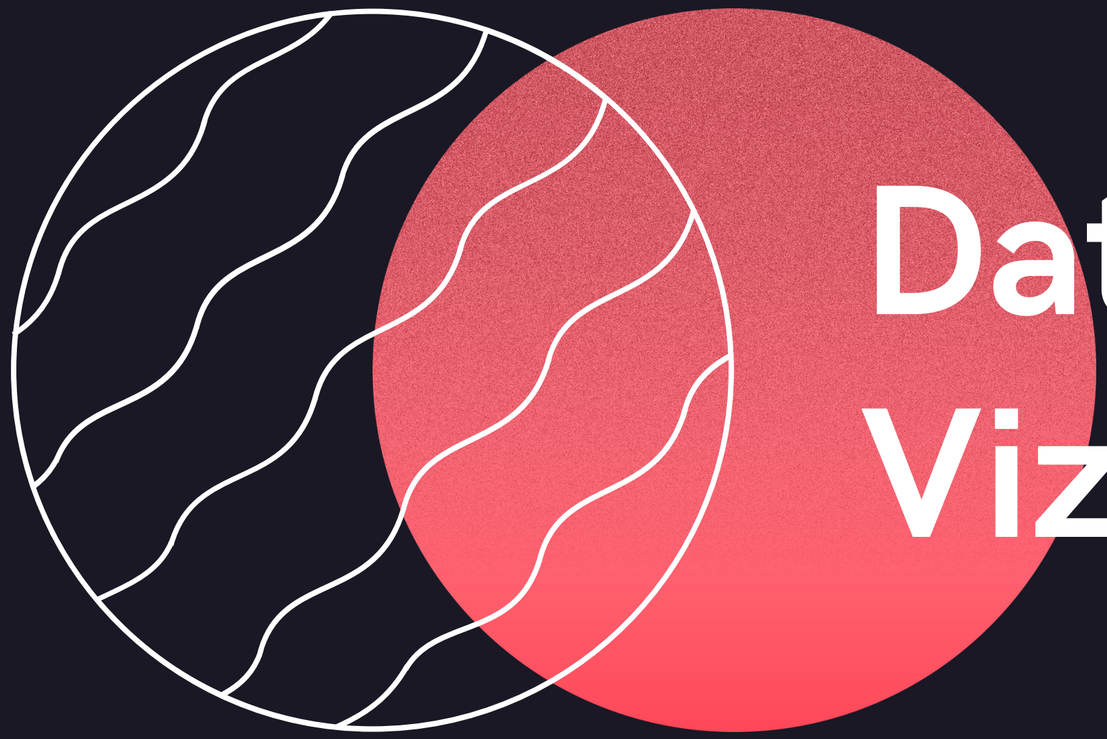
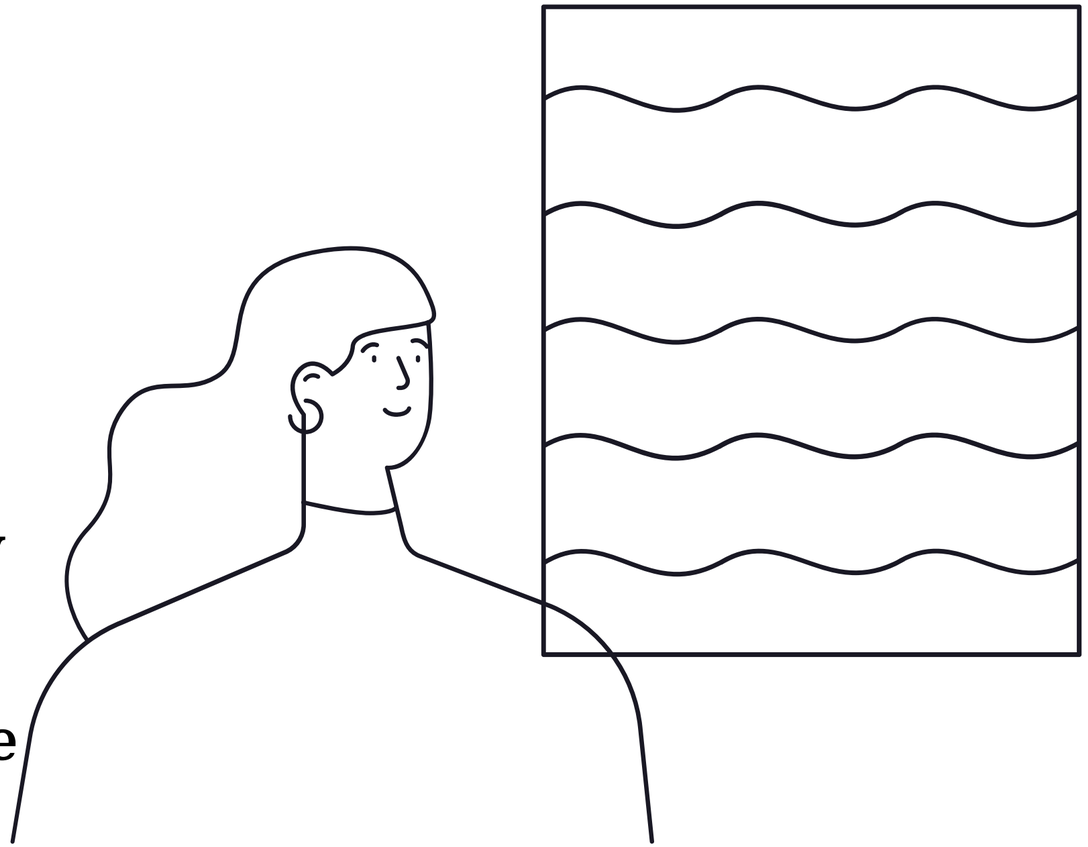


Data Analysis & Vizualization in R

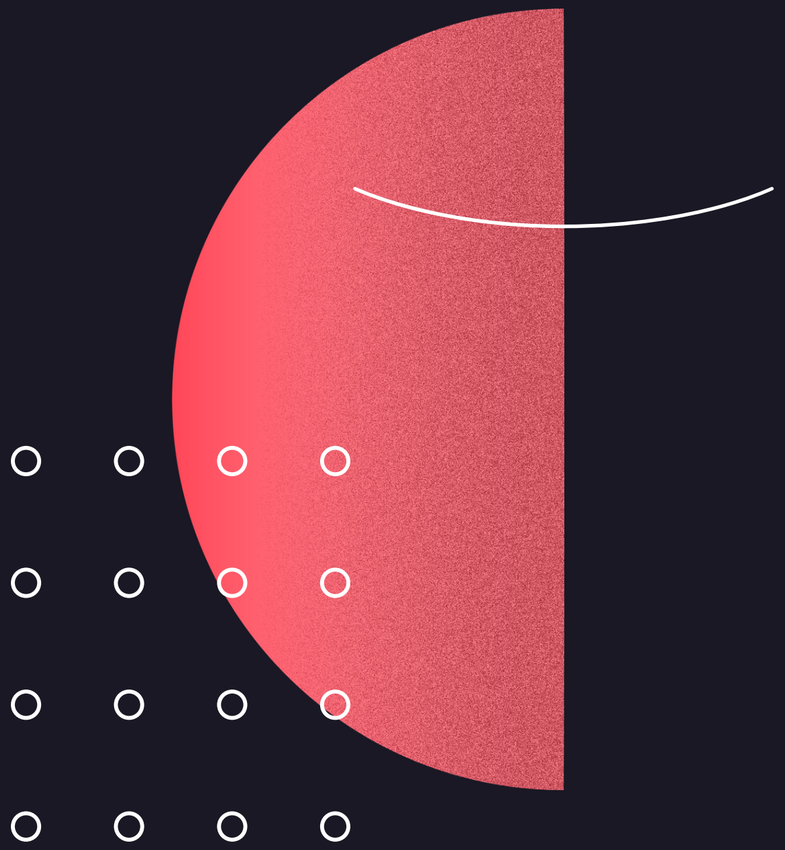
Team 142
Efthymia Kostak - 03740037
Ertugela Doçi - 03727843

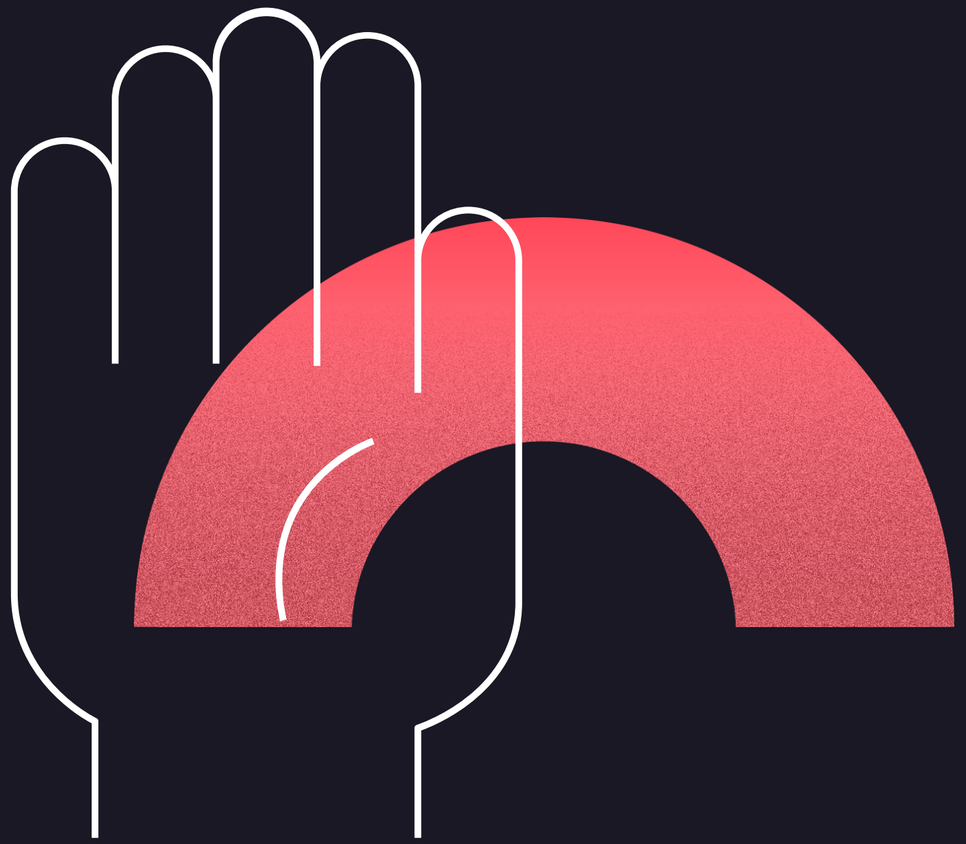
DISCLAIMER

The purpose of this presentation is to make research of the COVID-19 cases in South Korea published by Kaggle.
Each of our claim is worked independently by us and using only the basic knowledge from the script.
Our findings are based on our methodology and in order to prove each claim is supported, definitely some more knowledge and several methodologies are needed.



Contents:





WE'RE NOT JUST FIGHTING A PANDEMIC; WE'RE FIGHTING AN INFODEMIC*

Covid-19 affects the lives of everyone in the world. In order to design effective countermeasures for the pandemic, we need to analyze massive amounts of data to extract meaningful knowledge.

*https://govinsider.asia/health/aws-amazon-web-services-cloud-the-importance-of-data-in-the-fight-against-covid-19/

GOALS

1. Apply theoretical knowledge from the course by extracting meaningful knowledge from visualization techniques and statistical analysis

2. Assess the effectiveness of governmental measures in South Korea

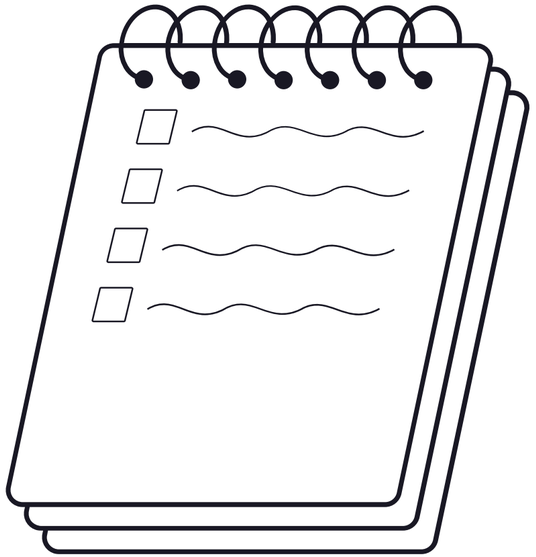
3. Investigate the understanding of South Korea's population of COVID-19

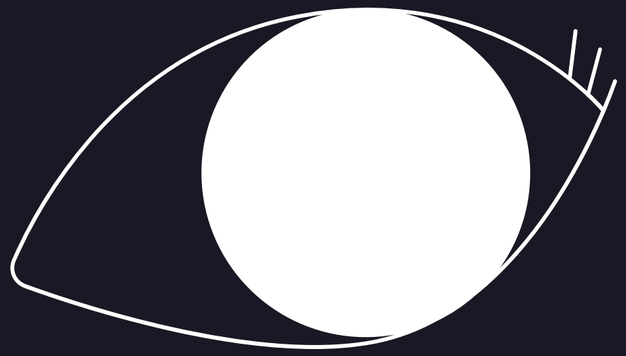
4. Analyze how the virus affected different population groups and areas

DATA PREPARATION FRAMEWORK

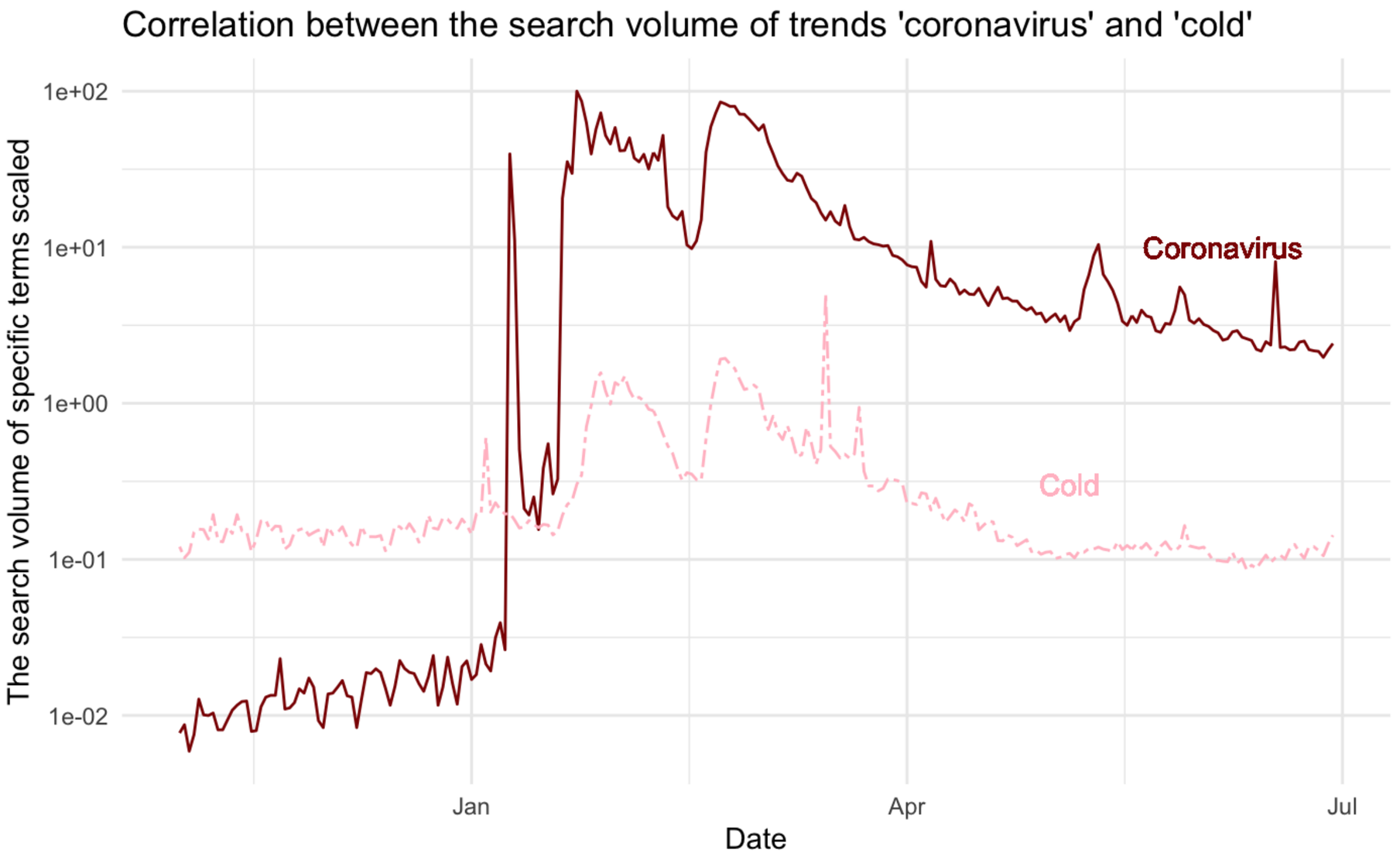
Not many data preparation steps were needed since the rating regarding the usability of the dataset in Kaggle is 10/10. We followed the following steps in our analysis:

1. Data retrieval: Used Kaggle dataset
2. Treatment of missing data: Subset the data table each time by disregarding NAs and empty cells.
3. Data aggregation: use a sufficient amount of data in terms of a number of observations and quality of each observation. Investigate data tables using str() and summary() functions, create meaningful plots using ggplot.





BUT FIRST, HOW DID PEOPLE PERCEIVE COVID-19?


Correlation between the search volume of trends 'coronavirus' and 'cold'
The search volume of specific terms scaled
1e+02
1e+01
1e+00
1e-01
1e-02
Coronavirus
Cold
Jan
Apr
Jul
Date

STATISTICAL TESTING → CLAIM SIGNIFICANT

09

Method:

- Pearson Correlation Test
- Spearman Correlation Test

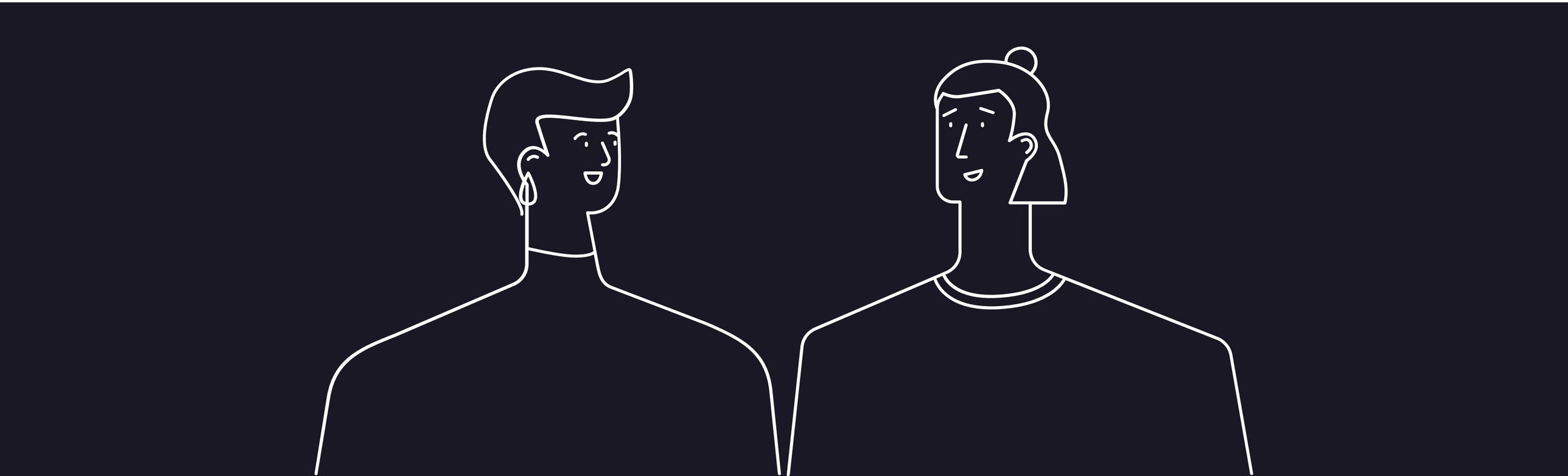
Possible Reasons:

- People perceive the symptoms of coronavirus and cold similarily
- Social level: coronavirus is nothing more than a cold

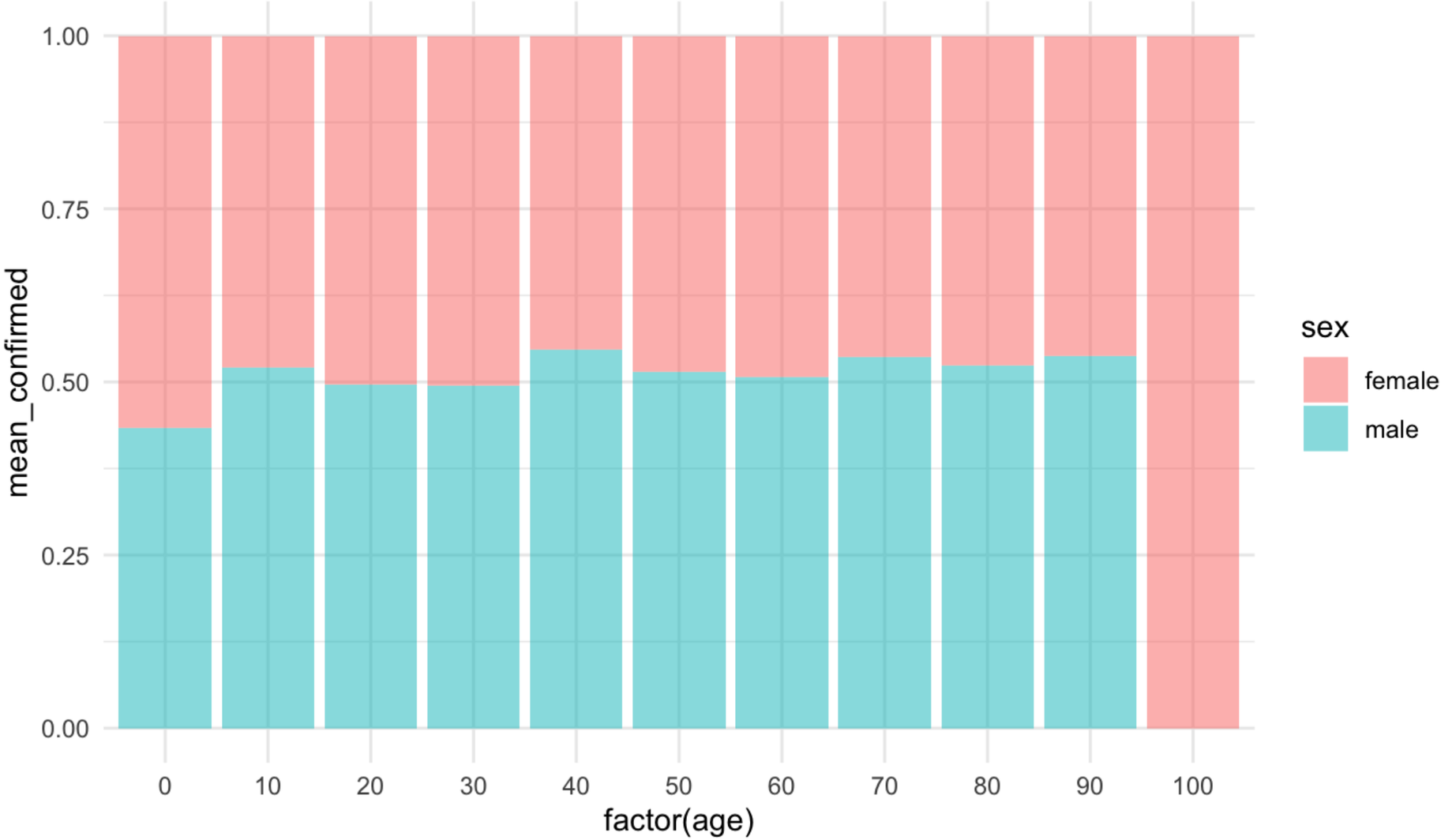
Limitations:

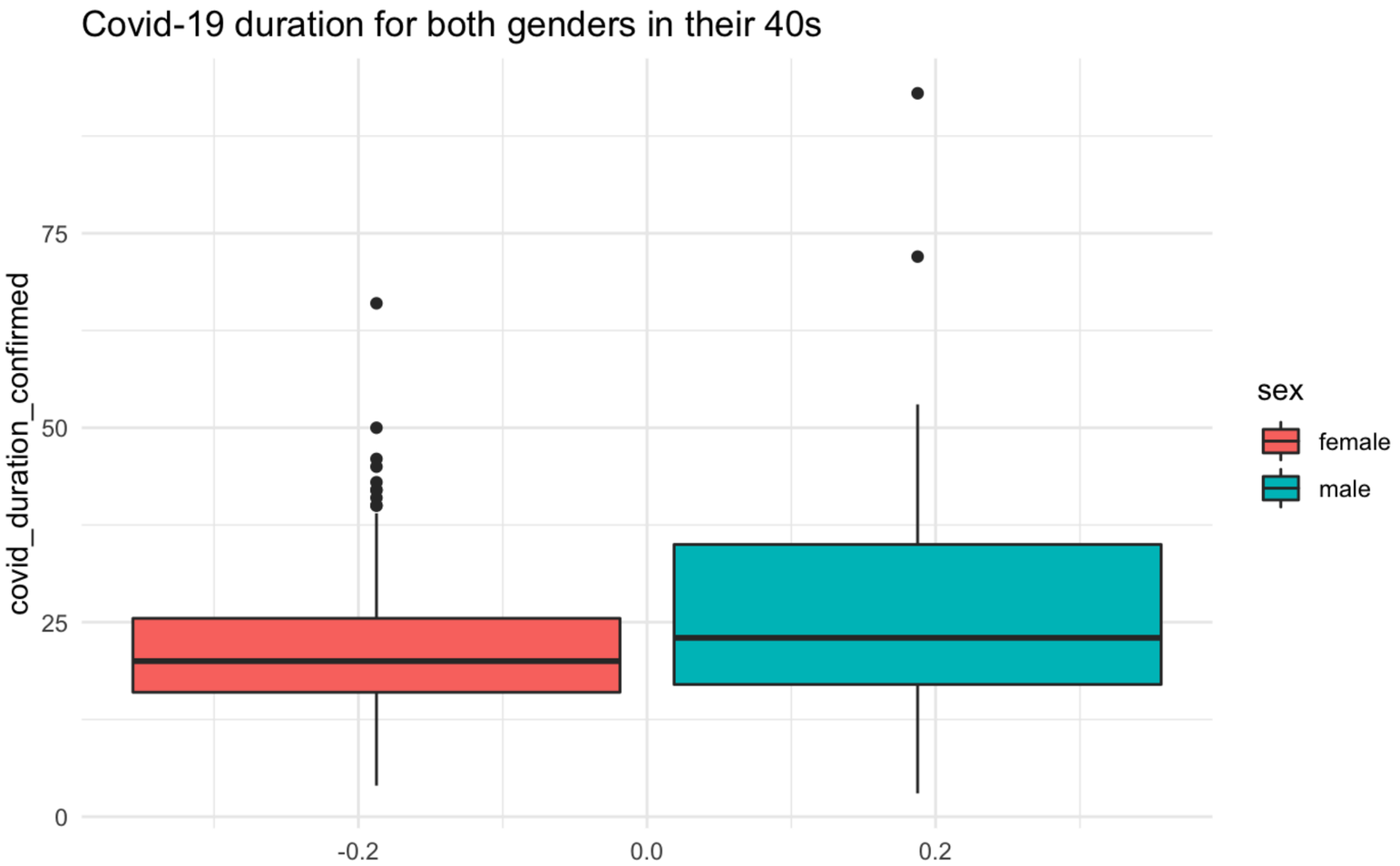
- Coronavirus didn't exist before the end of 2019
- Time series analysis was not applicable on this claim

BUT, WHICH GROUPS OF THE POPULATION WERE MOST AFFECTED AND WHERE?



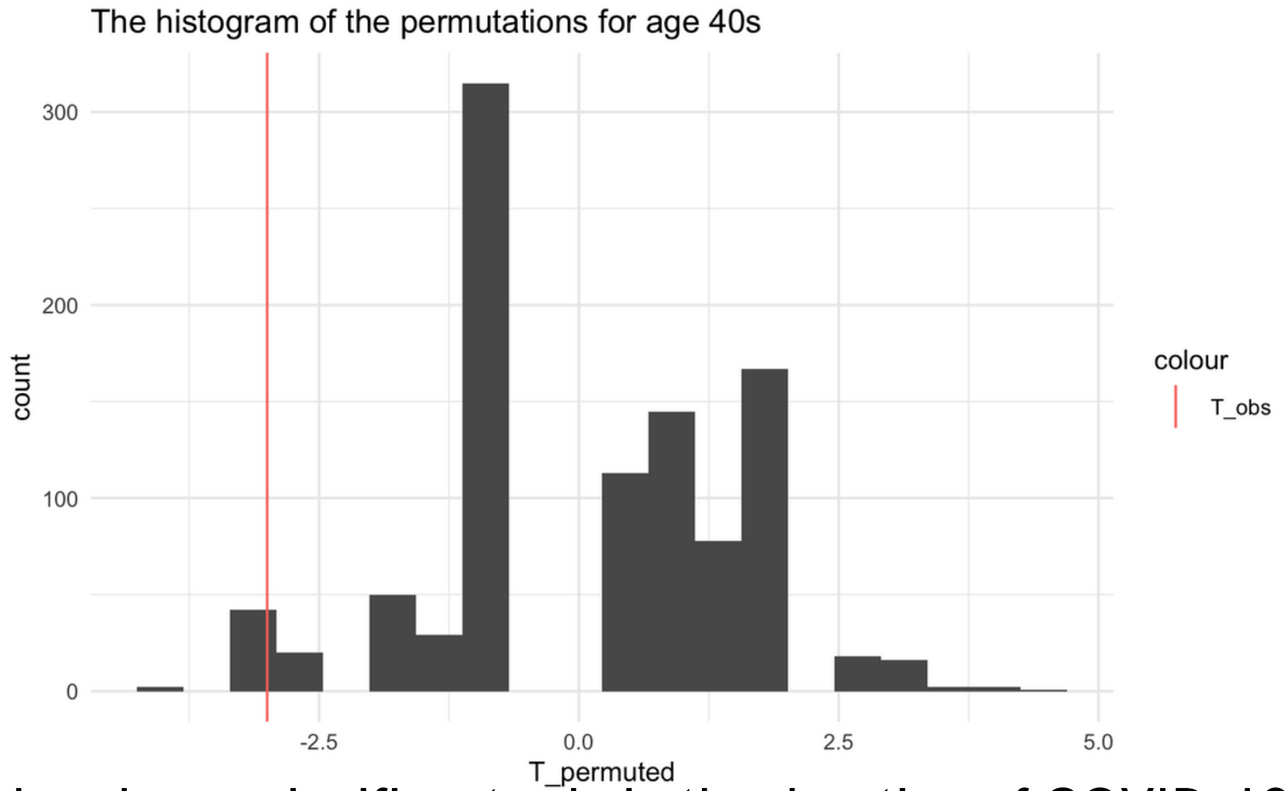
Lower proportion of duration of COVID for women in 40s, 70s and 80s





The duration for males was slightly higher than females in the dataset, but was this difference significant in the real population?

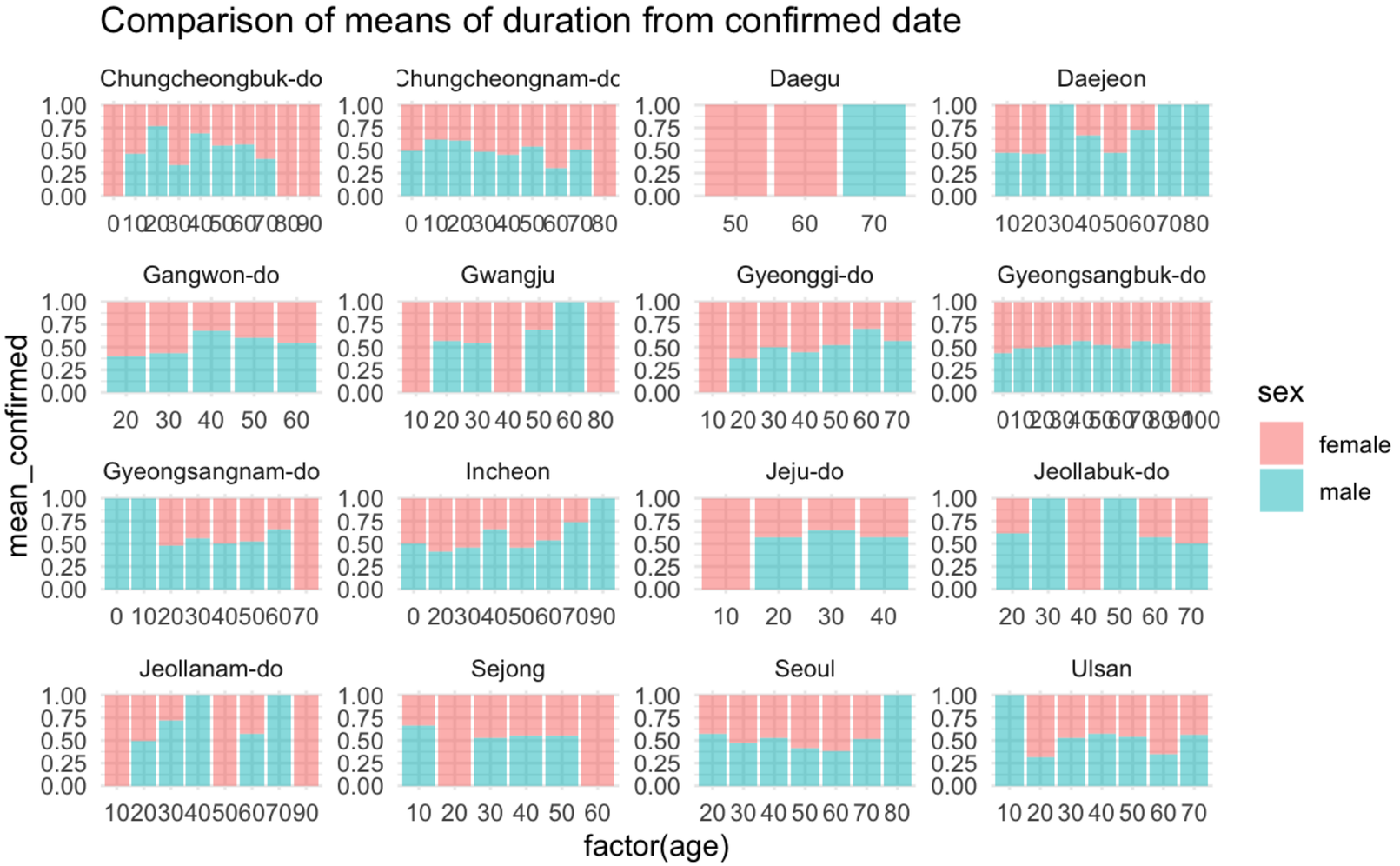
STATISTICAL TESTING → CLAIM SIGNIFICANT

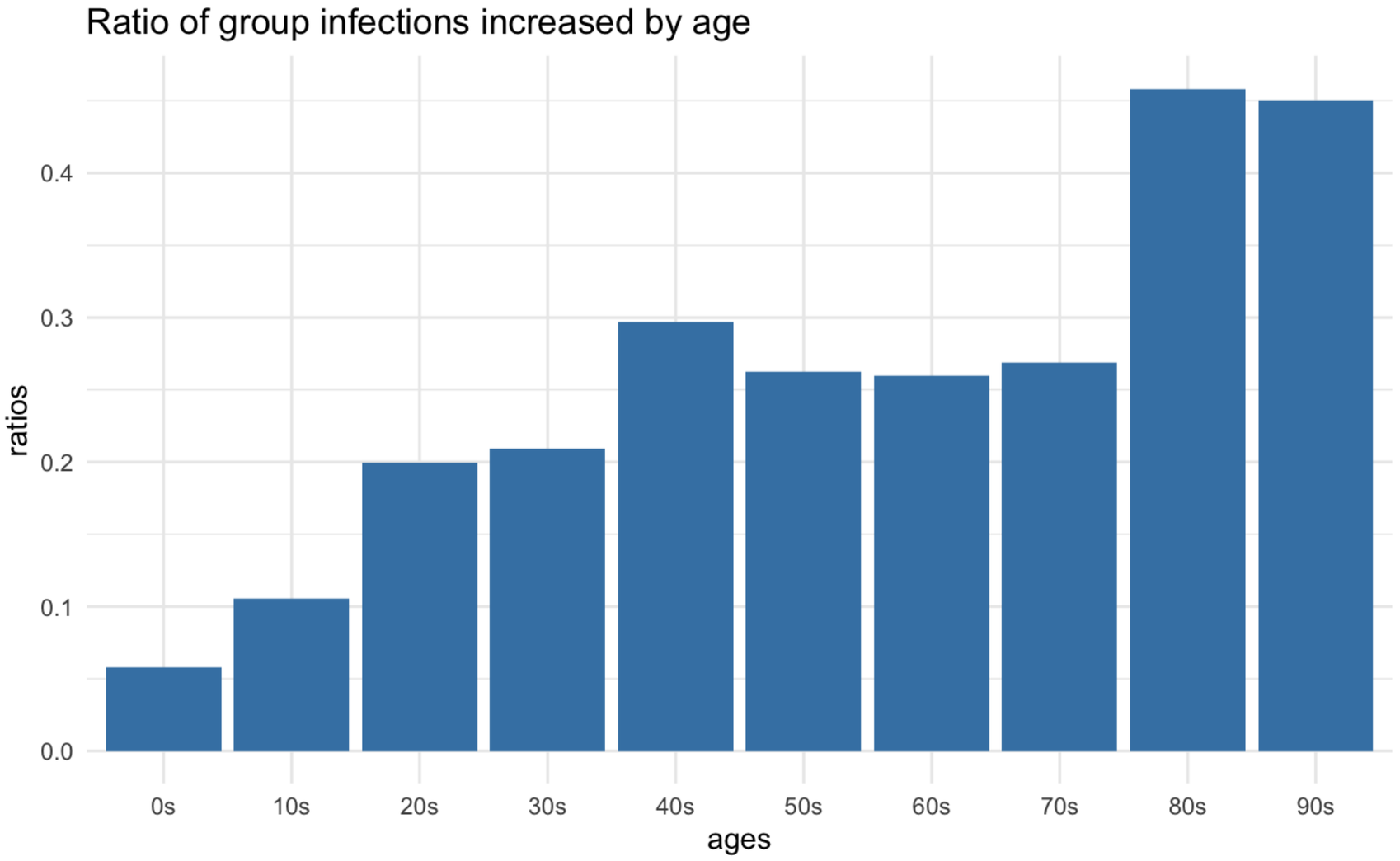


p-value < 0.05: Gender plays a significant role in the duration of COVID-19 for people in their 40s who were released with women being infected with the virus on average less time compared to men.

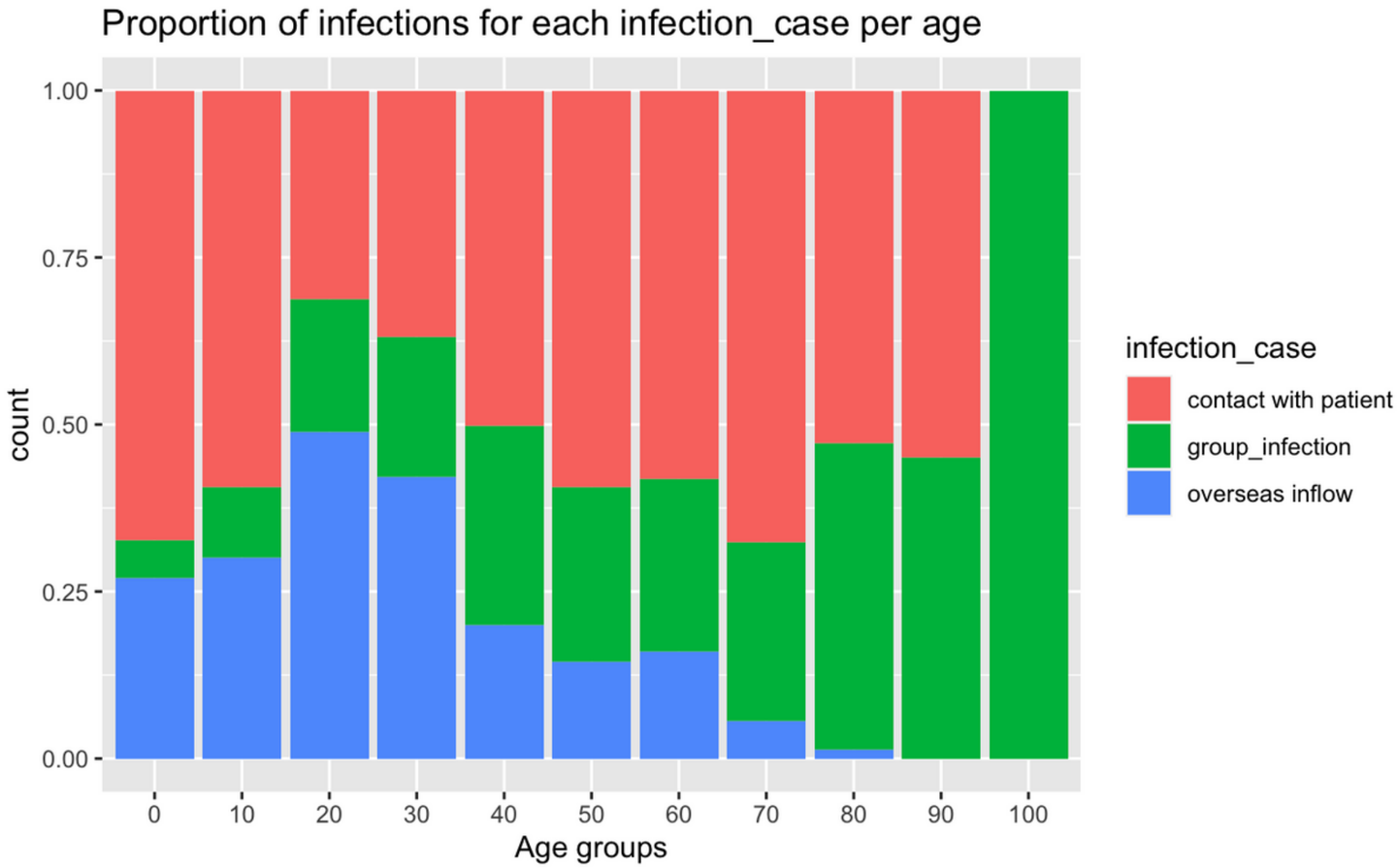
Further steps: Q-Q plots

PROVINCE IS NOT A CONFOUNDING FACTOR[14]




Ratio of group infections increased by age
ratios
0.0
0.1
0.2
0.3
0.4
0s
10s
20s
30s
40s
50s
60s
70s
80s
90s
ages

INFECTION CASE IS NOT A CONFOUNDING FACTOR



STATISTICAL TESTING → CLAIM SIGNIFICANT

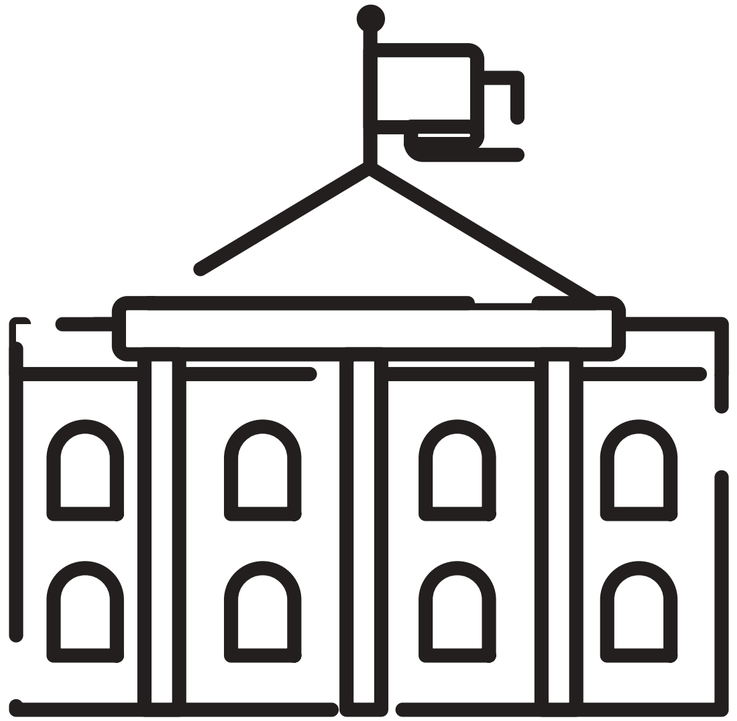
Method:

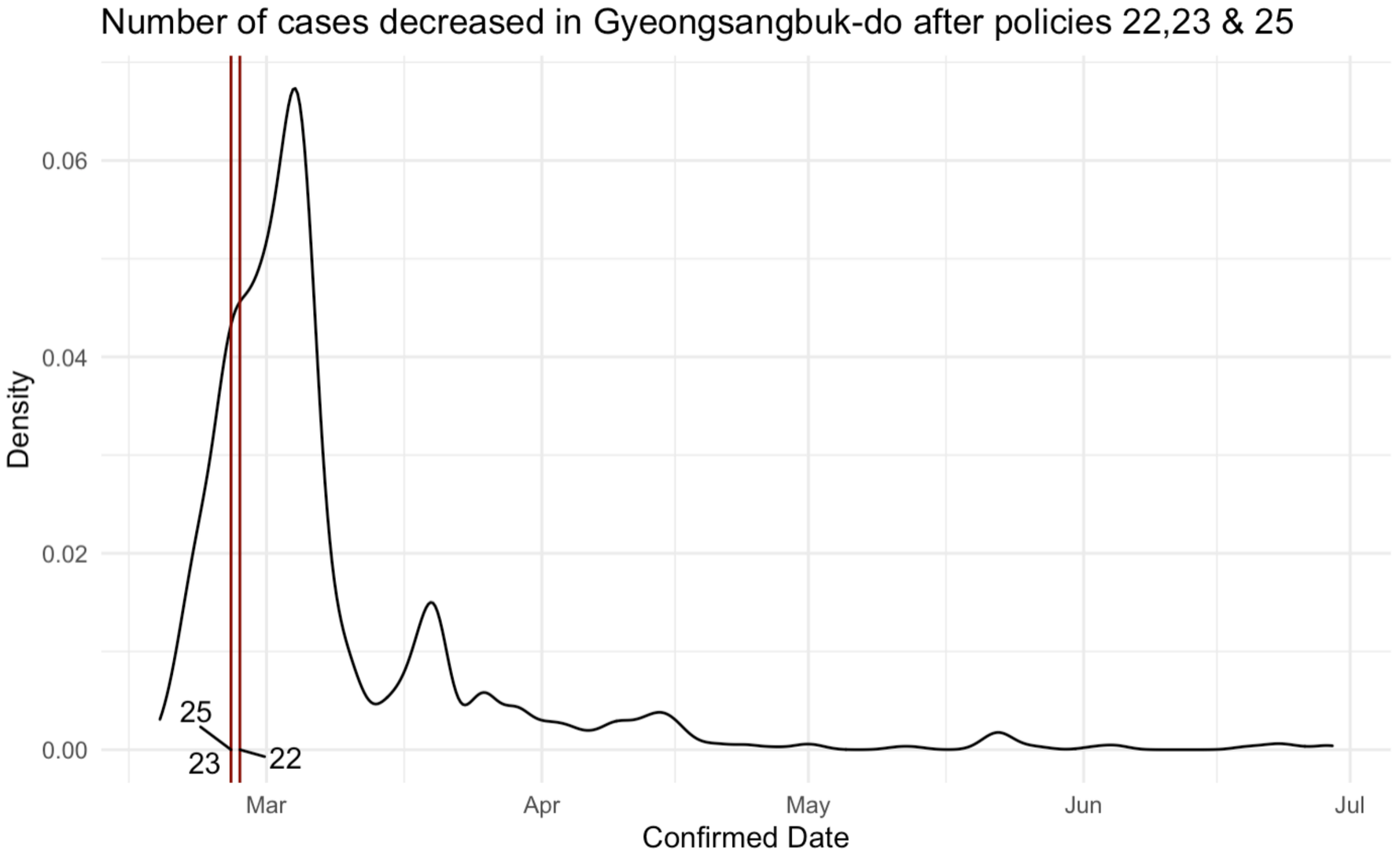
- Fisher's Test
- Created 3 age groups: young, middle, and old
- Conducted 3 tests for each combination

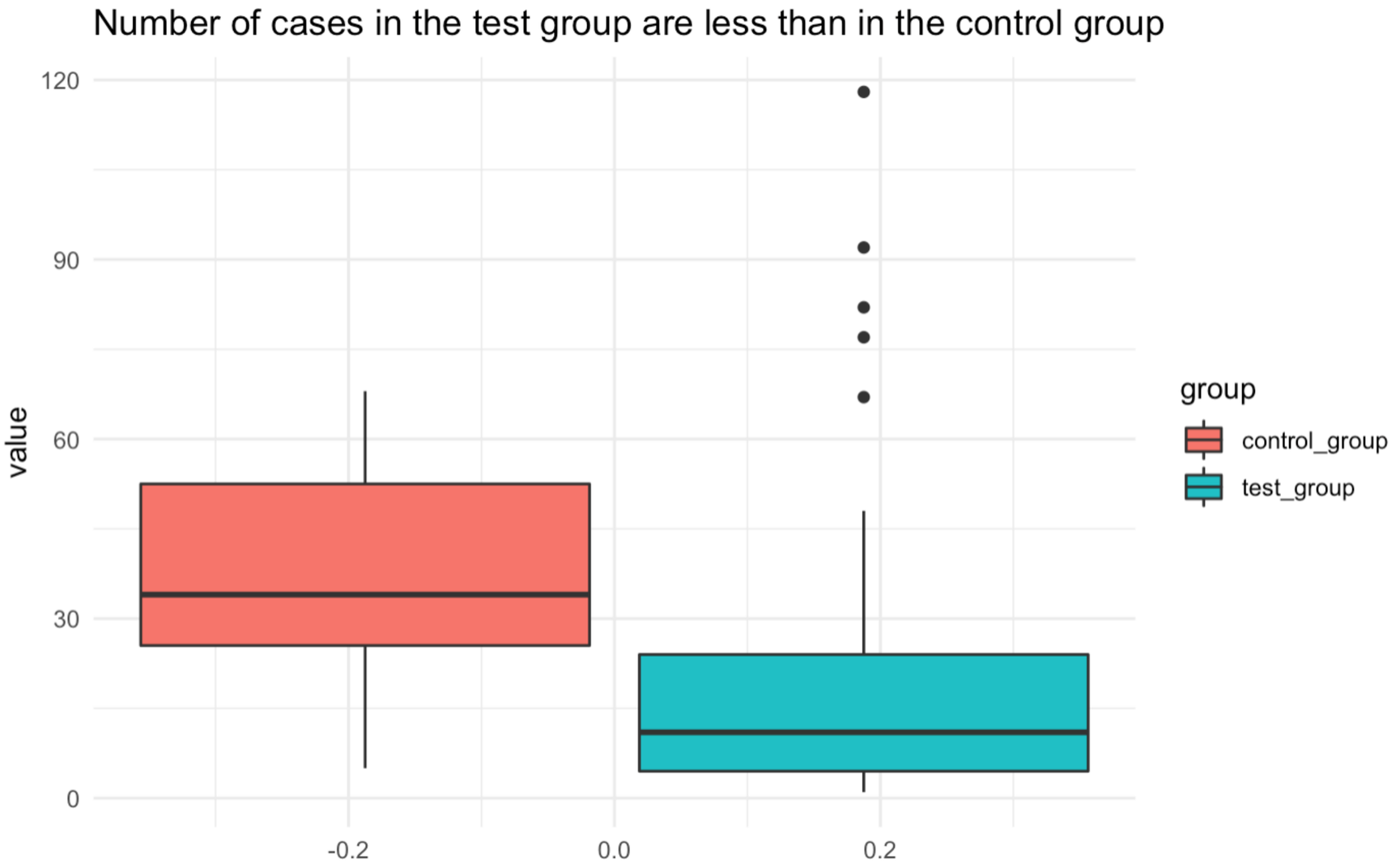
Possible Reasons:

- Older people in South Korea might be more involved with public gatherings such as going to the church.

BUT, DID THE GOVERNMENT OF SOUTH KOREA DEAL EFFECTIVELY WITH THE VIRUS?




Number of cases decreased in Gyeongsangbuk-do after policies 22,23 & 25
Density
0.06
0.04
0.02
0.00
25
23
22
Mar
Apr
May
Jun
Jul
Confirmed Date


Number of cases in the test group are less than in the control group
value
120
90
60
30
0
-0.2
0.0
0.2
group
control_group
test_group

STATISTICAL TESTING ↷ CLAIM SIGNIFICANT

Method:

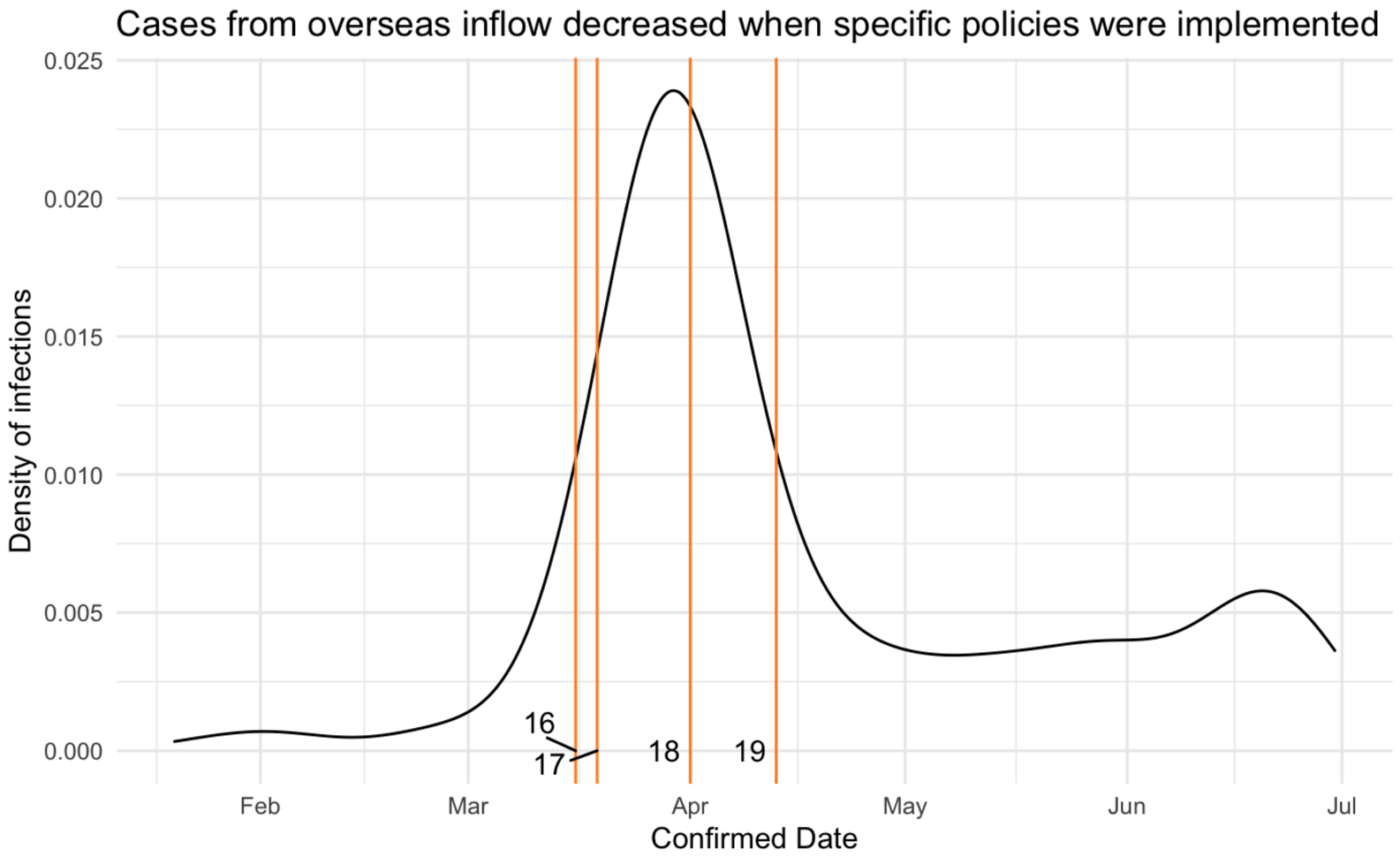
- Wilcoxon's Test
- Used control group: number of cases before the implementation of policy
- Used test group: number of cases after the implementation of policy

Possible Reasons:

- Policies investigated might have lead to a reduction of cases in the region
- Other reasons: implementations of other policies or reduction of overseas inflow population etc.

Limitations:

- We didn't perform a time series analysis such as ANOVCA



Same methodology: we used Wilcoxon's test and the claim proved significant

MAIN FINDINGS:

1. The population of South Korea understood that there some similarities between "cold" and "coronavirus" even at the beginning of the pandemic.

2. The trend of the infections of older people was in groups compared to the younger population.

3. Males in their 40s suffered from the virus longer compared to the females in their 40s.

4. The policies taken to have a special immigration procedure for all countries seem to have been effective.

5. Policies related to Health seem to have had an impact on the reduction of the number of cases in the province "Gyeongsangbuk-do".

It is easy to lie with statistics. It is hard to tell the truth without it.

Andrejs Dunkels

Thank you!