Final COVID 19 - Norway

2022-09-02

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

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

Introduction

Questions of interest

I want to explore how temperature affects incidence of new Covid 19 cases in Oslo, Norway. My expectations before analyzing the data is that until certain point, lower temperatures will increase the incidence of new cases of Covid 19, but further decrease will result in less contact between people (avoiding going out at very cold weather) and less infections.

I am aware that the number of new infections depends on many other factors.

Data description

For this purpose I am going to use two datasets:

1. A confirmed cases dataset from Johns Hopkins University
2. A dataset from the Meteorologisk Institutt about average temperature in Oslo, Norway

Import data

Brief description:
For this analysis I am using data from the CSSE github repository Time Series about Confirmed cases -

Global https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series

Temperature statistics has been retrieved by a web service provided by Meteorologisk Institutt https://seklima.met.no/observations/ and uploaded to my own github repo

```
url_in <- "https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_covid_19_time_s
global_cases <- read_csv(url_in, show_col_types = FALSE)

url_in_met <- "https://raw.githubusercontent.com/novembererfin/DTSA5301/main/2022-09-02_temperature_Oslo
oslo_temp_read <- read_csv(url_in_met, show_col_types = FALSE)
```

Some Data Definitions

Cases: Number of cumulative cases since the start of the registration
Temperature: Average daily temperature in Celsius registered in Blindern Station, Oslo, Norway
New_Cases: Differences between number of cases that date and the day before
Rel_New_Cases: relation between New_Cases and the sum of New_Cases the preceding 10 days (an aproximate measure of active contagious cases)

Reformat and clean up data

```
norway_cases <- global_cases %>%
  filter(global_cases$`Country/Region` == "Norway")

norway_cases <- norway_cases  %>%
  select(-c("Province/State", "Long", "Lat", "Country/Region"))

norway_cases <- pivot_longer(norway_cases,cols = everything())
colnames(norway_cases) <- c("Dates", "Cases")
norway_cases <- norway_cases %>%
  mutate(Dates = mdy(Dates))

oslo_temp <- oslo_temp_read  %>% select(-c("Navn", "Stasjon"))
oslo_temp <- oslo_temp %>%
  mutate(`Tid(norsk normaltid)` = dmy(`Tid(norsk normaltid)`))
colnames(oslo_temp) <- c("Dates", "Temperature")


cases_temp <- merge(x = norway_cases, y = oslo_temp, by = "Dates")
cases_temp <- cases_temp %>%
  mutate(New_Cases = Cases - lag(Cases))

cases_temp$New_Cases[is.na(cases_temp$New_Cases)] <- 0

cases_temp <- cases_temp %>%
  mutate(Rel_New_Cases = New_Cases / (Cases - lag(Cases, n=10)))


cases_temp$Rel_New_Cases[is.na(cases_temp$Rel_New_Cases)] <- 0
```

```
summary(cases_temp)
```

```
##      Dates                Cases          Temperature       New_Cases
##  Min.   :2020-01-22   Min.   :      0   Min.   :-11.300   Min.   :    0.00
##  1st Qu.:2020-09-16   1st Qu.:  12534   1st Qu.:  2.400   1st Qu.:   99.25
##  Median :2021-05-12   Median : 118155   Median :  8.150   Median :  313.50
##  Mean   :2021-05-12   Mean   : 386795   Mean   :  8.397   Mean   : 1530.66
##  3rd Qu.:2022-01-05   3rd Qu.: 426766   3rd Qu.: 15.000   3rd Qu.:  716.25
##  Max.   :2022-09-01   Max.   :1460246   Max.   : 24.600   Max.   :26109.00
##  Rel_New_Cases
##  Min.   :0.00000
##  1st Qu.:0.07180
##  Median :0.09686
##  Mean   :0.10102
##  3rd Qu.:0.12408
##  Max.   :1.00000
```

Analysis

Data at a glance

Some key values:

```
max_new_cases <- cases_temp[which.max(cases_temp$New_Case),]
paste("Average number of case per day: ", mean(cases_temp$Cases))
```

```
## [1] "Average number of case per day:  386795.44129979"
```

```
paste("Average number of new cases per day: ", mean(cases_temp$New_Cases))
```

```
## [1] "Average number of new cases per day:  1530.65618448637"
```

```
paste("Maximum number of new cases per day: ", max(cases_temp$New_Cases))
```

```
## [1] "Maximum number of new cases per day:  26109"
```

```
paste("Date with maximum number of cases: ", max_new_cases$Dates)
```

```
## [1] "Date with maximum number of cases:  2022-02-08"
```

```
paste("Highest average temperature measured in this period: ", max(cases_temp$Temperature))
```

```
## [1] "Highest average temperature measured in this period:  24.6"
```

```
paste("Average temperature measured in this period: ", mean(cases_temp$Temperature))
```

```
## [1] "Average temperature measured in this period:  8.39716981132075"
```

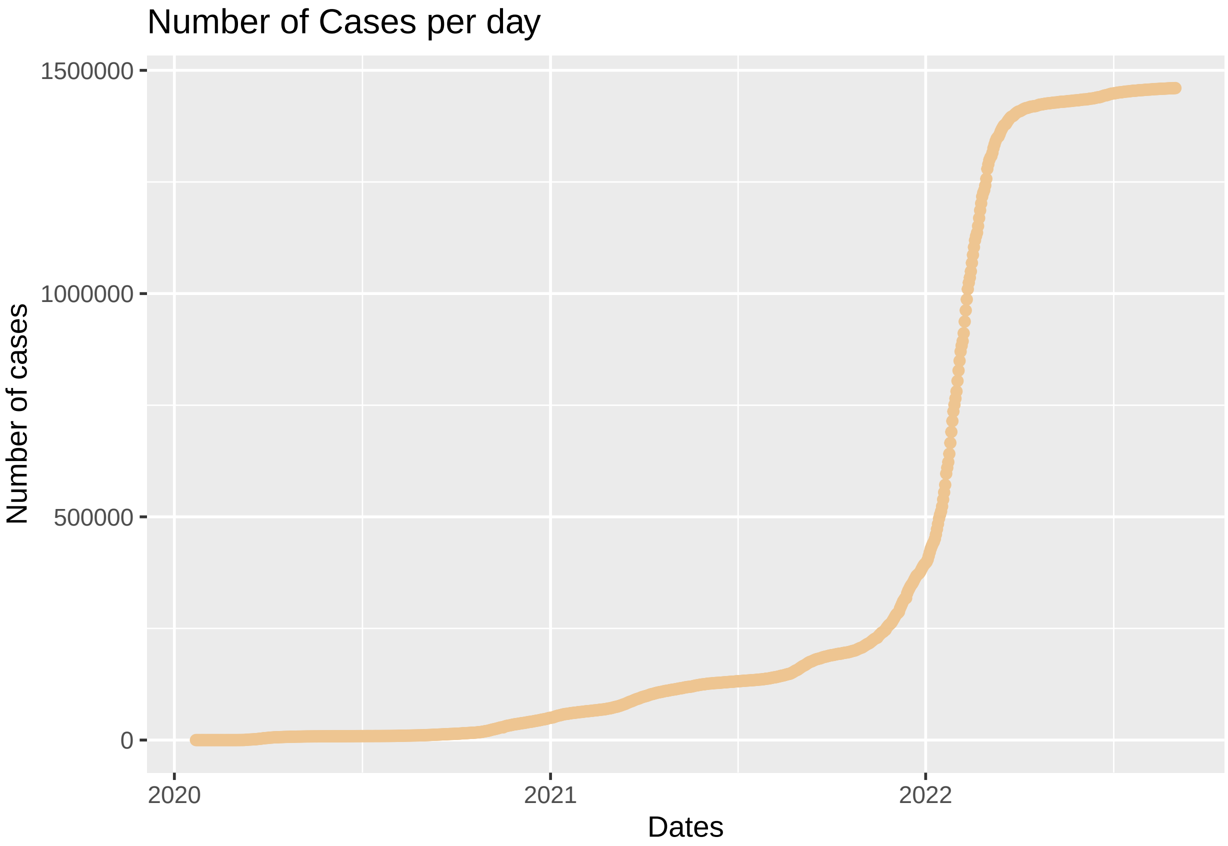
```
paste("Minimum average temperature measured in this period: ", min(cases_temp$Temperature))
```

```
## [1] "Minimum average temperature measured in this period:  -11.3"
```

Visualizations

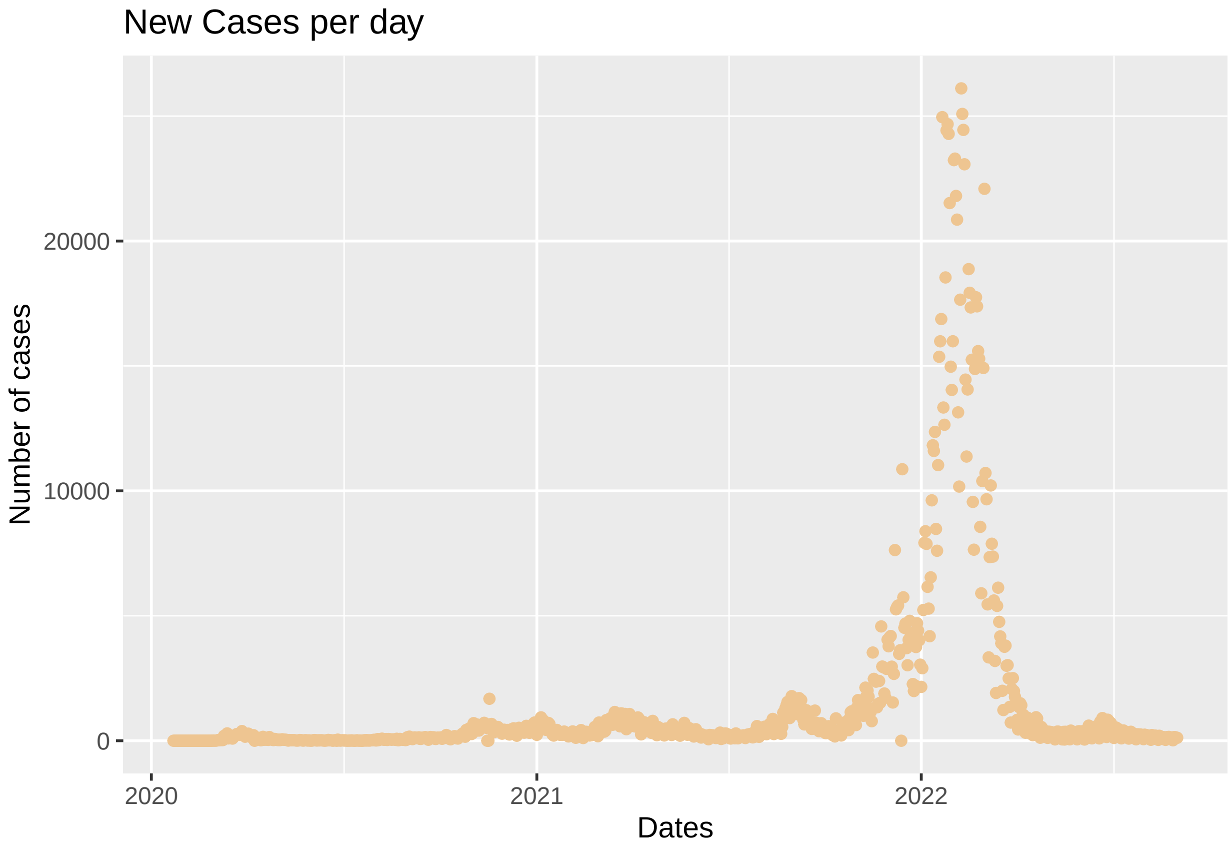
Confirmed cases per day (cumulative)

```
plot1 <- ggplot(cases_temp, mapping = aes(x = Dates)) +
  geom_point(mapping = aes(y = Cases), color = "burlywood2") +
  labs(title = "Number of Cases per day", x = "Dates", y = "Number of cases")
plot1
```



Confirmed new cases per day

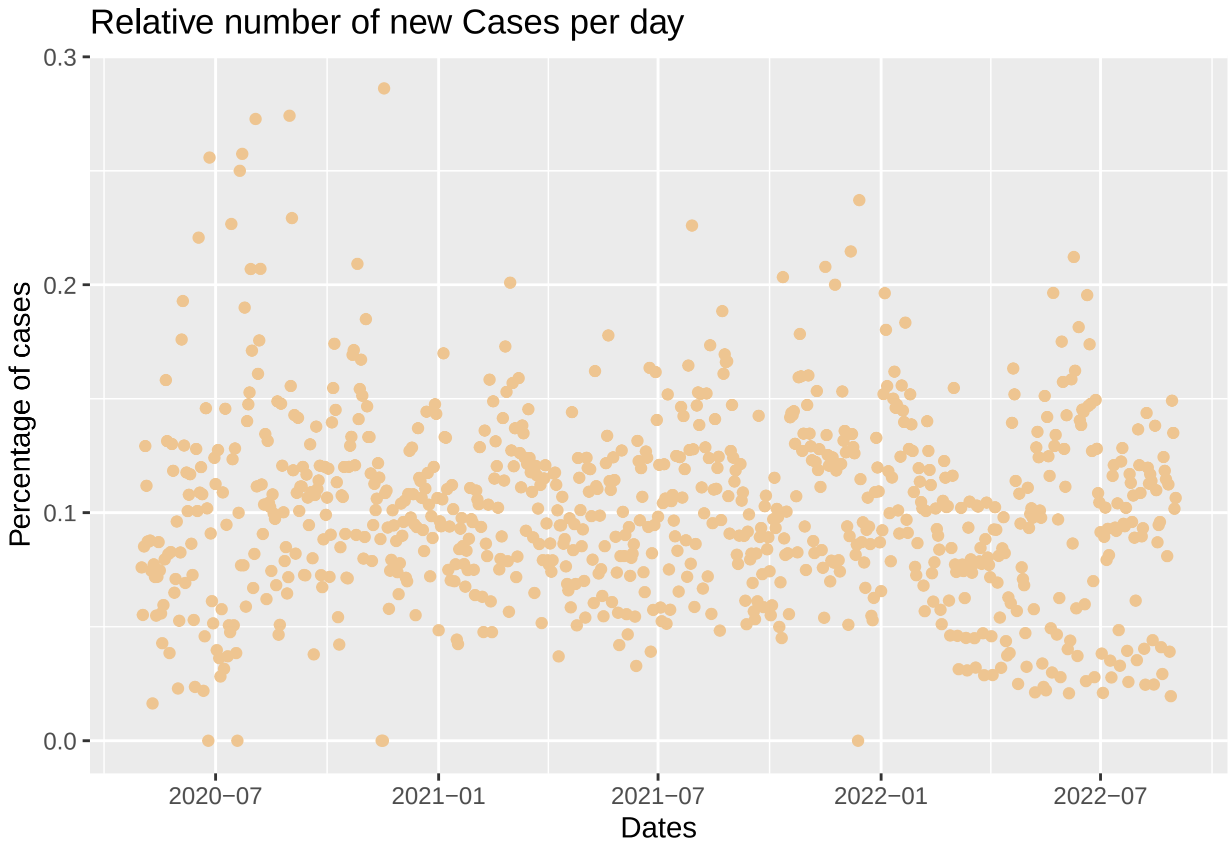
```
plot2 <- ggplot(cases_temp, mapping = aes(x = Dates)) +
  geom_point(mapping = aes(y = New_Cases), color = "burlywood2") +
  labs(title = "New Cases per day", x = "Dates", y = "Number of cases")
plot2
```



Note that Norway removed all Covid-related measures on February 12, 2022. This may help explain the major spike during the beginning of 2022.

Confirmed new cases per day relative to the total number of new cases the last 10 days

```
plot_cases_temp <- cases_temp %>%
  filter(cases_temp$Dates >= "2020-05-01")
plot3 <- ggplot(plot_cases_temp, mapping = aes(x = Dates)) +
  geom_point(mapping = aes(y = Rel_New_Cases), color = "burlywood2") +
  labs(title = "Relative number of new Cases per day", x = "Dates", y = "Percentage of cases")
plot3
```



Making a model

My hypothesis is that temperature is an important factor in promoting new cases. I use a linear model between new cases and temperature.

```
model <- lm(Rel_New_Cases ~ Temperature, data = plot_cases_temp)
summary(model)
```

```
##
## Call:
## lm(formula = Rel_New_Cases ~ Temperature, data = plot_cases_temp)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.101579 -0.025900 -0.002368  0.023017  0.185506
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.001e-01  2.131e-03   46.98   <2e-16 ***
## Temperature 6.614e-05  1.787e-04    0.37    0.711
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0413 on 852 degrees of freedom
## Multiple R-squared:  0.0001607,  Adjusted R-squared:  -0.001013
## F-statistic: 0.137 on 1 and 852 DF,  p-value: 0.7114
```

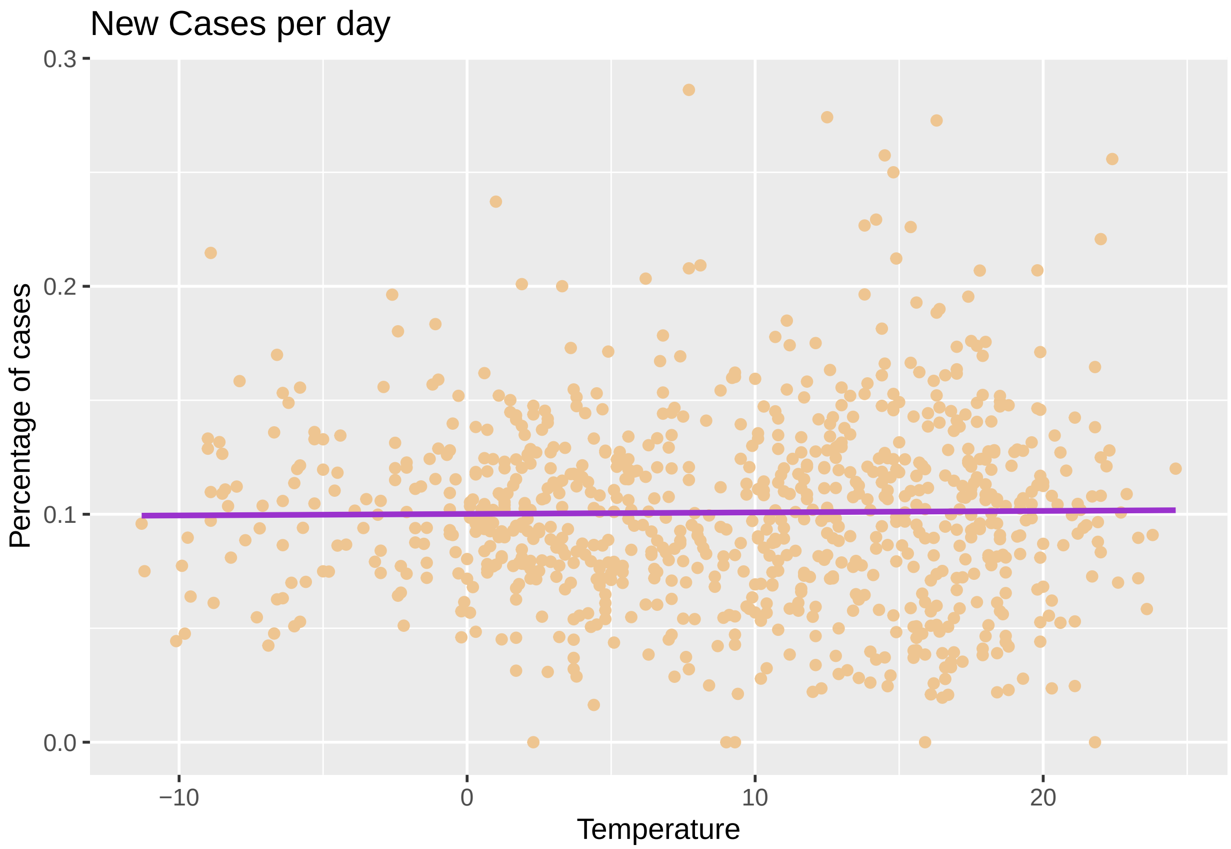
Predictions: correlation (purple) vs real observations (light orange)

```
paste("Correlation coefficient: ", round(cor(plot_cases_temp$Temperature, plot_cases_temp$Rel_New_Cases
```

```
## [1] "Correlation coefficient:  0.01"
```

```
plot4 <- ggplot(plot_cases_temp, mapping = aes(x = Temperature)) +
  geom_point(mapping = aes(y = Rel_New_Cases), color = "burlywood2") +
  geom_smooth(mapping = aes(y = Rel_New_Cases), method=lm, se = FALSE, color = "darkorchid3") +
  labs(title = "New Cases per day", x = "Temperature", y = "Percentage of cases")
plot4
```

```
## `geom_smooth()` using formula 'y ~ x'
```



From the graph, we can see that temperature is not correlated to the relative number of new cases

Because temperature was not correlated I want to explore another factor - Mobility changes. I am using Region Mobility Reports from Google for Norway for 2020, 2021 and 2022.

I will be using changes in mobility in transit stations and workplaces

```
url_in_mob <- "https://github.com/novembererfin/DTSA5301/raw/main/2020_NO_Region_Mobility_Report.csv"
no_mob_read_2020 <- read_csv(url_in_mob, show_col_types = FALSE)

url_in_mob <- "https://github.com/novembererfin/DTSA5301/raw/main/2021_NO_Region_Mobility_Report.csv"
no_mob_read_2021 <- read_csv(url_in_mob, show_col_types = FALSE)

url_in_mob <- "https://github.com/novembererfin/DTSA5301/raw/main/2022_NO_Region_Mobility_Report.csv"
no_mob_read_2022 <- read_csv(url_in_mob, show_col_types = FALSE)

no_mob_read <- rbind(no_mob_read_2020, no_mob_read_2021)
no_mob_read <- rbind(no_mob_read, no_mob_read_2022)
```

Some more definitions:
Mob_External: Mean of transit station mobility changes and workplaces mobility changes

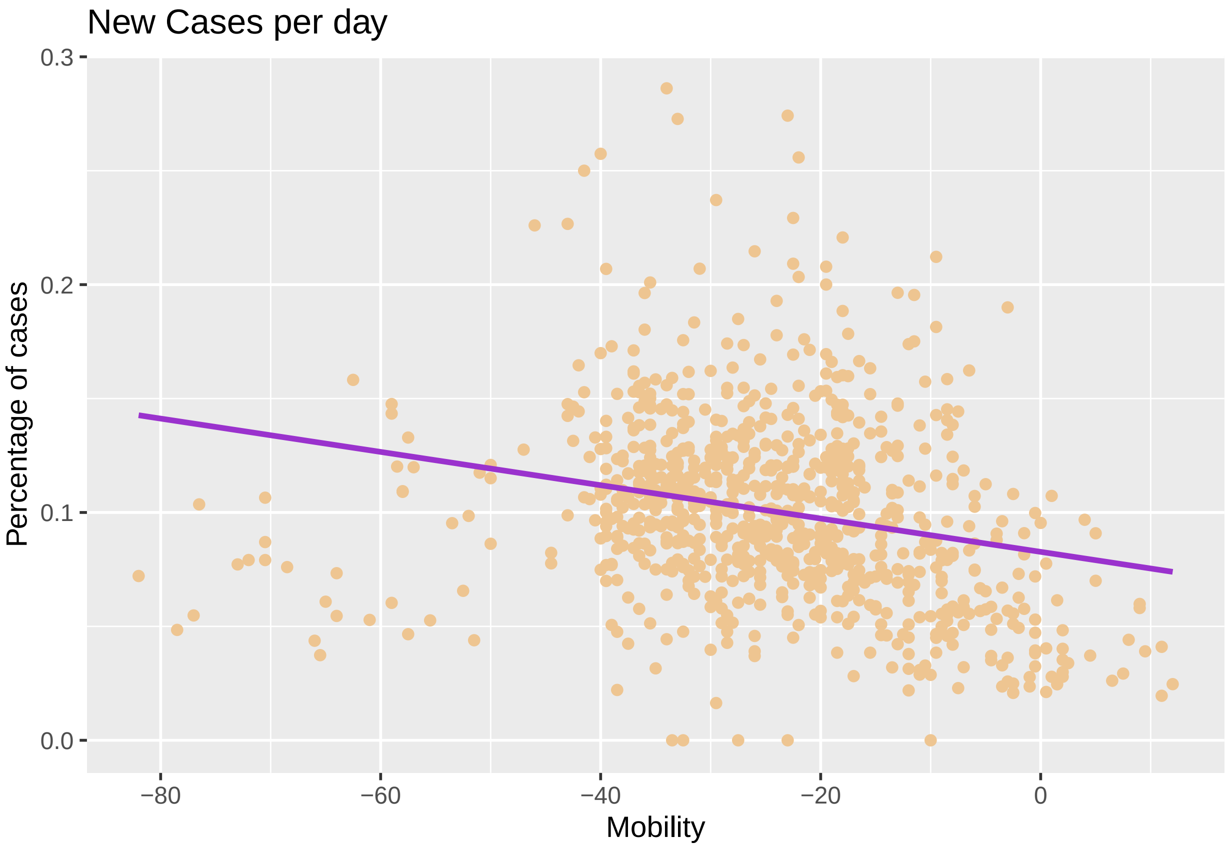
```
no_mob <- no_mob_read %>%
  filter(is.na(sub_region_1))
no_mob <- no_mob %>%
  mutate(date = ymd(date))
no_mob <- no_mob %>%
  select(c("date", "transit_stations_percent_change_from_baseline","workplaces_percent_change_from_basel

colnames(no_mob) <- c("Dates", "Mob_Transit", "Mob_Workplaces")

no_mob <- no_mob %>%
  mutate(Mob_External = (Mob_Workplaces + Mob_Transit)/2)

plot_cases_temp_mob <- merge(plot_cases_temp, no_mob, by = "Dates")
```

```
paste("Correlation coefficient: ", round(cor(plot_cases_temp_mob$Mob_External, plot_cases_temp_mob$Rel_
```

```
## [1] "Correlation coefficient:  -0.24"
```

```
plot5 <- ggplot(plot_cases_temp_mob, mapping = aes(x = Mob_External)) +
  geom_point(mapping = aes(y = Rel_New_Cases), color = "burlywood2") +
  geom_smooth(mapping = aes(y = Rel_New_Cases), method=lm, se = FALSE, color = "darkorchid3") +
  labs(title = "New Cases per day", x = "Mobility", y = "Percentage of cases")
plot5
```

```
## `geom_smooth()` using formula 'y ~ x'
```



There is a weak negative correlation between mobility in transit stations and workplaces and new cases. Initially this was not expected, but more can be explained in the Bias section.

Bias

Some bias that I have identified:

1. The temperature data is from one station in Oslo while the rest of the data is about the whole country. Even Oslo is the biggest city in Norway, still can be other weather conditions in other cities. This is a result of constrains in the data sources used. In real life one could have used a dataset from the Norwegian Folkehelse Institutt (fhi.no).
2. I haven't taken into account the time that goes between the person is infected, the appearance of the symptoms, testing and getting the results. In best case scenario it could be around 7 days, but more realistic, maybe 10 days.
3. In my alternative model with mobility, I used changes in workplace and transit station mobility as proxy measures for social contact. This may ignore the fact that usually higher mobility during summer months is often related to outdoors activities with lower risk of infection while mobility during the winter could result on the opposite.
4. Last: when analyzing the effect of temperature I should have chosen a period of time that starts and ends the same day and month. My data has more samples from the summer months than winter months.

Conclusions

My conclusion is that temperature has not been a significant factor for the relative changes in the number of new cases of Covid 19 in Norway. This seems to be counter-intuitive since we can expect similar behavior than other respiratory diseases (influenza, colds, pnumonia, etc) that usually thrives during winter months. One possible explanation may be that sanitary regulations, social distancing, pubs/bars lockdown has a much bigger role are more relevant for infection that temperature/air conditions alone.

Other explanation can be that data scope is too wide and I should divide it between periods with strict sanitary measures i.e. from March-June 2020, November 2020 - March 2021 and periods without them as from February 2022-Today.

Last I could have taking into account the dominant virus variant since epidemiologically they behaved different from the original Wuhan-variant, Delta and then Omicron.