MERS Korea 2015 - Marta Smith

2025-12-01

Introduction

In this R Markdown, I analyze the `mers_korea_2015` dataset from the "outbreaks" R package and answer questions about the MERS-CoV outbreak that happened in South Korea in 2015. Source: https://github.com/luisfmontemayor/learn-compbio/tree/main

Loading and inspecting the data

```r
data("mers_korea_2015")
```

Quick questions

1. What object class is `mers_korea_2015` in R? Describe what this class is like, and how is it different from a simple atomic type in R (like a `string` like "hello world!" or a `logical` value like `true`) or a vector type defined with `c()`

`mers_korea_2015` is a list. A list is a collection of data which is ordered and changeable (source).

2. What sub-items exist within this item?

It contains 2 sub-items that are dataframes: `linelist` and `contacts`. `linelist`: A dataframe of MERS-CoV cases and their attributes (source). `contacts`: A dataframe describing the relationship between MERS Co-V cases (source).

```r
# inspecting the raw data
class(mers_korea_2015)
```

```
## [1] "list"
```

```r
class(mers_korea_2015$linelist)
```

```
## [1] "data.frame"
```

```r
class(mers_korea_2015$contacts)
```

```
## [1] "data.frame"
```

```r
length(mers_korea_2015) # number of elements the list contains
```

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

```r
names(mers_korea_2015) # names of the elements of the list
```

```
## [1] "linelist" "contacts"
```

```r
str(mers_korea_2015) # structure function
```

```
## List of 2
##  $ linelist:'data.frame':    162 obs. of  15 variables:
##   ..$ id            : chr [1:162] "SK_1" "SK_2" "SK_3" "SK_4" ...
##   ..$ age           : int [1:162] 68 63 76 46 50 71 28 46 56 44 ...
##   ..$ age_class     : chr [1:162] "60-69" "60-69" "70-79" "40-49" ...
##   ..$ sex           : Factor w/ 2 levels "F","M": 2 1 2 1 2 2 1 1 2 2 ...
##   ..$ place_infect  : Factor w/ 2 levels "Middle East",..: 1 2 2 2 2 2 2 2 2 2 ...
##   ..$ reporting_ctry: Factor w/ 2 levels "China","South Korea": 2 2 2 2 2 2 2 2 2 1 ...
##   ..$ loc_hosp      : Factor w/ 13 levels "365 Yeollin Clinic, Seoul",..: 10 10 10 10 1 10 10 13 10
##   ..$ dt_onset      : Date[1:162], format: "2015-05-11" "2015-05-18" ...
##   ..$ dt_report     : Date[1:162], format: "2015-05-19" "2015-05-20" ...
##   ..$ week_report   : Factor w/ 5 levels "2015_21","2015_22",..: 1 1 1 2 2 2 2 2 2 2 ...
##   ..$ dt_start_exp  : Date[1:162], format: "2015-04-18" "2015-05-15" ...
##   ..$ dt_end_exp    : Date[1:162], format: "2015-05-04" "2015-05-20" ...
##   ..$ dt_diag       : Date[1:162], format: "2015-05-20" "2015-05-20" ...
##   ..$ outcome       : Factor w/ 2 levels "Alive","Dead": 1 1 2 1 1 2 1 1 1 1 ...
##   ..$ dt_death      : Date[1:162], format: NA NA ...
##  $ contacts:'data.frame':    98 obs. of  4 variables:
##   ..$ from         : chr [1:98] "SK_14" "SK_14" "SK_14" "SK_14" ...
##   ..$ to           : chr [1:98] "SK_113" "SK_116" "SK_41" "SK_112" ...
##   ..$ exposure     : Factor w/ 5 levels "Contact with HCW",..: 2 2 2 2 2 2 2 2 2 2 ...
##   ..$ diff_dt_onset: int [1:98] 10 13 14 14 15 15 15 16 16 16 ...
```

3. How can one access the sub-items within the `mers_korea_2015`?

```r
# separating the data sets
# different ways of accessing the sub-items: either with a $ or [1]
mers_korea_2015$linelist
mers_korea_2015[1]

# creating new variables for each sub-item
linelist <- mers_korea_2015$linelist
contacts <-mers_korea_2015$contacts
```

4. The "sub-items" within `mers_korea_2015` are different data sets within our project. Play with them! Open them, plot graphs, get your hands dirty. The questions won't always tell you what data set to use! So know what you're starting with, to make sure that you can

```r
# cleaning the data
linelist <- linelist %>%
  mutate(
    loc_hosp = str_replace(
      string = loc_hosp,
      pattern = ",.*",
      replacement = ""
    ) ) %>%
  mutate(
    loc_hosp = str_trim(loc_hosp)
  )
unique(linelist$loc_hosp)
```

```
##  [1] "Pyeongtaek St. Mary"
##  [2] "365 Yeollin Clinic"
##  [3] "Seoul Clinic"
##  [4] "Konyang University Hospital"
##  [5] "Dae Cheong Hospital"
##  [6] "Samsung Medical Center"
##  [7] "Hallym University Medical Center"
##  [8] "Hallym University Dongtan Sacred Heart Hospital"
##  [9] "Pyeongtaek goodmorning hospital"
## [10] "Pyeongtaek fraternity hospital"
## [11] "Konkuk University hospital"
## [12] "Gangdong Kyuang Hee University Hospital"
```

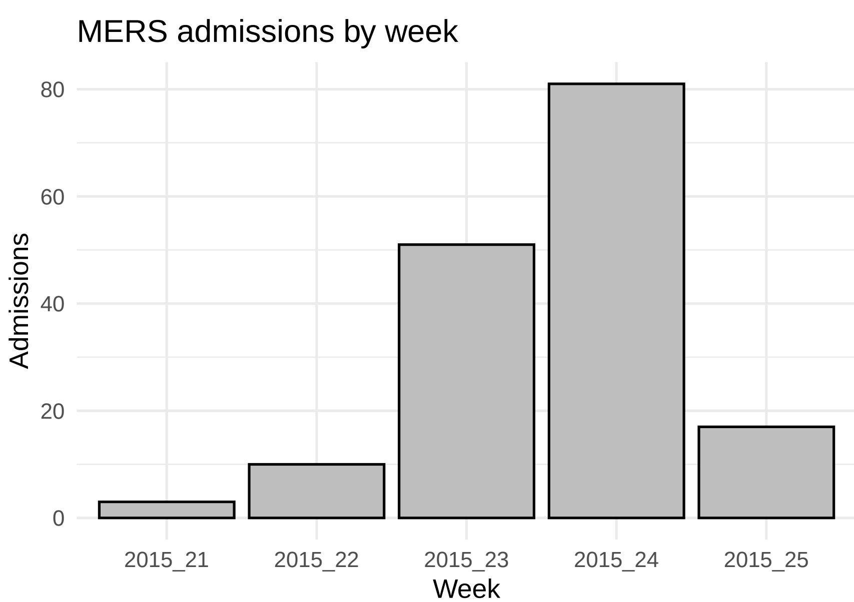
Analyses

Getting my hands dirty with the data

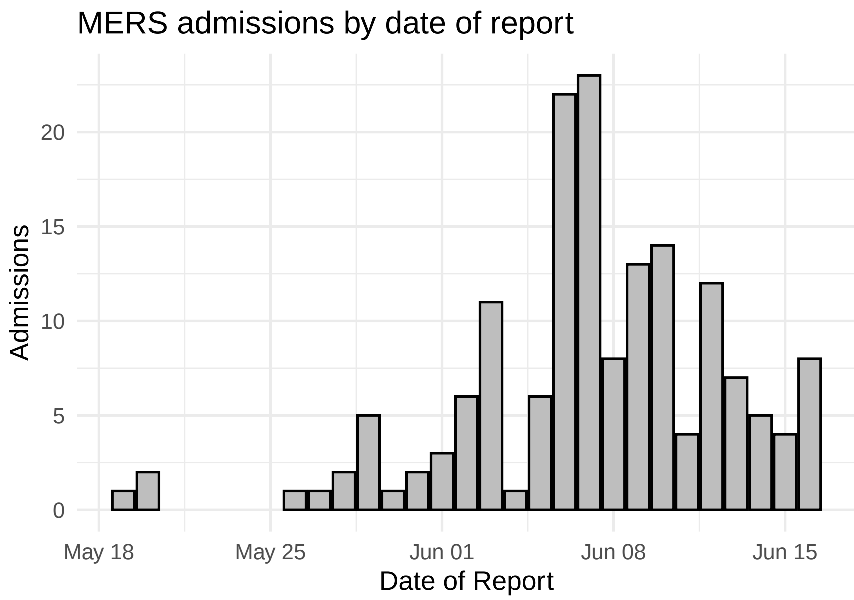
```r
# admissions grouped by sex and date of report
admissions_summary <- linelist %>%
  group_by(dt_report, sex) %>%
  summarise(
    count = n(),
    mean_age = mean(age, na.rm= TRUE),
    .groups = 'drop'
  )

# Pivot table by ages per date of report
pivot_age <- admissions_summary %>%
  pivot_wider(
    id_cols = dt_report, # columns
    names_from = sex, # names of the column
    values_from = mean_age, # values
    names_prefix = "mean_age_" # optional
  )

# count of admissions by the number of week
admissions_by_week <- ggplot(linelist) +
```

```r
  geom_bar(aes(x=week_report), fill="gray", color="black") + # color is used for the outline
  theme_minimal() +
  labs(
    title="MERS admissions by week",
    x="Week",
    y="Admissions"
  )

# admissions by date of report
admissions_by_dt_report <- ggplot(linelist) +
  geom_bar(aes(x=dt_report), fill="gray", color="black") + # color is used for the outline
  theme_minimal() +
  labs(
    title="MERS admissions by date of report",
    x="Date of Report",
    y="Admissions"
  )

# admissions by sex
admissions_by_sex <- ggplot(admissions_summary, aes(x=dt_report, y=count, color=sex)) +
  geom_line(linewidth =0.7, alpha=0.7) + geom_point(size=2, alpha=0.7)

# first approach to understand the average age of the two populations
age_by_outcome <- ggplot(linelist) +
  geom_violin(aes(x=outcome, y=age, fill=outcome)) +
  theme_minimal() +
  labs(title="",
      x="Outcome", y="Age",fill="Outcome")
```

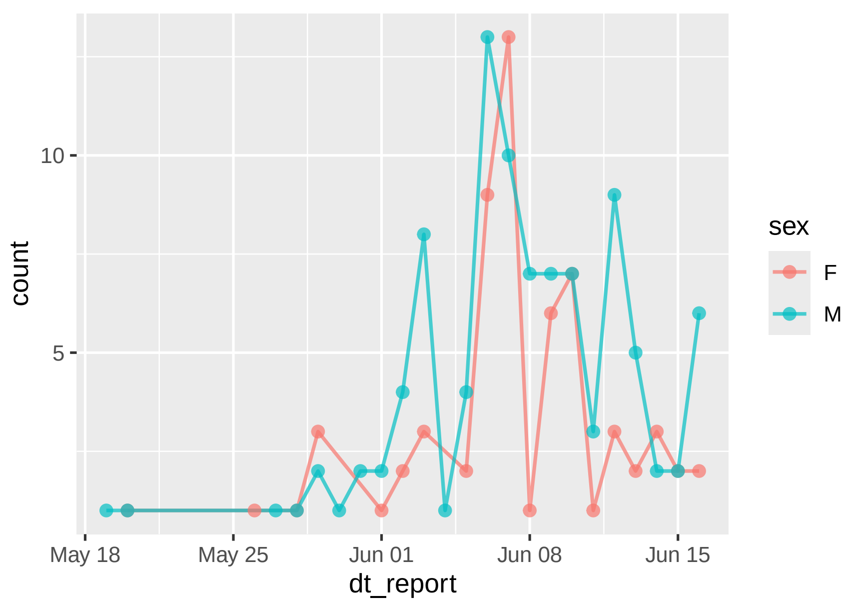
```r
admissions_by_week
```



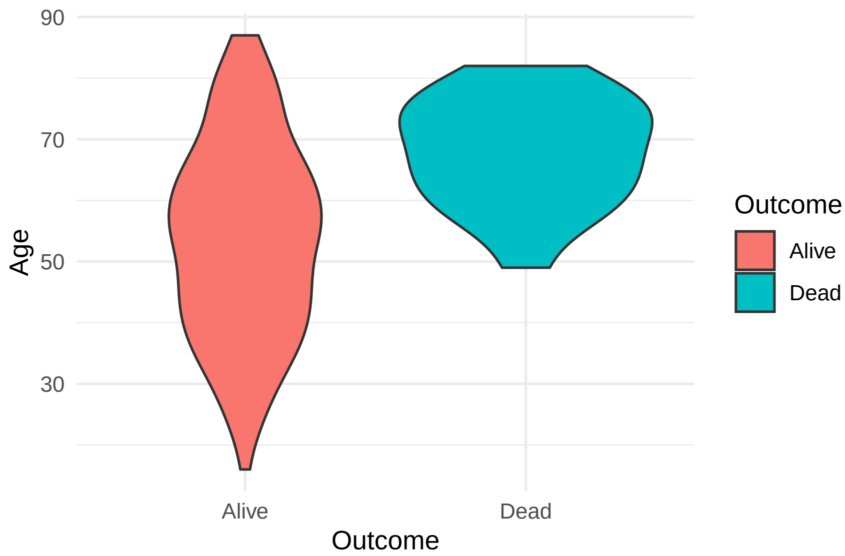
```
admissions_by_dt_report
```




```
admissions_by_sex
```




```
age_by_outcome
```



1. Demographics

1. Calculate the average age for patients with the outcome "Dead" and patients with the outcome "Alive."

Average ages: Dead: 68 years old; Alive: 53.6 years old. The average age of this sample was 55 years old.

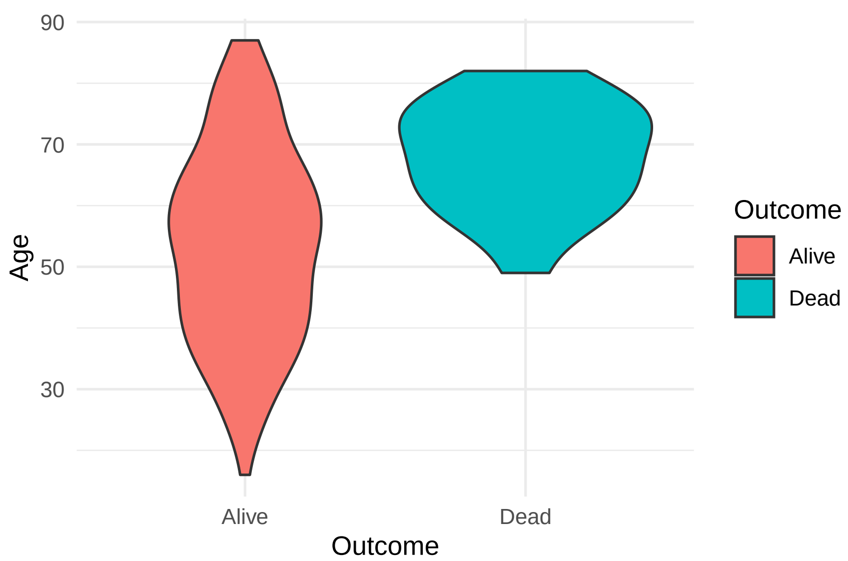
```
# age distribution (not grouped)
age_stats <- linelist %>%
  summarise(
    count = n(),
    mean_age = mean(age, na.rm= TRUE),
    median_age = median(age, na.rm = TRUE),
    sd_age = sd(age, na.rm = TRUE),
    .groups = 'drop'
  )
age_stats
```

```
##   count mean_age median_age sd_age
## 1   162 55.32099         56 15.814
```

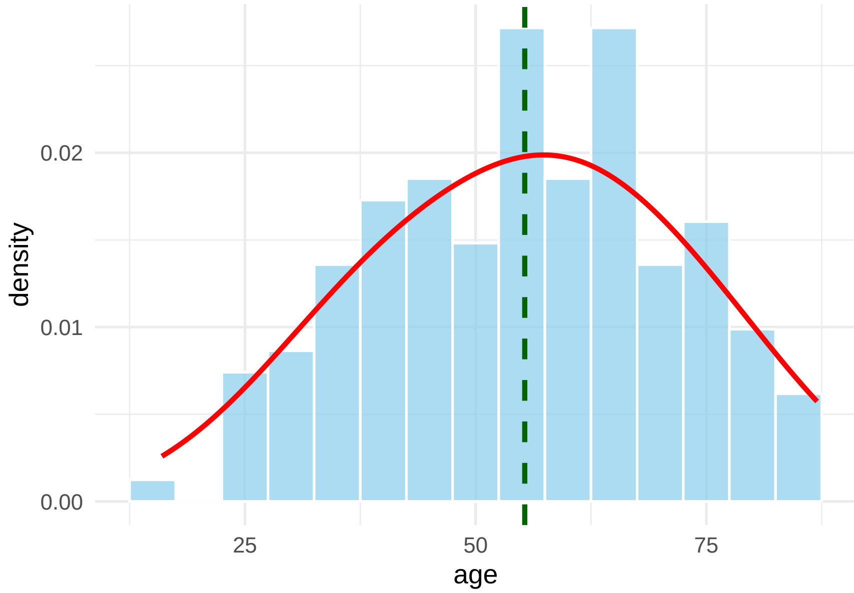
```
# summary table (grouped by outcome)
outcome_stats <- linelist %>%
  group_by(outcome) %>%
  summarise(
    count = n(),
    mean_age = mean(age, na.rm= TRUE),
    median_age = median(age, na.rm = TRUE),
    sd_age = sd(age, na.rm = TRUE),
    .groups = 'drop'
  )
outcome_stats
```

```
## # A tibble: 2 x 5
##   outcome count mean_age median_age sd_age
##   <fct>   <int>    <dbl>      <int>  <dbl>
## 1 Alive     143     53.6         55  15.8
## 2 Dead       19     68.1         68   8.90
```

```r
# ggplot for the average age of the two populations
age_by_outcome <- ggplot(linelist) +
  geom_violin(aes(x=outcome, y=age, fill=outcome)) +
  theme_minimal() +
  labs(title="",
       x="Outcome", y="Age",fill="Outcome")
age_by_outcome
```




```r
# Age histogram + normal distribution + mean age
age_histogram <- ggplot(linelist) +
  geom_histogram(
    aes(x = age, y = after_stat(density)), # y = density for overlay
    binwidth = 5, fill = "skyblue", color = "white", alpha = 0.7 ) +
  geom_density(
    aes(x = age),
    color = "red",
    linewidth = 1,
    adjust = 2
  ) +
  geom_vline( # Average age as vertical line
    xintercept = age_stats$mean_age,
    color = "darkgreen",
    linetype = "dashed",
    linewidth = 1 ) +
  theme_minimal()
age_histogram
```



2. Based on first impressions, does age seem to be a determinant of mortality in this outbreak? How might the social status of the elderly in South Korea contribute to exposure risks (e.g., care homes, hospital frequency)?

In South Korea, the elderly population is comprised of individuals aged 65 and over (https://en.wikipedia.org/wiki/Aging_of_

The deceased group represents 11.7% of the dataset, and age seems to be a determinant of mortality in the 2015 MERS outbreak in South Korea.

The average age of the Dead group is 68 years old, 14.4 years higher than the Alive group (53.6). Also, the standard deviation for the Dead group is 56% lower in the Dead group with respect to the Alive group (8.9 vs 15.8). This shows that the data points in the Dead group are closer to the mean.

Analyzing the types of exposures in this outbreak, 47% of all infections came from Hospital Room visits (27%) and Emergency Rooms (20%). "The index case was a returning traveler from the Middle East. The infection had spread within the hospital, and subsequently to other hospitals because of patient movement, resulting in nosocomial transmission at 16 clinics and hospitals." (Source: https://pmc.ncbi.nlm.nih.gov/articles/PMC5840604/).

```
hospital_count <- linelist %>%
  group_by(loc_hosp) %>%
  summarise(
    count = n(),
    percentage = count / nrow(linelist) * 100,
    mean_age = mean(age, na.rm= TRUE),
    .groups = 'drop'
  ) %>%
  arrange(desc(count))
head(hospital_count)
```

```
## # A tibble: 6 x 4
##   loc_hosp                              count percentage mean_age
##   <chr>                                 <int>      <dbl>    <dbl>
## 1 Samsung Medical Center                   85       52.5     53.9
```

```
## 2 Pyeongtaek St. Mary                37       22.8      51.7
## 3 Dae Cheong Hospital                12        7.41     67.1
## 4 Konyang University Hospital        11        6.79     69.6
## 5 Hallym University Medical Center    4        2.47     59.5
## 6 Pyeongtaek goodmorning hospital     4        2.47     73.2
```

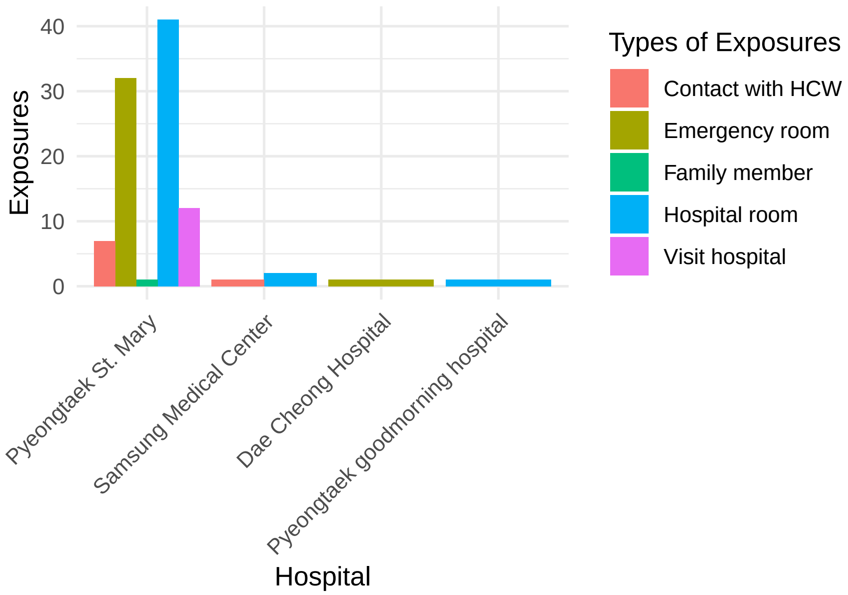
```
exposure_count <- contacts %>%
  group_by(exposure) %>%
  summarise(
    count = n(),
    percentage = count / nrow(linelist) * 100,
    .groups = 'drop'
  ) %>%
  arrange(desc(count))
head(exposure_count)
```

```
## # A tibble: 5 x 3
##   exposure           count percentage
##   <fct>              <int>      <dbl>
## 1 "Hospital room"       44     27.2
## 2 "Emergency room"      33     20.4
## 3 "Visit hospital"      12      7.41
## 4 "Contact with HCW"     8      4.94
## 5 "Family member "       1      0.617
```

```
hospital_dist <- contacts %>%
  rename(id = from) %>%
  select(id, exposure) %>%
  left_join(linelist, by="id") %>%
  group_by(loc_hosp, exposure) %>%
  summarise(
    count = n(),
    percentage = count / nrow(linelist) * 100,
    mean_age = mean(age, na.rm=TRUE),
    .groups = 'drop'
  ) %>%
  arrange(desc(count))
(hospital_dist)
```

```
## # A tibble: 9 x 5
##   loc_hosp                        exposure           count percentage mean_age
##   <chr>                           <fct>              <int>      <dbl>    <dbl>
## 1 Pyeongtaek St. Mary             "Hospital room"       41     25.3       47.0
## 2 Pyeongtaek St. Mary             "Emergency room"      32     19.8       36.3
## 3 Pyeongtaek St. Mary             "Visit hospital"      12      7.41      68
## 4 Pyeongtaek St. Mary             "Contact with HCW"     7      4.32      53.9
## 5 Samsung Medical Center          "Hospital room"        2      1.23      70
## 6 Dae Cheong Hospital             "Emergency room"       1      0.617     78
## 7 Pyeongtaek St. Mary             "Family member "       1      0.617     68
## 8 Pyeongtaek goodmorning hospital "Hospital room"        1      0.617     67
## 9 Samsung Medical Center          "Contact with HCW"     1      0.617     75
```

```r
ggplot(hospital_dist, aes(x = fct_infreq(loc_hosp), y = count, fill=exposure)) +
  geom_bar(stat="identity", position="dodge") +
  labs (
    x="Hospital",
    y="Exposures",
    fill = "Types of Exposures"
  ) +
  theme_minimal() + theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



3. Name two social factors or behaviours you would expect from the sampled demographic that would increase the likelihood of exposure to MERS-CoV (particularly for the elderly population). An example could be "hospital hopping" where patients visit multiple hospitals to see different medical professionals.

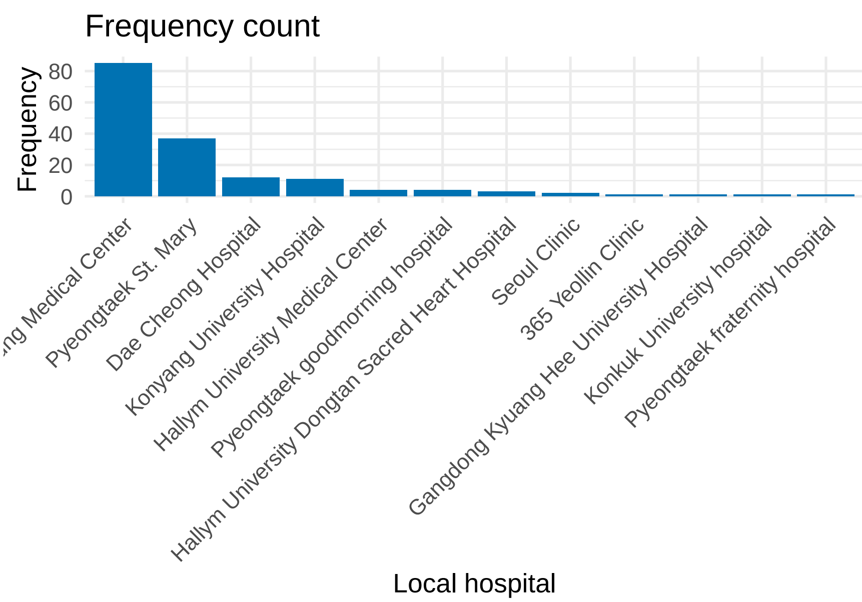
The main hospitals where MERS spread were Samsung Medical Center (52.5%) and Pyeongtaek St. Mary's Hospital (PTSM) (22.8%). These two hospitals accounted for 74% of all infections.

Hospital Room visits where family members and HCW could come and go could increase the likelihood of exposure to MERS-CoV. Another behavior that could increase the likelihood of becoming exposed would be to go to the emergency room for another medical condition/emergency.

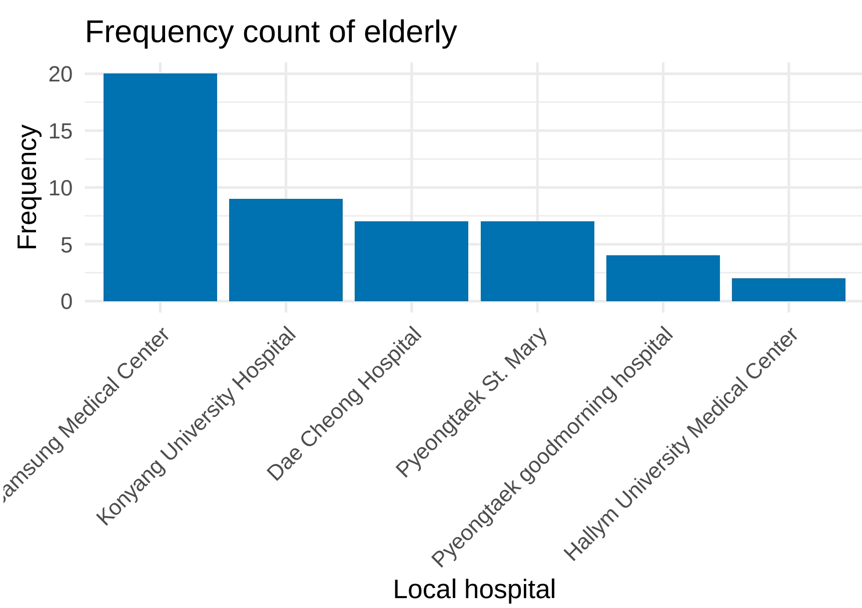
```r
# understanding the elderly population's behavior
elderly <- linelist %>%
  filter(age >= 65)

hospital_elderly <- elderly %>%
  group_by(loc_hosp) %>%
  summarise(
    count = n(),
    percentage = count / nrow(linelist) * 100,
    mean_age = mean(age, na.rm= TRUE),
    .groups = 'drop'
  ) %>%
  arrange(desc(count))
head(hospital_elderly)
```

```
## # A tibble: 6 x 4
##   loc_hosp                            count percentage mean_age
##   <chr>                               <int>      <dbl>    <dbl>
## 1 Samsung Medical Center                 20      12.3      72.4
## 2 Konyang University Hospital             9       5.56     74.1
## 3 Dae Cheong Hospital                     7       4.32     76.9
## 4 Pyeongtaek St. Mary                     7       4.32     74.3
## 5 Pyeongtaek goodmorning hospital         4       2.47     73.2
## 6 Hallym University Medical Center        2       1.23     71
```

```
elderly_dist <- ggplot(elderly) +
  geom_histogram(aes(x=age, fill=sex), binwidth=5) +
  labs(
    x="Age",
    y="Frequency",
    fill="Sex"
  ) +
  geom_vline( # Average age as vertical line
    xintercept = mean(elderly$age),
    color = "darkgreen",
    linetype = "dashed",
    linewidth = 1 ) +
  theme_minimal()

# distribution across hospitals
ggplot(linelist) +
  geom_bar(
    aes(x = fct_infreq(loc_hosp) ),
    fill = "#0072B2" ) +
  labs(
    title = "Frequency count",
    x = "Local hospital",
    y = "Frequency"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # changing the text's angle
```




```r
# distribution across hospitals (elderly)
ggplot(elderly) +
  geom_bar(
    aes(x = fct_infreq(loc_hosp) ),
    fill = "#0072B2" ) +
  labs(
    title = "Frequency count of elderly",
    x = "Local hospital",
    y = "Frequency"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # changing the text's angle
```



d. Synthesize your statistical finding with your identified social factors and behaviours to explain how **social**

status and norms acted as an **amplifier** of the biological risk posed by age. Maybe it's a good time to look into Korean culture too!

e. Repeat task (d) but with the following points: saturated emergency rooms and family caregiving, common in Korea, where there is a cultural expectation for family members (usually younger women or older teens) to stay in the patient's room providing simple emotional and practical support.

```
exposures <- contacts %>%
  group_by(exposure) %>%
  summarise (
    count = n(),
    percentage = count / nrow(linelist) * 100) %>%
  arrange(desc(count))
exposures

exposure_types <- ggplot(contacts) +
  geom_bar(aes(x = fct_infreq(exposure)), fill="lightgreen") +
  labs(
    x="Exposure Type",
    y="Cases" ) +
  theme_minimal()
```

2. Institutional Analysis

1. Calculate the median reporting lag from the original variables provided. On average, how many days were people sick and potentially interacting with their social networks before the institution 'captured' them?

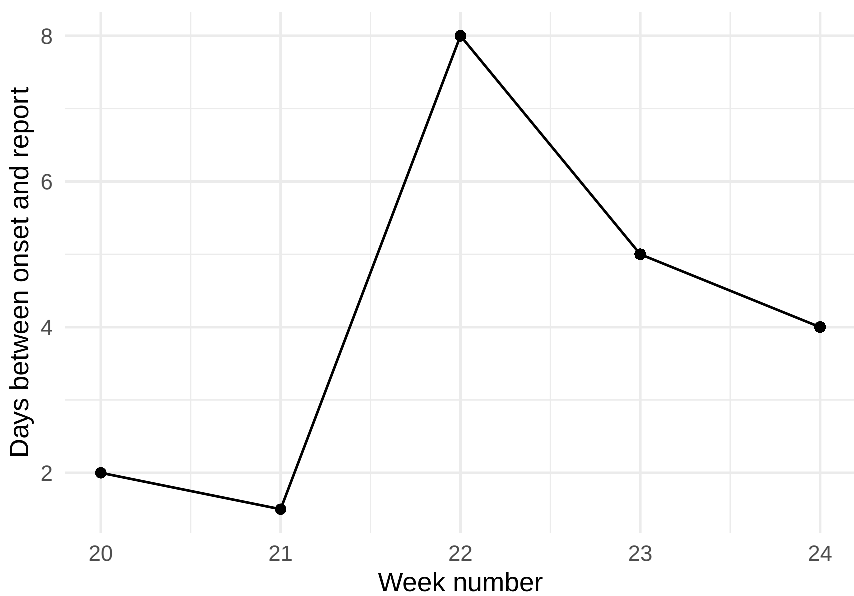
The median reporting between the date of onset and the date of reporting is 5 days. For 5 days, cases were interacting with other people, potentially transmitting MERS-CoV. SK_1, one of the super-spreaders and responsible for 26 infections, had a lag of 16 days between the onset of symptoms and their report.

```
median_reporting <- linelist %>%
  mutate (
    lag_report = dt_report - dt_onset,
    lag_exposures = dt_end_exp - dt_start_exp,
    lag_diagnosis = dt_diag - dt_onset,
    week_report = week(dt_report)) %>%
  filter(
    !is.na(lag_report),
    !is.na(lag_exposures),
    !is.na(lag_diagnosis)) %>%
  group_by(week_report) %>%
  reframe(
    week_report,
    lag_report = lag_report,
    lag_exposures = lag_exposures,
    lag_diagnosis = lag_diagnosis,
    median_report = median(lag_report, na.rm = TRUE),
    median_exposures = median(lag_exposures, na.rm = TRUE),
    median_diagonsis = median(lag_diagnosis, na.rm = TRUE),
    .groups = 'drop'
  )
head(median_reporting)
```

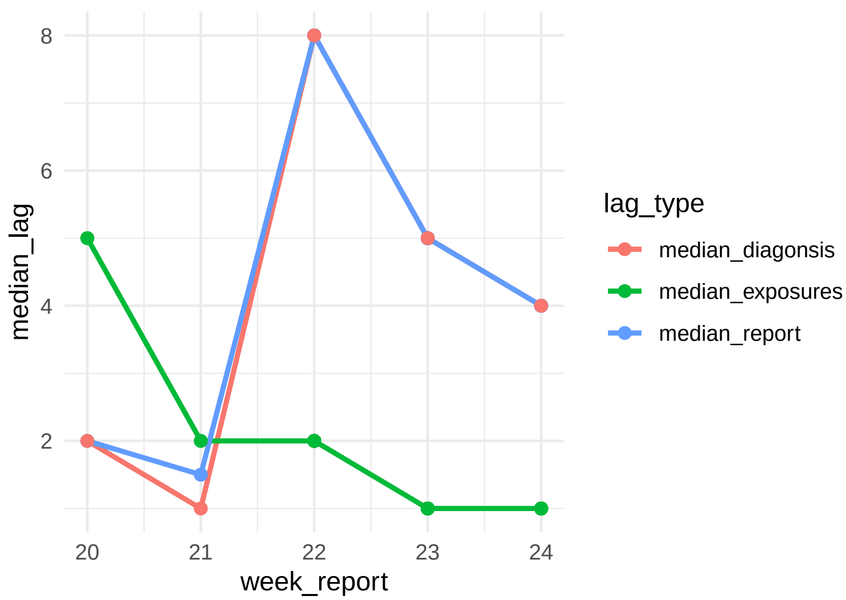
```
## # A tibble: 6 x 8
##   week_report lag_report lag_exposures lag_diagnosis median_report
##         <dbl> <drtn>     <drtn>        <drtn>        <drtn>
## 1          20 8 days     16 days       9 days        2.0 days
## 2          20 2 days      5 days       2 days        2.0 days
## 3          20 0 days      0 days       1 days        2.0 days
## 4          21 1 days      4 days       1 days        1.5 days
## 5          21 2 days      0 days       1 days        1.5 days
## 6          22 4 days      2 days       4 days        8.0 days
## # i 3 more variables: median_exposures <drtn>, median_diagonsis <drtn>,
## #   .groups <chr>
```

2. Analyse your new variable as a function of time. What does the trend tell you about the institution's ability to 'learn' and mobilize resources over time?

```r
# analyzing different lags
lag_by_week_long <- median_reporting %>%
  pivot_longer(
    cols = starts_with("median"),
    names_to = "lag_type",
    values_to = "median_lag" )

# median lag across time
ggplot(median_reporting, aes(x=week_report, y=median_report)) +
  geom_line() + geom_point() + theme_minimal() +
  labs(x="Week number", y="Days between onset and report")
```




```r
ggplot(lag_by_week_long,
       aes(x = week_report,
           y = median_lag,
           color = lag_type,
           group = lag_type) ) +
  geom_line(linewidth = 1) +
  geom_point(size = 2) + theme_minimal()
```



3. In MERS outbreaks in the Middle East, men were infected much more often. In this South Korean outbreak, is there a gender imbalance? What could this imply about the gender demographics of caregiving in Korean hospitals?

Bonus - investigate why there were more male than female cases in the middle east spread of MERS. Why would this not apply in Korea?

3. The Super-Spreader

a) Using the appropriate dataset, calculate the number of infections attributed to each case. Calculate the mean and the maximum of this variable. What is the ratio between the maximum "infector" and the average?

For this analysis, I will categorize the super-spreader as those ids that infected 5 or more other cases.

Based on the Contacts data set and the previous assumption, we can identify 10 spreaders in this outbreak, and 3 super-spreaders: SK_14, SK_1, and SK_16. These 3 super-spreaders are males with an average age of 35 years old and located in the Pyeongtaek St. Mary Hospital (PTSM).

The 3 super-spreaders were responsible for 85 infections (52%). The super-spreader infected 4 times more than the average spreaders.

```
# count of infections caused by each spreader
all_cases <- nrow(linelist)

spreader_count <- contacts %>%
  group_by(from) %>%
  summarise(
    infections = n(),
    percentage_of_infections = infections / all_cases * 100,
    .groups = 'drop' ) %>%
```

```
  arrange(desc(infections))
spreader_count
```

```
## # A tibble: 11 x 3
##    from   infections percentage_of_infections
##    <chr>       <int>                    <dbl>
##  1 SK_14          38                   23.5
##  2 SK_1           26                   16.0
##  3 SK_16          21                   13.0
##  4 SK_15           4                    2.47
##  5 SK_6            2                    1.23
##  6 SK_76           2                    1.23
##  7 SK_11           1                    0.617
##  8 SK_118          1                    0.617
##  9 SK_12           1                    0.617
## 10 SK_123          1                    0.617
## 11 SK_87           1                    0.617
```

```
# how did these spreaders infect the other cases?
superspreaders <- contacts %>%
  group_by(from,
           exposure) %>%
  summarise(
    infections = n(),
    percentage_of_infections = infections / all_cases * 100,
    .groups = 'drop') %>%
  arrange(desc(infections))
superspreaders
```

```
## # A tibble: 20 x 4
##    from   exposure           infections percentage_of_infections
##    <chr>  <fct>                   <int>                    <dbl>
##  1 SK_14  "Emergency room"           30                   18.5
##  2 SK_16  "Hospital room"            20                   12.3
##  3 SK_1   "Visit hospital"           12                    7.41
##  4 SK_1   "Hospital room"             9                    5.56
##  5 SK_14  "Hospital room"             6                    3.70
##  6 SK_1   "Contact with HCW"          4                    2.47
##  7 SK_15  "Hospital room"             3                    1.85
##  8 SK_14  "Contact with HCW"          2                    1.23
##  9 SK_1   "Family member "            1                    0.617
## 10 SK_11  "Hospital room"             1                    0.617
## 11 SK_118 "Hospital room"             1                    0.617
## 12 SK_12  "Hospital room"             1                    0.617
## 13 SK_123 "Hospital room"             1                    0.617
## 14 SK_15  "Contact with HCW"          1                    0.617
## 15 SK_16  "Emergency room"            1                    0.617
## 16 SK_6   "Emergency room"            1                    0.617
## 17 SK_6   "Hospital room"             1                    0.617
## 18 SK_76  "Contact with HCW"          1                    0.617
## 19 SK_76  "Hospital room"             1                    0.617
## 20 SK_87  "Emergency room"            1                    0.617
```

```r
superspreaders_pivot <- superspreaders %>%
  filter(infections >= 5) %>%
  pivot_wider(
    id_cols = from,
    names_from = exposure,
    values_from = infections
  )
superspreaders_pivot
```

```
## # A tibble: 3 x 4
##   from  `Emergency room` `Hospital room` `Visit hospital`
##   <chr>            <int>           <int>            <int>
## 1 SK_14               30               6               NA
## 2 SK_16               NA              20               NA
## 3 SK_1                NA               9               12
```

```r
# What is the ratio between the maximum "infector" and the average?
# formula for a ratio = max / average

# calculating the max
max_infector_count <- max(spreader_count$infections, na.rm = TRUE)
max_infector_count
```

```
## [1] 38
```

```r
# calculating the average
avg_spreader_count <- mean(spreader_count$infections, na.rm = TRUE)
avg_spreader_count
```

```
## [1] 8.909091
```

```r
# calculating the ratio
ratio <- max_infector_count / avg_spreader_count
ratio
```

```
## [1] 4.265306
```

b) Visualise the distribution of secondary cases. What does the shape of your plot imply about how “democratic” or “equal” disease transmission was in this specific sample?

c) Summarise the demographic profile of the “super spreader” using both datasets. Based on where they were infected and their demographic profile, hypothesize a social reason why this specific individual had such a high number of secondary cases.

```r
# left join to dig deeper into the superspreaders' details
superspreaders_deep <- superspreaders %>%
  rename(id = from) %>%   # rename 'from' to 'id'
  left_join(linelist, by = "id") %>%
  select(id,
         age,
         age_class,
```

```
         sex,
         infections,
         percentage_of_infections,
         place_infect,
         reporting_ctry,
         exposure,
         loc_hosp,
         outcome) %>%
  filter(infections >= 5)
superspreaders_deep
```

Bonus - Test the "Pareto Principle" (the 20/80 rule) on this data. Calculate what percentage of the total infections were caused by the top 20% of infectors. Does this outbreak fit the rule?

Appendix

Attempting to graph a transmission chain