Case Study Group11

1.Import packages

```
library(data.table)
library(bit64)
library(ggplot2)
library(magrittr)
library(car) #Levene's test
library(tidyverse)
library(gghighlight)
```

2.Import datasets

```
setwd("/Users/luki/Dropbox/5.Semester_tum/Data_Visualization_and_R/Case_study/data")
PInfo <- fread("PatientInfo.csv", na.strings = "")
Policy <- fread('Policy.csv')
```

3.Data exploration

```
summary(PInfo)
```

```
##    patient_id            sex                age              country
##  Min.   :1000000001   Length:5165        Length:5165        Length:5165
##  1st Qu.:1000001292   Class :character   Class :character   Class :character
##  Median :2000000370   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2863634561
##  3rd Qu.:6001000116
##  Max.   :7000000019
##
##    province             city           infection_case     infected_by
##  Length:5165        Length:5165        Length:5165        Length:5165
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  contact_number     symptom_onset_date   confirmed_date
##  Length:5165        Min.   :2020-01-19   Min.   :2020-01-20
##  Class :character   1st Qu.:2020-02-29   1st Qu.:2020-03-04
##  Mode  :character   Median :2020-03-20   Median :2020-03-27
##                     Mean   :2020-04-05   Mean   :2020-04-10
##                     3rd Qu.:2020-05-23   3rd Qu.:2020-05-27
##                     Max.   :2020-06-28   Max.   :2020-06-30
```

```
##                     NA's   :4476         NA's   :3
##  released_date        deceased_date           state
##  Min.   :2020-02-05   Min.   :2020-02-19   Length:5165
##  1st Qu.:2020-03-20   1st Qu.:2020-03-02   Class :character
##  Median :2020-03-28   Median :2020-03-09   Mode  :character
##  Mean   :2020-04-03   Mean   :2020-03-17
##  3rd Qu.:2020-04-14   3rd Qu.:2020-03-30
##  Max.   :2020-06-28   Max.   :2020-05-25
##  NA's   :3578         NA's   :5099
```

```
str(PInfo)
```

```
## Classes 'data.table' and 'data.frame':   5165 obs. of  14 variables:
##  $ patient_id        :integer64 1000000001 1000000002 1000000003 1000000004 1000000005 1000000006 100
##  $ sex               : chr  "male" "male" "male" "male" ...
##  $ age               : chr  "50s" "30s" "50s" "20s" ...
##  $ country           : chr  "Korea" "Korea" "Korea" "Korea" ...
##  $ province          : chr  "Seoul" "Seoul" "Seoul" "Seoul" ...
##  $ city              : chr  "Gangseo-gu" "Jungnang-gu" "Jongno-gu" "Mapo-gu" ...
##  $ infection_case    : chr  "overseas inflow" "overseas inflow" "contact with patient" "overseas infl
##  $ infected_by       : chr  NA NA "2002000001" NA ...
##  $ contact_number    : chr  "75" "31" "17" "9" ...
##  $ symptom_onset_date: IDate, format: "2020-01-22" NA ...
##  $ confirmed_date    : IDate, format: "2020-01-23" "2020-01-30" ...
##  $ released_date     : IDate, format: "2020-02-05" "2020-03-02" ...
##  $ deceased_date     : IDate, format: NA NA ...
##  $ state             : chr  "released" "released" "released" "released" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

It's obvious that the data types of age and contact_number are wrong, so we have to change them. Also, they both have many missing values, so we will temporarily remove them for the sake of upcoming data visualization.

```
#make age a data type of factor
PInfo[,age:= factor(age, levels = c('0s', '10s', '20s', '30s', '40s', '50s', '60s', '70s', '80s', '90s'

#make contact_number a data type of integer
PInfo[,contact_number := as.integer(contact_number)]
```

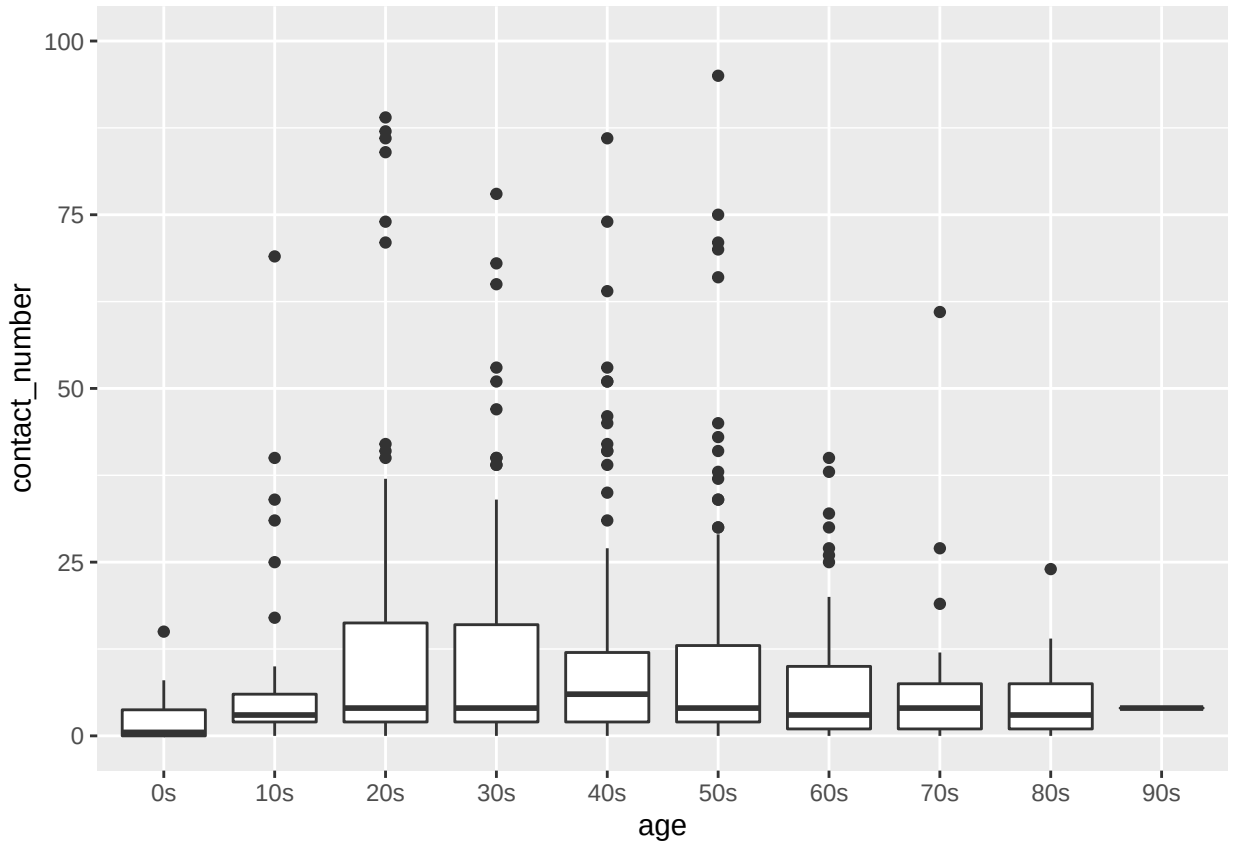
```
## Warning in eval(jsub, SDenv, parent.frame()): NAs introduced by coercion
```

```
## Warning in eval(jsub, SDenv, parent.frame()): NAs introduced by coercion to
## integer range
```

```
#remove missing values in the columns of age and contact_number
patient_rm_na <- PInfo[!is.na(age) & !is.na(contact_number)]
```

4.Data visualization

```
#create a boxplot to obeserve the association between age and contact_number
ggplot(patient_rm_na, aes(age, contact_number))+
  geom_boxplot()+
  ylim(0,100)
```



On the plot, we notice that age could play a role in the number of contact. Also, it appears that young people, which include college students, have a greater number of contact than the elderly. So we further claim that young people have a greater number of contact than the elderly. Therefore, we will statistically test: 1. the dependence between age and contact_number. 2. people aged at 20s have significantly higher number of contact than the rest.

5.Statistical testing - Kruskal-Wallis rank sum test

Because age is organized into various groups, we can use one-way ANOVA test to examine if there is any significant difference between the average contact_numbers in the various age groups.

```
res.aov <- aov(contact_number ~ factor(age), data = PInfo)
summary(res.aov)
```

```
##              Df  Sum Sq Mean Sq F value Pr(>F)
## factor(age)   9   25862    2874   0.618  0.783
## Residuals   761 3540108    4652
## 4394 observations deleted due to missingness
```

Nevertheless, we did not check ANOVA assumptions for our data,which are: 1. The variance across groups are homogeneous 2. The data of each factor level are normally distributed So let's check the homogenity of variances first.

```
leveneTest(contact_number ~ age, data = PInfo)
```

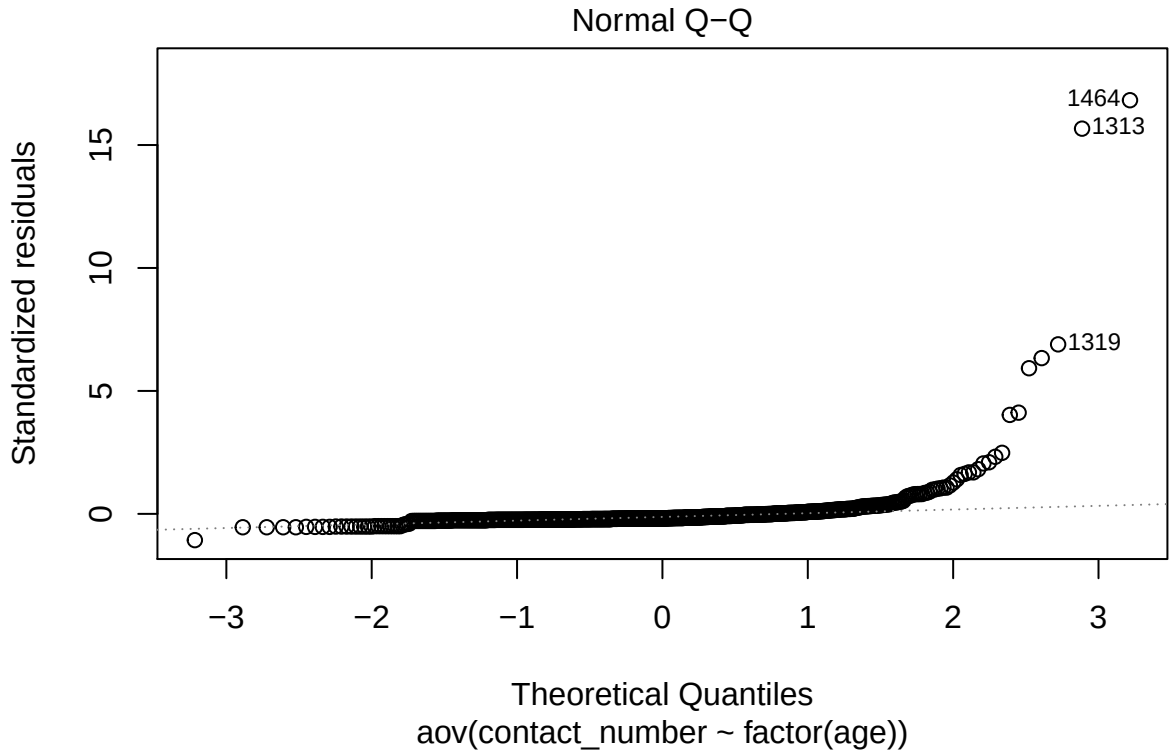
```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   9  0.6123 0.7872
```

```
##       761
```

From the output above we can see that the p-value is not less than the significance level of 0.05. This means that there is no evidence to suggest that the variance across groups is statistically significantly different. Therefore, we can assume the homogeneity of variances in the different treatment groups.

Then check the assumption of normality. We can check it by plotting Q-Q plot, whose y-xais is the quantiles of the residuals against x-axis of the quantiles of the normal distribution.

```
plot(res.aov, 2)
```



The normal probability plot of residuals is used to check the assumption that the residuals are normally distributed. It should approximately follow a dotted line.

To double validate the normality of our data, we will run hapiro-Wilk test.

```
# Extract the residuals
aov_residuals <- residuals(object = res.aov )

# Run Shapiro-Wilk test
shapiro.test(x = aov_residuals )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.22026, p-value < 2.2e-16
```

From the output above we can see that the p-value is less than the significance level of 0.05. This means that the null hypothesis that each age group is normally distributed is rejected. Therefore, we can confirm that the normality of each factor level does not exist.

In the case of data without normality, we can turn to another statistical test with less strict assumption. A non-parametric alternative to one-way ANOVA is Kruskal-Wallis rank sum test, which can be used when

ANOVA assumptions are not met.

```
kruskal.test(contact_number ~ age, data = PInfo)
```

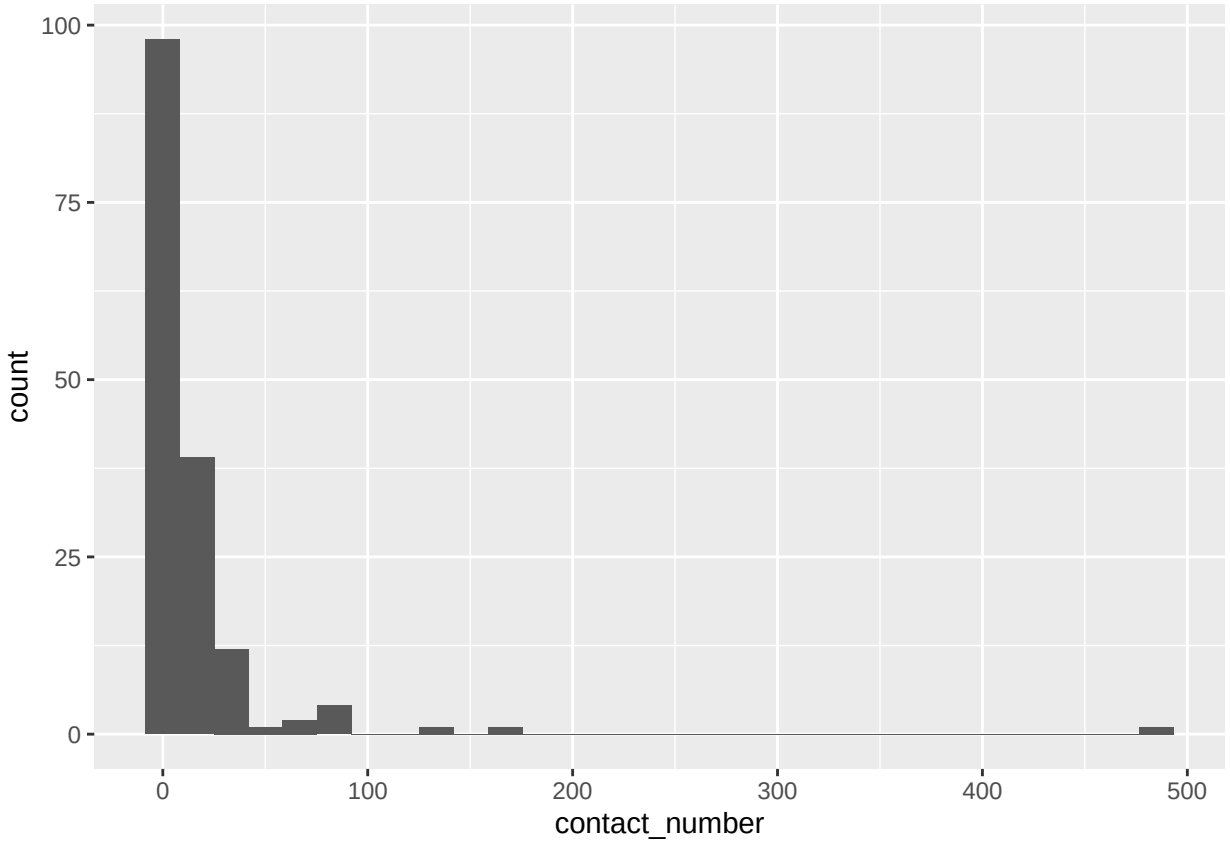
```
##
##  Kruskal-Wallis rank sum test
##
## data:  contact_number by age
## Kruskal-Wallis chi-squared = 24.454, df = 9, p-value = 0.003638
```

The Kruskal test suggests that there is a difference in the number of contact among the age groups since the P-value is lower than 0.05.

Now we will statistically test if people aged at 20s have significantly higher number of contact than the rest.

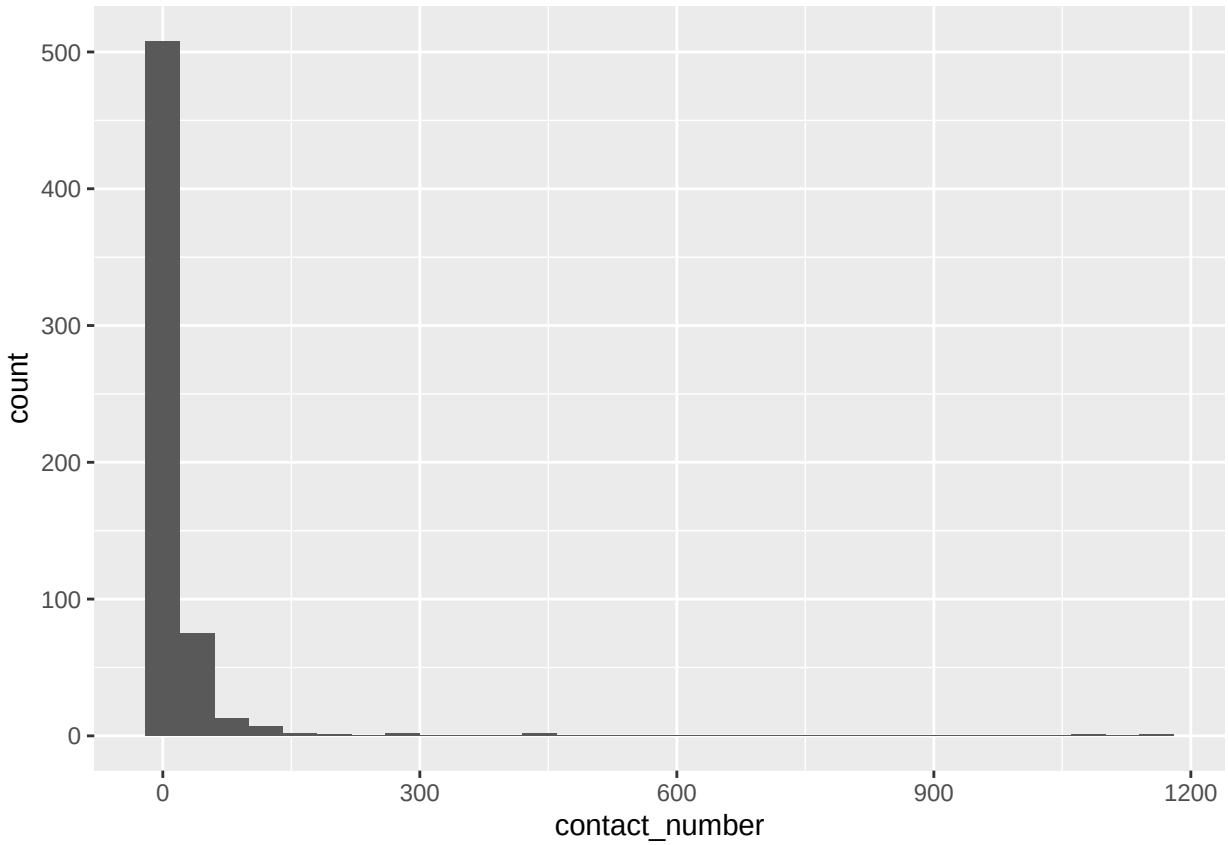
```
# the distribution of contact_number of patients aged 20s
ggplot(PInfo[age == "20s"], aes(contact_number))+
  geom_histogram()
```

```
## Warning: Removed 740 rows containing non-finite values (stat_bin).
```



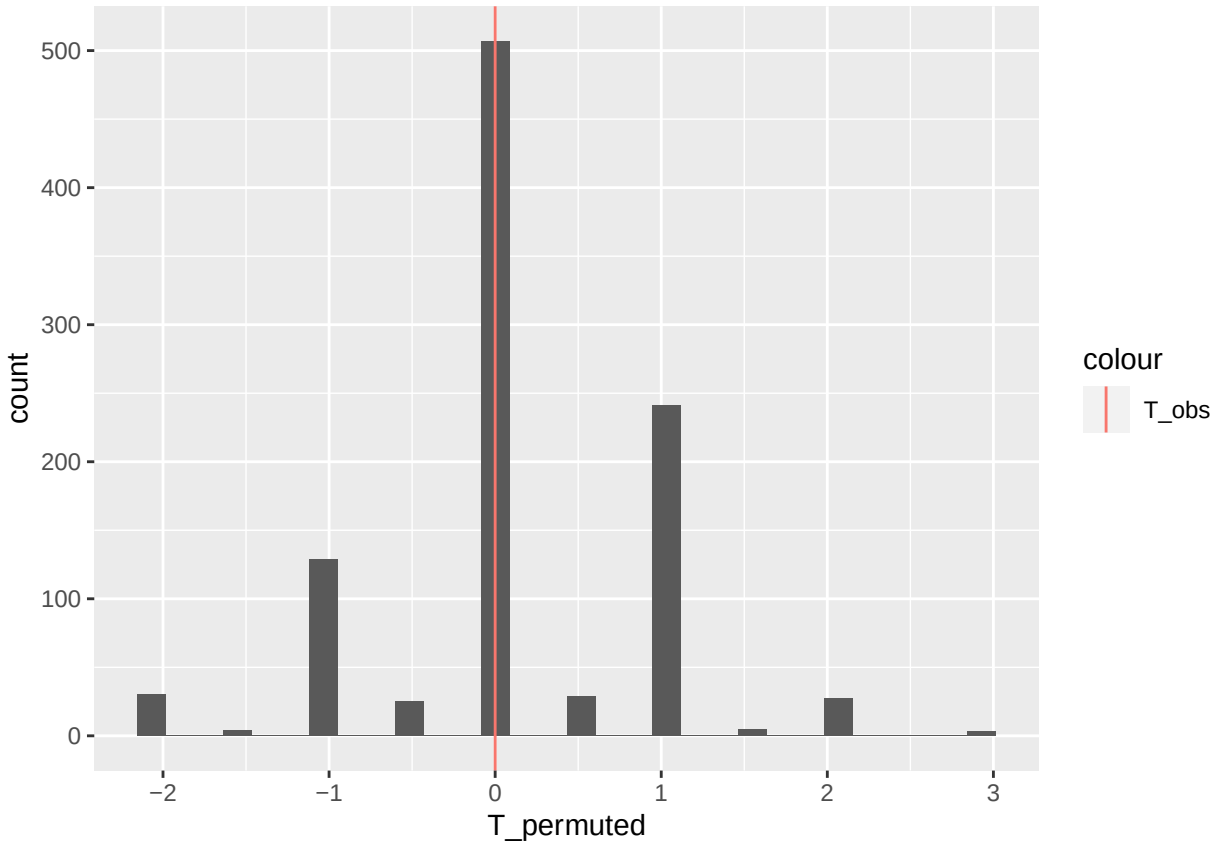

```
# the distribution of contact_number of patients aged 20s
ggplot(PInfo[age != "20s"], aes(contact_number))+
  geom_histogram()
```

```
## Warning: Removed 2274 rows containing non-finite values (stat_bin).
```



We notice that the two groups do not follow Gaussian distribution, so we will use permutation to test dependence.

```
#create a function of difference in median
median_dif_age <- function(dt, alt){
  dt[age == alt, median(contact_number, na.rm = T)]-
    dt[age != alt, median(contact_number, na.rm = T)]
}

#test statistics of data
T_obs <- median_dif_age(PInfo, "20s")

#create a vector of test statistics of 1000 permutated data
dt_permuted <- copy(PInfo)
set.seed(0)
T_permuted <- rep(NA, 1000)
for(i in 1:1000){
  # permute the genotype column in place
  dt_permuted[ , age:=sample(age)]
  # store the difference of medians in the i-th entry of T_permuted
  T_permuted[i] <- median_dif_age(dt_permuted, "20s")
}

#plot the test statistics of permutated data
ggplot( data.table(T_permuted), aes(x = T_permuted) ) +
  geom_histogram() +
  geom_vline( aes(xintercept=T_obs, color = "T_obs") )
```




```
#P-value
p_val_20 <- (sum(T_permuted>=T_obs)+1)/1001
```

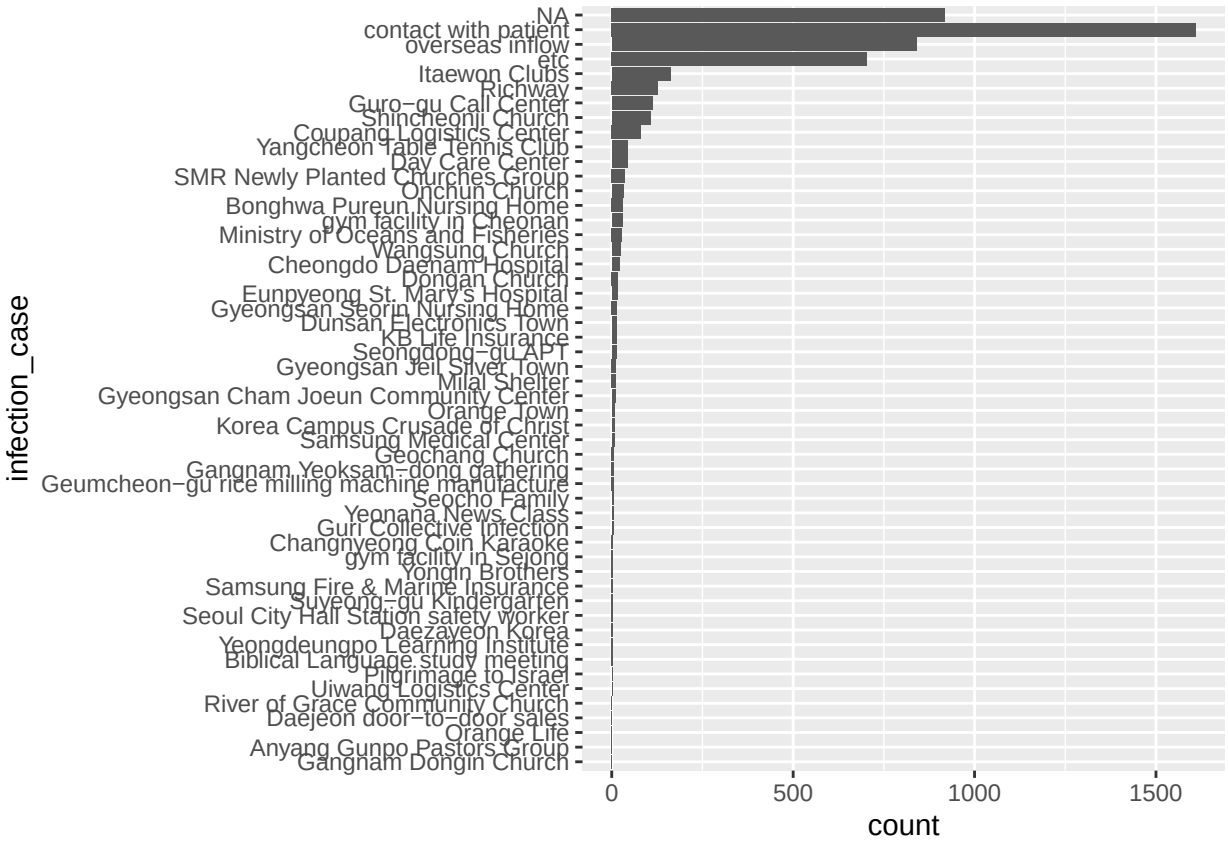
In the permutation test, the P-value is more than 0.05, so we can't claim patients aged at 20s are more likely to have a greater number of contact than the rest groups.

In conclusion, we prove our claim that age plays a role in the difference in contact_number. However, we cannot validate the fact that young patients(aged at 20s tested) have a greater number of contact than the rest.

6.Controlling for the effect of a third variable - infection_case to support a previously stated claim

Infection_case is like patients behavior indicating how they contact others. So we assume this is an indicator affecting the distribution of contact_number among age groups. Let's observe the categories within infection_case first.

```
g <- PInfo[, .(count = .N), by=infection_case]
g$infection_case <- factor(g$infection_case, levels = g$infection_case[order(g$count)])
ggplot(g, aes(infection_case, count))+
  geom_bar(stat = "identity")+
  coord_flip()
```



If the purpose of the analysis is to see where patients were infected, it would make sense to consider the use of all values. However, the purpose of this analysis is to find out what behavior of infection was at risk of death. Then, rather than labeling specific places, it is necessary to use this variable to know how infection happened and use it in the model. The values,'contact with patient','overseas inflow' remain the same, while 'unknown(etc)' & NA will be removed. The rest represents detailed cases and will be grouped together, labeling the value of 'group'.

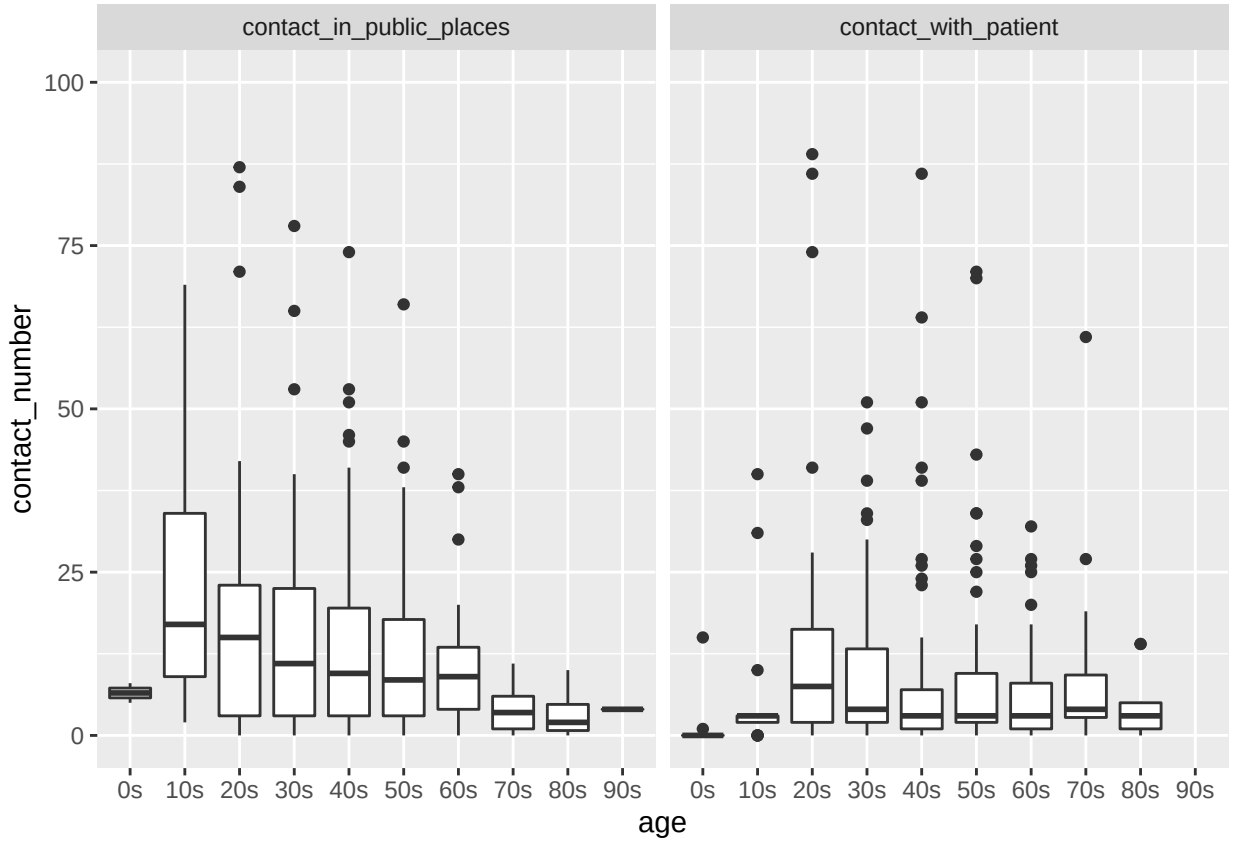
```
#remove infection case = etc or NA
patient_rm_na <- patient_rm_na[infection_case != "etc" | !is.na(infection_case)]

#create a function of categorizing infection_case
categorize <- function(X){
  if(X == 'overseas inflow'){
    X = "overseas_inflow"
  }else if(X == "contact with patient"){
    X = "contact_with_patient"
  }else{
    X = "contact_in_public_places"
  }
}

#categorize the data
patient_categorized <- copy(patient_rm_na)[, infection_case := sapply(infection_case, categorize)]

#the distribution of infection_case
table(patient_categorized$infection_case)
```

```
##
```

```
## contact_in_public_places     contact_with_patient          overseas_inflow
##                       262                      297                      156
```

Observe the distribution of contact_number of various age groups in the three scenarios of infection_case on the plots

```
ggplot(patient_categorized[infection_case != "overseas_inflow"], aes(age, contact_number))+
  geom_boxplot()+
  ylim(0,100)+
  facet_wrap(~infection_case)
```



The plot seems to show age is associated with contact_number when patients have contacted others in the public places or contacted with patients. So we will test both associations, with infection_case as a third variable.

```
#check ANOVA assumptions to see if we can use ANOVA or not
leveneTest(contact_number ~ age, data = patient_categorized[infection_case == "contact_in_public_places
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   9  2.2061 0.02219 *
##       252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
res.aov.public <- aov(contact_number ~ factor(age), data = patient_categorized[infection_case == "conta
aov_residuals.public <- residuals(object = res.aov.public )
shapiro.test(x = aov_residuals.public)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  aov_residuals.public
## W = 0.30844, p-value < 2.2e-16
```

```
#it turns out that the categorized data does not meet the requirement of Gaussian distribution, so we w

kruskal.test(contact_number ~ age, data = patient_categorized[infection_case == "contact_in_public_place
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  contact_number by age
## Kruskal-Wallis chi-squared = 22.343, df = 9, p-value = 0.007853
```

Since P-value is less than 0.05, we can claim that age is associated with contact_number when patients have contacted others in the public places.

Then test the association between age and contact_number in the case that patients have ever contacted with patients before.

```
#check ANOVA assumptions to see if we can use ANOVA or not
leveneTest(contact_number ~ age, data = patient_categorized[infection_case == "contact_with_patient"])
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   8   0.871 0.5413
##       288
```

```
res.aov.contact.patient <- aov(contact_number ~ factor(age), data = patient_categorized[infection_case =
aov_residuals.contact.patient <- residuals(object = res.aov.contact.patient )
shapiro.test(x = aov_residuals.contact.patient )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals.contact.patient
## W = 0.40661, p-value < 2.2e-16
```

```
#it turns out that the categorized data does not meet the requirement of Gaussian distribution, so we w

kruskal.test(contact_number ~ age, data = patient_categorized[infection_case == "contact_with_patient"]
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  contact_number by age
## Kruskal-Wallis chi-squared = 20.103, df = 8, p-value = 0.009954
```

Since P-value is less than 0.05, we can claim that age is associated with contact_number when patients have contacted with patients before.

So we conclude that infection_case as a third variable can support the association between age & contact_number.

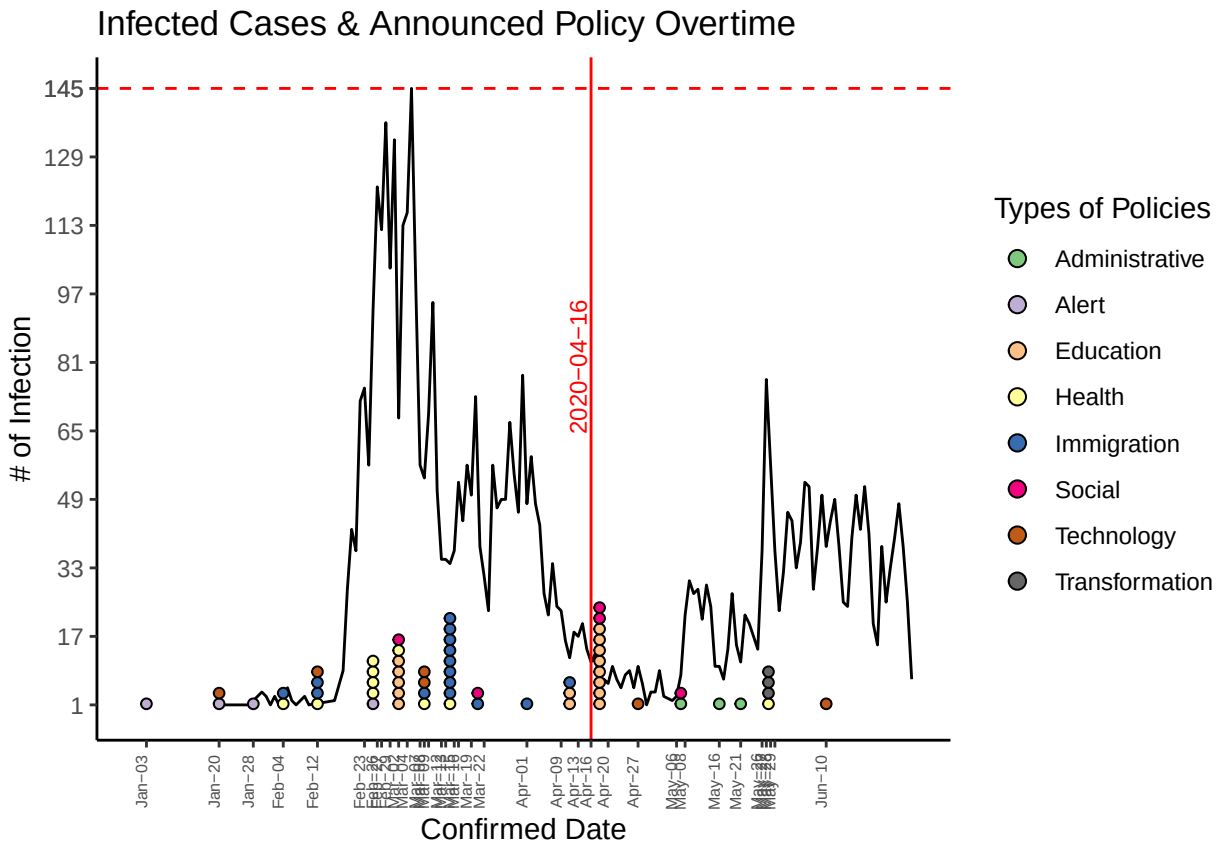
7.Policy

Infected Cases and Announced Policies Overtime

From the plot we can have an overview of number of infected cases and the amount of policies overtime. The largest number of cases increased was 145 cases per day at the beginning of March. Policies were announced between Feb and April to take control over Covid-19. 16th of April is the date the government announced the most policies.

```
amount <- PInfo[,.N,by=confirmed_date]
amount <- amount %>% arrange(confirmed_date)

ggplot()+
  geom_line(amount, mapping=aes(x=confirmed_date,y =N))+
  ggtitle("Infected Cases & Announced Policy Overtime")+
  labs(x="Confirmed Date", y="# of Infection", fill="Types of Policies")+
  geom_dotplot(Policy, mapping = aes(x=start_date, fill = type), binwidth=5, dotsize=.5,stackgroups = T
  scale_x_date(breaks = unique(Policy$start_date),date_labels="%b-%d")+
  scale_fill_brewer(palette = "Accent")+
  geom_hline(yintercept=max(amount$N), colour="red",linetype="dashed")+
  scale_y_continuous(breaks = sort(c(seq(min(amount$N),max(amount$N), length.out = 10),max(amount$N))))+
  geom_vline(xintercept = as.Date("2020-04-16"),colour="red")+
  geom_text(aes(x=as.Date("2020-04-16")-3,y=80,label="2020-04-16"),colour="red",size=3,angle=90)+
  theme_classic()+
  theme(axis.text.x=element_text(angle=90,size=6,vjust=0.05))
```

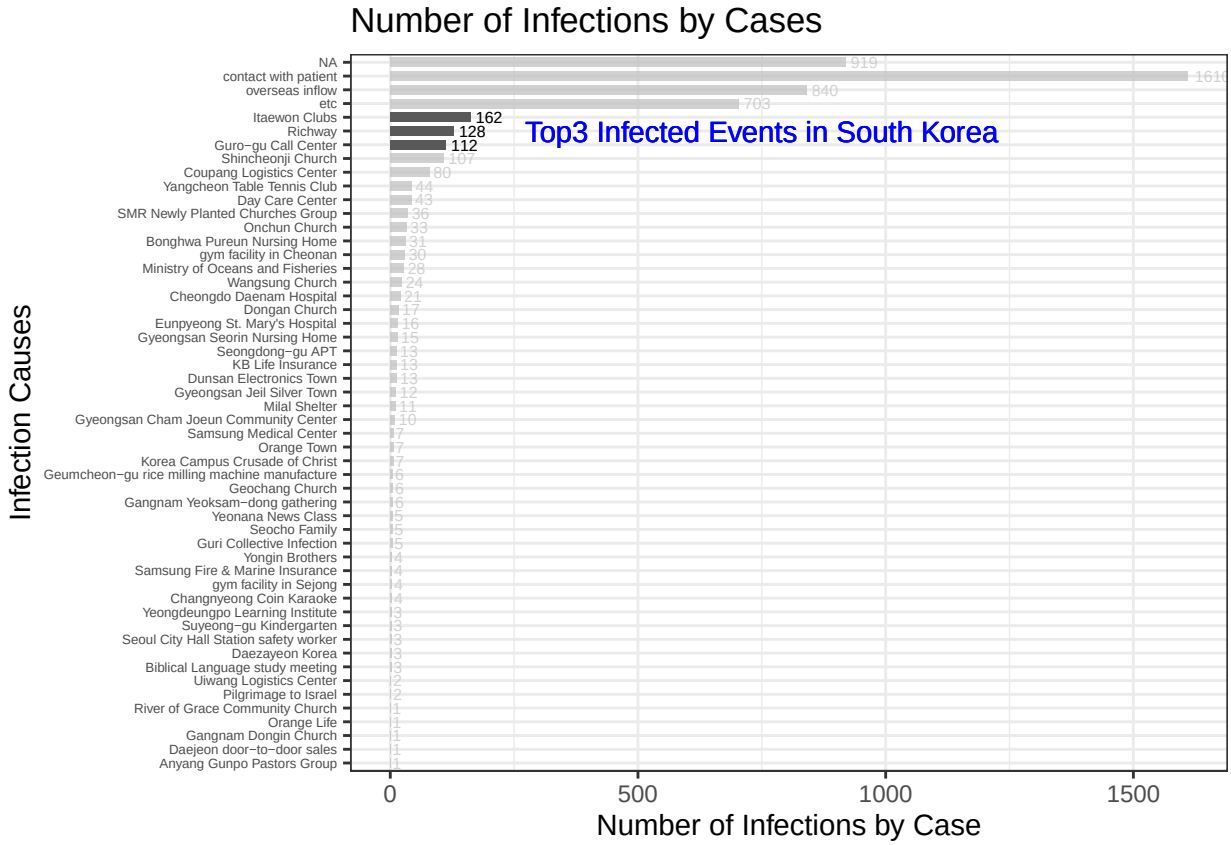
```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```



Infections by Cases

```
NInfectionCase <- PInfo[,.N,by=infection_case]

ggplot(NInfectionCase,aes(x=N, y = reorder(infection_case,N)))+
```

```
  geom_bar(stat = 'identity',width=.7)+
  theme_bw()+
  theme(axis.text.y = element_text(size = 4.5))+
  geom_text(aes(label=N),size = 2,hjust =(-.2) ,colour= "black")+
  labs(x="Number of Infections by Case", y ="Infection Causes")+
  ggtitle("Number of Infections by Cases")+
  gghighlight(N>110, N<700)+
  geom_text(aes(x=750,y=47,label='Top3 Infected Events in South Korea'),colour="Blue")
```



"Itaewon Clubs", "Richway" and "Guro-gu Call Center" are top3 infected events in South Korea apart from "contact with patient", "overseas inflow" and "etc".

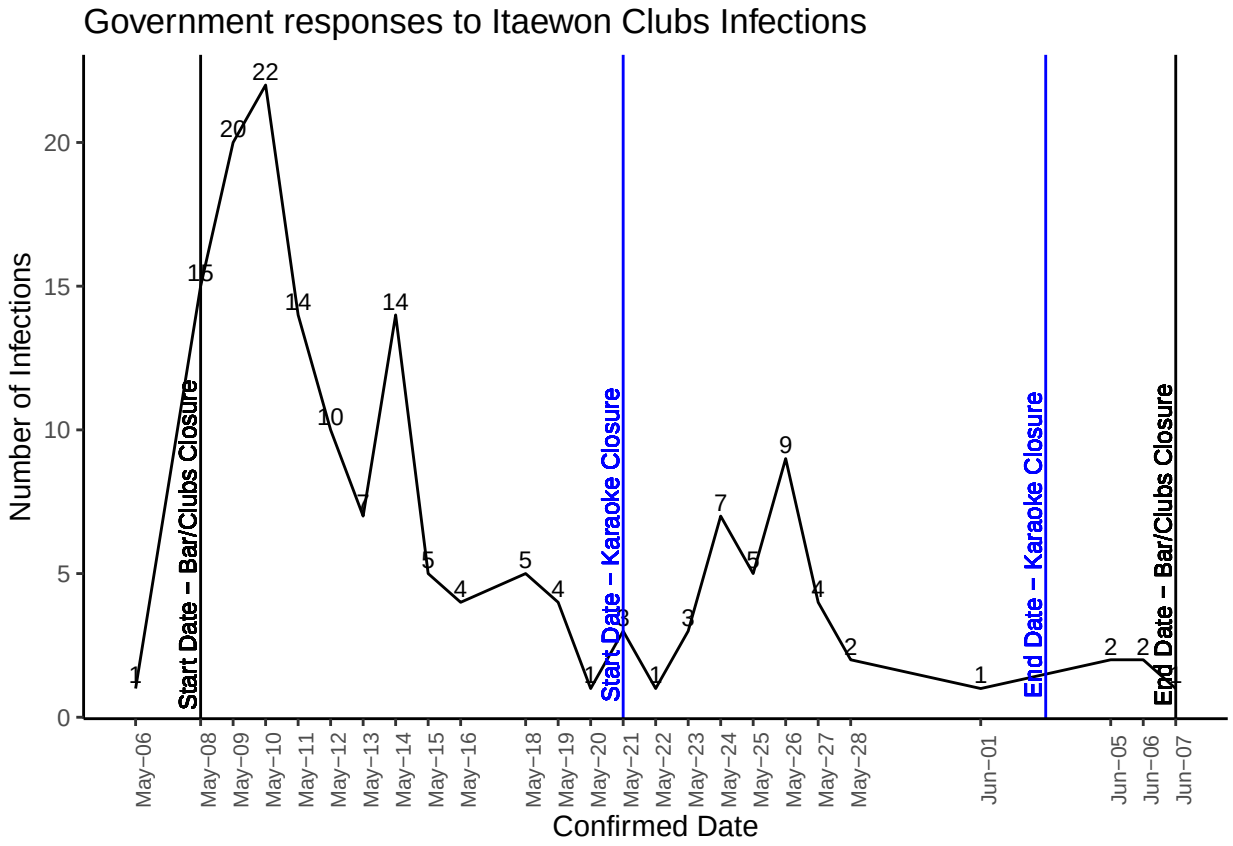
8.Itaewon Clubs

As soon as the case happened, government of South Korea have taken the action immediately (2020-05-08) by closing bars and clubs.

Following with the time line of Itaweon Clubs Infection Case:

```
Cases_Category <- unique(PInfo$infection_case)

Itaewon_Clubs <- PInfo[infection_case == "Itaewon Clubs", .N, by=confirmed_date]

Itaewon <- ggplot(Itaewon_Clubs, aes(x=confirmed_date, y = N))+geom_line()
Itaewon +
  geom_vline(xintercept = c(Policy[gov_policy == "Close bars and clubs"]$start_date,Policy[gov_policy ==
  ggtitle("Government responses to Itaewon Clubs Infections")+
```

```
  geom_text(aes(label = N),size = 3,vjust=(-.3),colour = "black")+
  labs(x="Confirmed Date", y="Number of Infections")+
  scale_x_date(breaks=Itaewon_Clubs$confirmed_date,date_labels = "%b-%d")+
  geom_text(aes(x=as.Date("2020-05-08")-0.4,y=6,label="Start Date - Bar/Clubs Closure"),colour="Black",a
  geom_text(aes(x=as.Date("2020-06-07")-0.4,y=6,label="End Date - Bar/Clubs Closure"),colour="Black",ang
  geom_vline(xintercept=c(Policy[gov_policy == "Close karaoke"]$start_date, Policy[gov_policy == "Close
  geom_text(aes(x=as.Date("2020-05-21")-0.4,y=6,label="Start Date - Karaoke Closure"),colour="Blue",angl
  geom_text(aes(x=as.Date("2020-06-03")-0.4,y=6,label="End Date - Karaoke Closure"),colour="Blue",angle=
  theme_classic()+
  theme(axis.text.x=element_text(angle=90,size=8))
```



As we saw the trend, we doubt that will the trend of decreasing is the result of announced policies. We are going into detail of the case to see if policies have significant influences on infected numbers. Assume that people are pub-lovers and would still go to bars if government did not close all the bars and clubs.

Do Policies actually influenced the confirmed cases in SK?

Picking out the policies category of Administrative, including "Close bars and clubs"... which are more relative to the public.

```
dt_PAdministrative <- Policy[type=="Administrative"]

#create a column of 14 days before
dt_PAdministrative[, FT_before := start_date-14]
#create a column of 14 days after
dt_PAdministrative[,FT_after:= start_date+14]
```

```
amount[,CumSum := cumsum(N)]
```

```
## Warning in `[.data.table`(amount, , `:=`(CumSum, cumsum(N))):
## Invalid .internal.selfref detected and fixed by taking a (shallow) copy of the
## data.table so that := can add this new column by reference. At an earlier point,
## this data.table has been copied by R (or was created manually using structure()
## or similar). Avoid names<- and attr<- which in R currently (and oddly) may
## copy the whole data.table. Use set* syntax instead to avoid copying: ?set, ?
## setnames and ?setattr. If this message doesn't help, please report your use case
## to the data.table issue tracker so the root cause can be fixed or this message
## improved.
```

```
#calculate sum of infections in duration
#close bars and clubs - CumSum 14days before
calculation <- amount[confirmed_date == (dt_PAdministrative$start_date[1]-1)]$CumSum-
  amount[confirmed_date == (dt_PAdministrative$FT_before[1]-1)]$CumSum
dt_PAdministrative[policy_id == 54,FT_before_CumSum := calculation]

#local gov order - CumSum 14 days before
calculation <- amount[confirmed_date == (dt_PAdministrative$start_date[2]-1)]$CumSum-
  amount[confirmed_date == (dt_PAdministrative$FT_before[2]-1)]$CumSum
dt_PAdministrative[policy_id == 55,FT_before_CumSum := calculation]

#close karaoke - CumSum 14 days before
calculation <- amount[confirmed_date == (dt_PAdministrative$start_date[3]-1)]$CumSum-
  amount[confirmed_date == (dt_PAdministrative$FT_before[3]-1)]$CumSum
dt_PAdministrative[policy_id == 56,FT_before_CumSum := calculation]

#close bars and clubs - CumSum 14days after
calculation <- amount[confirmed_date == (dt_PAdministrative$FT_after[1]+14)]$CumSum-
  amount[confirmed_date == (dt_PAdministrative$FT_after[1])]$CumSum

dt_PAdministrative[policy_id == 54,FT_after_CumSum := calculation]

#local gov order - CumSum 14days after
calculation <- amount[confirmed_date == (dt_PAdministrative$FT_after[2]+14)]$CumSum-
  amount[confirmed_date == (dt_PAdministrative$FT_after[2])]$CumSum

dt_PAdministrative[policy_id == 55,FT_after_CumSum := calculation]

#close karaoke - CumSum 14 days after
calculation <- amount[confirmed_date == (dt_PAdministrative$FT_after[3]+14)]$CumSum-
  amount[confirmed_date == (dt_PAdministrative$FT_after[3])]$CumSum

dt_PAdministrative[policy_id == 56,FT_after_CumSum := calculation]
```

H0: Administrative policies has no effect on confirmed cases –> H0: policy announced before - policy announced after = 0

t-test

```
diff <- dt_PAdministrative$FT_before_CumSum[1] - dt_PAdministrative$FT_after_CumSum[1]
dt_PAdministrative[policy_id == 54, diff := diff ]
```

```r
diff <- dt_PAdministrative$FT_before_CumSum[2] - dt_PAdministrative$FT_after_CumSum[2]
dt_PAdministrative[policy_id == 55, diff := diff ]

diff <- dt_PAdministrative$FT_before_CumSum[3] - dt_PAdministrative$FT_after_CumSum[3]
dt_PAdministrative[policy_id == 56, diff := diff ]
```

```r
a <- dt_PAdministrative$FT_before_CumSum
b <- dt_PAdministrative$FT_after_CumSum
d <- a-b

t.test(a,b,alternative ="two.sided" ,paired=T)
```

```
##
##  Paired t-test
##
## data:  a and b
## t = -8.1926, df = 2, p-value = 0.01457
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -571.4369 -177.8964
## sample estimates:
## mean of the differences
##               -374.6667
```

Result: reject H0

There is no statistical evidence to show that implementation of targeted policies do not effect infection numbers.