

Analysis of the MERS-CoV outbreak in Korea (2015) using R tools

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 in South Korea in 2015. Source: <https://github.com/luisfmontemayor/learn-compbio/tree/main>

Loading and inspecting the data

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

Quick questions

1. What object class is `mers_korea_2015` in R?

`mers_korea_2015` is a list, a collection of data that is ordered and changeable (source: https://www.w3schools.com/r/r_lists.asp).

2. What sub-items exist within this item?

It contains 2 sub-items that are dataframes: `linelist` and `contacts` (source: <https://www.rdocumentation.org/packages/outbreaks/versions/1.1.0/topics/mers.korea.2015>).

`linelist`: A dataframe of MERS-CoV cases and their attributes.

`contacts`: A dataframe describing the relationship between MERS Co-V cases.

```
# inspecting the raw data
class(mers_korea_2015)
class(mers_korea_2015$linelist)
class(mers_korea_2015$contacts)
```

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

```
length(mers_korea_2015) # number of elements the list contains
names(mers_korea_2015) # names of the elements of the list
summary(mers_korea_2015)
str(mers_korea_2015)
```

```
## [1] 2
## [1] "linelist" "contacts"
##           Length Class      Mode
## linelist 15      data.frame list
## contacts  4      data.frame list
## 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 ...
## ..$ reporting_ctr: Factor w/ 2 levels "China","South Korea": 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 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 ...
## ..$ 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 ...
```

```
describe(mers_korea_2015$linelist)
describe(mers_korea_2015$contacts)
```

```
summary(mers_korea_2015$linelist)
summary(mers_korea_2015$contacts)
```

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

```
# different ways of accessing the sub-items
mers_korea_2015$linelist
mers_korea_2015[1]
```

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

```
# linelist
linelist <- mers_korea_2015$linelist

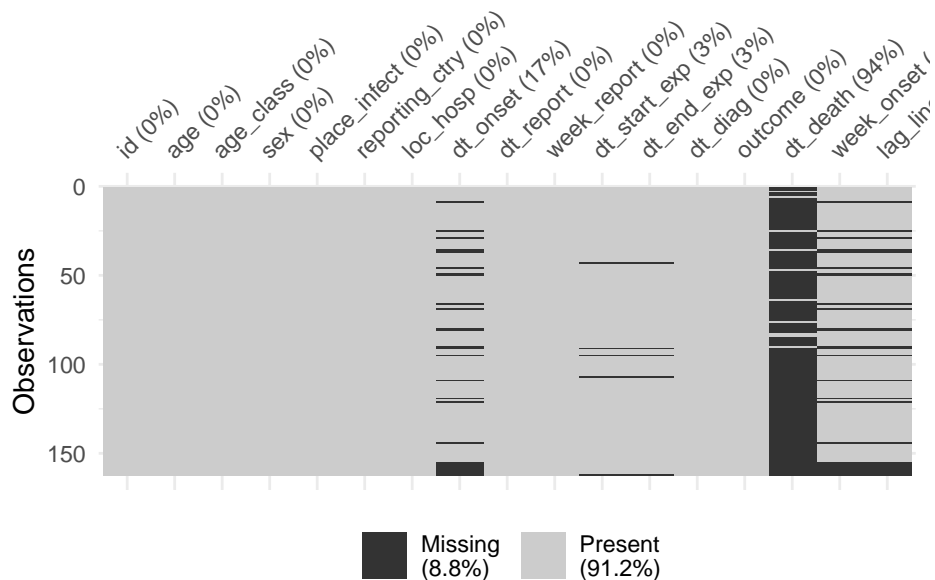
# cleaning the linelist
linelist <- linelist %>%
  mutate(
    loc_hosp = str_replace(
      string = loc_hosp,
      pattern = ",.*",
      replacement = ""
    ) ) %>%
```

```
mutate(
  loc_hosp = str_trim(loc_hosp),
  week_report = week(dt_report),
  week_onset = week(dt_onset),
  lag_linelist = as.numeric(dt_report - dt_onset)
)

# missing values
linelist %>%
  summarise(across(everything(), ~ sum(is.na(.)))) %>%
  pivot_longer(cols = everything(),
    names_to = "variable",
    values_to = "NA_count") %>%
  mutate(percentage = round((NA_count / nrow(linelist)) * 100, 2)) %>%
  filter(NA_count > 0)
```

```
## # A tibble: 6 x 3
##   variable    NA_count percentage
##   <chr>      <int>      <dbl>
## 1 dt_onset        27        16.7
## 2 dt_start_exp     5         3.09
## 3 dt_end_exp       5         3.09
## 4 dt_death       152        93.8
## 5 week_onset      27        16.7
## 6 lag_linelist    27        16.7
```

```
# library(naniar) to check the missing values in the dataset
linelist %>%
  vis_miss() +
  theme(axis.text.x = element_text(vjust = 0) )
```



```
# contacts
contacts <- mers_korea_2015$contacts
```

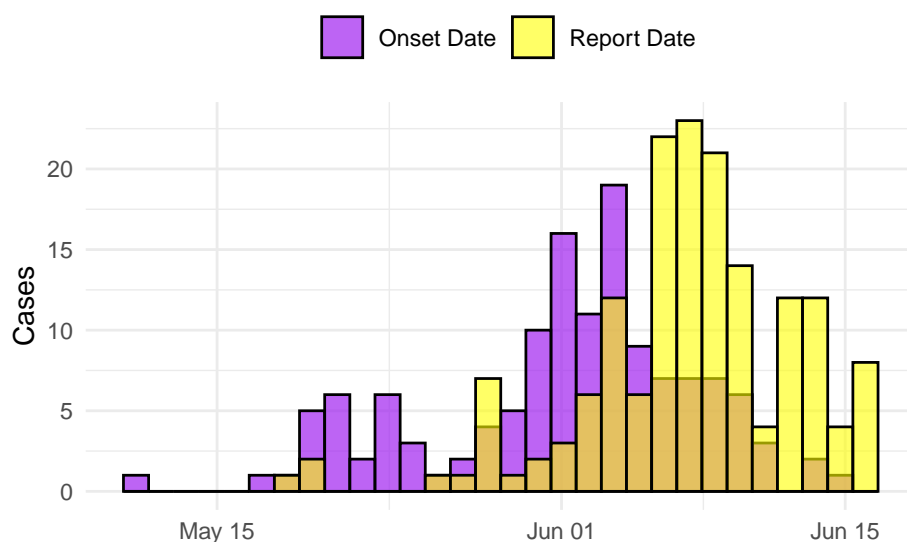
```
# cleaning the contacts
contacts <- mers_korea_2015$contacts %>%
  rename(lag_contacts = diff_dt_onset) %>%
  mutate(
    lag_contacts = as.numeric(lag_contacts) )
```

Analyses

Getting my hands dirty with the data

```
# distribution of cases by onset and report dates
ggplot(linelist) +
  geom_histogram(aes(x=dt_onset, fill="Onset Date"), alpha = 0.7, color="black") +
  geom_histogram(aes(x=dt_report, fill="Report Date"), alpha=0.6, color="black") +
  scale_fill_manual(values = c("Onset Date" = "purple",
                                "Report Date" = "yellow")) +
  theme_minimal() + labs(x="", y="Cases",
                        title="Distribution of cases by Onset and Report Dates",
                        fill="") +
  theme(legend.position="top")
```

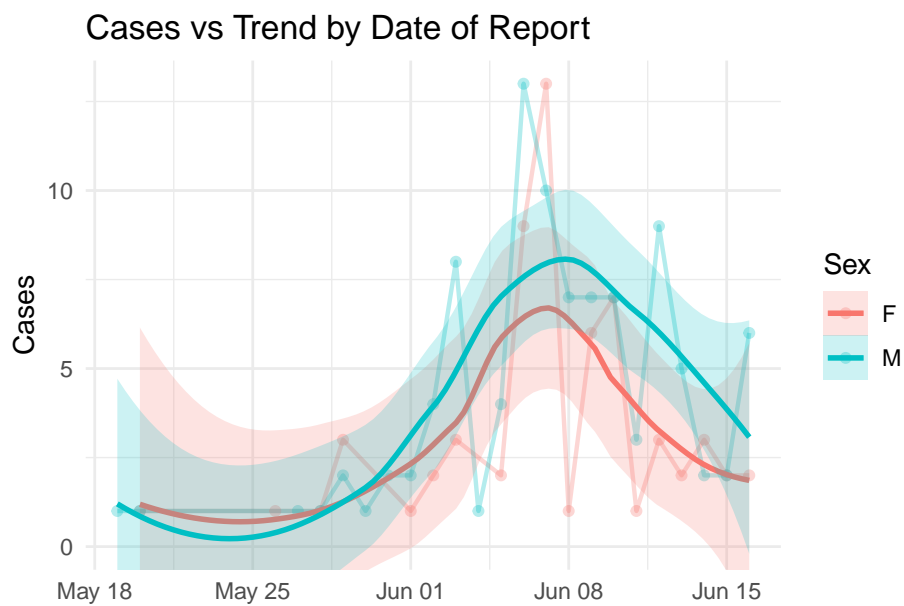
Distribution of cases by Onset and Report Dates



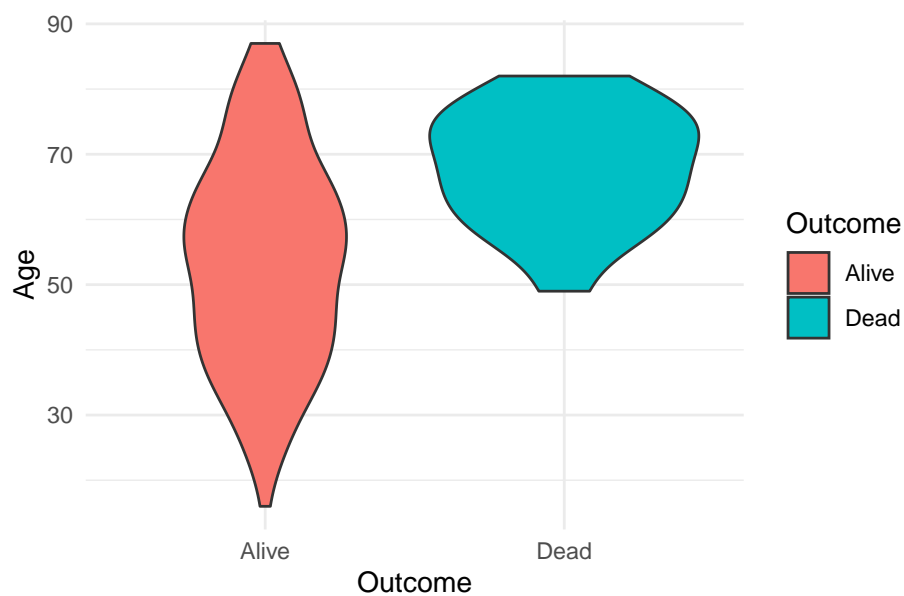
```
# admissions by sex
linelist %>%
  group_by(dt_report, sex) %>%
  summarise(count = n(), .groups = 'drop') %>%

  ggplot(aes(x = dt_report, y = count, color = sex, fill = sex)) +
  geom_line(alpha = 0.3, size = 0.8) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "loess", level=0.9, alpha = 0.2) +
  coord_cartesian(ylim = c(0, NA)) +
```

```
theme_minimal() +
labs(title = "Cases vs Trend by Date of Report",
     x = "", y = "Cases", color="Sex", fill="Sex")
```



```
# average age by outcome
ggplot(linelist) +
  geom_violin(aes(x=outcome, y=age, fill=outcome)) +
  theme_minimal() +
  labs(title="",
       x="Outcome", y="Age", fill="Outcome")
```



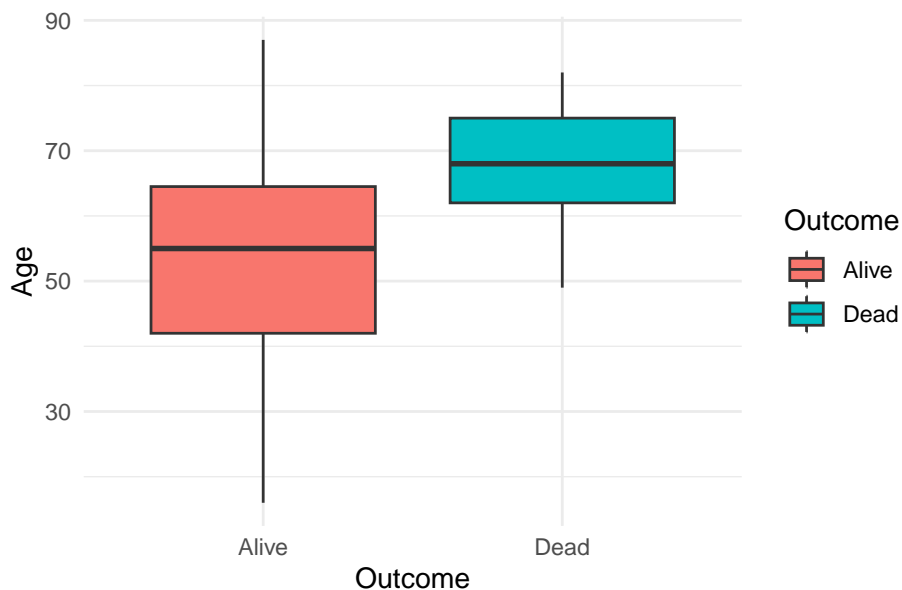
1. Demographics

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

```
# age stats grouped by outcome
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'
  )
```

```
## # 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
```

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

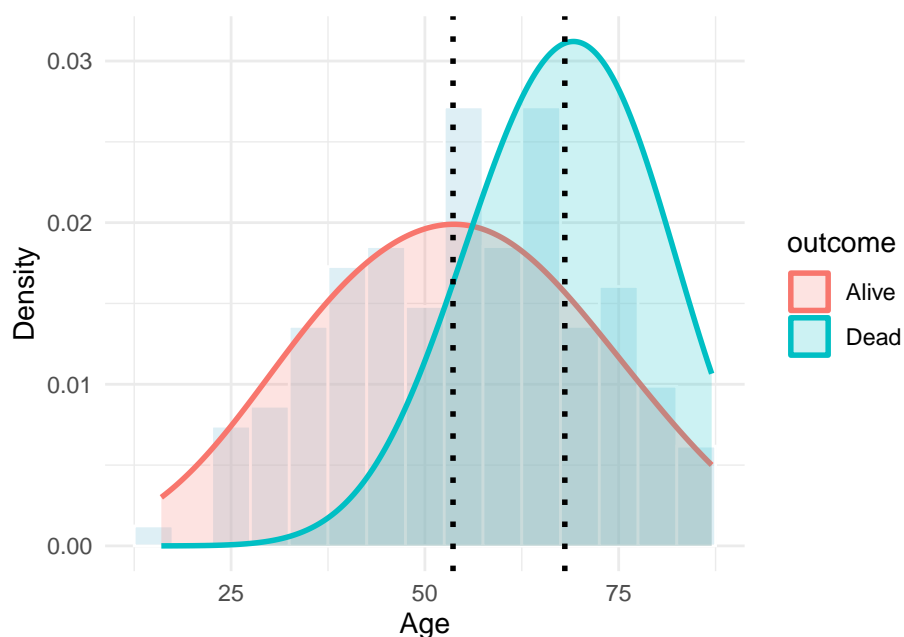


```
# distribution of age by outcome
mean_age <- linelist %>%
  group_by(outcome) %>%
  summarise(
    mean_age = mean(age, na.rm=TRUE) )
```

```
ggplot(linelist, aes(x=age, fill = outcome, color = outcome)) +
  geom_histogram(
    aes(y = after_stat(density)), # y = density for overlay
    binwidth = 5, fill = "lightblue", color = "white", alpha = 0.4 ) +

  geom_density(aes(fill = outcome), alpha = 0.2, linewidth = 1, adjust = 2) +

  geom_vline( # Average age as vertical line
    data = mean_age, aes(xintercept = mean_age),
    show.legend = FALSE,
    linetype = "dotted",
    size = 2, linewidth = 1 ) +
  theme_minimal() +
  labs(y="Density", x="Age")
```



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 older (source: <https://en.wikipedia.org/wiki/Aging>). The deceased group represents 11.7% of the dataset and the average age for this group was 68 years old. The standard deviation for this group was lower with respect to the Alive group (8.9 vs 15.8), showing that the data points in the Dead group were closer to the mean. This hints that age might be a determinant of mortality during the MERS-CoV outbreak.

47% of all infections came from nosocomial transmission (Hospital Room Visits, Emergency Rooms) (source: <https://pmc.ncbi.nlm.nih.gov/articles/PMC5840604/>). Hospital room visits where family members, healthcare workers and other infected patients could interact 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.

```
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 <- linelist %>%
  rename(to = id) %>%
  left_join(contacts, by="to") %>%
  rename(id = to) %>%
  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))
head(hospital_dist)

```

```

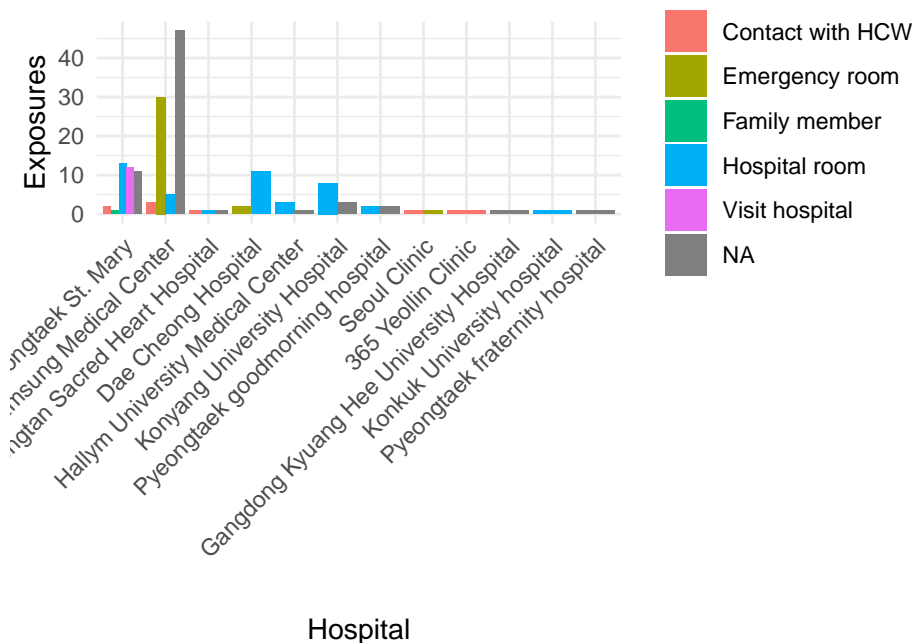
## # A tibble: 6 x 5
##   loc_hosp          exposure          count percentage mean_age
##   <chr>          <fct>          <int>     <dbl>     <dbl>
## 1 Samsung Medical Center <NA>          47      29.0      53.5

```



```
## 2 Samsung Medical Center Emergency room 30 18.5 55.0
## 3 Pyeongtaek St. Mary Hospital room 13 8.02 56.3
## 4 Pyeongtaek St. Mary Visit hospital 12 7.41 48.2
## 5 Dae Cheong Hospital Hospital room 11 6.79 67.4
## 6 Pyeongtaek St. Mary <NA> 11 6.79 55.4
```

```
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))
```



- 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-CoV 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. In the case of the elderly population (49 cases), 20 of them were infected at Samsung Medical Center.

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.

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

```
## [1] 49
```

```

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)

```

```

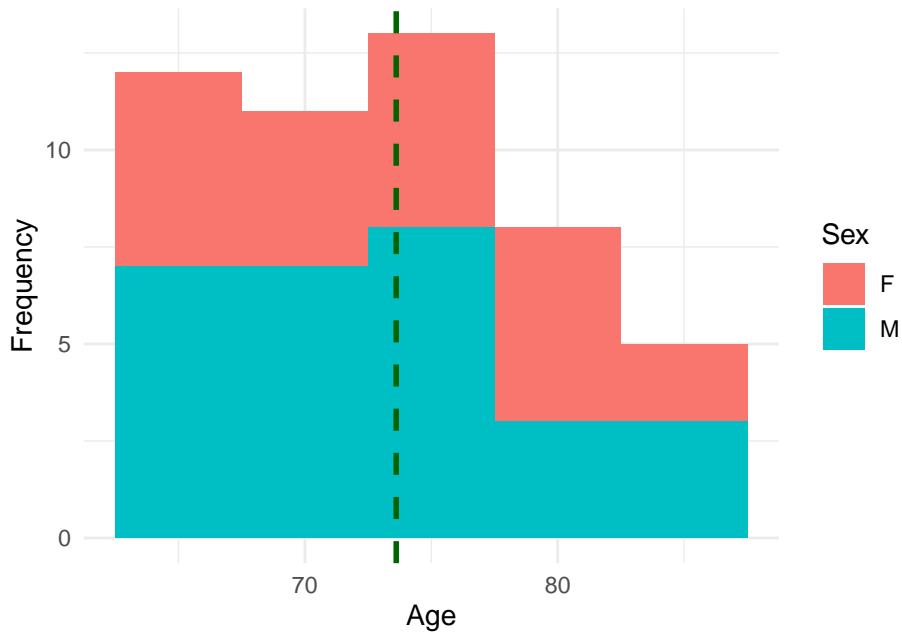
## # A tibble: 6 x 4
##   loc_hosp                count percentage mean_age
##   <chr>                <int>      <dbl>    <dbl>
## 1 Samsung Medical Center      20      12.3     72.4
## 2 Konyang University Hospital    9       5.56     74.1
## 3 Dae Cheong Hospital          7       4.32     76.9
## 4 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

```

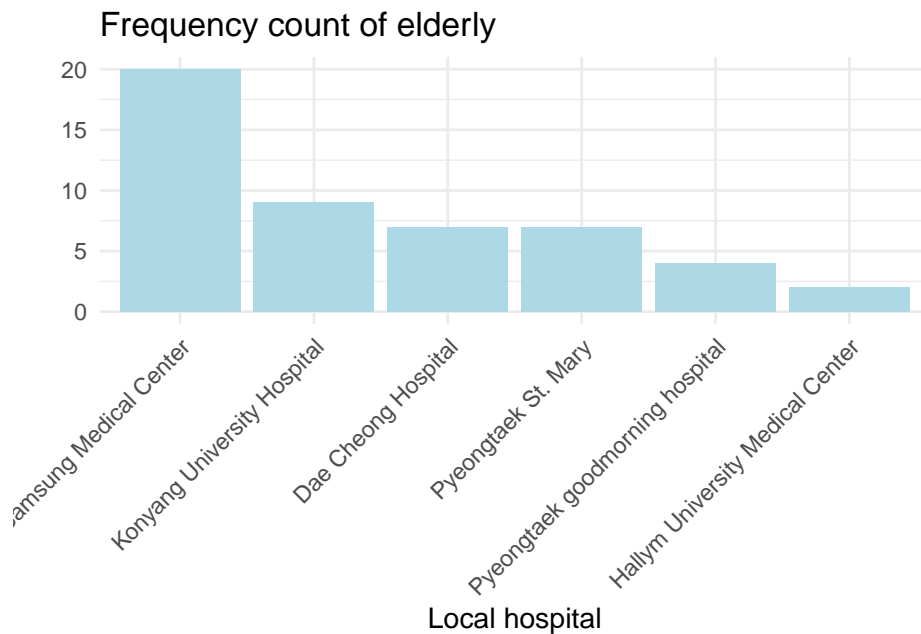
```

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()
elderly_dist

```



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



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

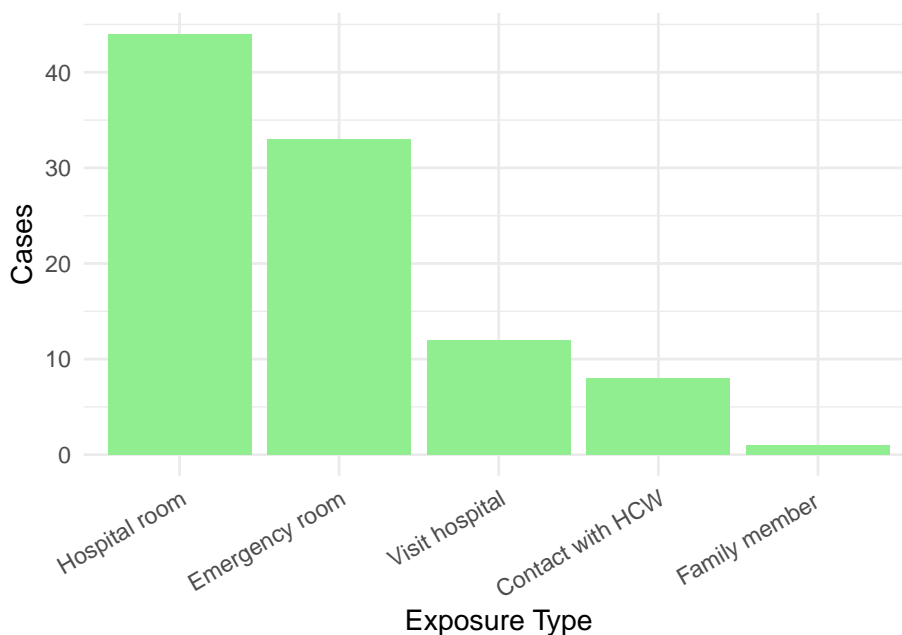
status and norms acted as an **amplifier** of the biological risk posed by age. Maybe it's a good time to look into Korean culture too!

- 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))
head(exposures)
```

```
## # 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
```

```
exposure_types <- ggplot(contacts) +
  geom_bar(aes(x = fct_infreq(exposure)), fill="lightgreen") +
  labs(
    x="Exposure Type",
    y="Cases" ) +
  theme_minimal() + theme(axis.text.x = element_text(angle = 30, hjust = 1))
exposure_types
```



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?

On a first approach, the median reporting lag in the `linelist` dataset is 5 days while the median lag in the `contacts` dataset is 14 days. Digging deeper into the data, the primary cases of the `contacts` dataset represent 6.7% of the `linelist`, while the secondary cases represent 58%. Also, 16% of the cases from the `linelist` do not have an onset date.

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.

The reporting lag of the `contacts` dataset is nearly 3 times higher than the `linelist`. This suggests that the secondary cases took longer to trace and report.

```
# to see the number of onset dates that are missing
```

```
no_onset_date <- linelist %>%  
  summarise(  
    count = n(),  
    yes_dt_onset = sum(!is.na(dt_onset)),  
    no_dt_onset = sum(is.na(dt_onset)),  
    percentage = (no_dt_onset / count) * 100  
  )  
no_onset_date
```

```
##   count yes_dt_onset no_dt_onset percentage  
## 1   162           135           27   16.66667
```

```
median_contacts <- median(contacts$diff_dt_onset)  
median_linelist <- median(as.numeric(linelist$dt_report - linelist$dt_onset), na.rm=TRUE)
```

```
median_contacts  
median_linelist
```

```
# difference between both datasets
```

```
median_contacts - median_linelist
```

```
## NULL  
## [1] 5  
## numeric(0)
```

```
# what is the proportion of the contacts dataset
```

```
n_distinct(contacts$from) / nrow(linelist) * 100 # primary cases  
n_distinct(contacts$to) / nrow(linelist) * 100 # secondary cases
```

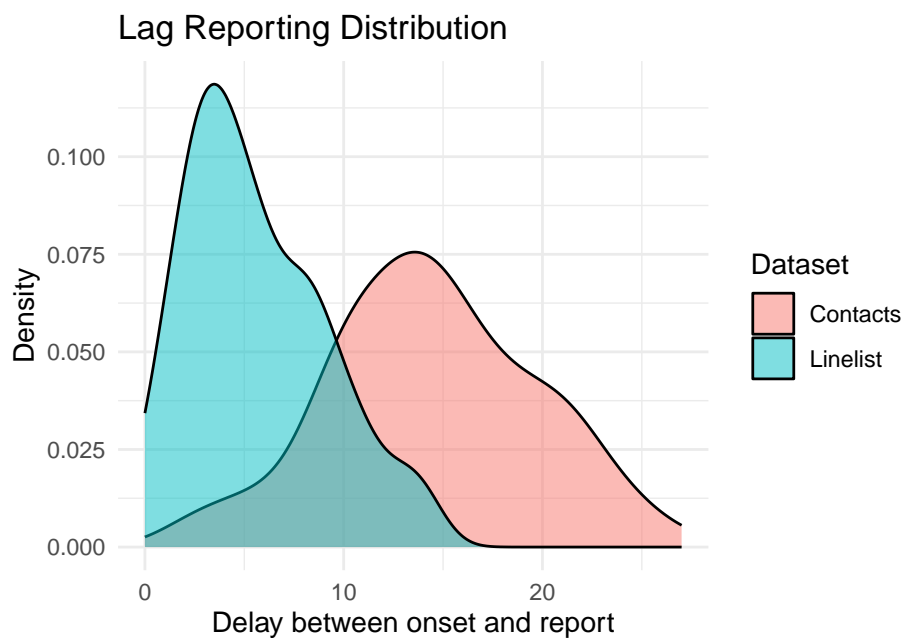
```
## [1] 6.790123  
## [1] 58.64198
```

```
# all lag observations across linelist and contacts
```

```
lag_comparison <- data.frame(  
  delay = c(linelist$lag_linelist,  
            contacts$lag_contacts),  
  dataset = c(rep("Linelist", nrow(linelist)),  
              rep("Contacts", nrow(contacts)))
```

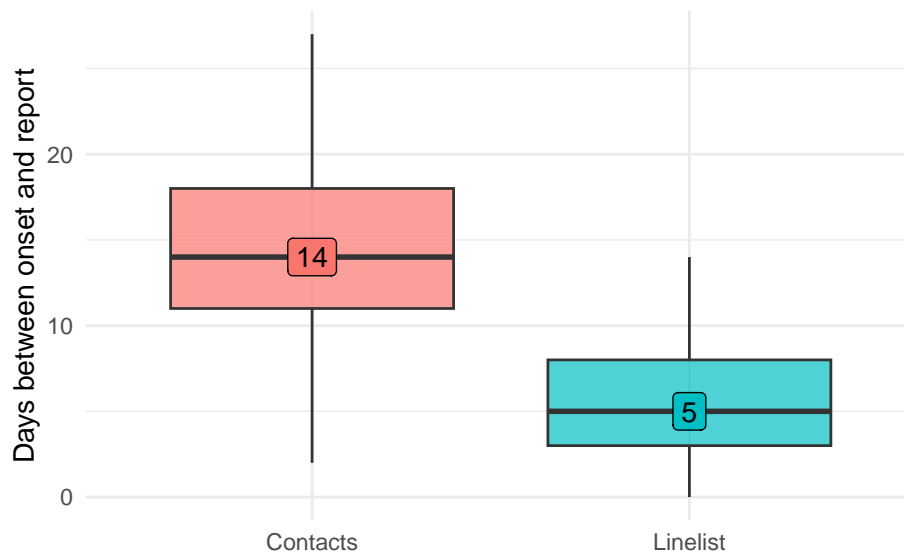
```
)

# comparing the delay distributions across datasets
ggplot(lag_comparison, aes(x = delay, fill = dataset)) +
  geom_density(alpha = 0.5) +
  theme_minimal() +
  labs(title = "Lag Reporting Distribution",
       x="Delay between onset and report",
       y="Density",
       fill="Dataset")
```



```
# comparing the median reporting lags across datasets
ggplot(lag_comparison, aes(x = dataset, y = delay, fill = dataset)) +
  geom_boxplot(alpha = 0.7) +
  stat_summary(fun = median, geom = "label", aes(label = after_stat(y)), vjust = 0.5) +
  theme_minimal() +
  labs(title = "Median Reporting Lag: Linelist vs Contact Chain",
       y = "Days between onset and report",
       x = "") +
  guides(fill = "none")
```

Median Reporting Lag: Linelist vs Contact Chain



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?

In week 22, the median reporting lag between the onset of symptoms and the report was the highest recorded (8 days). After week 22, the Korea CDC learned and the cases reported in weeks 23 and 24 decreased their lags to 5 and 4 days respectively.

3. In MERS-CoV 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?

In South Korea, men were infected 1.5 times higher than women.

```
# looking up the gender distribution
gender_stats <- linelist %>%
  filter(
    place_infect != "Middle East"
  ) %>%
  group_by(place_infect, sex) %>%
  summarise(
    count = n(),
    percentage = count / nrow(linelist) * 100,
    median_age = median(age, na.rm = TRUE))
gender_stats
```

```
## # A tibble: 2 x 5
## # Groups:   place_infect [1]
##   place_infect      sex  count percentage median_age
##   <fct>          <fct> <int>      <dbl>      <dbl>
## 1 Outside Middle East F      63      38.9       57
## 2 Outside Middle East M      98      60.5       55
```

```
gender_ratio <- nrow(subset(linelist, sex == "M")) / nrow(subset(linelist, sex == "F"))
gender_ratio
```

```
## [1] 1.571429
```

```
# distribution by age and sex
ggplot(linelist, aes(x = age, fill = sex)) +
  geom_histogram(data = subset(linelist, sex == "F"), aes(y = ..count.. * -1)) +
  geom_histogram(data = subset(linelist, sex == "M"), aes(y = ..count..)) +
  coord_flip() +
  theme_minimal() +
  labs(title = "Distribution by Age and Gender", y = "Cases", x = "Age") +
  scale_y_continuous(labels = abs)
```



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

3. The Super-Spreader

- 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))
head(spreader_count)

```

```

## # A tibble: 6 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

```

```

# 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))
head(superspreaders)

```

```

## # A tibble: 6 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

```

```

# 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
```

```

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

```

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

```
# 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.

```
# 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_ctype,
         exposure,
         loc_hosp,
         outcome) %>%
  filter(infections >= 5)
superspreaders_deep
```

```
## # A tibble: 5 x 11
##   id      age age_class sex   infections percentage_of_infections place_infect
##   <chr> <int> <chr>    <fct>      <int>                <dbl> <fct>
## 1 SK_14   35 30-39    M          30                18.5 Outside Middl~
## 2 SK_16   40 40-49    M          20                12.3 Outside Middl~
## 3 SK_1    68 60-69    M          12                 7.41 Middle East
## 4 SK_1    68 60-69    M           9                 5.56 Middle East
## 5 SK_14   35 30-39    M           6                 3.70 Outside Middl~
## # i 4 more variables: reporting_ctype <fct>, exposure <fct>, loc_hosp <chr>,
## #   outcome <fct>
```

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

Infection Network

```
library(igraph)
links <- data.frame(
  source= contacts$from,
```

```

target = contacts$to
)
network <- graph_from_data_frame(
  d=links,
  directed=F
)

plot(network,
  vertex.size = 10,
  vertex.label.cex = 0.5,
  vertex.color="lightblue")

```

