FluSurvNET EDA and Questions

Evan L. Ray

2020-11-17

```r
library(tidyverse)
library(cdcfluview)
```

This document contains an exploratory data analysis of FluSurvNET data with some questions that came up.

Reading in data

```r
# load data
all_hosp <- surveillance_areas() %>%
  purrr::pmap_dfr(hospitalizations) %>%
  # subset to regions to forecast
  dplyr::filter(
    !(region %in% c("Idaho", "Oklahoma", "Rhode Island", "South Dakota")),
    !(region == "Entire Network" & surveillance_area %in% c("EIP", "IHSP"))
  ) %>%
  # add variables for epi week and week of season starting on epi week 31
  dplyr::mutate(
    epiweek = paste0(year, sprintf("%02d", year_wk_num)),
    season_week = cdcfluutils::mmwr_week_to_season_week(year_wk_num, year)
  )
```

Investigate missing data

Missing age labels

```r
# How many rows have a non-missing age label?
sum(!is.na(all_hosp$age_label))
```

```
## [1] 40026
```

```r
# First few rows with non-missing age label
all_hosp %>%
  dplyr::filter(!is.na(age_label)) %>%
  head() %>%
  dplyr::select(region, epiweek, weeklyrate, age, age_label)
```

```
## # A tibble: 6 x 5
##   region         epiweek weeklyrate   age age_label
##   <chr>          <chr>        <dbl> <int> <fct>
## 1 Entire Network 201040         0       3 18-49 yr
## 2 Entire Network 201041         0       3 18-49 yr
## 3 Entire Network 201042         0       3 18-49 yr
## 4 Entire Network 201043         0       3 18-49 yr
## 5 Entire Network 201044         0.1     3 18-49 yr
## 6 Entire Network 201045         0       3 18-49 yr
```

```
# Unique values of age, among rows with non-missing age label
all_hosp %>%
  dplyr::filter(!is.na(age_label)) %>%
  dplyr::pull(age) %>%
  unique()
```

```
## [1] 3 2 6 5 4 1
```

```
# How many rows have a missing age label?
sum(is.na(all_hosp$age_label))
```

```
## [1] 45022
```

```
# First few rows with missing age label
all_hosp %>%
  dplyr::filter(is.na(age_label)) %>%
  head() %>%
  dplyr::select(region, epiweek, weeklyrate, age, age_label)
```

```
## # A tibble: 6 x 5
##   region         epiweek weeklyrate   age age_label
##   <chr>          <chr>        <dbl> <int> <fct>
## 1 Entire Network 201840         0.1    12 <NA>
## 2 Entire Network 201841         0.1    12 <NA>
## 3 Entire Network 201842         0.1    12 <NA>
## 4 Entire Network 201843         0.1    12 <NA>
## 5 Entire Network 201844         0.2    12 <NA>
## 6 Entire Network 201845         0.1    12 <NA>
```

```
# Unique values of age, among rows with missing age label
all_hosp %>%
  dplyr::filter(is.na(age_label)) %>%
  dplyr::pull(age) %>%
  unique()
```

```
## [1] 12  9  8 11 10  7
```

Questions: What is the age variable? Is it safe to ignore the rows with ages but no age labels?

In everything below, we subset to the data without any missing values for the age label.

```
all_hosp <- all_hosp %>% dplyr::filter(!is.na(age_label))
```

Missing values for some age categories

```
# Counts of age label categories -- note fewer for some age categories
table(all_hosp$age_label)
```

```
##
##   0-4 yr  5-17 yr 18-49 yr 50-64 yr   65+ yr  Overall
##     7006     7006     6336     6336     6336     7006
```

```
# Which observations have missing values for the
# 18-49, 50-64, and 65+ categories?
hosp_wide <- all_hosp %>%
  dplyr::select(-rate, -age) %>%
  tidyr::pivot_wider(names_from = "age_label", values_from = "weeklyrate")

age_cols <- as.character(unique(all_hosp$age_label[!is.na(all_hosp$age_label)]))
rows_missing <- hosp_wide %>%
  apply(1, function(hosp_row) { any(is.na(hosp_row[age_cols])) })

# Which locations?
hosp_wide %>%
  dplyr::slice(which(rows_missing)) %>%
  dplyr::distinct(region)
```

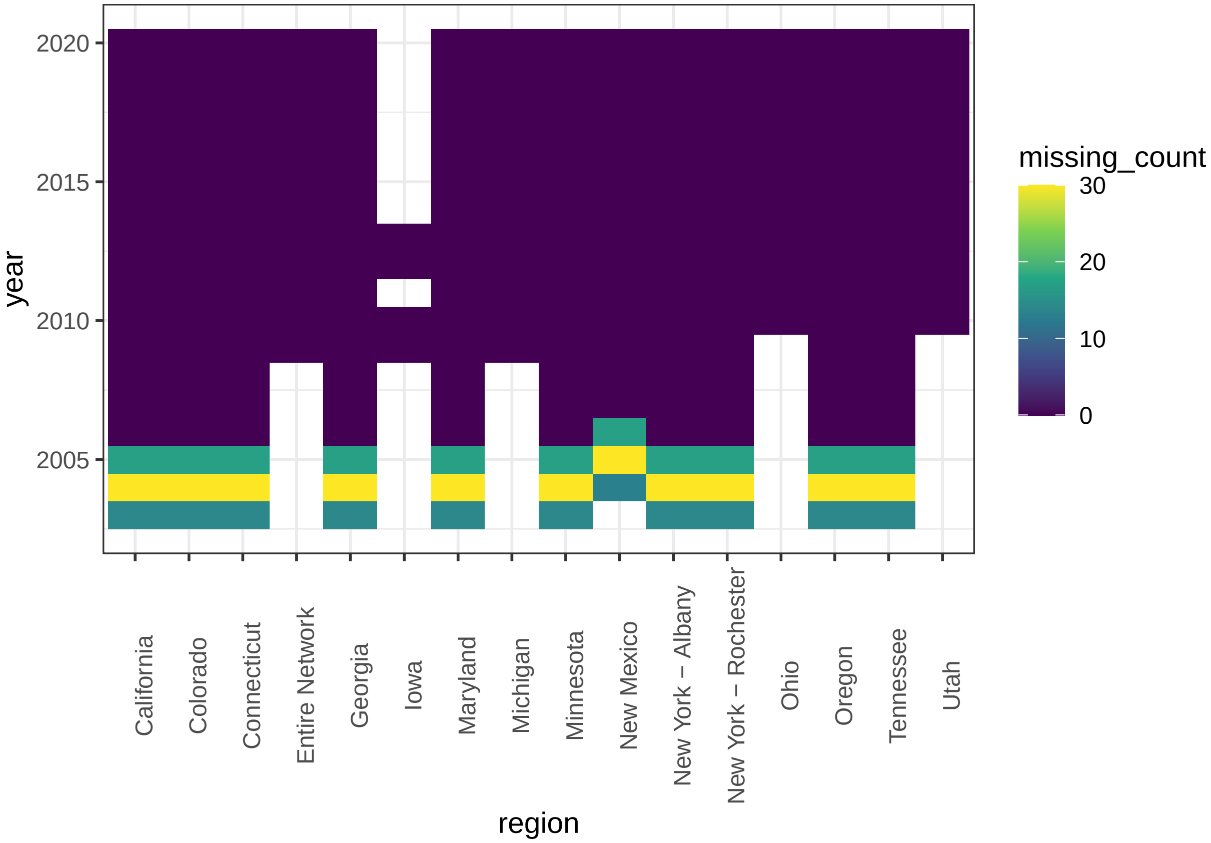
```
## # A tibble: 11 x 1
##    region
##    <chr>
##  1 California
##  2 Colorado
##  3 Connecticut
##  4 Georgia
##  5 Maryland
##  6 Minnesota
##  7 New Mexico
##  8 New York - Albany
##  9 New York - Rochester
## 10 Oregon
## 11 Tennessee
```

```
# Which seasons?
hosp_wide %>%
  dplyr::slice(which(rows_missing)) %>%
  dplyr::distinct(year)
```

```
## # A tibble: 4 x 1
##    year
##   <int>
## 1  2003
## 2  2004
## 3  2005
## 4  2006
```

```
# Heat map
hosp_wide %>%
  dplyr::mutate(missing_one = rows_missing) %>%
  dplyr::group_by(region, year) %>%
  dplyr::summarise(missing_count = sum(missing_one)) %>%
  ggplot() +
    geom_raster(mapping = aes(x = region, y = year, fill = missing_count)) +
    scale_fill_viridis_c() +
    theme_bw() +
  theme(axis.text.x = element_text(angle = 90))
```

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

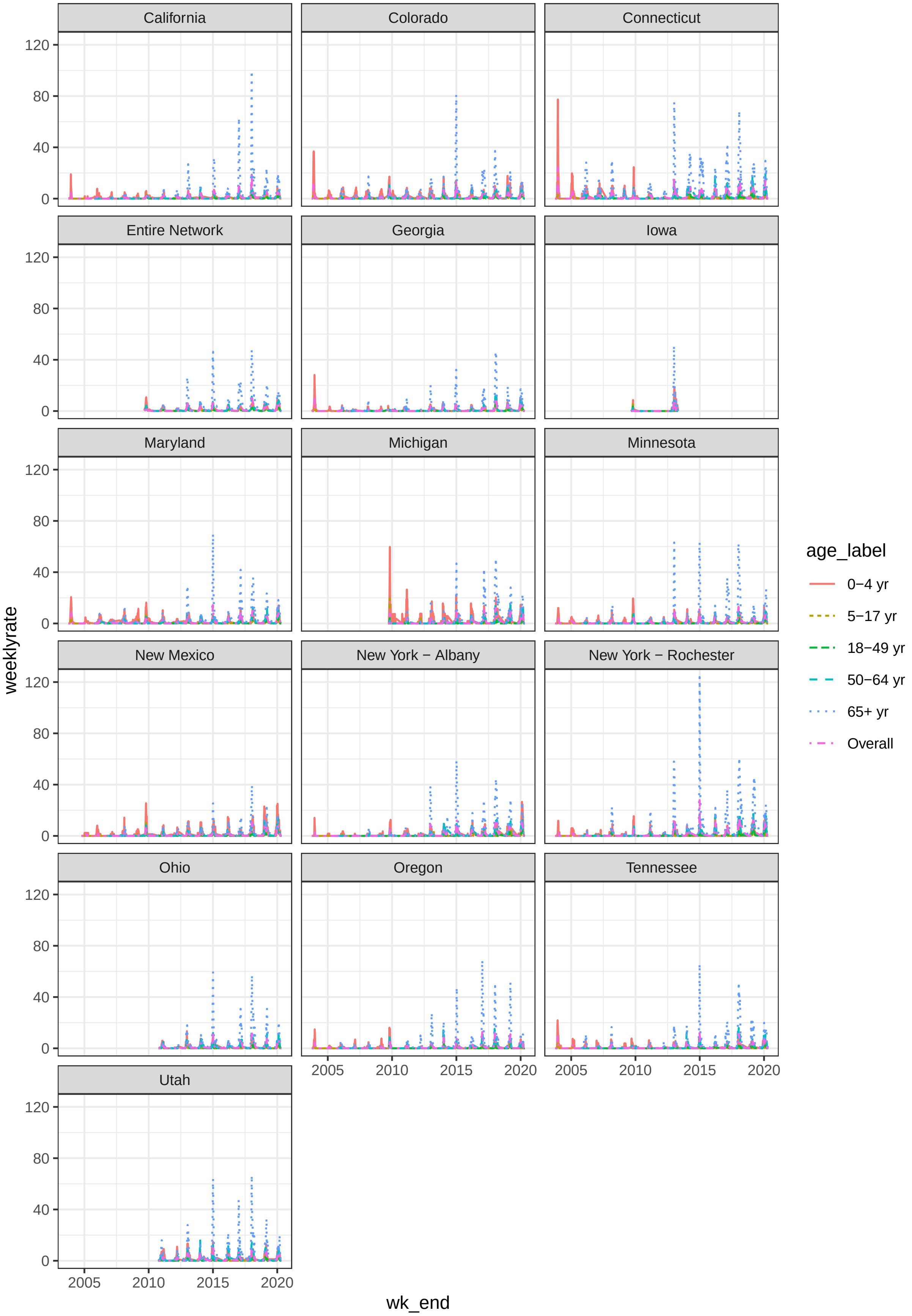


Questions: Was the full set of age categories introduced in 2005/2006? Should we be worried that the data for NM is off by a year?

Plots of data for all seasons and locations

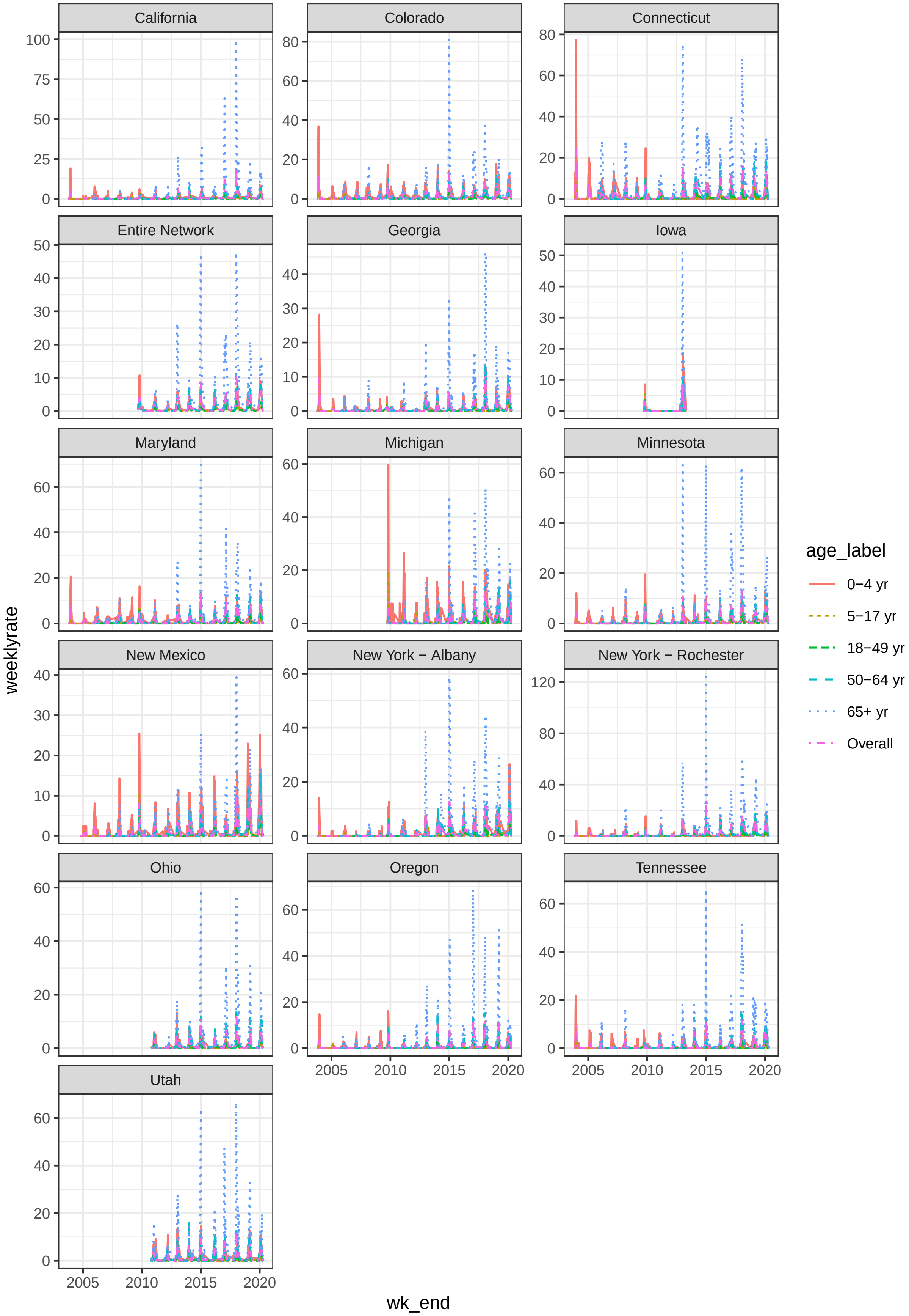
Same vertical axis scale:

```
ggplot(data = all_hosp,
  mapping = aes(x = wk_end, y = weeklyrate, color = age_label, linetype = age_label)) +
  geom_line() +
  facet_wrap( ~ region, , ncol = 3) +
  theme_bw(base_size = 10)
```


California
Colorado
Connecticut
Entire Network
Georgia
Iowa
Maryland
Michigan
Minnesota
New Mexico
New York − Albany
New York − Rochester
Ohio
Oregon
Tennessee
Utah
120
80
40
0
2005
2010
2015
2020
weeklyrate
wk_end
age_label
0−4 yr
5−17 yr
18−49 yr
50−64 yr
65+ yr
Overall

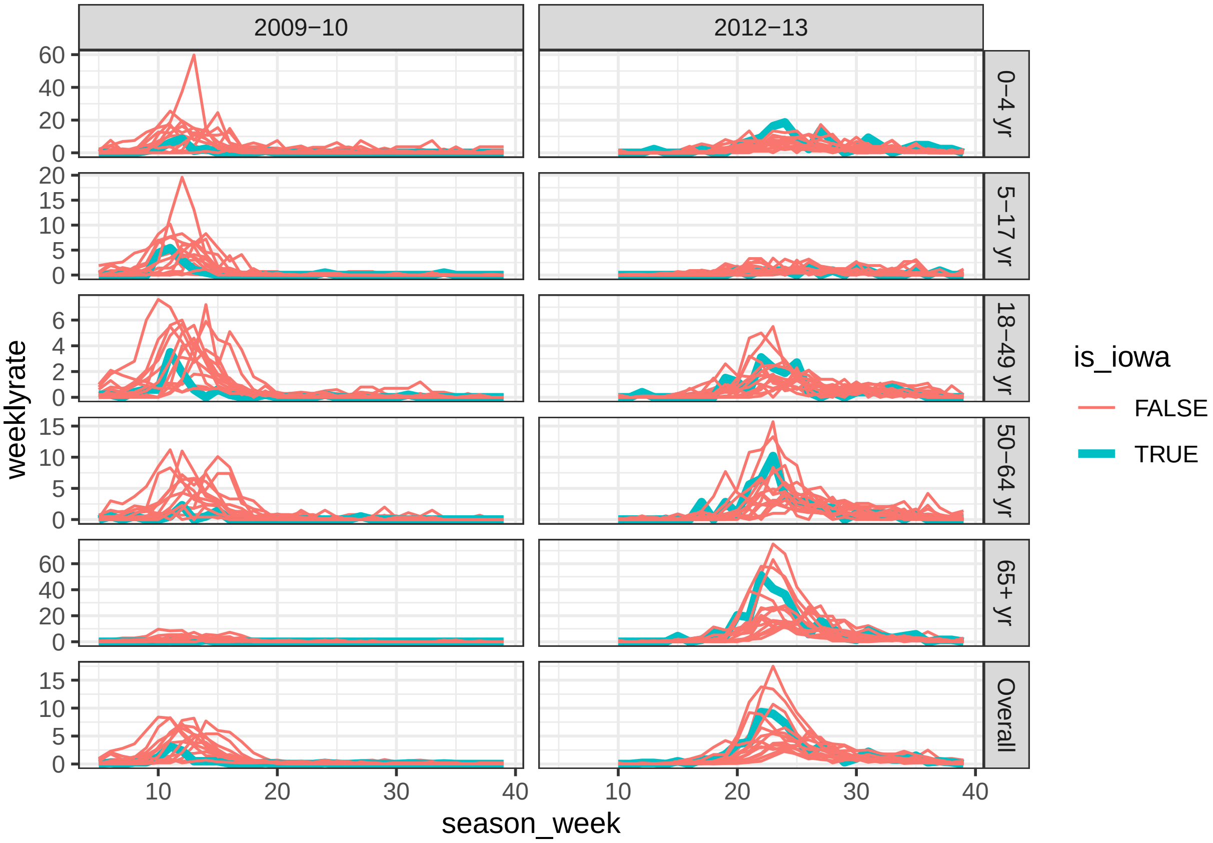
Different vertical axis scales:

```r
ggplot(data = all_hosp,
  mapping = aes(x = wk_end, y = weeklyrate, color = age_label, linetype = age_label)) +
  geom_line() +
  facet_wrap( ~ region, scales = "free_y", ncol = 3) +
  theme_bw(base_size = 10)
```


California
Colorado
Connecticut
Entire Network
Georgia
Iowa
Maryland
Michigan
Minnesota
New Mexico
New York − Albany
New York − Rochester
Ohio
Oregon
Tennessee
Utah
weeklyrate
wk_end
2005
2010
2015
2020
age_label
0−4 yr
5−17 yr
18−49 yr
50−64 yr
65+ yr
Overall

Comparing Iowa vs other regions for the 2 seasons with data for Iowa:

```r
all_hosp %>%
  dplyr::filter(sea_label %in% c("2009-10", "2012-13")) %>%
  dplyr::mutate(is_iowa = (region == "Iowa")) %>%
  ggplot() +
    geom_line(
      mapping = aes(x = season_week, y = weeklyrate, color = is_iowa, group = region, size = is_iowa)
    ) +
    scale_size_manual(values = c(0.5, 1.5)) +
    facet_grid(age_label ~ sea_label, scales = "free_y") +
    theme_bw()
```



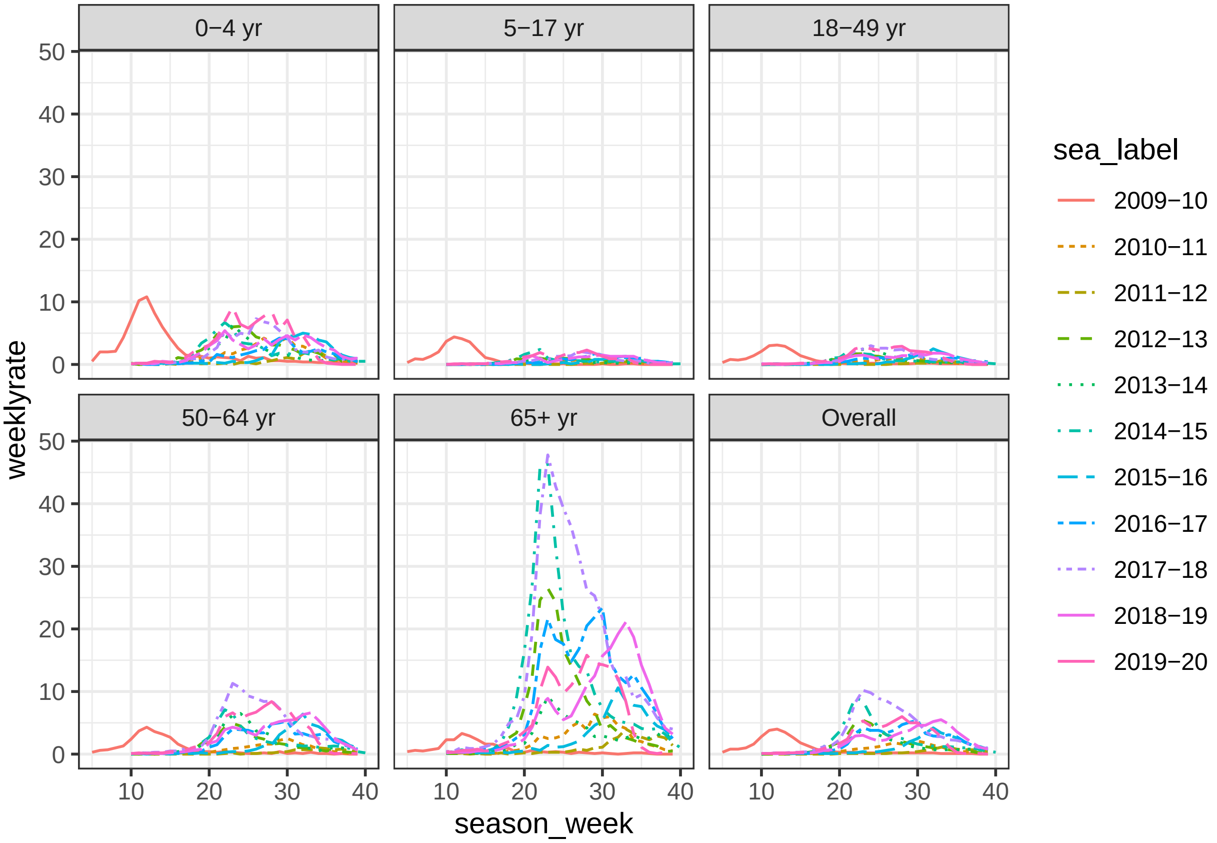
Question: For many locations, it appears that there is a systematic trend in relative age category breakdowns: before about 2010-2012 or so, there were relatively more cases in the 0-4 age group; after that, there were relatively more cases in the 65+ age group. Do you have a sense of whether these apparent trends are due to systematic changes in reporting or measurement at that time so that we would expect to continue seeing relatively higher prevalence in the elderly age category? Or is this just chance, and we could anticipate a change back to seeing higher prevalence in the younger age category? (Would it be reasonable to build into the model that we expect higher prevalence in the elderly category, or not?)

Question: Related to the above, it seems like in some locations there was systematically less reporting during the first few years (e.g. New York). Do you know if there really was less ili in those locations at that time, or if there have been changes in the measurement/reporting process since then?

Question: Iowa was on the list of locations for which we should generate forecasts, but there has not been any data for Iowa in recent years. Is that program just getting started up again?

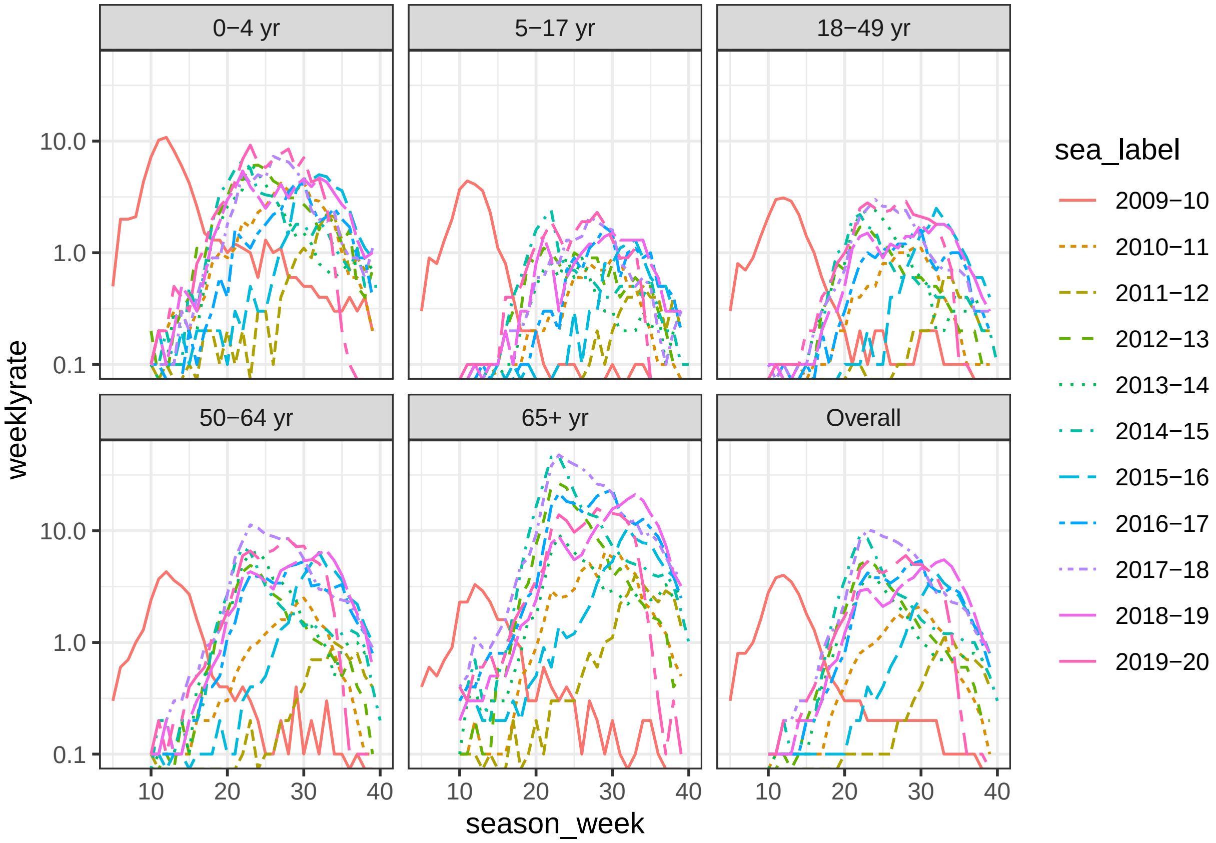
Other exploratory plots below; no further questions

```
ggplot(data = all_hosp %>%
  dplyr::filter(
    region == "Entire Network",
    !is.na(age_label)
  ),
  mapping = aes(x = season_week, y = weeklyrate, color = sea_label, linetype = sea_label)) +
  geom_line() +
  facet_wrap( ~ age_label) +
  theme_bw()
```



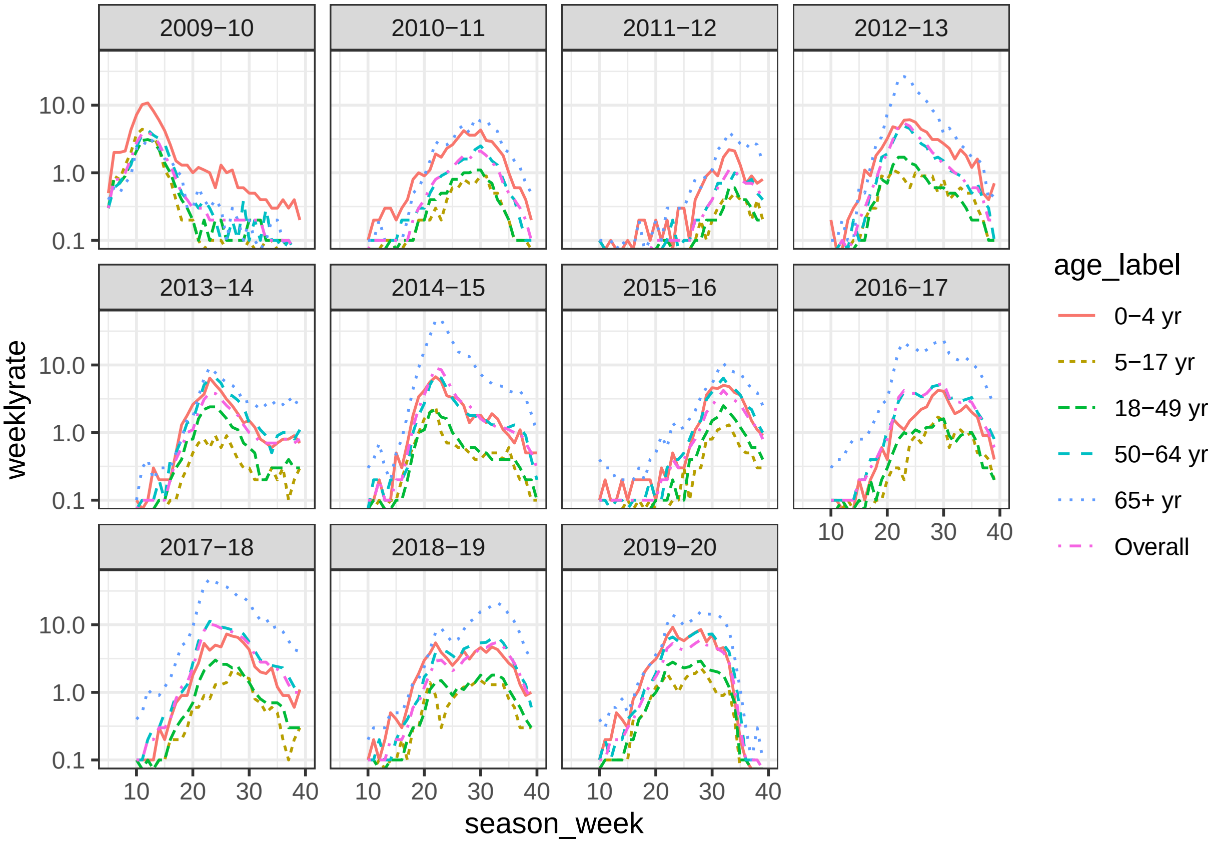
```
ggplot(data = all_hosp %>%
  dplyr::filter(
    region == "Entire Network",
    !is.na(age_label)
  ),
  mapping = aes(x = season_week, y = weeklyrate, color = sea_label, linetype = sea_label)) +
  geom_line() +
  scale_y_log10() +
  facet_wrap( ~ age_label) +
  theme_bw()
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```



```
ggplot(data = all_hosp %>%
  dplyr::filter(
    region == "Entire Network",
    !is.na(age_label)
  ),
  mapping = aes(x = season_week, y = weeklyrate, color = age_label, linetype = age_label)) +
  geom_line() +
  scale_y_log10() +
  facet_wrap( ~ sea_label) +
  theme_bw()
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```


2009−10
2010−11
2011−12
2012−13
2013−14
2014−15
2015−16
2016−17
2017−18
2018−19
2019−20
10.0
1.0
0.1
weeklyrate
10
20
30
40
season_week
age_label
0−4 yr
5−17 yr
18−49 yr
50−64 yr
65+ yr
Overall