Data Analysis and Visualization in R - Bonus Project by team 142

This is an R Markdown Notebook for the bonus project of the course "Data Analysis and Visualization in R".

Authors: Efthymia Kostaki (03740037) & Ertugela Doçi (03727843)

We start by importing the necessary libraries.

```
library(data.table)
library(ggrepel)
```

```
## Loading required package: ggplot2
```

```
library(magrittr)
library(tidyr)
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:magrittr':
##
##     extract
```

```
library(dplyr)
```

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

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

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

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

```
library(ggplot2)
library(gapminder)
library(ggbeeswarm)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(ggrepel)
library(zoo)
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(dslabs)
```

```
##
## Attaching package: 'dslabs'
```

```
## The following object is masked from 'package:gapminder':
##
##     gapminder
```

```
library(ggbeeswarm)
library(xts)
```

```
##
## Attaching package: 'xts'
```

```
## The following objects are masked from 'package:dplyr':
##
##     first, last
```

```
## The following objects are masked from 'package:data.table':
##
##     first, last
```

Read files

```
patientInfo<- fread("archive/PatientInfo.csv")
```

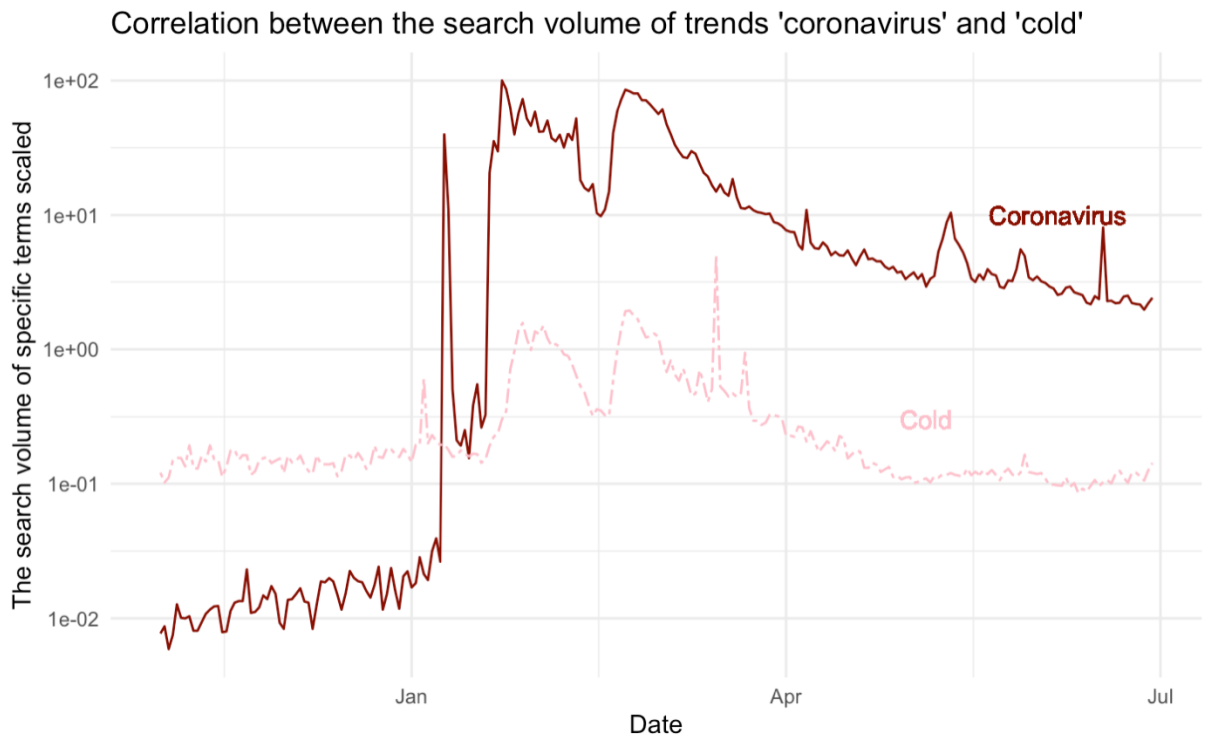
Claim 1: The distribution of the number of searches for the terms “coronavirus” and “cold” are correlated.

```
# Load the dataset for the search trends where each observation is the search
volume of each term.
searchTrends <- fread("archive/SearchTrend.csv")
# Limit the date search volume to end of 2019
intermediate<- searchTrends[date >= '2019-11-01']
# Take an initial look into the correlation between the search terms "coronav
irus" and "cold"
intermediate[, cor(coronavirus, cold)]
```

```
## [1] 0.6842782
```

This positive correlation implies that the two variables under consideration vary in the same direction, i.e., if a variable like “coronavirus” term increases, the other one “cold” increases and if one decreases the other one decreases as well. So, these variables are both dependent to each other.

```
# Plot the search volume of these two terms in a ggplot

ggplot(searchTrends[date >= '2019-11-01'], aes(x=date)) +
  geom_line(aes(y = coronavirus), color = "darkred") +
  geom_line(aes(y = cold), color="pink", linetype="twodash")   +scale_y_log10
() + labs(x="Date", y="The search volume of specific terms scaled")    +
        ggtitle("Correlation between the search volume of trends 'coronavirus
' and 'cold'") + theme_minimal()  +
geom_text(aes(x = as.Date("2020-06-06"), y = 10, label = "Coronavirus"), colo
r = "darkred") +
geom_text(aes(x = as.Date("2020-05-05"), y = 0.3, label = "Cold"),color= "pin
k")
```




```
# There is a correlation between  the plots of the terms "Coronavirus" and "C
old" which is more visible when coronavirus was spreading in all countries.
```

Statistical Testing on this claim:

```
# Correlation test between the two variables
# Transform data in log scales for better comparison in the correlation tests
```

```r
trends_logged<- searchTrends[date >= '2019-11-01', `:=` (cold=log10(cold), co
ronavirus = log10(coronavirus))]

str(trends_logged)
```

```
## Classes 'data.table' and 'data.frame':   1642 obs. of  5 variables:
##  $ date       : IDate, format: "2016-01-01" "2016-01-02" ...
##  $ cold       : num  0.117 0.134 0.149 0.175 0.172 ...
##  $ flu        : num  0.0559 0.1714 0.2232 0.1863 0.1507 ...
##  $ pneumonia  : num  0.157 0.208 0.193 0.29 0.246 ...
##  $ coronavirus: num  0.00736 0.0089 0.00845 0.01145 0.01381 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```r
# The pearson correlation test makes more assumptions compared to the spearma
n correlation test.
# More specifically it assumes that the two variables compared follow a linea
r relationship.
# Even though in this claim this seems to be the issue, it is not correct to
assume this relationship beforehand.
# Therefore, we decided to conduct both tests, with focusing more on the resu
lts of the second spearman test.
cor.test(trends_logged[,cold], trends_logged[,coronavirus], method="pearson")
# reject null
```

```
##
##  Pearson's product-moment correlation
##
## data:  trends_logged[, cold] and trends_logged[, coronavirus]
## t = 4.1695, df = 1640, p-value = 3.212e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.05431124 0.15004843
## sample estimates:
##      cor
## 0.102417
```

```r
cor.test(trends_logged[,cold], trends_logged[,coronavirus], method="spearman"
) # reject null
```

```
## Warning in cor.test.default(trends_logged[, cold], trends_logged[,
## coronavirus], : Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  trends_logged[, cold] and trends_logged[, coronavirus]
## S = 571650490, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2252483
```

Conclusion: Since both tests yielded a small p-value, there's a significant correlation between the two variables supported by the Spearman's test primarily and the Pearson's test secondly.

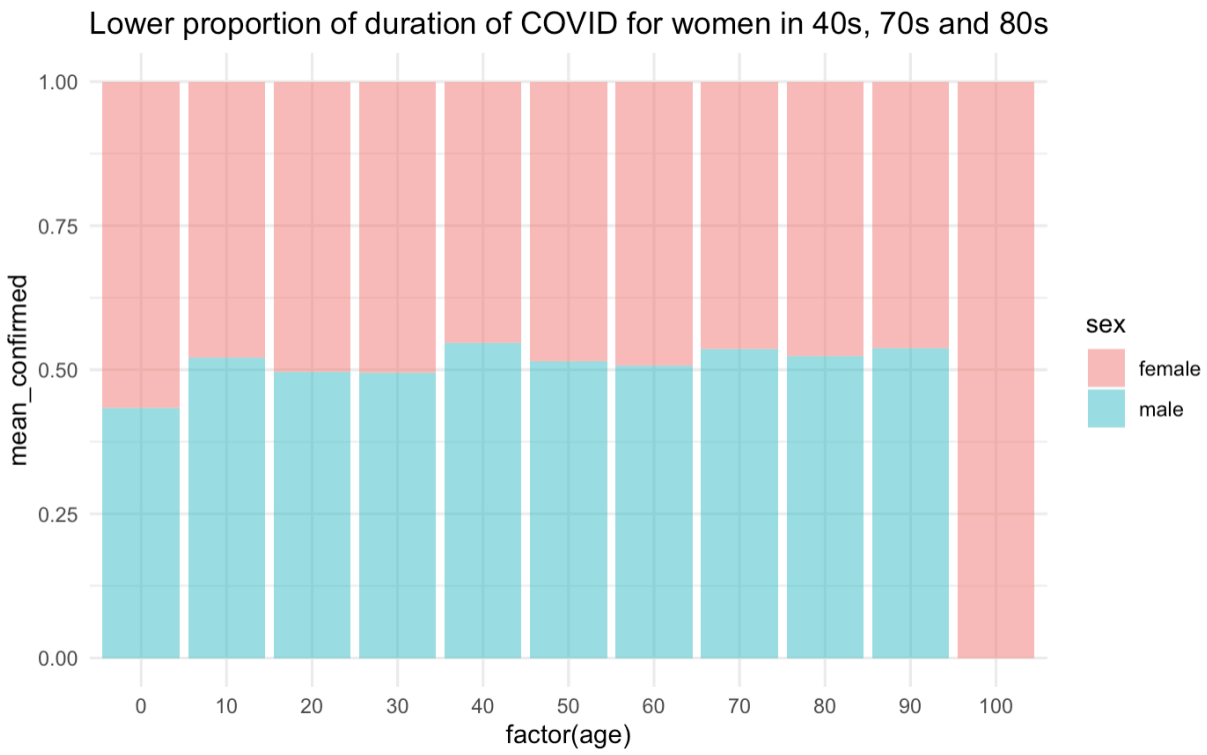
The reason behind this correlation could be some specific arguments like the symptoms of coronavirus and cold are similar. Before coronavirus was " known" as a term to people and before the spread of the virus was announced, people considered it as a flu or cold in their search. Another possible reason of this correlation in the macro-sociological level can be that eventually coronavirus is nothing more than a cold. For this conclusion we must consider the missing data to compare search trends for the previous years for "coronavirus" term. Due to this fact, there cannot be time series analysis, nor seasonality because the data is available only for 8 months and not a full year. However, we do have data from previous years for the other terms apart from coronavirus.

Claims related with patients:

Claim 2: The duration of COVID-19 after the people had been confirmed with COVID-19 in the age groups [40,49), [50,59) and [70,79) was lower for women compared to men belonging to the same age group.

For this claim first we start with patientinfo file where each observation shows the data for a certain patient.

```
# Load patient info dataset
patientInfo<- fread("archive/PatientInfo.csv")

# Subset the dataset for the observations- patients whose age, sex and releas
e date are known.
dt_check_only_test_mean <- patientInfo[!is.na(released_date) & !is.na(age) &
age !='' & sex!='']
dt_check_only_test_mean<- dt_check_only_test_mean[, `:=` ( covid_duration_con
firmed= released_date-confirmed_date)]
dt_check_only_test_mean<- dt_check_only_test_mean[, age:= gsub("s","",age) %>
% as.numeric]
means<- dt_check_only_test_mean[,.(mean_confirmed= mean(covid_duration_confir
med), countN= .N),by=c('age','sex')]
ggplot(means, aes(x=factor(age),fill=sex)) +
  geom_bar(stat='identity', position=position_fill(),aes(y = mean_confirmed),
linetype="twodash", alpha = 0.5)   +
  ggtitle("Lower proportion of duration of COVID for women in 40s, 70s and 80
s")  + theme_minimal()
```



From the plot made above we observe significant difference for 40s, 70s, 80s because for 90s and 0s we have less than 20 observationsand therefore we do not take them into account for statistical testing due to lack of enough observations.

If we open from the Rstudio environment the means data table we investigate the following results for the age groups of interest.

age sex mean_duration number of observations

40 male 26.58140 86

40 female 22.02581 155

70 male 36.20000 25

70 female 31.34483 58

80 male 37.05556 18

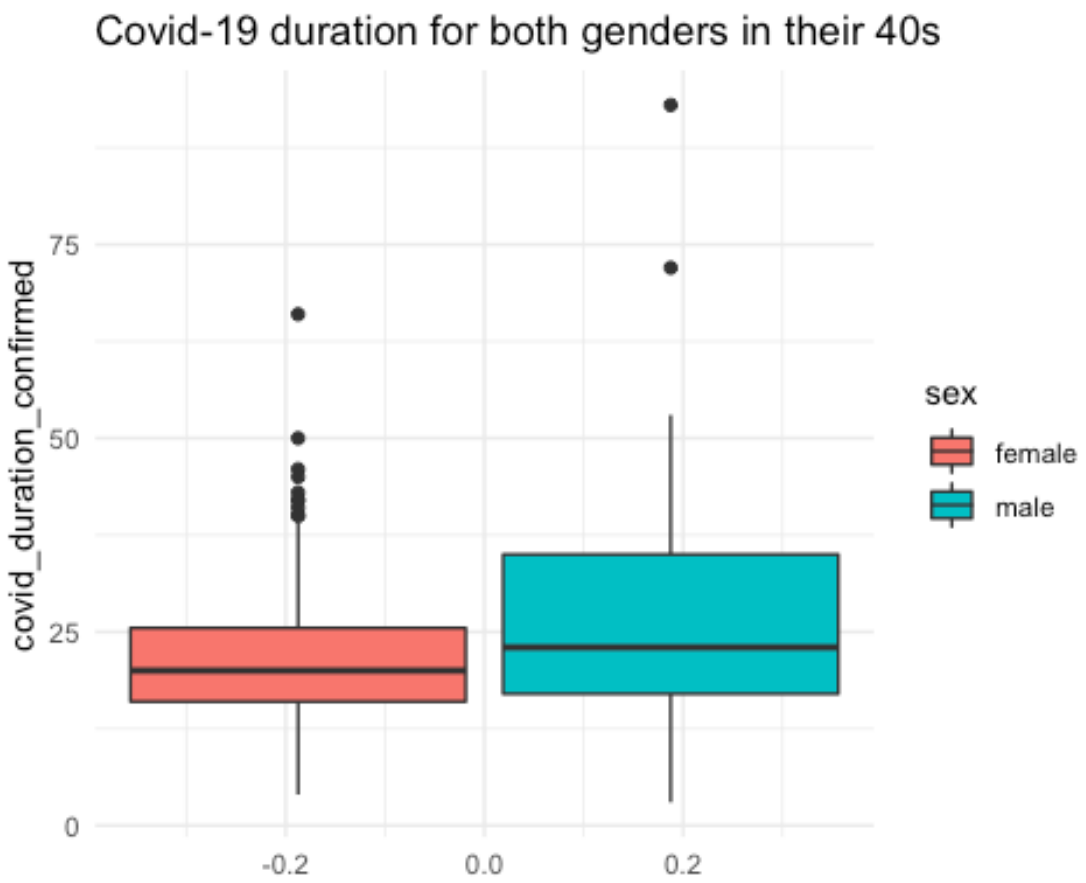
80 female 33.73810 42

```
# Function used to find the mean difference of duration of COVID-19 between w
omen and men of a specific age group
median_diff <- function(tab){
  tab[sex == 'female', median(covid_duration_confirmed, na.rm=T)] -
```

```
    tab[sex == 'male', median(covid_duration_confirmed, na.rm=T)]
}
```

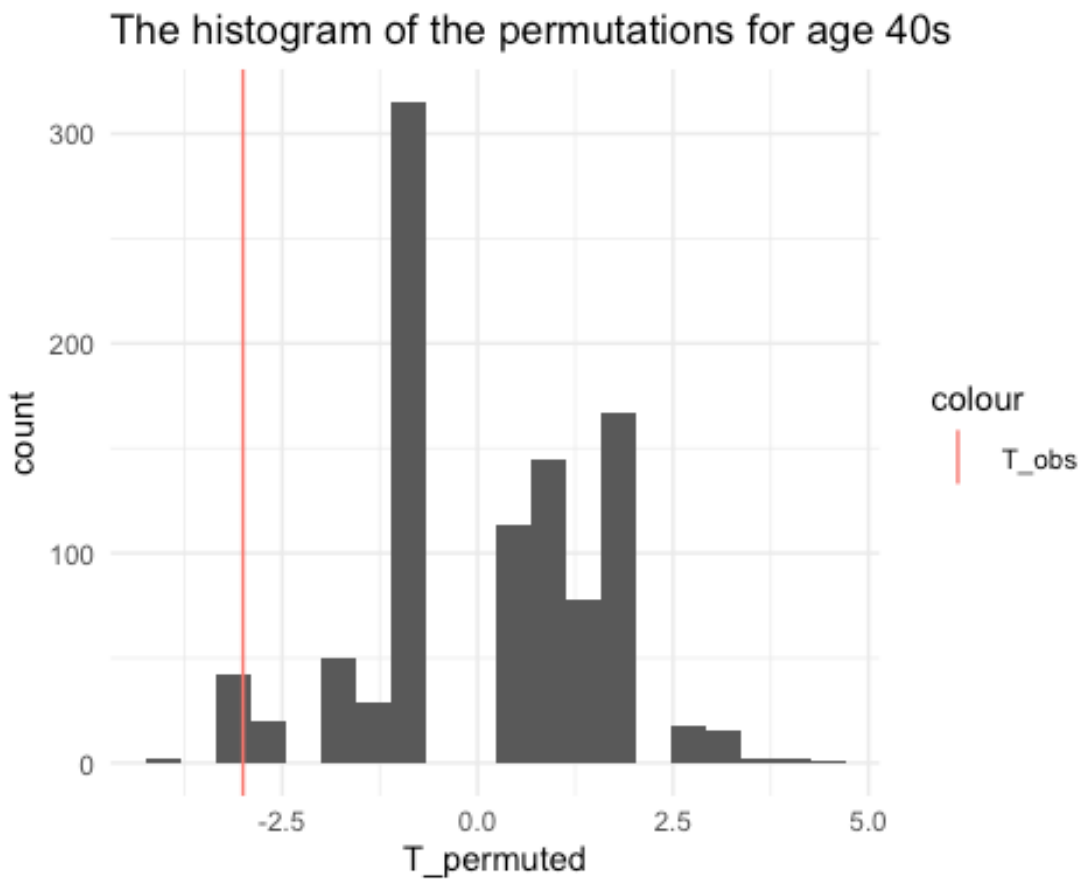
1st permutation testing for the 40s age group: H0: mean duration women >= mean duration men. Gender doesn't play a significant role in the duration of COVID-19 for people in their 40s who were released. H1: mean duration women < mean duration men. Gender plays a significant role in the duration of COVID-19 for people in their 40s who were released with women being infected with the virus on average less time compared to men.

```
# Create a boxplot for the 40s group
ggplot(dt_check_only_test_mean[age==40], aes(y=covid_duration_confirmed, fill
=sex)) +
  geom_boxplot() + theme_minimal()  +
        ggtitle("Covid-19 duration for both genders in their 40s")
```




```
T_obs <- median_diff(dt_check_only_test_mean[age==40])
T_obs

## [1] -3

# mean female - mean man gives negative result
# meaning: mean female < mean male
dt_permuted <- copy(dt_check_only_test_mean[age==40])
set.seed(0) # the seed of the random number generator
```

```r
# number of permutations
m <- 1000

# initialize T_permuted with missing values
# (safer than with 0's)
T_permuted <- rep(NA, m)

# iterate for i=1 to m
for(i in 1:m){
  # permute the sex column in place
  dt_permuted[ , sex:=sample(sex)]
  # stores the median difference in the i-th entry of T_permuted
  T_permuted[i] <- median_diff(dt_permuted)
}

# Plot the histogram of the permutations
ggplot( data.table(T_permuted), aes(x = T_permuted) ) +
  geom_histogram(bins = 20) +
  geom_vline( aes(xintercept=T_obs, color = "T_obs") ) + theme_minimal()  +
        ggtitle("The histogram of the permutations for age 40s")
```



```r
# Count the number of times the permuted value was less than the observed val
ue.
p_val <- (sum(T_permuted < T_obs) + 1) / (m + 1)
p_val

## [1] 0.002997003
```

Since p_val < 0.05 confidence level there's evidence to reject the null hypothesis. The fact that the duration of COVID-19 for women in their 40s is less on average than the duration of COVID-19 for men in their 40s didn't occur by chance but there's indication that it exists in the general population. Further steps could be to investigate the gaps in the histogram using Q-Q plots.

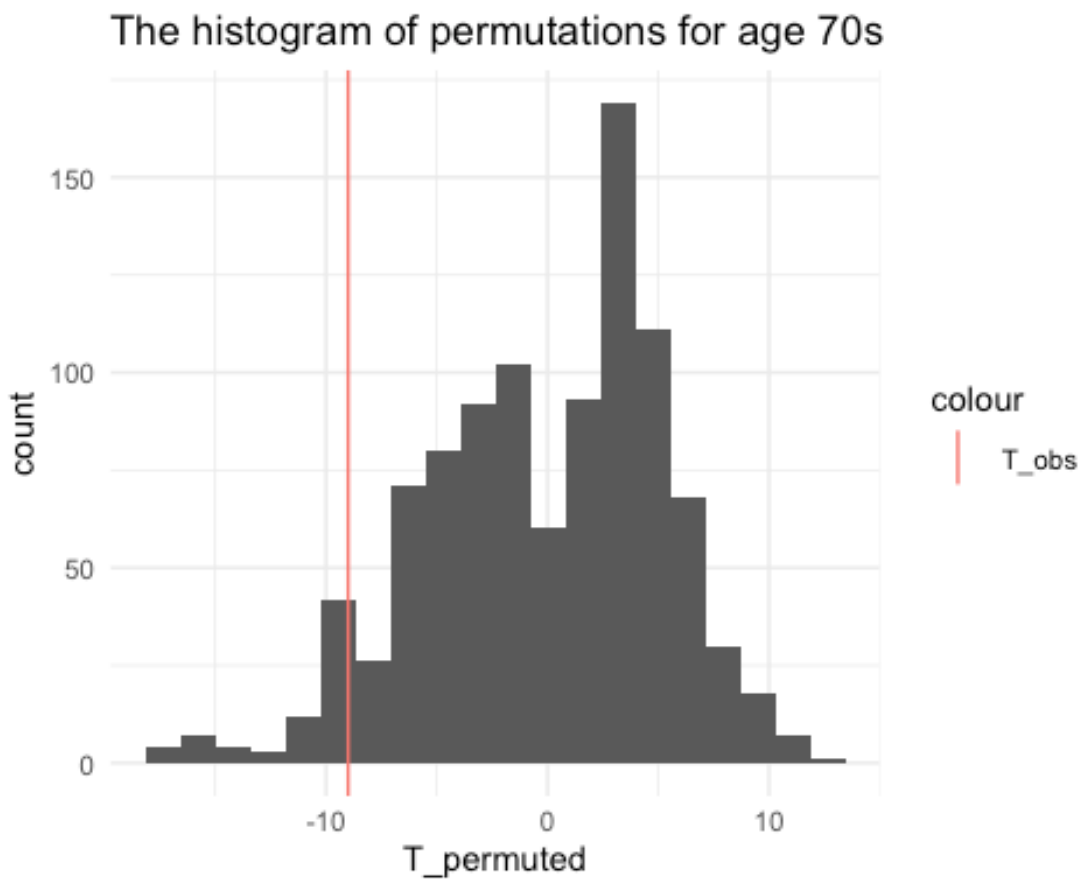
2nd permutation testing for the 70s age group: H0: mean duration women >= mean duration men. Gender doesn't play a significant role in the duration of COVID-19 for people in their 70s who were released. H1: mean duration women < mean duration men. Gender plays a significant role in the duration of COVID-19 for people in their 70s who were released with women being infected with the virus on average less time compared to men.

```r
ggplot(dt_check_only_test_mean[age==70], aes(y=covid_duration_confirmed, fill
= sex)) +
  geom_boxplot() + theme_minimal()  +
        ggtitle("The Covid-19 duration for both genders in their 70s")
```

```r
T_obs <- median_diff(dt_check_only_test_mean[age==70])
T_obs
```

```
## [1] -9
```

```r
dt_permuted <- copy(dt_check_only_test_mean[age==70])
set.seed(0) # the seed of the random number generator

# number of permutations
m <- 1000

# initialize T_permuted with missing values
# (safer than with 0's)
T_permuted <- rep(NA, m)

# iterate for i=1 to m
for(i in 1:m){
  # permute the genotype column in place
  dt_permuted[ , sex:=sample(sex)]
  # store the median difference in the i-th entry of T_permuted
  T_permuted[i] <- median_diff(dt_permuted)
}

ggplot( data.table(T_permuted), aes(x = T_permuted) ) +
  geom_histogram(bins = 20) +
  geom_vline( aes(xintercept=T_obs, color = "T_obs")) + theme_minimal()  +
        ggtitle("The histogram of permutations for age 70s")
```

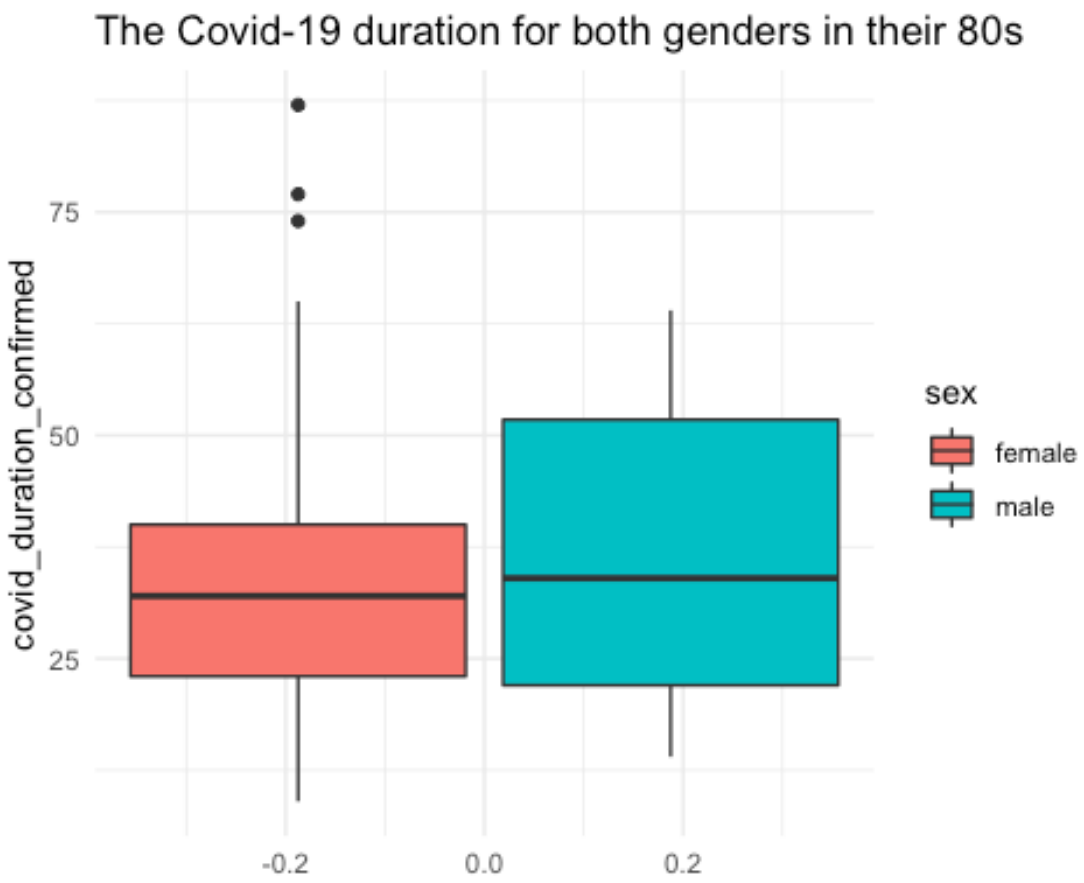



```
p_val <- (sum(T_permuted < T_obs) + 1) / (m + 1)
p_val

## [1] 0.05494505

# p_val > 0.05 -> fail to reject the null hypothesis
```

Since, the p_val > 0.05 we fail to reject the null hypothesis. If we used a higher significance level such as 10% we would reject the null hypothesis.

3rd permutation testing for the 80s age group: H0: mean duration women >= mean duration men. Gender doesn't play a significant role in the duration of COVID-19 for people in their 80s who were released. H1: mean duration women < mean duration men. Gender plays a significant role in the duration of COVID-19 for people in their 80s who were released with women being infected with the virus on average less time compared to men.

```
ggplot(dt_check_only_test_mean[age==80], aes(y= covid_duration_confirmed, fil
l=sex)) +
  geom_boxplot() + theme_minimal()  +
        ggtitle("The Covid-19 duration for both genders in their 80s")
```

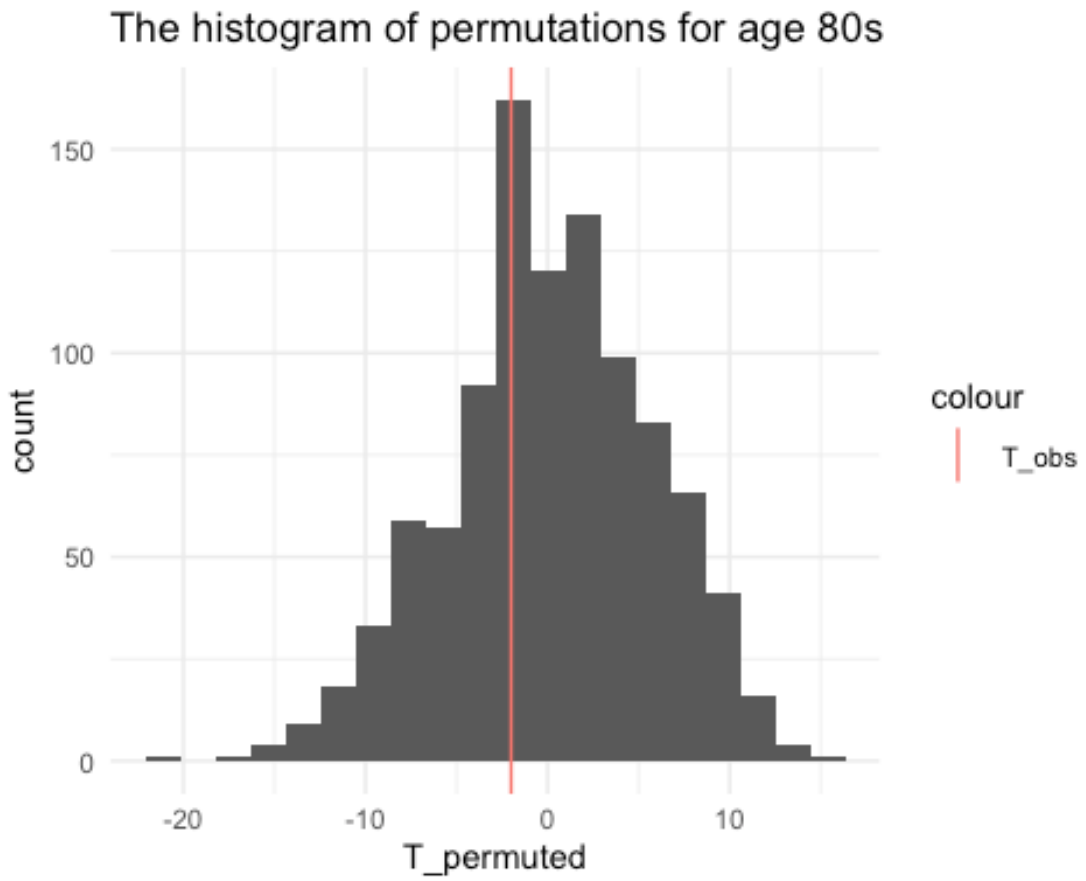



```r
T_obs <- median_diff(dt_check_only_test_mean[age==80])
T_obs
```

```
## [1] -2
```

```r
dt_permuted <- copy(dt_check_only_test_mean[age==80])
set.seed(0) # the seed of the random number generator

# number of permutations
m <- 1000

# initialize T_permuted with missing values
# (safer than with 0's)
T_permuted <- rep(NA, m)

# iterate for i=1 to m
for(i in 1:m){
  # permute the genotype column in place
  dt_permuted[ , sex:=sample(sex)]
  # store the median difference in the i-th entry of T_permuted
  T_permuted[i] <- median_diff(dt_permuted)
}
```

```
ggplot( data.table(T_permuted), aes(x = T_permuted) ) +
  geom_histogram(bins = 20) +
  geom_vline( aes(xintercept=T_obs, color = "T_obs") ) + theme_minimal()  +
        ggtitle("The histogram of permutations for age 80s")
```




```
p_val <- (sum(T_permuted < T_obs) + 1) / (m + 1)
p_val
```

```
## [1] 0.3036963
```

Since the p_val > 0.05 we fail to reject the null hypothesis.

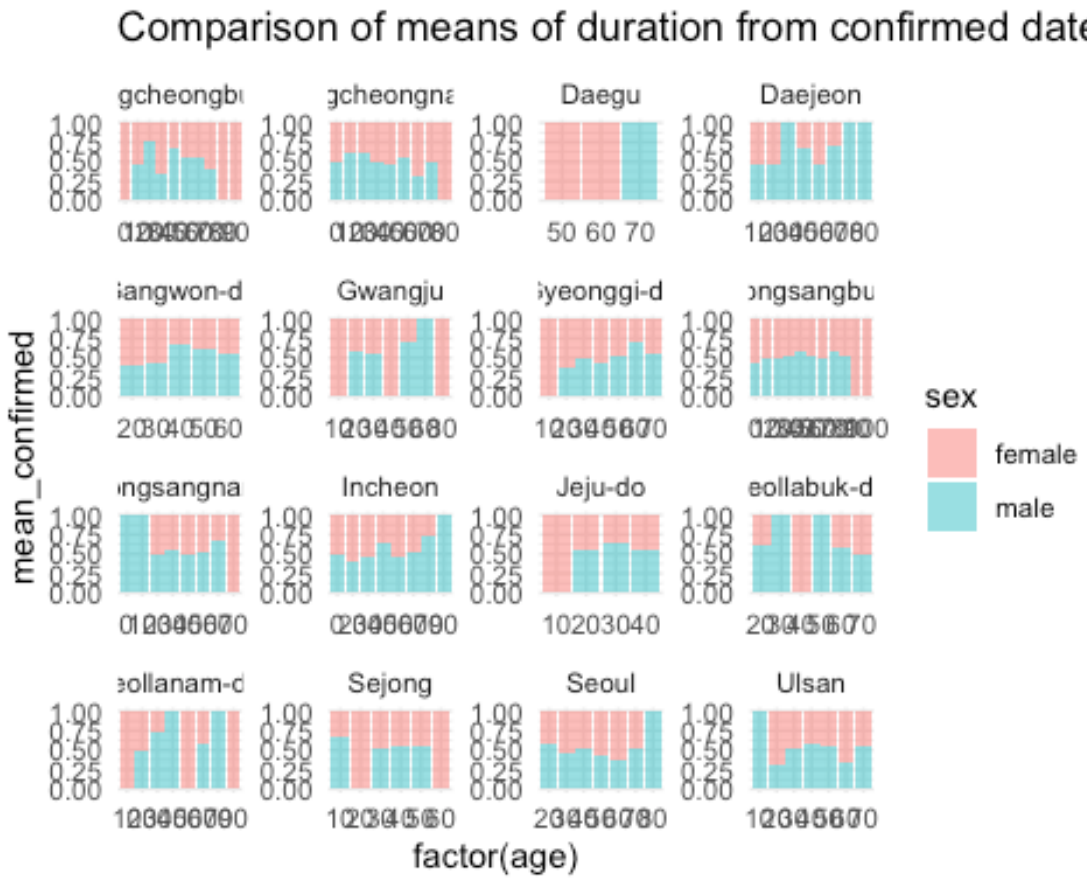
Confounding with a 3rd variable:

```
# confounding with province
means_province<- dt_check_only_test_mean[,.(mean_confirmed= mean(covid_durati
on_confirmed), countN= .N),by=c('province','age','sex')]
means_province
```

```
##       province age    sex mean_confirmed countN
##    1:    Seoul  50   male       13.75000      8
##    2:    Seoul  30   male       14.50000      8
##    3:    Seoul  20   male       17.14286      7
##    4:    Seoul  20 female       13.00000      9
```

```
##    5:    Seoul  50 female       19.00000      3
##   ---
## 190:  Jeju-do  40 female       13.00000      1
## 191:  Jeju-do  40   male       17.00000      1
## 192:  Jeju-do  30   male       44.00000      2
## 193:  Jeju-do  10 female       14.00000      2
## 194:  Jeju-do  30 female       24.00000      2
```

```
ggplot(means_province, aes(x=factor(age),fill=sex)) +
  geom_bar(stat='identity', position=position_fill(),aes(y = mean_confirmed),
linetype="twodash", alpha = 0.5)   +
  ggtitle("Comparison of means of duration from confirmed date") + facet_wrap
(~ province, scales = "free") + theme_minimal()
```



By taking a look to the plot above, we can see that the relationship suggested in the alternative hypothesis for the age group 40 does not hold for every province but it seems to hold for the most important ones for example Daejeon, Gyeongsangbuk-do, Chungcheongbuk-do, Gangwon-do etc.

To investigate further this relationship in the provinces which have the highest number of COVID-19 patients, we find that the province with the highest number of observations is Gyeongsangbuk-do: But if we look at the highest number of women and men in the table we see that:

```
means_province[age==40,sum(countN),by=province][max(V1)==V1]
```

```
##            province V1
## 1: Gyeongsangbuk-do 94
```

Meaning that the group with the highest total number of women and men, which probably leads the hypothesis of this age group.

```
m<-means_province[age==40,sum(countN),by=province]
m[order(-V1)]
```

```
##              province V1
##  1:  Gyeongsangbuk-do 94
##  2: Chungcheongnam-do 48
##  3:            Sejong 20
##  4:           Incheon 17
##  5:  Gyeongsangnam-do 12
##  6:       Gyeonggi-do  9
##  7:        Gangwon-do  7
##  8:             Seoul  6
##  9:           Daejeon  6
## 10: Chungcheongbuk-do  6
## 11:             Ulsan  5
## 12:           Gwangju  4
## 13:      Jeollanam-do  4
## 14:           Jeju-do  2
## 15:      Jeollabuk-do  1
```

We see that the next province with half number of observations is Chungcheongnam-do in which the relationship still holds and the relationship holds for the third in ordering Sejong as well. Therefore province instead isn't a confounding factor since the relationship of the age and sex for the proportion of COVID-19 duration still holds after we faceted with province.

Claim 3: Older people tend to get infected more in groups than younger people

First import the PatientInfo.csv dataset where once again we observe each patient data and use it for further analysis.

```
patientInfo <- fread("archive/PatientInfo.csv")
```

Remove unnecessary columns and ignore the rows with unknown infection cases.

```
dt = subset(patientInfo, select = -c(state,city,country,province,infected_by,
contact_number, confirmed_date,released_date,deceased_date,symptom_onset_date
) )
dt = dt[infection_case!= "" & infection_case != "etc" & age != ""]
```

Filter Group Infections

Group infections are taken as all infections that happened in a specific location (not contact with patient or from overseas)

```
gi = dt[infection_case != "overseas inflow" & infection_case != "contact with patient"]
```

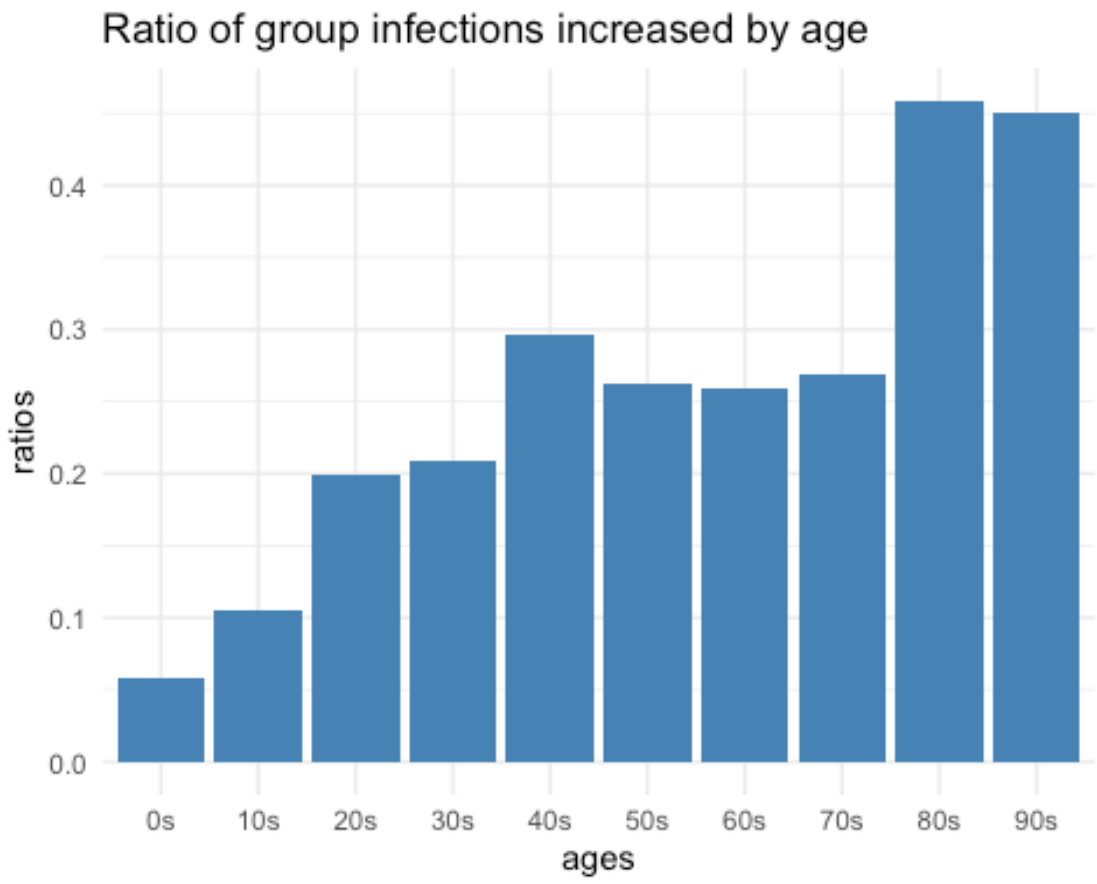
Plot

Calculating the ratio by dividing group infections by total infections.

```
M <- c("0s","10s","20s","30s","40s","50s","60s","70s","80s","90s")
# Divide group infections with all infections to get the ratio
H <- c()
for (i in M) {
  H <- c(H, count(gi[age == i]) / count(dt[age == i]))
}
H <- unlist(H, use.names=FALSE)
plotData <- data.table(ages = M, ratios = H)
```

Calculating the ratio by dividing group infections by total infections.

```
ggplot(plotData, aes(x=ages,y=ratios)) +
  geom_bar(stat="identity", fill="steelblue")+
  ggtitle("Ratio of group infections increased by age")+
  theme_minimal()
```



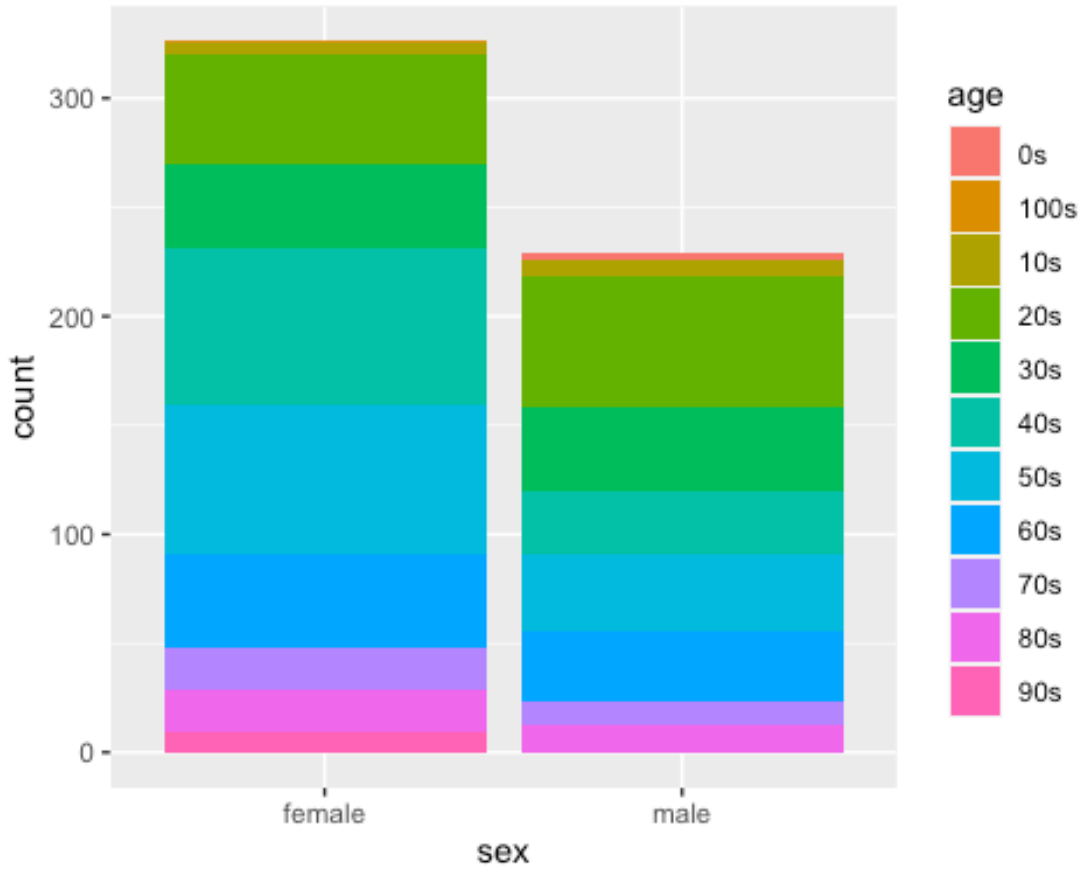
Confounding Variables:

We can investigate relationships between age, gender and type of infection (group vs individual). We can add population and find the ratio of infected people per total population.

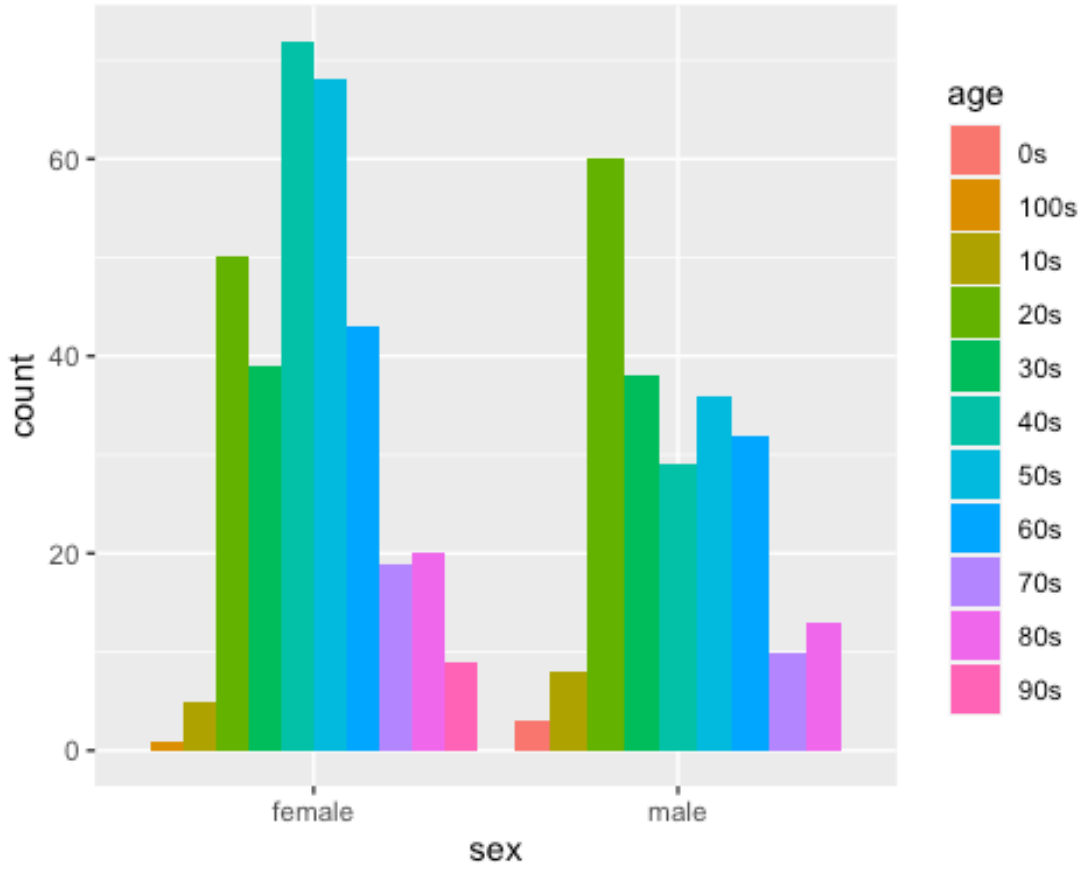
```
head(gi)
```

```
##    patient_id    sex age                infection_case
## 1: 1000000015   male 70s              Seongdong-gu APT
## 2: 1000000020 female 70s              Seongdong-gu APT
## 3: 1000000022   male 30s Eunpyeong St. Mary's Hospital
## 4: 1000000023   male 50s             Shincheonji Church
## 5: 1000000025   male 60s Eunpyeong St. Mary's Hospital
## 6: 1000000028 female 70s Eunpyeong St. Mary's Hospital
```

```
ggplot(gi, aes(sex, fill=age)) + geom_bar()
```



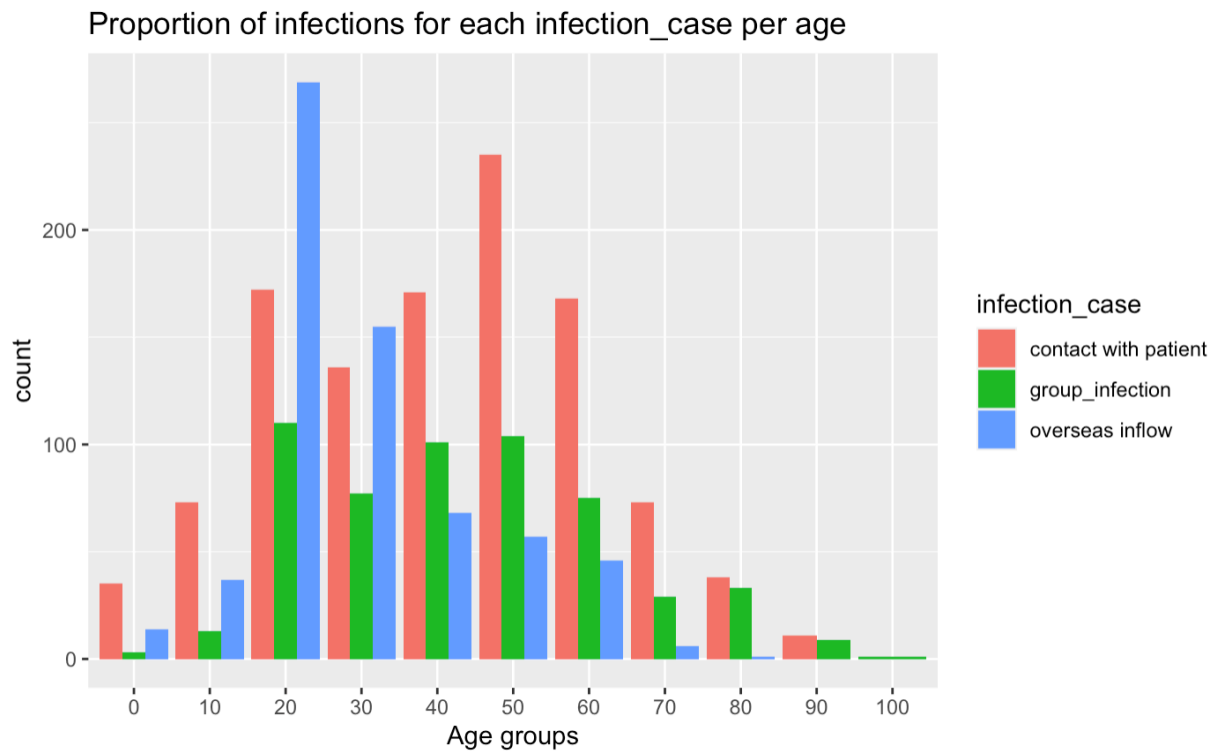

```
ggplot(gi, aes(sex, fill=age)) +
        geom_bar(position='dodge')
```



Confounding investigation (further): dt dataset While for the group infection this condition that the proportion of the specific type of infection compared to total infections increases over time this doesn't hold true for the other data -> overseas inflow and contact with patient. We can also see that the data seems to be gaussian in the case of group infection and contact with patient more than the overseas inflow.

```
`%notin%` <- Negate(`%in%`)
df_confounding <- copy(dt)
df_confounding[,infection_case:=  ifelse(infection_case %notin% c("overseas i
nflow","contact with patient"),"group_infection",infection_case)]
ggplot(df_confounding, aes(factor(gsub("s","",age) %>% as.numeric), fill=infe
ction_case)) +
        geom_bar(position='fill') +xlab("Age groups")+
  ggtitle("Proportion of infections for each infection_case per age")
```


```
ggplot(df_confounding, aes(factor(gsub("s","",age) %>% as.numeric), fill=infe
ction_case)) +
        geom_bar(position='dodge') +xlab("Age groups")+
  ggtitle("Proportion of infections for each infection_case per age")
```




```
head(gi)
```

```
##    patient_id    sex age               infection_case
## 1: 1000000015   male 70s             Seongdong-gu APT
## 2: 1000000020 female 70s             Seongdong-gu APT
## 3: 1000000022   male 30s Eunpyeong St. Mary's Hospital
## 4: 1000000023   male 50s            Shincheonji Church
## 5: 1000000025   male 60s Eunpyeong St. Mary's Hospital
## 6: 1000000028 female 70s Eunpyeong St. Mary's Hospital
```

```
summary(gi)
```

```
##    patient_id            sex                 age            infection_case
##  Min.   :1000000015   Length:555         Length:555         Length:555
##  1st Qu.:1000000315   Class :character   Class :character   Class :charact
er
##  Median :1700000014   Mode  :character   Mode  :character   Mode  :charact
er
##  Mean   :2868075304
##  3rd Qu.:6001000413
##  Max.   :7000000014
```

Statistical Testing: Fisher's test approach table structure but TEST 1 young middle age group infection X Y
non-group infection Z L

TEST 2 middle age old group infection X Y
non-group infection Z L

TEST 3 young old group infection X Y
non-group infection Z L

TEST 1 Null hypothesis H0: Age doesn't play a significant role in the observations for COVID-19 cases for all age groups. Observations are independent from age. Ratio of group infections of younger age group [0,30) is greater or equal to the ratio of group infections of middle age group [30,60).

H1: Age plays a significant role in the observations for COVID-19 cases for all age groups. Observations are dependent from age. Ratio of group infections of younger age group [0,30) is less than ratio of group infections of middle age group [30,60).

```
##Three groups
#young 0s, 10s, 20s [0,30)
#middle age  30s, 40s, 50s [30,60)
#old [60, 100)
young_group_infection <- gi[age %in% c("0s", "10s", "20s"), .N]
young_non_group_infection <- dt[age %in% c("0s", "10s", "20s"), .N] - young_g
roup_infection
middle_age_group_infection <-  gi[age %in% c("30s", "40s", "50s"), .N]
middle_age_non_group_infection <- dt[age %in% c("30s", "40s", "50s"), .N] - m
iddle_age_group_infection
old_group_infection <-  gi[age %in% c("60s", "70s", "80s","90s"), .N]
old_non_group_infection <- dt[age %in% c("60s", "70s", "80s","90s"), .N] - ol
d_group_infection

tbl = data.table(
  young = c(young_group_infection, young_non_group_infection),
  middle_age = c(middle_age_group_infection, middle_age_non_group_infection)
)
tbl

##    young middle_age
## 1:   126        282
## 2:   600        822

tst <- fisher.test(tbl,  alternative = "less")
tst

##
##  Fisher's Exact Test for Count Data
##
## data:  tbl
## p-value = 2.034e-05
## alternative hypothesis: true odds ratio is less than 1
## 95 percent confidence interval:
##  0.0000000 0.7501527
## sample estimates:
```

```
## odds ratio
##   0.612291
```

TEST 2

```
tbl = data.table(
  middle_age = c(middle_age_group_infection, middle_age_non_group_infection),
  old = c(old_group_infection, old_non_group_infection)
)
tbl
```

```
##    middle_age old
## 1:        282 146
## 2:        822 343
```

```
tst <- fisher.test(tbl,  alternative = "less")
tst
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tbl
## p-value = 0.04241
## alternative hypothesis: true odds ratio is less than 1
## 95 percent confidence interval:
##  0.0000000 0.9904872
## sample estimates:
## odds ratio
##   0.806098
```

TEST 3

```
tbl = data.table(
  young = c(young_group_infection, young_non_group_infection),
  old = c(old_group_infection, old_non_group_infection)
)
tbl
```

```
##    young old
## 1:   126 146
## 2:   600 343
```

```
tst <- fisher.test(tbl,  alternative = "less")
tst
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tbl
## p-value = 2.593e-07
## alternative hypothesis: true odds ratio is less than 1
## 95 percent confidence interval:
```

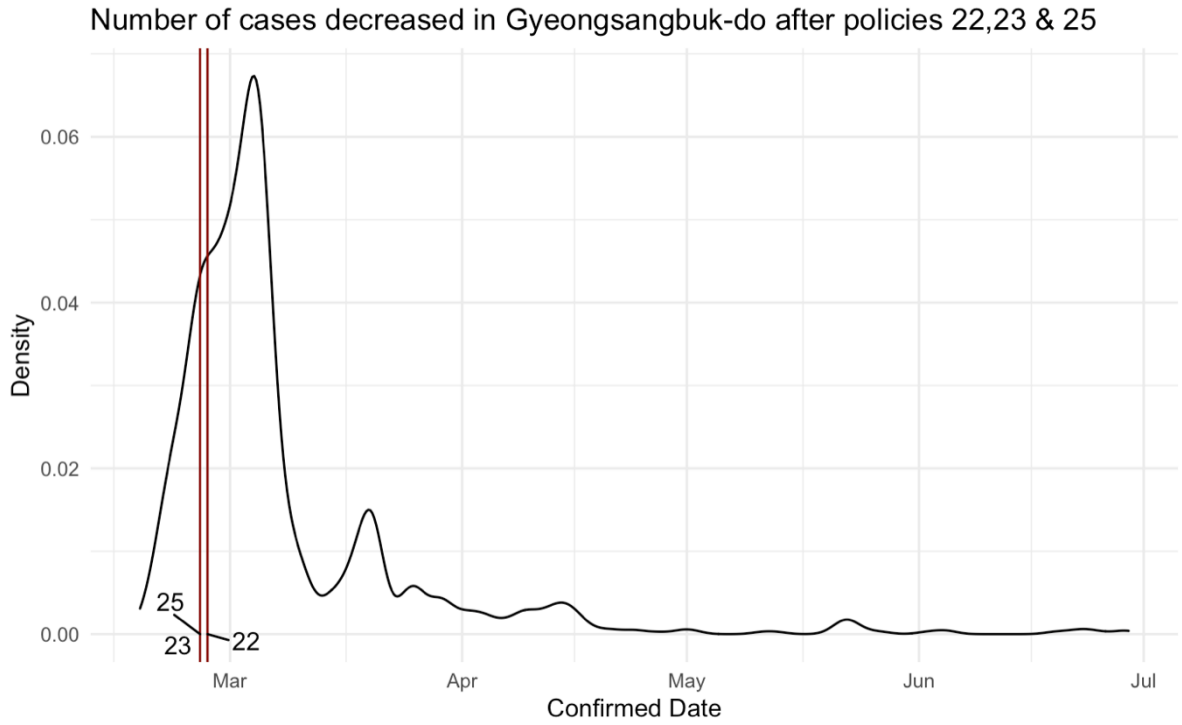
```
##  0.0000000 0.6263052
## sample estimates:
## odds ratio
##  0.4936486
```

From all the combinations of tests it turns out that the proportion of infection in groups compared to all the infections of the specific age group increase with age since all the tests we conducted yielded that the correlations are statistically significant. Possible reasons this occured could be that older people in South Korea are more involved with public gatherings such as going to the church.

Claims related with effectiveness of policies:

Claim 4: For the policies with policy ids 22,23 and 25 it is probable that they (either all of them or one or two of them) had an impact on the reduction of the density of cases in province “Gyeongsangbuk-do”. To reach a conclusion for this claim we select the policy dataset where we observe all the measures taken from the government during the first 6 months of 2020 after the coronavirus started spreading in the country of South Korea.

```
# Load policy dataset
policy<- fread("archive/Policy.csv")
# Convert start date as date
policy[,start_date:= as.Date(start_date,"%Y-%m-%d")]
## Exclude in policy table the policies we are not interested in and the ones
that are not in health category
policy<- policy[type=='Health' & policy_id %in% c(23,25,22)]
# PLot the density of cases compared to stat time of policies
ggplot(patientInfo[province %in% c("Gyeongsangbuk-do")], aes(x=confirmed_date
)) + geom_density()  + geom_vline(data=policy, mapping= aes(xintercept=as.num
eric(unlist(start_date))), colour="darkred")+geom_text_repel(data=policy, map
ping=aes(x=start_date, y=0, label=policy_id, vjust=-0.4, hjust=0))+ labs(x="C
onfirmed Date", y="Density") + theme_minimal()  +
        ggtitle("Number of cases decreased in Gyeongsangbuk-do after policies
22,23 & 25")

## Warning: Removed 3 rows containing non-finite values (stat_density).
```



As we can see the the density of cases started decreasing after one week these policies started being put in place.

Here are more details about these policies:

```
policy
```

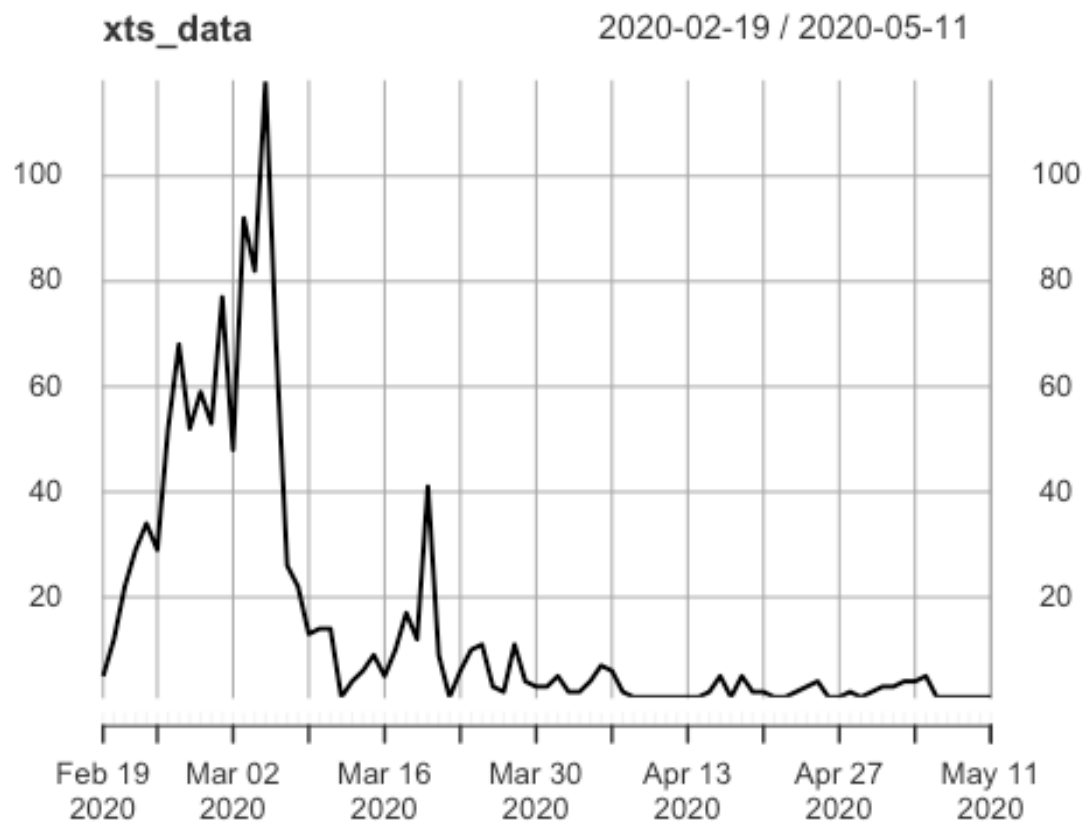
```
##    policy_id country   type                                        gov_policy
## 1:        22   Korea Health Emergency Use Authorization of Diagnostic Kit
## 2:        23   Korea Health Emergency Use Authorization of Diagnostic Kit
## 3:        25   Korea Health                Drive-Through Screening Center
##                 detail start_date end_date
## 1:              3rd EUA 2020-02-27     <NA>
## 2:              4th EUA 2020-02-27     <NA>
## 3: by Local Government 2020-02-26     <NA>
```

To support this claim we use additional help since here we have to do with time series and it might effect the final results. For this, we use XTS library, creating 2 groups (test and control group) to help with our analysis. In our analysis we incorporate the theory of control and test group. In theory, a control group is a statistically significant portion of participants in an experiment that are shielded from exposure to variables and the test group is a statistically significant portion of participants in an experiment that are exposed to certain variables (https://clevertap.com/blog/what-is-a-control-group/). Even though not 100% applicable in our study of the effect of policies to a certain portion of the population they can give indication for further statistical analysis. The control group is considered the number of confirmed cases per day some days before the implementation of the policies and the test group is the number of confirmed cases per day for some days

after the implementation of the policies. Then, we use the Wilcoxon test here to compare the two groups.

Statistical Testing:

```
# Select region of dataset of interest
ts_time<- patientInfo[province %in% c("Gyeongsangbuk-do") & !is.na(confirmed_
date),.N, by=confirmed_date]
# Create the sequence of dates included in the ts_time data table in order to
create an XTS object
dates <- seq(as.Date(ts_time[1,confirmed_date]), length=nrow(ts_time), by="da
ys")
# Create XTS X-time series object
xts_data <- xts(ts_time[,N], order.by=dates)
plot(xts_data)
```



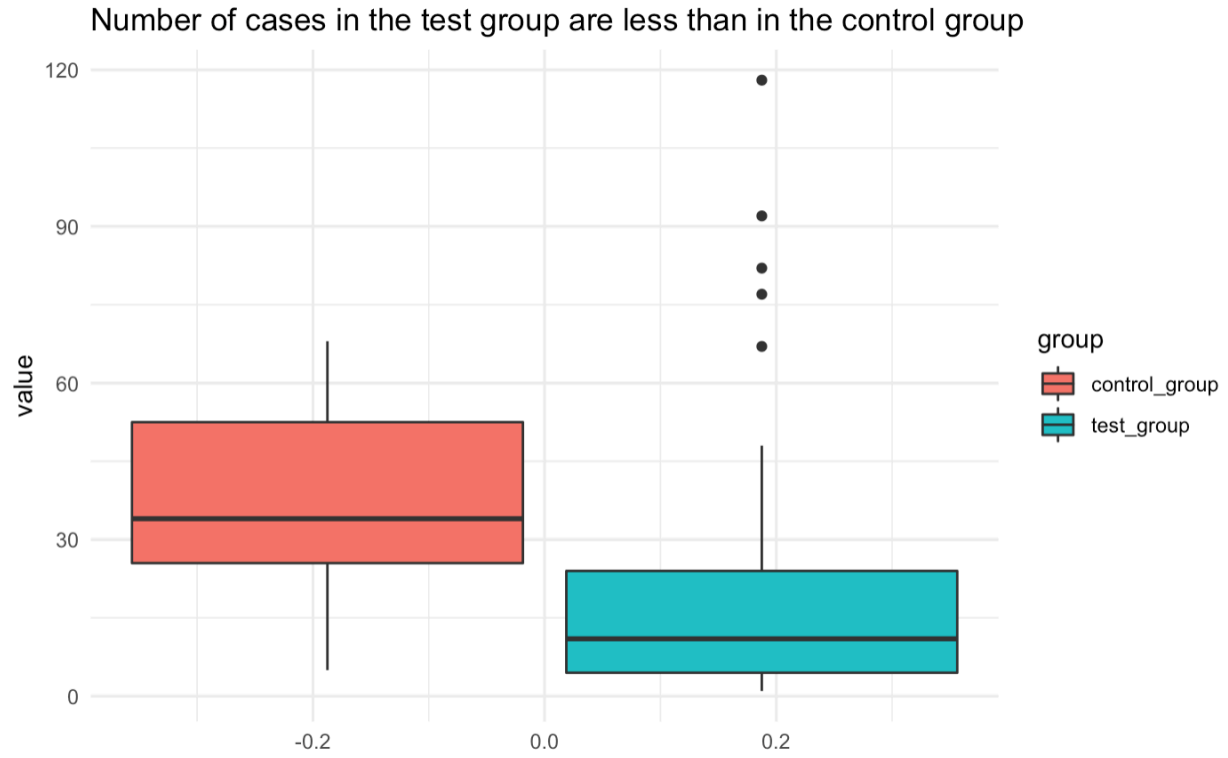

```
#creating control and test group
# Control group for the whole month of February
control_group<- as.numeric(xts_data["/2020-02"])
summary(control_group)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   25.50   34.00   37.73   52.50   68.00
```

```r
# Test group for March
test_group <- as.numeric(xts_data["2020-03/2020-3"])
summary(test_group)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0     4.5    11.0    23.9    24.0   118.0
```

```r
# Create a boxplot showing the result
# Boxplot shows a reduction in the number of cases distribution for the test
group compared to the control group
dt <- data.table(group=c(rep("control_group", times=length(control_group)), r
ep("test_group", times=length(test_group))), value=c(control_group, test_grou
p))
ggplot(data=dt, aes(y=value, fill=group)) + geom_boxplot() + theme_minimal()
+
        ggtitle("Number of cases in the test group are less than in the contr
ol group")
```




```r
## Wilcoxon rank-sum test
# H0: control_group <= test_group. The average number of cases in the control
group are less or equal than the average
# number of cases in the test group.
# H1: control_group > test_group. The average number of cases in the control
group are more than the average
# number of cases in the test group.
wilcox.test(control_group, test_group,  exact = FALSE, alternative = "greater
")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  control_group and test_group
## W = 251.5, p-value = 0.0106
## alternative hypothesis: true location shift is greater than 0
```

Conclusion: The result shows that the null hypothesis is rejected. By this, we can say that the effect of policies was real and the cases lowered due to the government's policies. However, there could also be additional reasons giving us this outcome. Therefore, we should interpret the result with cautiousness. Other possible reasons could be, the implementations of other policies, the reduction of overseas inflow population or other factors we haven't considered this far.

Claim 5: The implementation of immigration policies decreased confirmed cases from overseas.

Questions which this claim tries to answer: - What is the role of immigration policies on reducing the cases of coronavirus? - Were these policies effective from their start date to the latest data we have analyzed?

Start with reading data of policy and specifying it only for the Immigration part that we need in order to access to their details and start & end date. This can be done by observing the policy dataset, selecting only the immigration policies and doing further analysis.

```
policy <- fread("archive/Policy.csv")
```

It results on 15 poicies taken only for immigration for 15 countries or regions who were seen as risky ones.

```
policyov <- policy[type == "Immigration"]
nrow(policyov)

## [1] 15

policyov[, unique(detail)]

##  [1] "from China"             "from Hong Kong"         "from Macau"
##  [4] "from Japan"             "from Italy"             "from Iran"
##  [7] "from France"            "from Germany"           "from Spain"
## [10] "from U.K."              "from Netherlands"       "from Europe"
## [13] "from all the countries" "from U.S."
```

Since there is no end date for these policies (ALL END DATES ARE N/A) and we cannot use it on further analysis, we can delete it to make easier calculations and analysis in the next steps. On the next step, we can check the confirmed cases from overseas before and after these policies were put into place. First of all, let's remove the unneccessary columns like city-3, latitude-7 and longtitude-8, which for the overseas infection case was empty.

```
policyov[, end_date := NULL]
head(policyov)
```

```
##    policy_id country        type                    gov_policy         det
ail
## 1:         5   Korea Immigration Special Immigration Procedure     from Ch
ina
## 2:         6   Korea Immigration Special Immigration Procedure from Hong K
ong
## 3:         7   Korea Immigration Special Immigration Procedure     from Ma
cau
## 4:         8   Korea Immigration Special Immigration Procedure     from Ja
pan
## 5:         9   Korea Immigration Special Immigration Procedure     from It
aly
## 6:        10   Korea Immigration Special Immigration Procedure      from I
ran
##    start_date
## 1: 2020-02-04
## 2: 2020-02-12
## 3: 2020-02-12
## 4: 2020-03-09
## 5: 2020-03-12
## 6: 2020-03-12
```

```
case <- fread("archive/Case.csv")
caseov <- case[infection_case == "overseas inflow"]
caseov <- caseov[, -c("city" , "latitude","longitude")]
```

From this data, we see there are only 17 provinces, where we can find the most risky ones. Also, the tendency of everyone coming from overseas is that of an individual case for all provinces of South Korea.

The highest number of confirmed cases only from overseas infection cases are from province:

```
max(caseov[,confirmed])
```

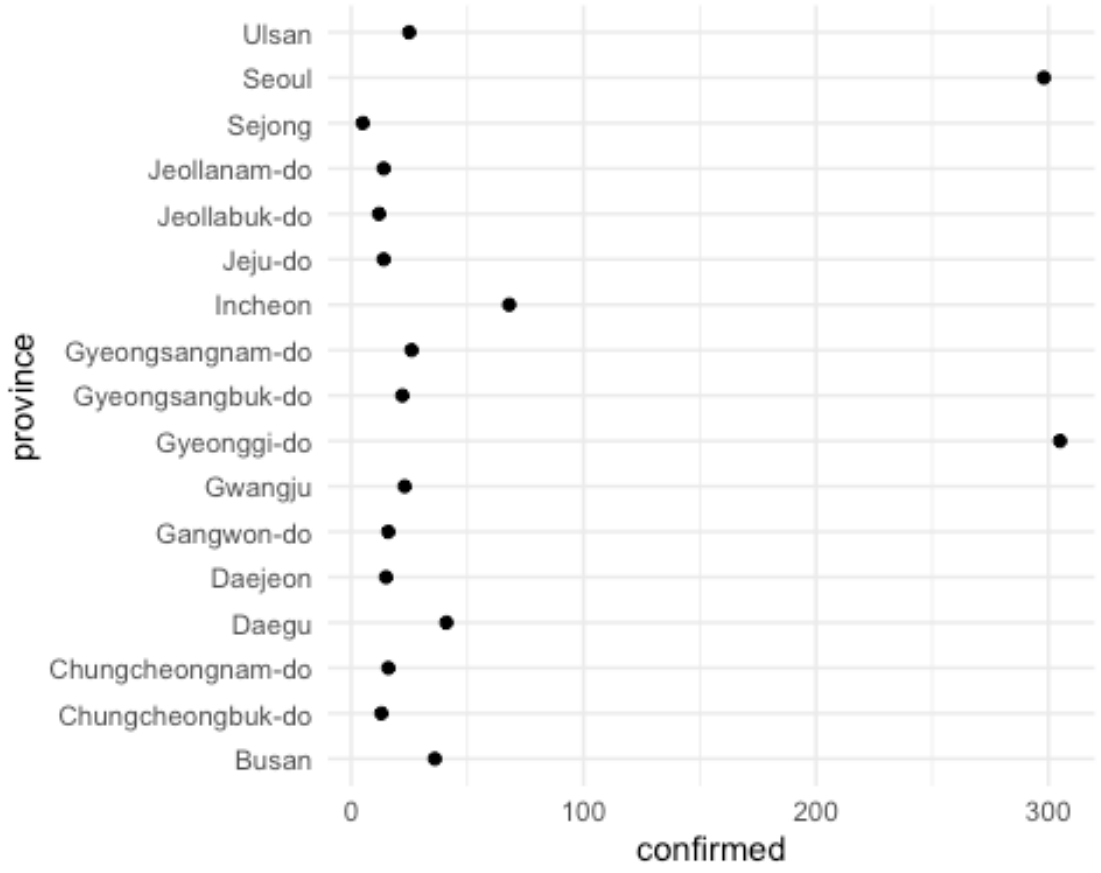
```
## [1] 305
```

```
confirmed_max_ov <- which.max(caseov[,confirmed])
caseov$province[confirmed_max_ov]
```

```
## [1] "Gyeonggi-do"
```

Showing from a density graph, the density of confirmed infection cases from overseas from all provinces, in the entire SOUTH KOREA. Overall view with more specific details for each province:

```
ggplot(caseov, aes(confirmed, province)) +
  geom_beeswarm() + theme_minimal()
```



From the analysis results until now, we can confirm that Seoul and Gyeonggi-do have the highest cases of infection from overseas, shown in the graph too.

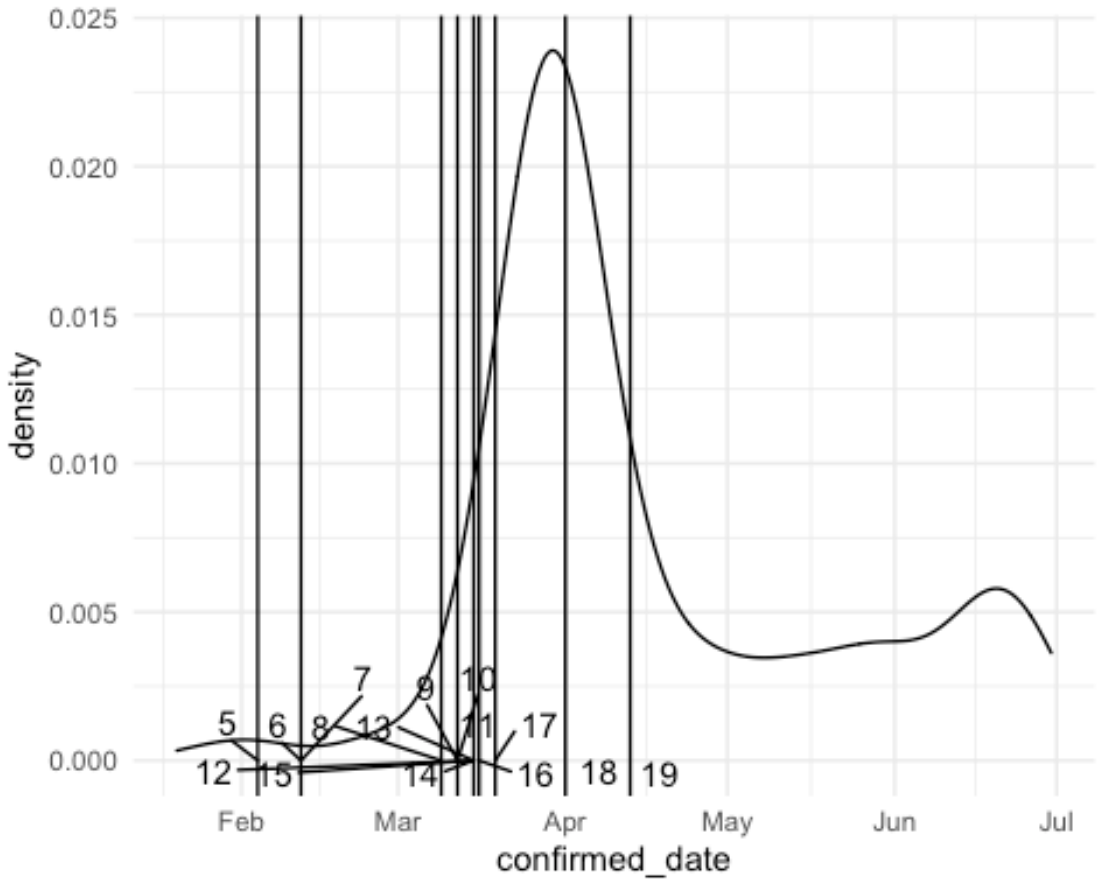
Next, we can see more in detail for each person, and analyze more specifically and follow cases overtime.

Filter the patientInfo with the infection case of overseas inflow too to see if there is any connection. We delete columns which are not needed like symptoms onset, released and deceased.

```
PatientInfo <- fread("archive/PatientInfo.csv")
PatientInfo <- PatientInfo[infection_case == "overseas inflow"]
patientinfo_ov <- PatientInfo[, -c("symptom_onset_date","deceased_date", "rel
eased_date")]
```

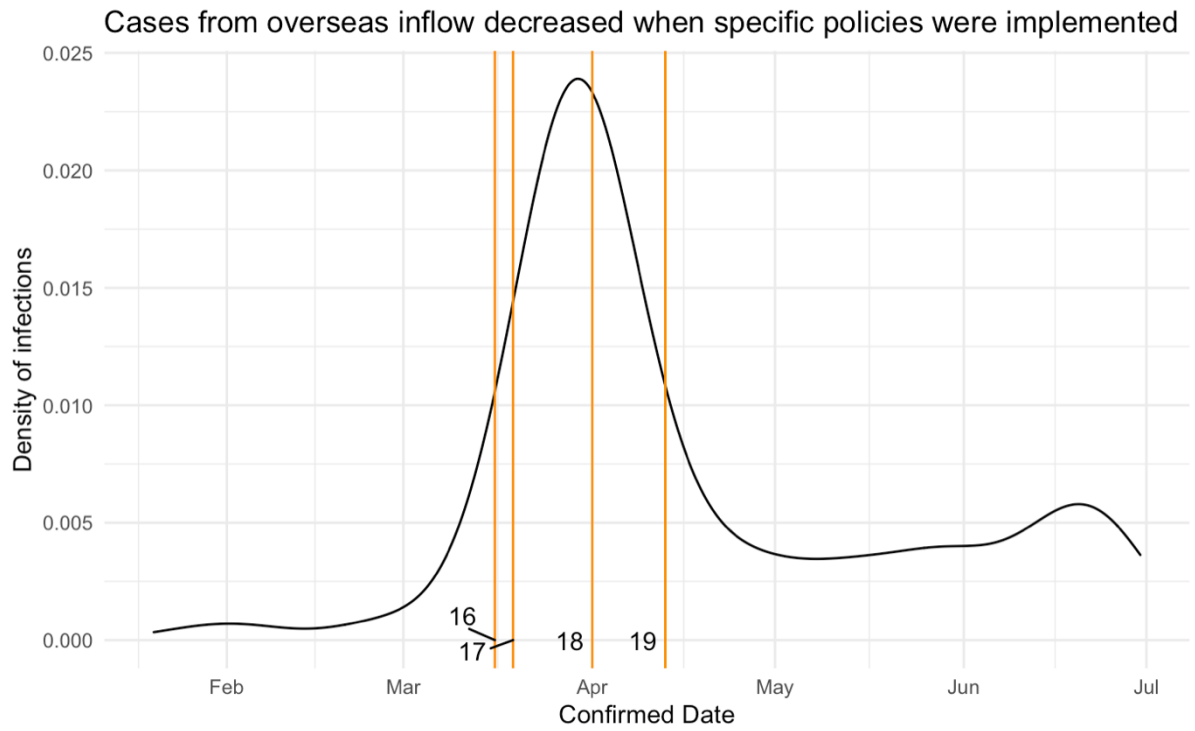
The highest infected people coming from overseas with the policies taken.

```
ggplot(patientinfo_ov, aes(x= confirmed_date)) + geom_density() + geom_vline(
data=policyov, mapping = aes(xintercept=as.numeric(start_date)))  + geom_text
_repel(data=policyov, mapping= aes(x=start_date, y=0, label=policy_id, vjust=
-0.4, hjust=0)) + theme_minimal()
```



As it is shown in this graph too, it can be tricky because the first policies were taken for certain states whereas the most effect in lowering the cases was when the policies of immigration included most countries at once. These policies were: 16-17-18-19 as the most effective ones in decreasing the infection rate. These policies were taken specifically as shown in the graph below.

```
policy_effect <- policyov[type == "Immigration" & policy_id %in% c(16, 17, 18
, 19)]
ggplot(patientinfo_ov, aes(x= confirmed_date)) + geom_density() + geom_vline(
data=policy_effect, mapping = aes(xintercept=as.numeric(start_date)), color=
'darkorange')  + geom_text_repel(data=policy_effect, mapping= aes(x=start_dat
e, y=0, label=policy_id, vjust=-0.4, hjust=0)) + theme_minimal() + labs(x='Co
nfirmed Date', y='Density of infections') + ggtitle("Cases from overseas infl
ow decreased when specific policies were implemented")
```



The specific details about these policies:

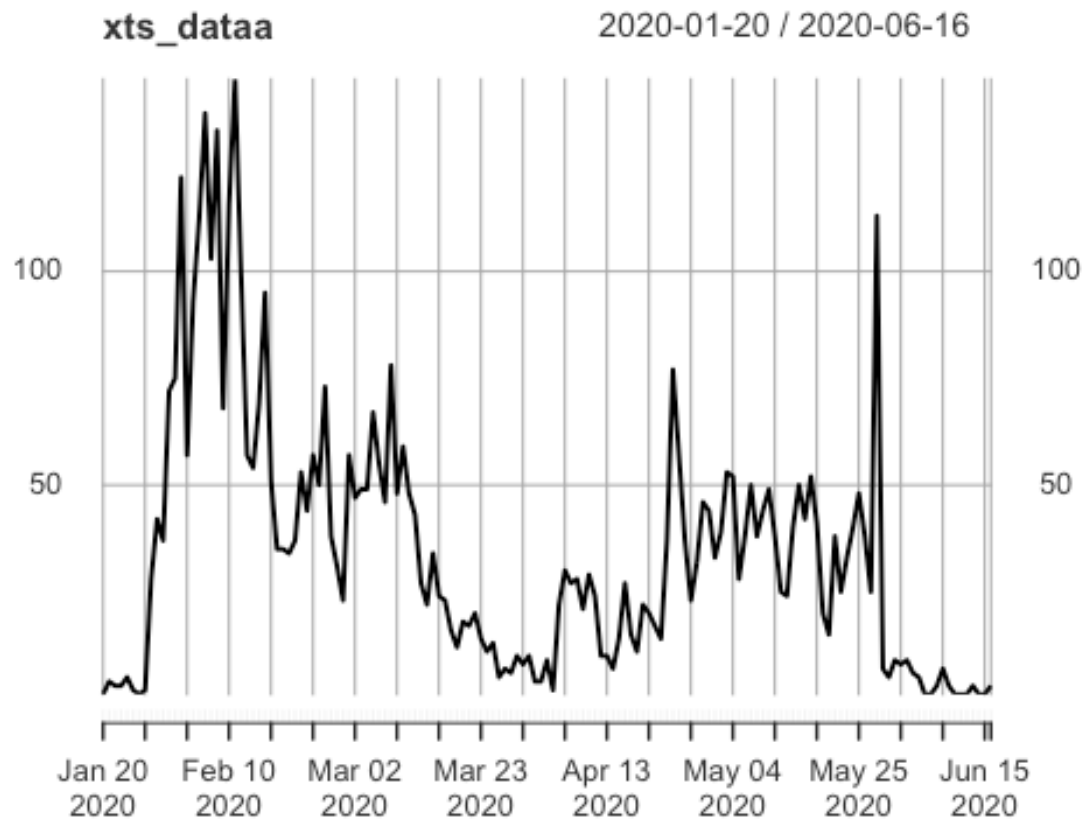
```
policy_effect[,c("policy_id","gov_policy","detail")]
```

```
##    policy_id                                    gov_policy
## 1:        16                Special Immigration Procedure
## 2:        17                Special Immigration Procedure
## 3:        18             Mandatory 14-day Self-Quarantine
## 4:        19 Mandatory Self-Quarantine & Diagonostic Tests
##                   detail
## 1:            from Europe
## 2: from all the countries
## 3: from all the countries
## 4:              from U.S.
```

The highest effect in lowering the confirmed cases coming from overseas, was almost after 2 weeks the policies 16 and 17 were taken. Due to these policies, on 16th POlicy it was Special Immigration Procedure from Europe taken on 16/03/2020 and on 17th Policy it was Special Immigration Procedure from all countries taken on 19/03/2020. Whereas the next policies, respectively policy 18 and 19 helped immediately in lowering the cases.

Same as the previous claim, this one too follows the same method; using Wilcoxon rank-sum test to compare control and test group just to be more correct in our final results.

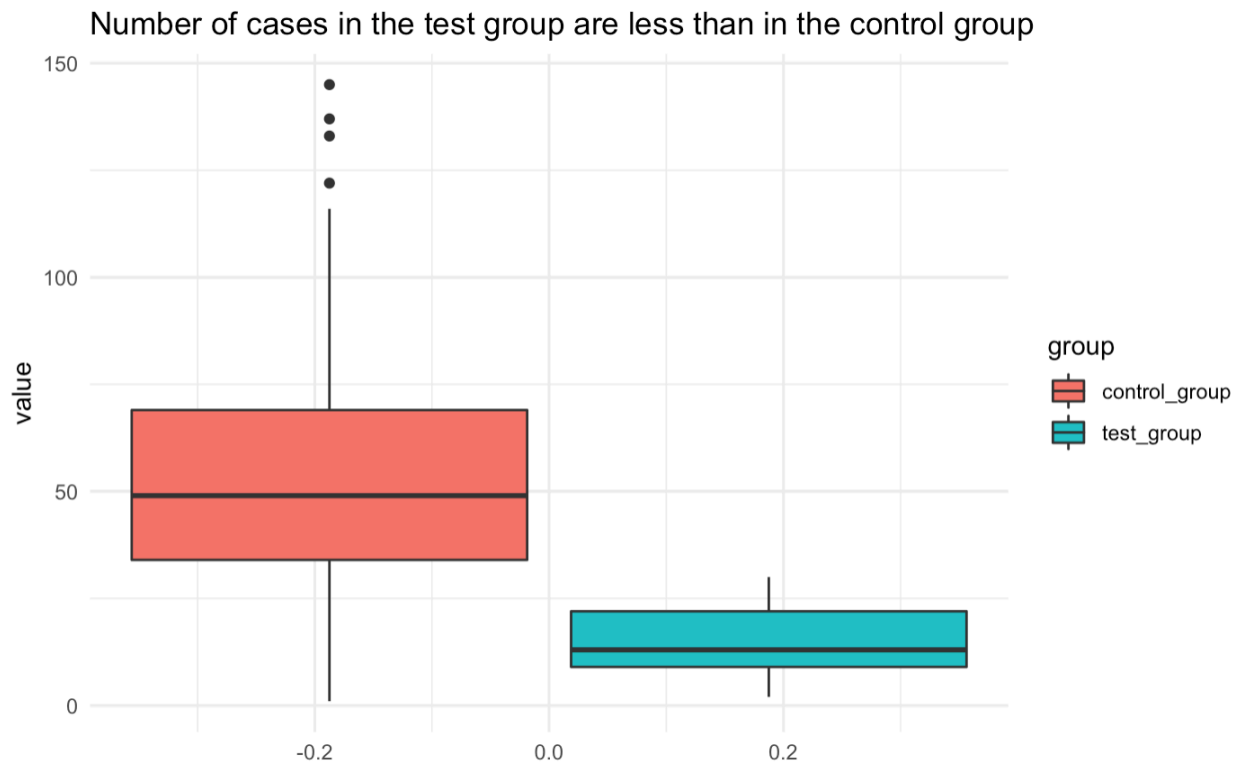
```
dates <- seq(as.Date("2020-01-20"), length=149, by="days")
ts_time<- patientInfo[,.N, by=confirmed_date]
xts_dataa <- xts(ts_time[,N], order.by=dates)
plot(xts_dataa)
```




```
control_group<- as.numeric(xts_dataa["/2020-03-16"])
summary(control_group)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   34.00   49.00   53.49   69.00  145.00

test_group <- as.numeric(xts_dataa["2020-03-16/2020-04-13"])
summary(test_group)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    9.00   13.00   14.97   22.00   30.00

dt <- data.table(group=c(rep("control_group", times=length(control_group)), r
ep("test_group", times=length(test_group))), value=c(control_group, test_grou
p))
print(dt)

##             group value
##  1: control_group     1
##  2: control_group     4
##  3: control_group     3
##  4: control_group     3
##  5: control_group     5
##  6: control_group     2
```

```
##  7: control_group     1
##  8: control_group     2
##  9: control_group    28
## 10: control_group    42
## 11: control_group    37
## 12: control_group    72
## 13: control_group    75
## 14: control_group   122
## 15: control_group    57
## 16: control_group    93
## 17: control_group   112
## 18: control_group   137
## 19: control_group   103
## 20: control_group   133
## 21: control_group    68
## 22: control_group   116
## 23: control_group   145
## 24: control_group    99
## 25: control_group    57
## 26: control_group    54
## 27: control_group    69
## 28: control_group    95
## 29: control_group    51
## 30: control_group    35
## 31: control_group    35
## 32: control_group    34
## 33: control_group    37
## 34: control_group    53
## 35: control_group    44
## 36: control_group    57
## 37: control_group    50
## 38: control_group    73
## 39: control_group    38
## 40: control_group    31
## 41: control_group    23
## 42: control_group    57
## 43: control_group    47
## 44: control_group    49
## 45: control_group    49
## 46: control_group    67
## 47: control_group    55
## 48: control_group    46
## 49: control_group    78
## 50: control_group    48
## 51: control_group    59
## 52: control_group    48
## 53: control_group    43
## 54: control_group    27
## 55: control_group    22
## 56: control_group    34
```

```
## 57: control_group    24
## 58:    test_group    24
## 59:    test_group    23
## 60:    test_group    16
## 61:    test_group    12
## 62:    test_group    18
## 63:    test_group    17
## 64:    test_group    20
## 65:    test_group    14
## 66:    test_group    11
## 67:    test_group    13
## 68:    test_group     5
## 69:    test_group     7
## 70:    test_group     6
## 71:    test_group    10
## 72:    test_group     8
## 73:    test_group    10
## 74:    test_group     4
## 75:    test_group     4
## 76:    test_group     9
## 77:    test_group     2
## 78:    test_group    22
## 79:    test_group    30
## 80:    test_group    27
## 81:    test_group    28
## 82:    test_group    21
## 83:    test_group    29
## 84:    test_group    24
## 85:    test_group    10
## 86:    test_group    10
##             group value

ggplot(data=dt, aes(y=value, fill=group)) + geom_boxplot() + theme_minimal()
+
        ggtitle("Number of cases in the test group are less than in the contr
ol group")
```




```
## Wilcoxon rank-sum test
# H0: control_group <= test_group. The average number of cases in the control
group are less or equal than the average
# number of cases in the test group.
# H1: control_group > test_group. The average number of cases in the control
group are more than the average
# number of cases in the test group.
wilcox.test(control_group, test_group,  exact = FALSE, alternative = "greater
")

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  control_group and test_group
## W = 1404.5, p-value = 6.588e-08
## alternative hypothesis: true location shift is greater than 0
```

Conclusion: The result shows that the null hypothesis is rejected for this claim too. The effect of policies was real and the cases had a decrease due to the government's policies and the main effect was showed 2 weeks later. However, there could also be additional reasons giving us this outcome.

Claim 6: The policies taken about education decreased the cases among children and teens. Questions related to claim: - Did the policies taken about education affect the confirmed cases among children?

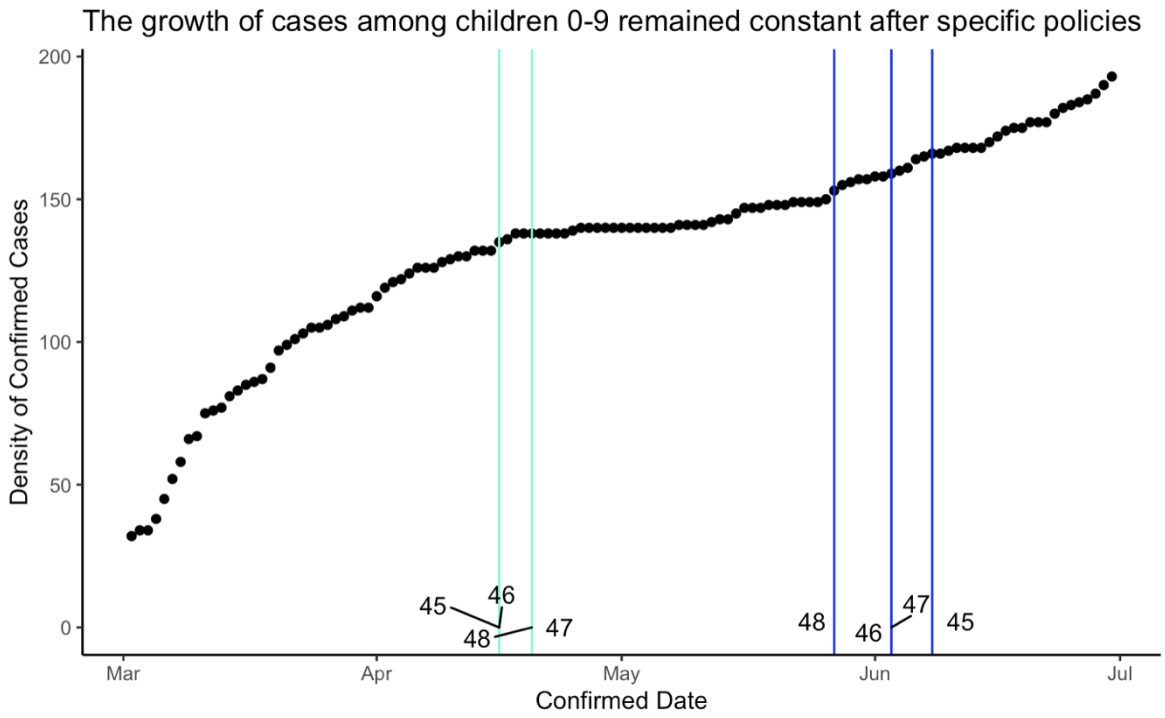
To check these, we need to find the start/end date of the education policies and we will find the cases for the age 0-20 and discover if the policies affecting Kindergarten, Elementary School, Middle School, High School and Children Daycare Center were effective enough. This can be done by observing the policy dataset, selecting only education policies and doing further analysis.

```
policy <- fread("archive/Policy.csv")
policy_edu <- policy[type == "Education"]
TimeAge <- fread('archive/TimeAge.csv')
TimeAge <- TimeAge[age == "0s" | age == "10s"]
```

Now that we have the age of children who attended school, cases in numbers and date of confirmed cases, we can follow the policies, how they affected for their respective age range during the peak of the corona virus. Specifically for each age group we have the policies taken for their respective school grades so we get the start and end date.

We compare both 0s and 10s in 2 different graphs with their specific start and end date of policies.

```
policy_children <- policy_edu[type== 'Education' & policy_id %in% c(45,46,47,
48)]
ggplot(TimeAge[age=='0s'], aes(x=date, y=confirmed)) + geom_point() +
  geom_vline(data = policy_children, mapping = aes(xintercept=as.numeric(star
t_date)), color='aquamarine') +
  geom_text_repel(data = policy_children, mapping = aes(x=start_date, y=0, la
bel=policy_id, vjust=-0.4, hjust=0)) + geom_vline(data = policy_children, map
ping = aes(xintercept=as.numeric(end_date)), color='blue')+
  geom_text_repel(data = policy_children, mapping = aes(x=end_date, y=0, labe
l=policy_id, vjust=-0.4, hjust=0)) + theme_classic() + labs(x='Confirmed Date
', y='Density of Confirmed Cases') + ggtitle("The growth of cases among child
ren 0-9 remained constant after specific policies")
```

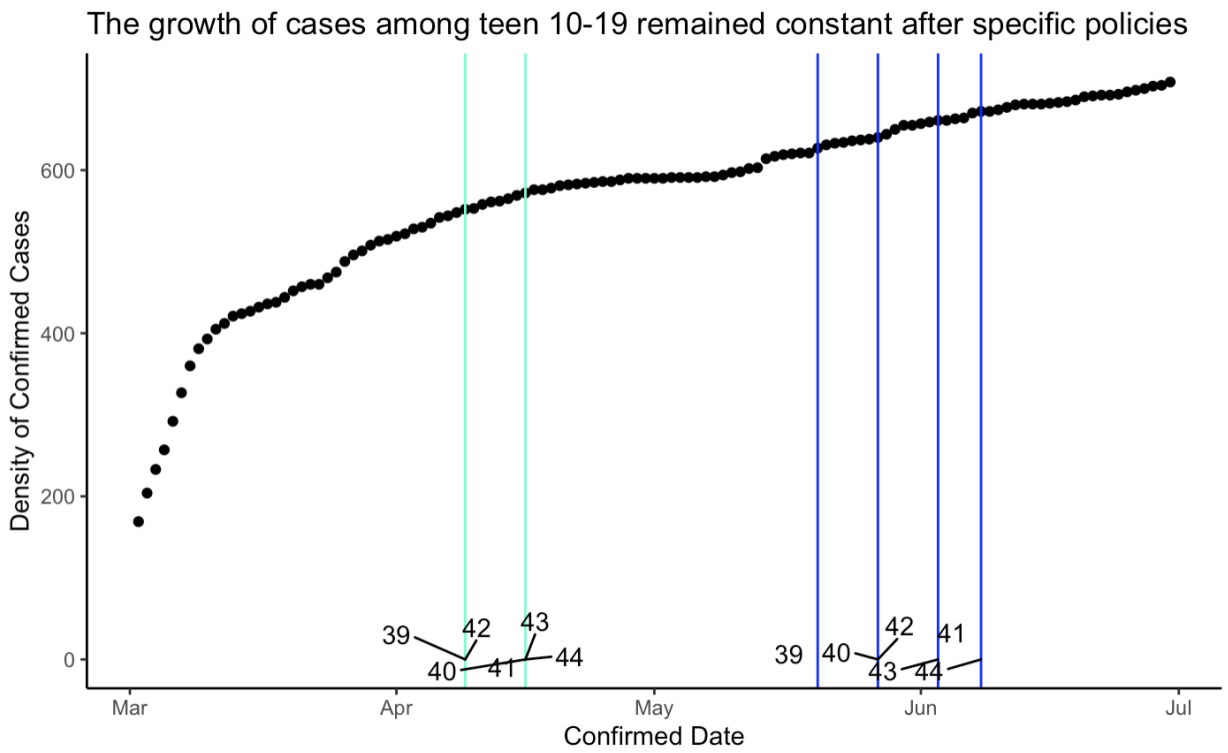


More details about these policies:

```
policy_children[,c("policy_id","gov_policy", "detail")]
```

```
##    policy_id                        gov_policy
## 1:        45 School Opening with Online Class
## 2:        46 School Opening with Online Class
## 3:        47 School Opening with Online Class
## 4:        48 School Opening with Online Class
##                                  detail
## 1: Elementary School (5th ~ 6th grade)
## 2:       Elementary School (4th grade)
## 3:       Elementary School (3rd grade)
## 4: Elementary School (1st ~ 2nd grade)
```

```
policy_teen <- policy_edu[type == 'Education' & policy_id %in% c(39,40,41,42,43,44)]
ggplot(TimeAge[age=='10s'], aes(x=date, y=confirmed)) + geom_point() +
  geom_vline(data = policy_teen, mapping = aes(xintercept=as.numeric(start_date)), color='aquamarine') +
  geom_text_repel(data = policy_teen, mapping = aes(x=start_date, y=0, label=policy_id, vjust=-0.4, hjust=0)) + geom_vline(data = policy_teen, mapping = aes(xintercept=as.numeric(end_date)), color='blue')+
  geom_text_repel(data = policy_teen, mapping = aes(x=end_date, y=0, label=policy_id, vjust=-0.4, hjust=0)) + theme_classic() + labs(x='Confirmed Date', y='Density of Confirmed Cases') + ggtitle("The growth of cases among teen 10-19 remained constant after specific policies")
```



This graph shows that the highest effect in lowering the confirmed cases did not happen. Since Policies taken about schools, apparently did not have any positive outcome in decreasing the confirmed cases among children up to 20years old, we can say that: Graphs show slightly effect. We can add a hypothesis that at least, these policies helped to maintain and keep a constant number of infections and prevented huge infections among children. For this, we continue with our analysis.

More details about the policies:

```
policy_teen[,c("policy_id","gov_policy", "detail")]
```

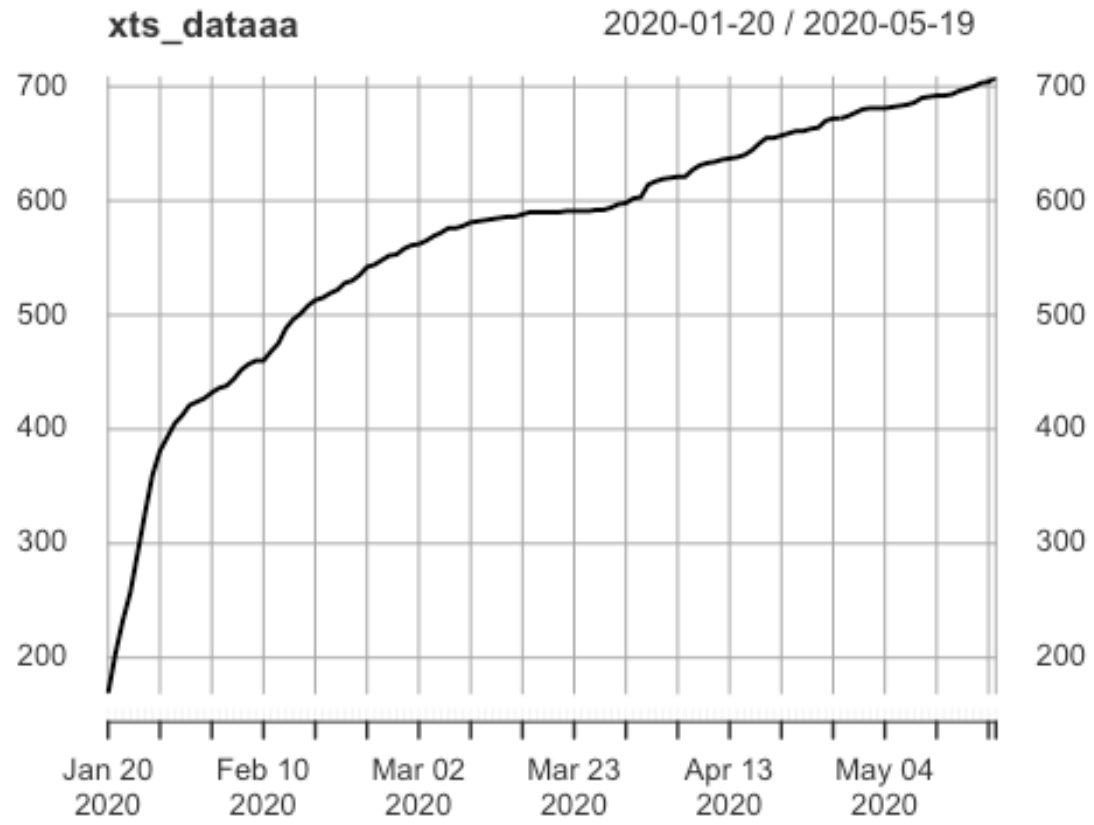
```
##    policy_id                        gov_policy                    detail
## 1:        39 School Opening with Online Class   High School (3rd grade)
## 2:        40 School Opening with Online Class   High School (2nd grade)
## 3:        41 School Opening with Online Class   High School (1st grade)
## 4:        42 School Opening with Online Class Middle School (3rd grade)
## 5:        43 School Opening with Online Class Middle School (2nd grade)
## 6:        44 School Opening with Online Class Middle School (1st grade)
```

Now, we can use the same method as on previous claims, by using XTS library, creating control and test group and having the Wilcoxon rank-sum test to give us the final result. We only conduct the statistical testing for the teens [10-19) but the methodology and implementation is exactly for the children [0,9).

```
ts_time<- TimeAge[ age=="10s", by=date]
```

```
## Warning in `[.data.table`(TimeAge, age == "10s", by = date): Ignoring by=
## because j= is not supplied
```

```r
dates <- seq(as.Date("2020-01-20"), length=nrow(ts_time), by="days")
xts_dataaa <- xts(ts_time[, confirmed], order.by=dates)
plot(xts_dataaa)
```




```r
control_group<- as.numeric(xts_dataaa["/2020-04-09"])
summary(control_group)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     169     460     558     517     590     631
```

```r
test_group <- as.numeric(xts_dataaa["2020-04-09/2020-06-08"])
summary(test_group)
```

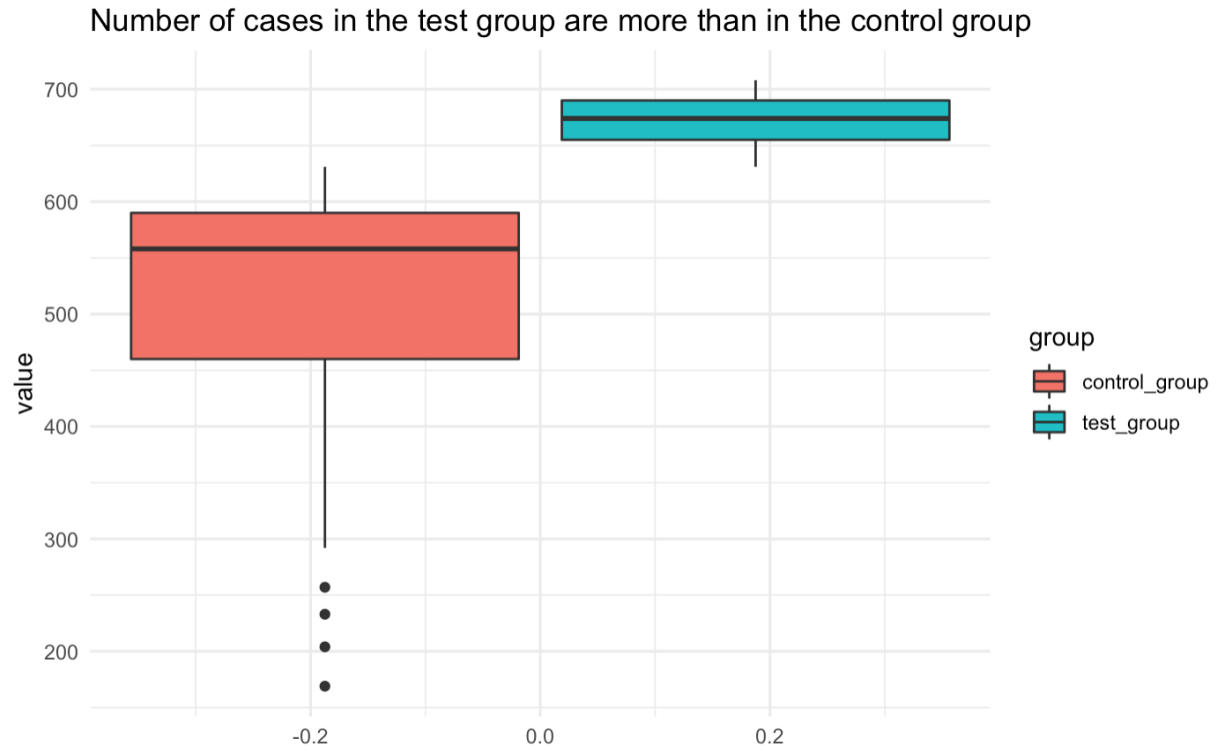
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   631.0   655.0   674.0   670.9   690.0   708.0
```

```r
dt <- data.table(group=c(rep("control_group", times=length(control_group)), r
ep("test_group", times=length(test_group))), value=c(control_group, test_grou
p))
print(dt)
```

```
##              group value
##   1: control_group   169
##   2: control_group   204
```

```
##   3: control_group   233
##   4: control_group   257
##   5: control_group   292
##  ---
## 118:    test_group   698
## 119:    test_group   700
## 120:    test_group   703
## 121:    test_group   704
## 122:    test_group   708
```

```
ggplot(data=dt, aes(y=value, fill=group)) + geom_boxplot() + theme_minimal()
+
        ggtitle("Number of cases in the test group are more than in the contr
ol group")
```



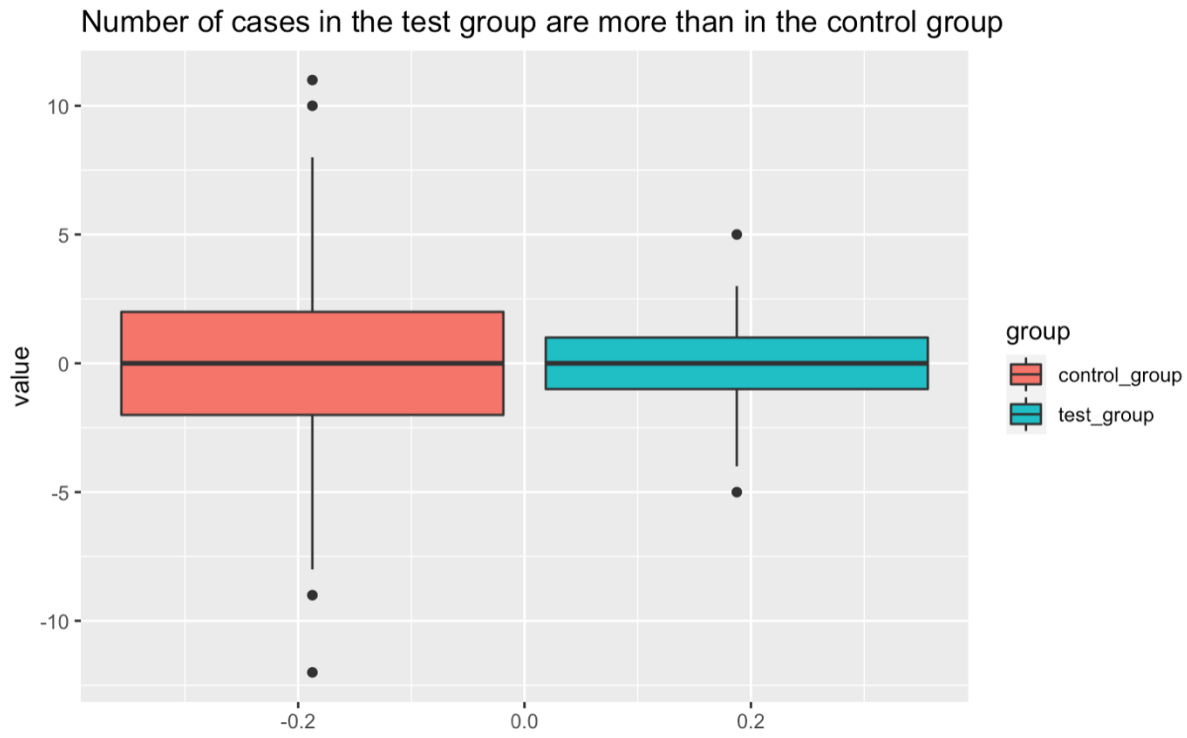

```
## Wilcoxon rank-sum test
# H0: control_group <= test_group. The average number of cases in the control
group are less or equal than the average
# number of cases in the test group.
# H1: control_group > test_group. The average number of cases in the control
group are more than the average
# number of cases in the test group.
wilcox.test(control_group, test_group,  exact = FALSE, alternative = "greater
")
```

```
##
##  Wilcoxon rank sum test with continuity correction
```

```
##
## data:  control_group and test_group
## W = 0.5, p-value = 1
## alternative hypothesis: true location shift is greater than 0
```

We perform a second alternative statistical test in which instead of using the number of cases per day as a determining factor, we use the growth of cases per pair of consecutive days. With this method we can investigate if the growth of cases increased, therefore increasing the slope of the cases curve, or if it decreased, therefore decreasing the slope of the curve.

```
control_group<- diff(control_group)
test_group<- diff(test_group)
dt <- data.table(group=c(rep("control_group", times=length(control_group)), r
ep("test_group", times=length(test_group))), value=c(control_group, test_grou
p))
print(dt)

##              group value
##   1: control_group    35
##   2: control_group    29
##   3: control_group    24
##   4: control_group    35
##   5: control_group    35
##  ---
## 116:    test_group     2
## 117:    test_group     2
## 118:    test_group     3
## 119:    test_group     1
## 120:    test_group     4

ggplot(data=dt, aes(y=value, fill=group)) + geom_boxplot()
```




```
## Wilcoxon rank-sum test
# H0: control_group <= test_group. The average daily increase of cases in the
control group are less or equal than the average
# daily increase of cases in the test group.
# H1: control_group > test_group. The average daily increase of cases in the
control group are more than the average
# daily increase of cases in the test group.
wilcox.test(control_group, test_group,  exact = FALSE, alternative = "greater
")

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  control_group and test_group
## W = 2114, p-value = 0.001937
## alternative hypothesis: true location shift is greater than 0
```

Conclusion: As it is shown also after Wilcoxon test took place, we fail to reject. There was no effect on lowering the cases of infection among children of age 10-19 even though the government took the policies 39,40,41,42,43,44 to follow online classes. Comparing before and after these policies were implemented, we can see that before the policies started, cases were increasing faster. Same thing can be said by viewing the graph of the cases after the policies ended. This can be seen because the policies might have kept the numbers constant, neither decreasing, nor increasing the cases. To investigate if the policies had an effect in keeping the numbers constant, it still requires a different methodology (like, analyzing the slope of per day cases) and it requires more time and knowledge to continue with this. The alternative test we conducted is similar to the methodology suggested in the

previous sentence but still fails to reject the null hypothesis. What is important an important finding, however, is the variance of the control group compared to the test group. The control group has by far more variance than the test group. Therefore, further analysis should be conducted to see if this difference in variance is statistically significant. The observations now show that the policies implemented seem to have decreased the fluctuations of daily cases for the period studied.