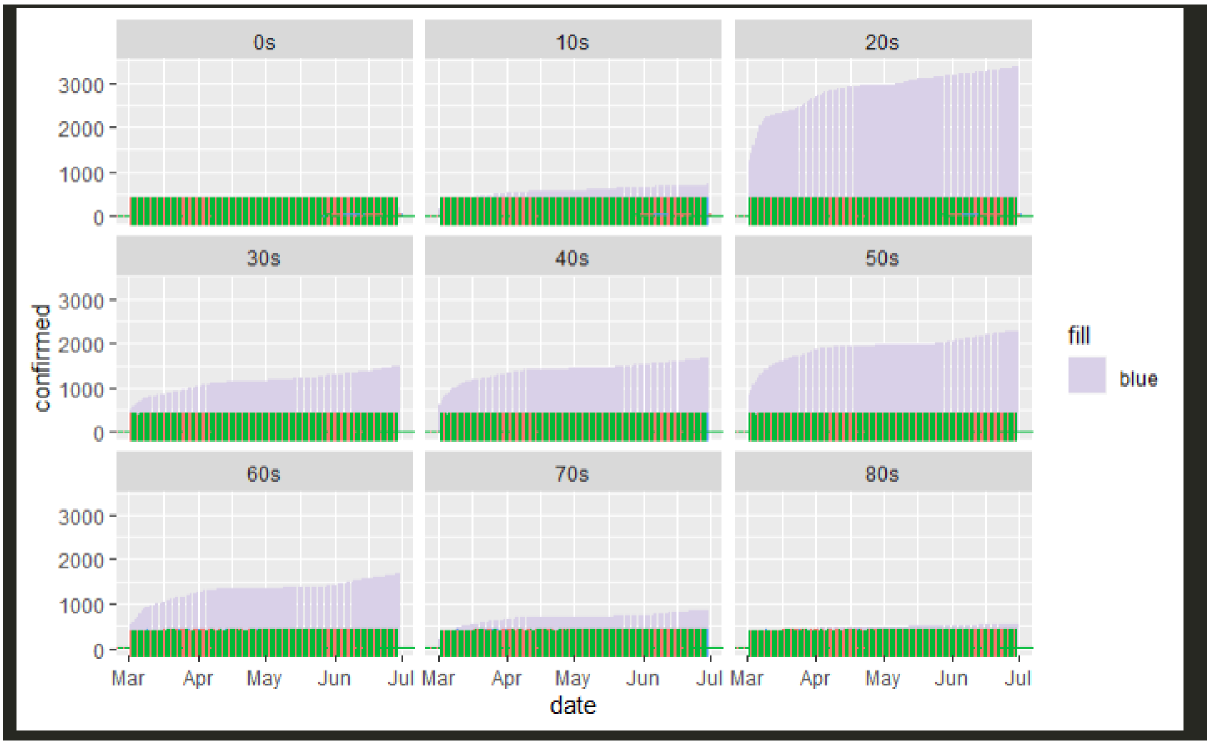
1st visualisation for assignment



The goal here was to identify how the virus is spreading its biggest cities and segregate them by group vs non-group infection – this was achieved by replacing Boolean values with the titles above. It is not clear what 'etc' is. Something that immediately stands out here is that infections from a church in Daegu vastly overtake any other mode of infection combined. Within the same province there seems to have been a break out in the hospital as its documented under group infection. Reviewing and improving protocols in said hospital would be ideal to prevent further spread.

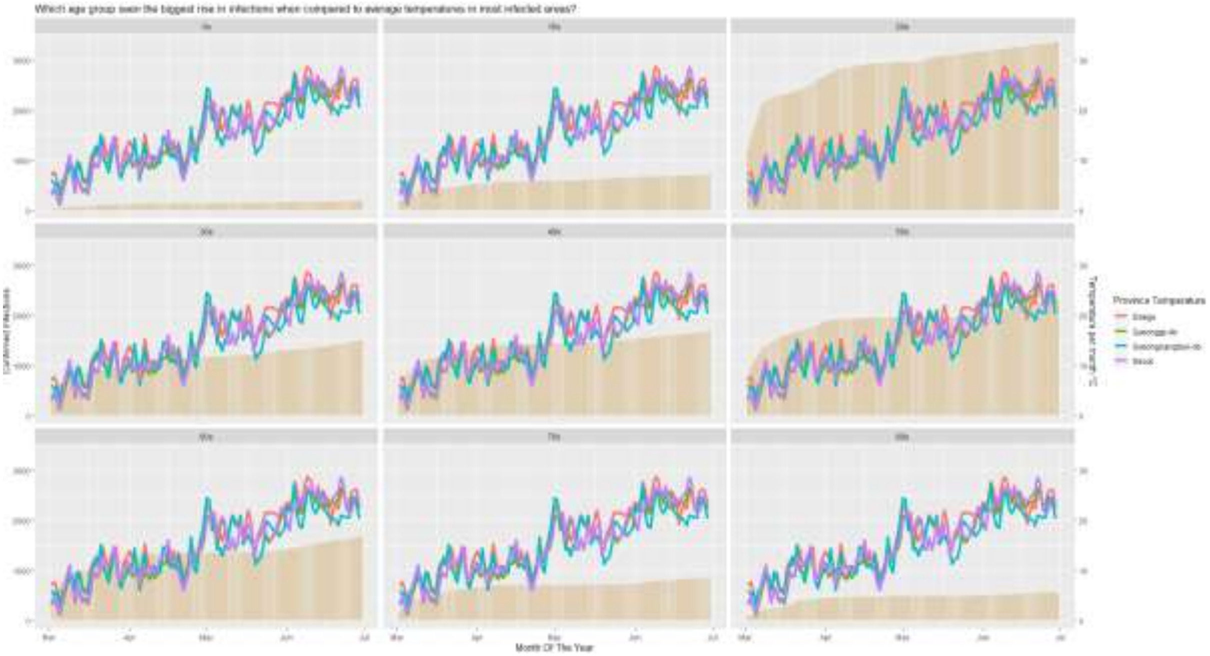
Limiting capacity if outright closure isn't possible would be an obvious recommendation. Leaning over to the non-group infections contact with patient seems to be the top mode of infection along with 'etc' ending with plane travel in Seoul which seems to be the least impacted province from the top 4 most infected, at least when compared to the other 2.

2. Drill down into age groups and temperatures (there isn't a way to map age groups back to specific regions unfortunately, hence a general overview of the country is used)
    - The first iteration didn't work out so well where I tried to map both weather records with cases per age group



After reviewing the data, I tried doing the same thing but rather for the 4 selected regions, this was then layered on top of the original graph where the temperature values were upscaled by 100 when compared to the infections. Some other early work was to use a library called patchwork that combines graphs together however I'm not sure if id loose marks for that because they're not combined into one visualisation like the one below but rather just on the same pane.

2nd visualisation for assignment



The idea here was to isolate data from each age group and map it back to each province and compare it to weather conditions over the 5/6-month period – this was not entirely possible - instead the overall confirmed case count for the whole country was segregated by age group and combined with a combination of line plots constructed from the accompanying weather dataset. They were then both filtered to have overlapping dates so they would share the x-axis.

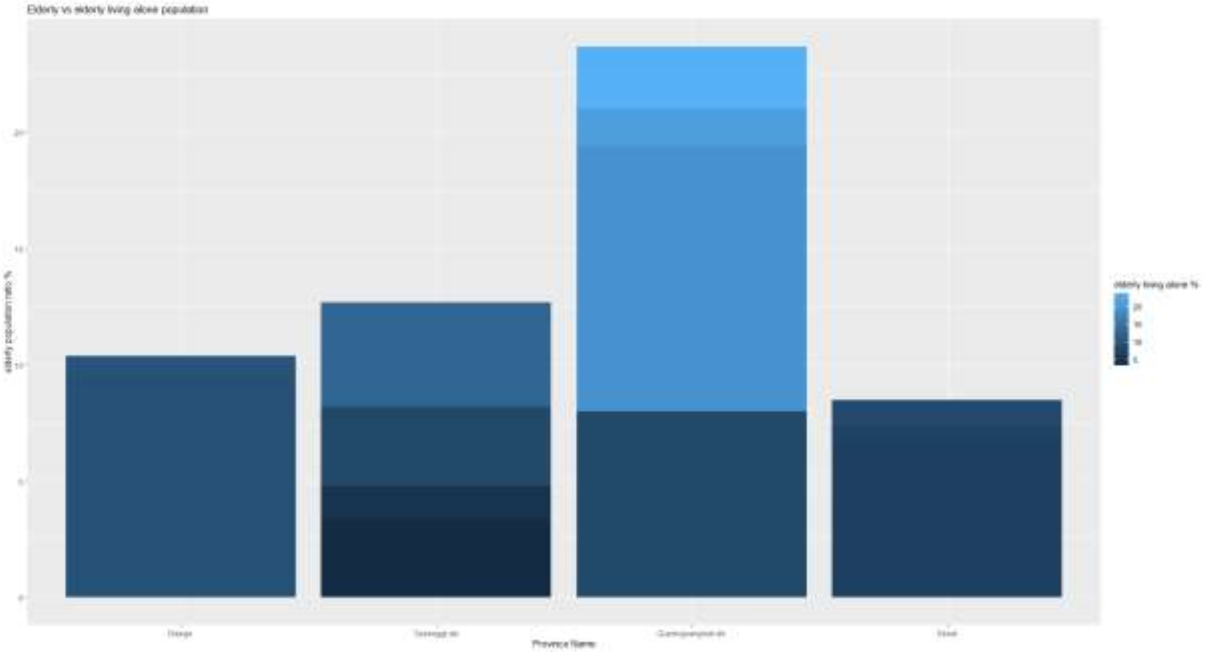
The purpose of this was to see which groups were carrying the infection the most but also if weather had any significant role to play as one would imagine. This would potentially unveil how did each age group respond to both information dispensed by their media and current conditions – the point I'm trying to allude to is, did the 20-year-olds have a linear uptick in infections, which they did or did any older groups show a rise and fall despite a variable that would stimulate the spread of the disease, that being weather in this case as it was the only one that offered an ability to be mapped to the other dataset.

Personally, this visualisation could have been better but due to the dataset offering limited features and little space for more engineered approaches. Even though there is a linear relationship between temperatures and covid cases rising across most groups this isn't the precise insight I was going for – albeit the same formula would be used however in order to properly discern the information a precise separation between age groups and regions would be required and this wasn't available. I wanted to attempt this regardless to examine the relationship.

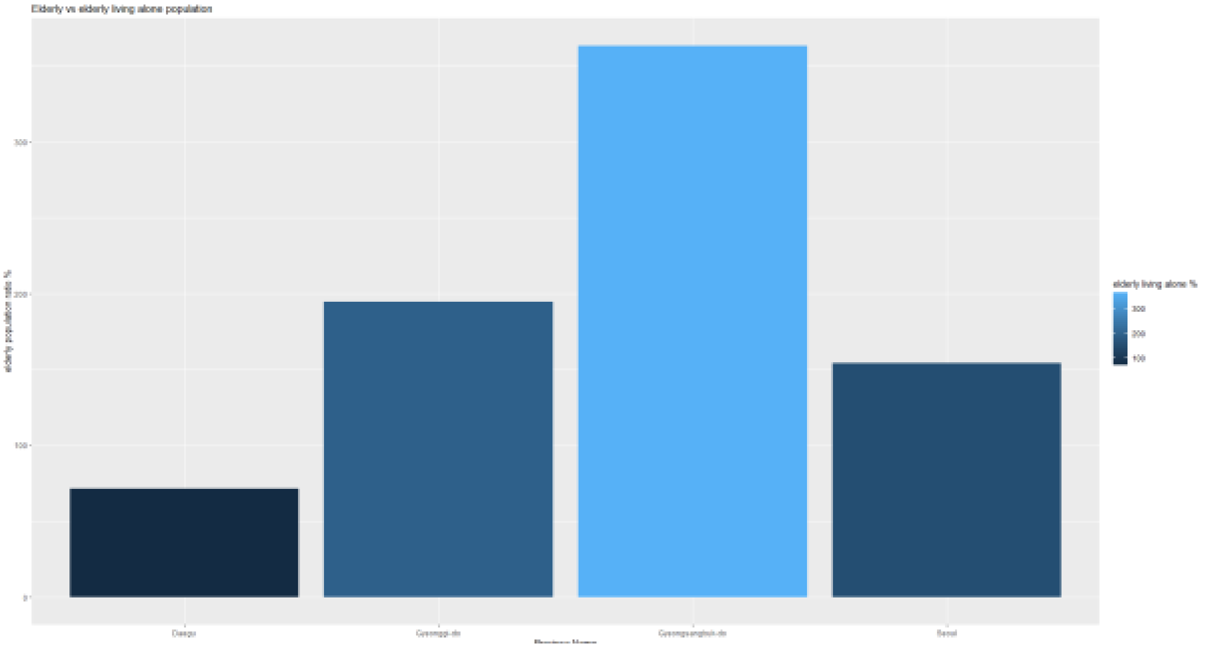
Furthermore, a similar analysis would have been conducted if there was a way to map age group infections to regions and compare the amount of early adulthood education facilities which are provided. This idea was scrapped because just as the visualisation above, it wouldn't have given the desired output due to dataset being limited.

3. Elderly population/Elderly living alone in most affected areas

3rd Visualisation for assignment



Alternative version



The rational behind this is that one uses a combination of values whereas the other maps the whole province per bar which shows a much better split between the size of chunks/groups of elderly living alone. As mentioned above – there isn't a way to map data back to a per province basis when looking at more interesting tables. The only one where this was possible to an extent, was the second visualisation because the date could be used as an x axis that could combine and overlay multiple graphs.

Code

```r
# Data setup
case_df = read.csv("/Users/Grim/Documents/GitHub/covid19/Case.csv", sep= ',' , header=T )
# subset(df_here, select = -c(cols_to_remove))
# pre-setup
# changing color palette setup
n_color = 22
set.seed(2643598)
palette = distinctColorPalette(n_color)
# test palette
# http://www.sthda.com/english/wiki/ggplot2-colors-how-to-change-colors-automatically-and-manually
# scale_fill_manual() for box plot, bar plot, violin plot, etc
# scale_color_manual() for lines and points
pie(rep(1, n_color),
    col=palette,
    radius=1,
    main="test palette")
# data prep
# find most infected areas
colnames(case_df)
most_infected_areas = subset(case_df, select = -c(case_id, latitude, longitude))
most_infected_areas
most_infected_areas_sum = sqldf("select province as Province, sum(confirmed) as 'Confrimed Infections' from most_infected_areas group by province")
most_infected_areas_sum_plot = ggplot(data = most_infected_areas_sum, aes(Province, `Confrimed Infections`, fill=as.factor(Province))) +
  geom_col(position="dodge") +
  labs(x="Province Name", y="Confrimed Infections Count", title="What are the most infected areas?", fill="Province")
most_infected_areas_sum_plot
# pick out top 4 hot spots
hotspots_df = sqldf("select * from most_infected_areas where province == 'Daegu' or province == 'Seoul' or province == 'Gyeonggi-di' or province == 'Gyeongsangbuk-do'")
hotspots_df = sqldf("select * from hotspots_df where confirmed >= 30")
hotspots_df
colnames(hotspots_df)[4] = "Infected_In_Location"
hotspots_df
# Group vs non group infections
pie(rep(1, n_color),
    col=palette,
    radius=1,
    main="test")
plot_for_assign_requirements_1 = ggplot(data=hotspots_df, aes(x=province, y=confirmed, fill=Infected_In_Location )) +
  scale_fill_manual(values = c(palette)) +
  geom_bar(stat="identity", color="black", position=position_dodge())+
  theme_minimal() +
  labs(x="Province Name", y="Confirmed Infections", title="Where are infections spreading the most?", fill="Reported area of infection")
plot_for_assign_requirements_1
```

```r
weather_df_filtered = sqldf("select * from weather_df where province == 'Daegu' or province == 'Seoul' or province == 'Gyeonggi-di' or province == 'Gyeongsangbuk-do'")

tmp_weather = drop_na(weather_df_filtered)

# filter the dates to match confrimed_cases_age_group_timeline_df
tmp_weather = subset(tmp_weather, date > "2020-03-01")
class(tmp_weather$date)

tmp_weather$date = as.Date(tmp_weather$date, format = "%Y-%m-%d")
tmp_weather

weather_plot_3_hotspots = ggplot() +
  geom_line(data=tmp_weather, aes(x=date, y=avg_temp, group=province, colour=province),
stat="identity") + labs(x="date", y="avg_temp",
                      title="which groups spread it the most and does weather have an effect?",
fill="province")

confrimed_cases_age_group_timeline_df =
read.csv("/Users/Grim/Documents/GitHub/covid19/TimeAge.csv", sep= ',' , header=T )

# whos spreading them?

colnames(confrimed_cases_age_group_timeline_df)[3] = "age_group"
confrimed_cases_age_group_timeline_df$date = as.Date(confrimed_cases_age_group_timeline_df$date,
format = "%Y-%m-%d")

plot_for_assign_requirements_2 = ggplot() +
  geom_bar(data=confrimed_cases_age_group_timeline_df, aes(x=date, y=confirmed,
color="#A4A4A4"), stat="identity", color="black", position=position_dodge())+
  facet_wrap(~age_group) +
  labs(x="Month of the year", y="Confirmed Infections", title="Which age group got infected the
most during the months of March to July?",
       fill="")

# patchwork functionality allows to add plots to display them side by side
# however im not sure if id loose marks for this
# library(patchwork)
# plot_for_assign_requirements_2 + weather_plot_3_hotspots

ggplot() +
  geom_bar(data=confrimed_cases_age_group_timeline_df, aes(x=date, y=confirmed, fill="blue" ),
stat="identity") +
  facet_wrap(~age_group) +
  scale_fill_manual(values = c(palette)) +
  geom_point(data=tmp_weather, aes(x=date, y=avg_temp, group=province, colour=province),
position = position_dodge(width = 0.9), shape=3, size=5, show.legend=FALSE)

# now combine iterations to 1 viz
colnames(tmp_weather)[2] = "Province Temperature"

pie(rep(1, n_color),
    col=palette,
    radius=1,
    main="test palette")

ggplot() +
  geom_bar(data=confrimed_cases_age_group_timeline_df, aes(x=date, y=confirmed, fill = "blue"),
size = 1, stat="identity", position="stack", show.legend=FALSE) +
  scale_fill_manual(values = c(palette)) +
  facet_wrap(~age_group) +
  labs(x="Month Of The Year", y="Confirmed Infections", title="In which age group did infections
increase when compared to average temperatures in most infected areas?", fill=" ") +
  geom_line(data=tmp_weather, aes(x=date, y=100*avg_temp, group=`Province Temperature`,
colour=`Province Temperature`), size=1.5)+
  scale_y_continuous(sec.axis = sec_axis(~./100, name = "Temperature per month °C"))
```

```
region_df = read.csv("/Users/Grim/Documents/GitHub/covid19/Region.csv", sep= ',' , header=T )
case_df = read.csv("/Users/Grim/Documents/GitHub/covid19/Case.csv", sep= ',' , header=T )

most_infected_areas_sum_plot = ggplot(data = most_infected_areas_sum, aes(Province, `Confrimed
Infections`, fill=as.factor(Province))) +
  geom_col(position="dodge") +
  labs(x="Province Name", y="Confrimed Infections Count", title="What are the most infected
areas?", fill="Province")
most_infected_areas_sum_plot

region_df
```

```
tmp_df = sqldf("select province, sum(elderly_population_ratio) as elderly_population_ratio,
sum(elderly_alone_ratio) as elderly_alone_ratio from region_df where province == 'Daegu' or
                                        province == 'Seoul' or province == 'Gyeonggi-do' or
province == 'Gyeongsangbuk-do' group by province")

tmp_df

tmp_df_2 = sqldf("select province, elderly_population_ratio, elderly_alone_ratio from region_df
where province == 'Daegu' or
                                        province == 'Seoul' or province == 'Gyeonggi-do' or
province == 'Gyeongsangbuk-do'")

tmp_df_2

living_alone_elderly_population = ggplot(data = tmp_df_2, aes(province, elderly_alone_ratio ,
fill=elderly_alone_ratio)) +
  geom_col(position="dodge") +
  labs(x="Province Name", y="elderly population ratio %", title="Elderly vs elderly living alone
population", fill="elderly living alone %")
living_alone_elderly_population
```