Group_25_CaseStudy_Final

Dilara Ademoğlu - 03722489
Aynur Süne - 03725788
Anna Illarionova - 03678137
Samira John - 03744714

```r
Case <- fread('./data/Case.csv')
PatientInfo <- fread('./data/PatientInfo.csv')
Policy <- fread('./data/Policy.csv')
Region <- fread('./data/Region.csv')
SearchTrend <- fread('./data/SearchTrend.csv')
SeoulFloating <- fread('./data/SeoulFloating.csv')
Time <- fread('./data/Time.csv')
TimeAge<- fread('./data/TimeAge.csv')
TimeGender <- fread('./data/TimeGender.csv')
TimeProvince <- fread('./data/TimeProvince.csv')
Weather <- fread('./data/Weather.csv')
```

What plays into the virus's hands? – An Introduction

Due to the rapid spread, the varying courses of the disease and the increased presence of the corona virus in everyday life, many people did already realize that each individual is confronted with the risk of a severe infection. Many people are currently asking themselves how high their own risk of contracting the disease is, in order to adapt their own behavior and, if necessary, to better understand the regulations of politics.

Through the German media we get an overview of the typical characteristics of a patient. For example, mainly older people are seriously affected by Covid-19, but also pre-existing conditions play a decisive role. The following Case Study aims to further investigate these characteristics and compare observations in the German population, mainly communicated by the media, with the data of the South Korean population.

We have paid special attention to characteristics that we typically use to begin to characterize a person, such as gender, age and origin/place of residence. So the following analysis deals with three questions:

1. If I am a man does it mean I have higher chances to die?
2. f I live in a province with a bigger population does it mean I have higher chances to get the virus because my province would have more confirmed cases?
3. If I am young does it mean I will get better sooner, than somebody who is old?

Data Preparation

We first start with necessary data manipulation and wrangling. For our analysis, we have collected additional population [1] data for each province.

```r
a <- length(TimeProvince[, province])
b <- length(unique(TimeProvince[,province]))
```

[1] *Data is obtained from here: **http://index.go.kr/potal/main/EachDtlPageDetail.do?idx_cd=1007#quick_01**

```
province_cases_dt <- TimeProvince[province != "",
                                  .(province, confirmed, released, deceased)] %>%
  group_by(province) %>%
  as.data.table(.) %>%
  .[(a-b+1):a] %>%
  .[, proportion_confirmed := confirmed / sum(confirmed) * 100]

### Province Population - external data
province_population_dt <- data.table(province_cases_dt[,province],
                                     population = c("9662", "3373", "2432", "2944", "1494",
                                                    "1509","1147", "331", "13238", "1517",
                                                    "1626", "2188", "1803", "1773", "2665",
                                                    "3350", "660")
                                     )

province_population_dt <- province_population_dt %>%
  rename(province = V1) %>%
  .[order(.$province)]

province_cases_dt <- province_cases_dt %>%
  .[order(.$province)] %>%
  merge(.,province_population_dt, by="province")

province_cases_dt$population <- as.integer(province_cases_dt$population)
province_cases_dt <- province_cases_dt %>%
   .[order(.$population)]

PatientInfo$contact_number <- gsub('-',NA,PatientInfo$contact_number)
PatientInfo$recovery_days[PatientInfo$recovery_days==""] <- NA
PatientInfo[, symptom_onset_date := as.Date(symptom_onset_date)]
PatientInfo[, confirmed_date := as.Date(confirmed_date)]
PatientInfo[, released_date := as.Date(released_date)]
PatientInfo[, deceased_date := as.Date(deceased_date)]
PatientInfo[, contact_number := as.numeric(contact_number)]
PatientInfo[, infected_by := as.numeric(infected_by)]
PatientInfo[, recovery_days := as.numeric(released_date - confirmed_date)]
PatientInfo$age <-factor(PatientInfo$age,
                         levels = c("0s", "10s", "20s", "30s", "40s", "50s", "60s",
                                    "70s", "80s", "90s", "100s"))

PatientInfo_age = copy(PatientInfo)
PatientInfo_age$age <- gsub('s', '', PatientInfo_age$age)
PatientInfo_age[, age := as.numeric(age)]
patient_recovery_age <- PatientInfo_age %>%
   filter(is.na(age)==F) %>%
   filter(is.na(recovery_days)==F) %>%
   select(recovery_days, age) %>%
   group_by(age) %>%
   summarise(mean = mean(recovery_days))
as.data.table(patient_recovery_age)
```

```
##     age     mean
##  1:   0 24.52632
##  2:  10 20.81538
##  3:  20 23.48166
##  4:  30 23.19431
##  5:  40 23.65145
##  6:  50 23.88000
##  7:  60 26.83237
##  8:  70 32.80723
##  9:  80 34.73333
## 10:  90 28.78571
## 11: 100 66.00000
```

```
patient_recovery_age_all <- PatientInfo_age %>%
   filter(is.na(age)==F) %>%
   filter(is.na(recovery_days)==F) %>%
   select(recovery_days, age)
as.data.table(patient_recovery_age_all)
```

```
##       recovery_days age
##    1:            13  50
##    2:            32  30
##    3:            20  50
##    4:            16  20
##    5:            24  20
##   ---
## 1572:            46  30
## 1573:            32  20
## 1574:            12  10
## 1575:            34  30
## 1576:            14  30
```

```
patient_recovery_age_all_categorical <- PatientInfo %>%
   filter(is.na(age)==F) %>%
   filter(is.na(recovery_days)==F) %>%
   select(recovery_days, age)
as.data.table(patient_recovery_age_all_categorical)
```

```
##       recovery_days age
##    1:            13 50s
##    2:            32 30s
##    3:            20 50s
##    4:            16 20s
##    5:            24 20s
##   ---
## 1572:            46 30s
## 1573:            32 20s
## 1574:            12 10s
## 1575:            34 30s
## 1576:            14 30s
```

```
PatientInfo[sex == "female", .N]
```

```
## [1] 2218
```

```
PatientInfo[sex == "male", .N]
```

```
## [1] 1825
```

```r
PatientInfo[, sex := as.factor(sex)]

patientState <- PatientInfo[sex != "", .(sex, state)]
patientState[state != "deceased", state:="alive"]
patientState[, state:=as.factor(state)]


path_dt <- Case %>%
  select(province, infection_case, confirmed) %>%
  as.data.table(path_dt)

path_dt2 <- Case %>% group_by (infection_case) %>%
  summarise(confirmed = sum(confirmed)) %>%
  arrange(desc(confirmed))

top_5_infection_cases <- path_dt2$infection_case[1:5]

path_dt3 <- data.table(path_dt[(province=='Daegu'| province=='Gyeongsangbuk-do') &
                                 (
                                infection_case == "Shincheonji Church" |
                                 infection_case == "contact with patient"|
                                 infection_case== "etc" |
                                 infection_case =="overseas inflow"|
                                 infection_case == "Itaewon Clubs"
                                ),])

sum(path_dt3[path_dt3$province == 'Daegu',]$confirmed)
```

```
## [1] 6218
```

```r
path_dt3 <- path_dt3 %>%
  add_row (infection_case='others', province = 'Daegu',
           confirmed = (
             province_cases_dt[province_cases_dt$province == 'Daegu',]$confirmed-
               sum(path_dt3[path_dt3$province == 'Daegu',]$confirmed)
             )) %>%
  add_row (infection_case='others', province = 'Gyeongsangbuk-do',
           confirmed = (
             province_cases_dt[province_cases_dt$province=='Gyeongsangbuk-do',]$confirmed-
               sum(path_dt3[path_dt3$province == 'Gyeongsangbuk-do',]$confirmed)
           ))

path_dt3 <-  path_dt3[,proportion_confirmed_total:=confirmed/sum(path_dt$confirmed)*100]
path_dt3 <-  path_dt3[1:5,
                      proportion_confirmed_for_this_province :=
                      confirmed /province_cases_dt[province_cases_dt$province ==
                                                     'Daegu',]$confirmed * 100]

path_dt3 <-  path_dt3[6:11,
                      proportion_confirmed_for_this_province :=
                        confirmed /province_cases_dt[province_cases_dt$province ==
                                                    'Gyeongsangbuk-do',]$confirmed * 100]
```

Claim 1: Male patients have a higher fatality rate compared to female patients.

According to Global Health 50/50, men die more compared to women from COVID-19 [2]. We have found out that this applies. State of patients were separated into 3 categories in the given data. We combined "recovered" and "isolated" together to separate the "deceased" category.

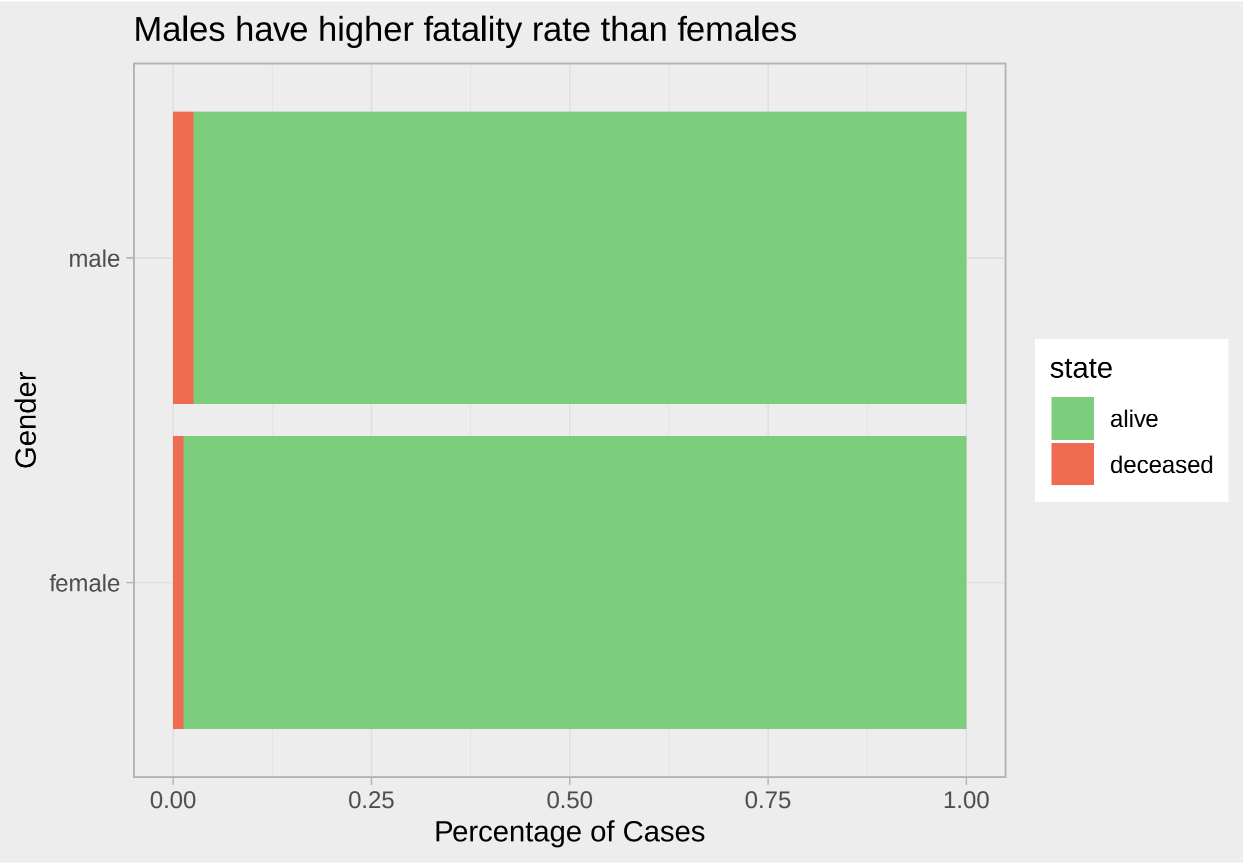
The distributions of states are different, therefore instead of plotting a count of each state plotting the ratios of them provide more information for comparison purposes.

We want to show that males have higher fatality rate than females, so we plotted them side by side together. With barplot's position attribute set to "fill" we have represented the state ratios for both genders.

```
patientState %>%
  group_by(sex, state) %>%
  summarise(count = n()) %>%
  ggplot(., aes(x = sex, y = count, fill = state)) +
  geom_bar(stat = "identity", position = "fill") +
  labs(
    x = "Gender",
    y = "Percentage of Cases",
    title = "Males have higher fatality rate than females"
  ) +
  scale_fill_manual("state",
                    values = c("deceased" = "coral2", "alive" = "palegreen3")) +
  coord_flip() + theme_light()+
  theme(panel.background = element_rect(fill = "#ededed",colour = NA),
        plot.background = element_rect(fill = "#ededed",colour = NA)) +
   ggsave('claim2.png',width = 9, height = 6)
```

```
## `summarise()` regrouping output by 'sex' (override with `.groups` argument)
```

[2]* Source: **https://globalhealth5050.org/the-sex-gender-and-covid-19-project/the-data-tracker/?explore=variable&variable=Deaths**



The plot has very small area for "deceased" patients, but that is because the datatable has a lot more data on the "released" or "isolated" patients.

Since the attributes we worked for this claim are both binary, we decided applying the Fisher's Exact Test would be appropriate. For that reason we created a contingency table:

```
table(patientState)
```

```
##         state
## sex      alive deceased
##              0        0
##   female  2190       28
##   male    1778       47
```

The contingency table also approves that number of "deceased" cases are not many compared to non-deceased ones.

Next we applied Fisher's test to our contingency table. As our null hypothesis, we assumed there is no correlation between the gender and state of patients.

```
test2 <- fisher.test(table(patientState))
test2
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(patientState)
## p-value = 0.002248
## alternative hypothesis: two.sided
```

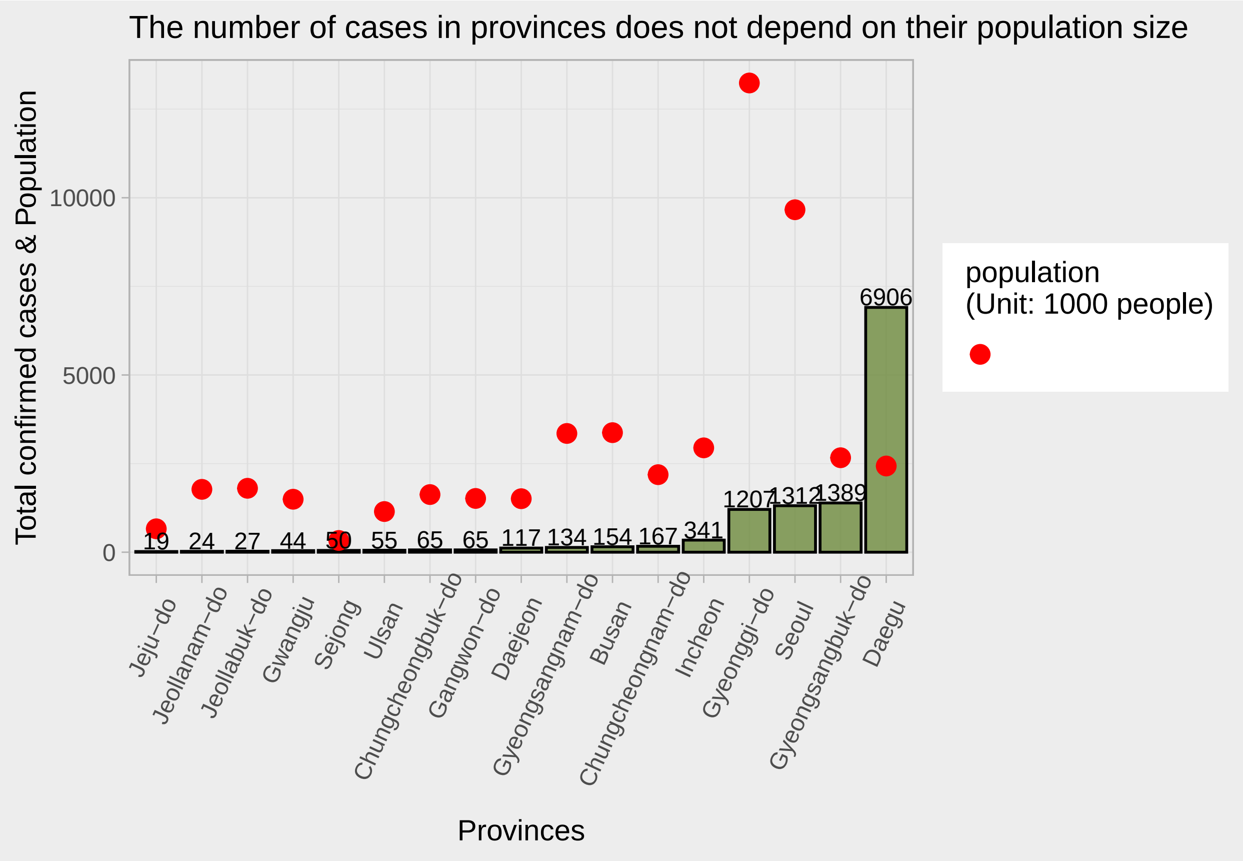
The alive vs deceased table's p-value is 0.002248 which is significant, since it is less than 0.05. We reject the null hypothesis and prove that the two variables are not independent. We can interpret that gender has an impact on death rate. It supports our observation from the plot as well as Global Health 50/50's statement that the male patients have a higher death rate than female's.

There may be other reasons that give rise to our observation of the relationship between gender and stage. For example, the age structure in the gender groups was not considered in this analysis. However, publications on the generally higher mortality rate among men support our hypothesis, citing for example health awareness and smoking as a risk factor.

Claim 2: Cases in provinces do not depend on their population

We suggested there is a connection between the number of confirmed cases in provinces and the population size of these provinces.

```
province_cases_plot <- ggplot(province_cases_dt, aes(x = reorder(province, confirmed),
                                                     y = confirmed)) +
  geom_bar(stat = "identity",
           color = "black",
           fill = "darkolivegreen4",
           alpha = 0.8) +
  geom_point(aes(y= population,  color=""), size=3)+
  geom_text(aes(label = confirmed, y = confirmed + 300), size = 3.1)+
  scale_color_manual(" population \n (Unit: 1000 people)",values = "red")+
  labs(title = "The number of cases in provinces does not depend on their population size",
       #subtitle = "With a referece to the population",
       x = "Provinces",
       y= "Total confirmed cases & Population")+
  theme_light()+
  theme(axis.text.x = element_text(angle = 65, vjust = 0.9, hjust=0.8),
        plot.title = element_text(size=12),
        panel.background = element_rect(fill = "#ededed",colour = NA),
        plot.background = element_rect(fill = "#ededed",colour = NA)) +
ggsave('claim2_1.png',width = 9, height = 6)
province_cases_plot
```

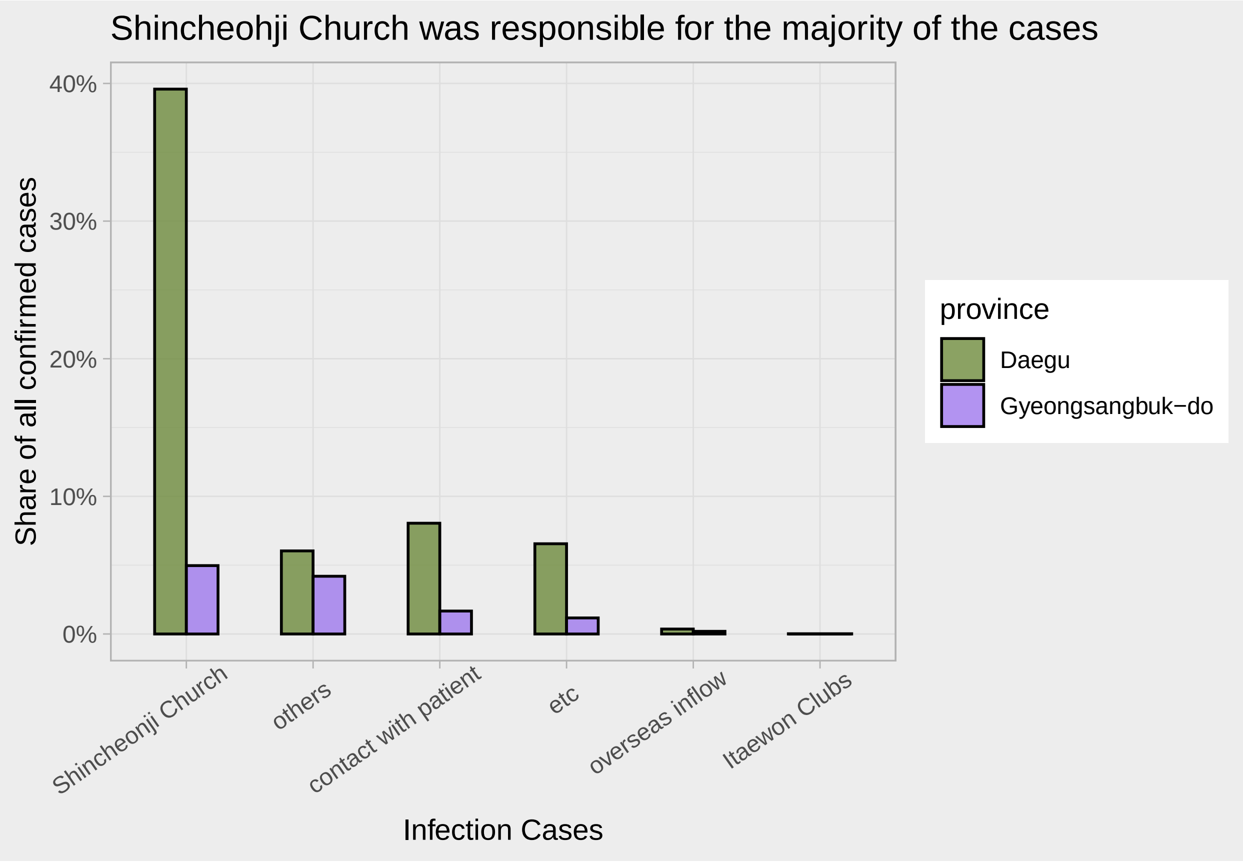


On the plot, we see that the provinces Daegu and Gyeongsangbuk-do are standing out from the rest of the provinces, as the number of confirmed cases are the highest while the population size is not.

The possible reasons for such distribution could be: -Infection Case Path - Cities' density - Elderly Population Ratio

After further data analysis, we found out that the number of cases in provinces is depending on the infection path as well, so in this case, we have the effect of the third variable: Infection Case. We found out that around 45 % of total confirmed cases are coming from the Shincheonji Church.

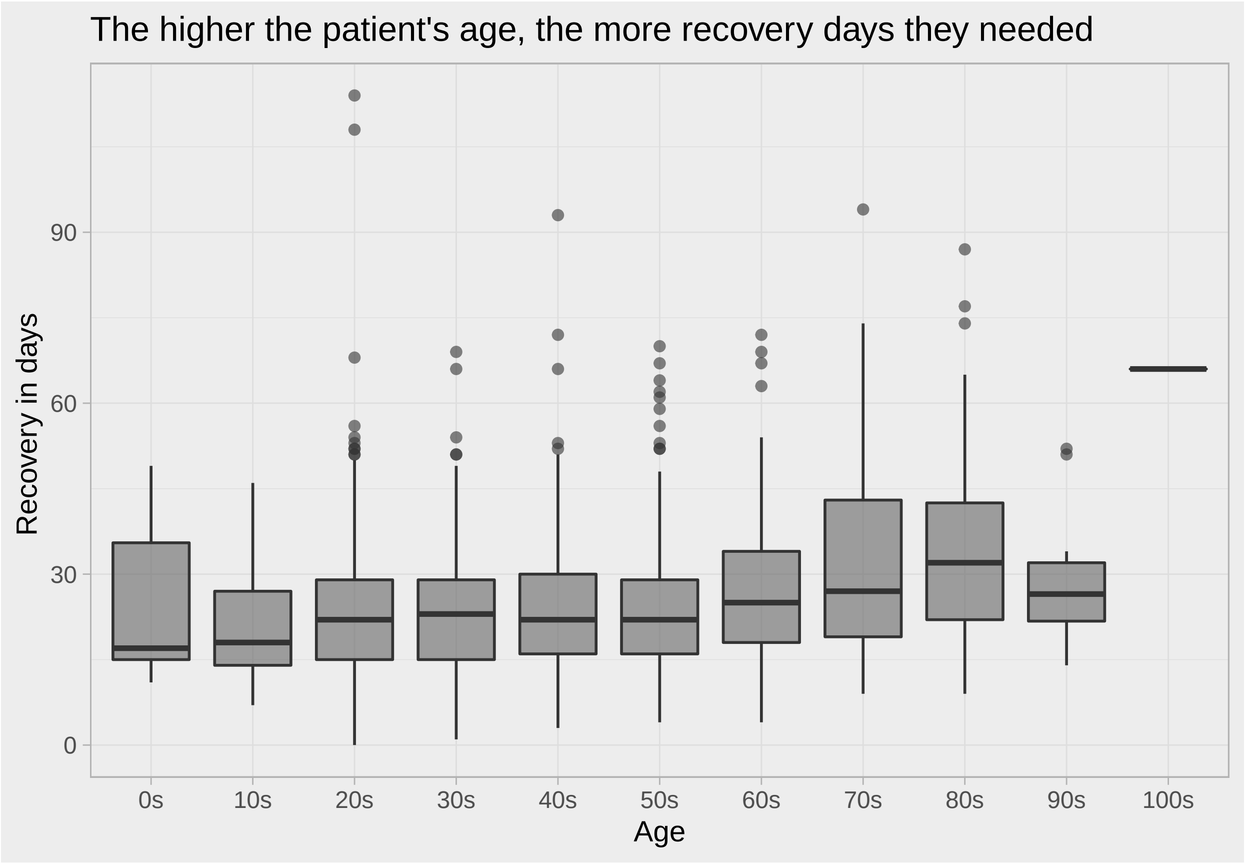
```
infection_path_plot1 <- ggplot(path_dt3,
                               aes(x = reorder(infection_case,desc(path_dt3$confirmed)),
                               y = proportion_confirmed_total,
                               label=paste0(proportion_confirmed_for_this_province,'%'),
                               fill = province)) +
  geom_bar(stat = "identity", position = "dodge",width = 0.5, color = "black", alpha=0.8)+
  labs(x = "Infection Cases",
       y = "Share of all confirmed cases",
title = "Shincheohji Church was responsible for the majority of the cases")+
  scale_y_continuous(labels = function(proportion_confirmed_for_this_province)
    paste0(proportion_confirmed_for_this_province, "%"))+
   theme_light()+
  theme(axis.text.x = element_text(angle = 35, vjust = 0.9, hjust=0.8),
        panel.background = element_rect(fill = "#ededed",colour = NA),
        plot.background = element_rect(fill = "#ededed",colour = NA))+
  scale_fill_manual(
    "province",
    values = c(
      "Daegu" = "darkolivegreen4",
      "Gyeongsangbuk-do" = "mediumpurple2")) +
ggsave('claim2_2.png',width = 9, height = 6)

infection_path_plot1
```



Claim 3: The higher the patient's age, the more recovery days they needed

Young people usually have stronger and immune system and they are expected to recover faster than elderly people. From the start of the pandemic, it is known that elderly people are at bigger risk. Therefore, it wouldn't be a surprise to see young people recovering faster while older people struggle. After examining the data set, we see that the situation is same for South Korea as well.

```
patient_recovery_age_all_categorical %>%
  group_by(age) %>%
  ggplot(., aes(x=age, y=recovery_days)) +
  geom_boxplot(color="grey20",fill="grey40", alpha=0.6) +
  labs(y="Recovery in days ", x="Age",
       title = "The higher the patient's age, the more recovery days they needed") +
   theme_light()+
  theme(panel.background = element_rect(fill = "#ededed",colour = NA),
        plot.background = element_rect(fill = "#ededed",colour = NA)) +
  ggsave('claim3.png',width = 9, height = 6)
```



To support our claim, we have conducted statistical testing. After converting age groups to ordinals, we checked if there is a correlation between number of days needed to recover and age. We performed a Pearson correlation test. As our null hypothesis, we assumed there is no correlation between number of days needed to recover and age.

```
cor.test(patient_recovery_age_all$age, patient_recovery_age_all$recovery_days, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  patient_recovery_age_all$age and patient_recovery_age_all$recovery_days
## t = 7.7219, df = 1574, p-value = 2.024e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1430229 0.2381820
## sample estimates:
##       cor
## 0.1910513
```

We obtained p-value = 2.024e-14 which is very significant. We reject the null hypothesis and prove our claim that there is a correlation between recovery days and age. Correlation coefficient is around 0.19 which shows a week correlation, but they are still correlated.

There might be other reasons that affect the recovery days needed; such as former diseases, genetics and immune system condition. Unfortunately, there is not enough data to check these conditions.

Conclusion

The results of the case study provide important information on Covid-19 patient characteristics. It coincides with the facts already disseminated by the media and further complement them.

It can be stated significantly (p-Value = 0.002248 < 0.05) that in South Korea gender has an impact on the death rate. This is in line with Global Health 50/50's statement about male's fatality rate.
As a supplement to the generally known fact that mainly older people die from the virus, however, the analysis also tackles the extent to which young people should nevertheless be prepared for a longer course of the disease. It was found that age is significantly (p-value = 2.024e-14 < 0.05) correlated with the number of recovery days, since elderly patients needed on average a longer recovery than the young patients.
What could be also proven with the analysis, however, is that the characteristics about the origin/place of residence play a secondary role. The number of Cases in the South Korean provinces do not depend on their population size. By taking into account the impact of an additional variable it is found that the infection number in those provinces can be heavily influenced by the infection case path. Highlighted in this analysis was the Shincheonji Church, which caused many infections, also outside of their own place of residence.
Future analysis can add additional patient characteristics or tackle given limitations stated in this case study.