Assessment 11 - Data Visualization Principles Part 3

Gabriele Mineo - Harvard Data Science Professional

Tile plot - measles and smallpox

The sample code given creates a tile plot showing the rate of measles cases per population. We are going to modify the tile plot to look at smallpox cases instead.

Instructions

- Modify the tile plot to show the rate of smallpox cases instead of measles cases.
- Exclude years in which cases were reported in fewer than 10 weeks from the plot.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

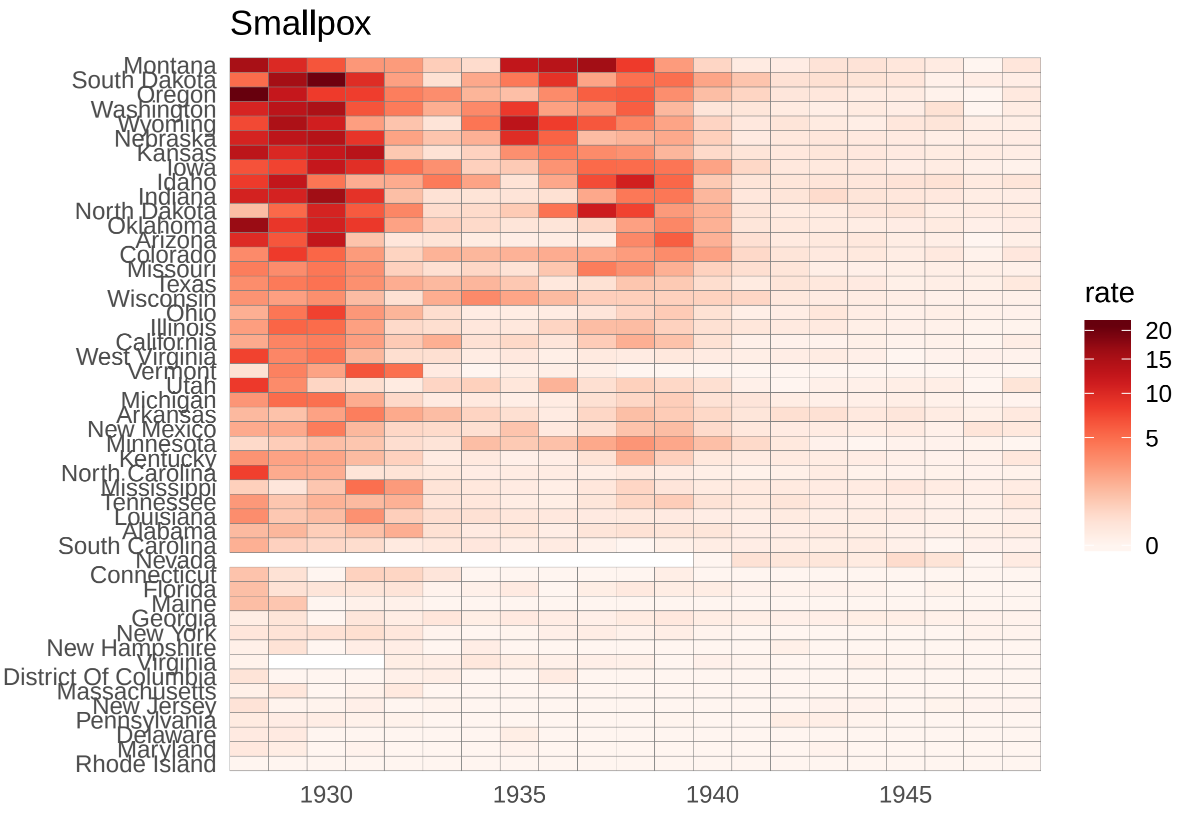
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(RColorBrewer)
library(dslabs)
data(us_contagious_diseases)

the_disease = "Smallpox"
dat <- us_contagious_diseases %>%
   filter(!state%in%c("Hawaii","Alaska") & disease == the_disease & weeks_reporting>=10) %>%
   mutate(rate = count / population * 10000) %>%
   mutate(state = reorder(state, rate))

dat %>% ggplot(aes(year, state, fill = rate)) +
  geom_tile(color = "grey50") +
  scale_x_continuous(expand=c(0,0)) +
  scale_fill_gradientn(colors = brewer.pal(9, "Reds"), trans = "sqrt") +
  theme_minimal() +
  theme(panel.grid = element_blank()) +
  ggtitle(the_disease) +
  ylab("") +
  xlab("")
```



Time series plot - measles and smallpox

The sample code given creates a time series plot showing the rate of measles cases per population by state. We are going to again modify this plot to look at smallpox cases instead.

Instructions

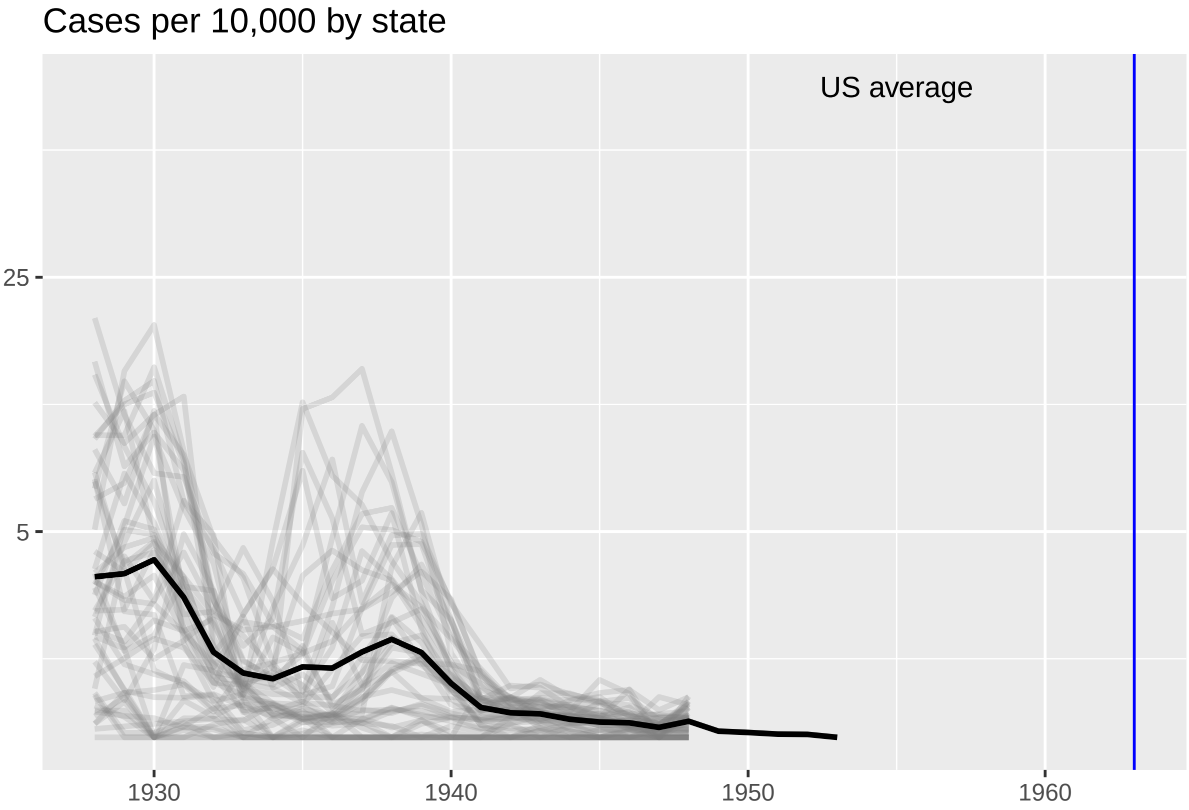
- Modify the sample code for the time series plot to plot data for smallpox instead of for measles.
- Once again, restrict the plot to years in which cases were reported in at least 10 weeks.

```
library(dplyr)
library(ggplot2)
library(dslabs)
library(RColorBrewer)
data(us_contagious_diseases)

the_disease = "Smallpox"
dat <- us_contagious_diseases %>%
   filter(!state%in%c("Hawaii","Alaska") & disease == the_disease & weeks_reporting>=10) %>%
   mutate(rate = count / population * 10000) %>%
   mutate(state = reorder(state, rate))

avg <- us_contagious_diseases %>%
  filter(disease==the_disease) %>% group_by(year) %>%
  summarize(us_rate = sum(count, na.rm=TRUE)/sum(population, na.rm=TRUE)*10000)

dat %>% ggplot() +
  geom_line(aes(year, rate, group = state),  color = "grey50",
            show.legend = FALSE, alpha = 0.2, size = 1) +
  geom_line(mapping = aes(year, us_rate),  data = avg, size = 1, color = "black") +
  scale_y_continuous(trans = "sqrt", breaks = c(5,25,125,300)) +
  ggtitle("Cases per 10,000 by state") +
  xlab("") +
  ylab("") +
  geom_text(data = data.frame(x=1955, y=50), mapping = aes(x, y, label="US average"), color="black") +
  geom_vline(xintercept=1963, col = "blue")
```



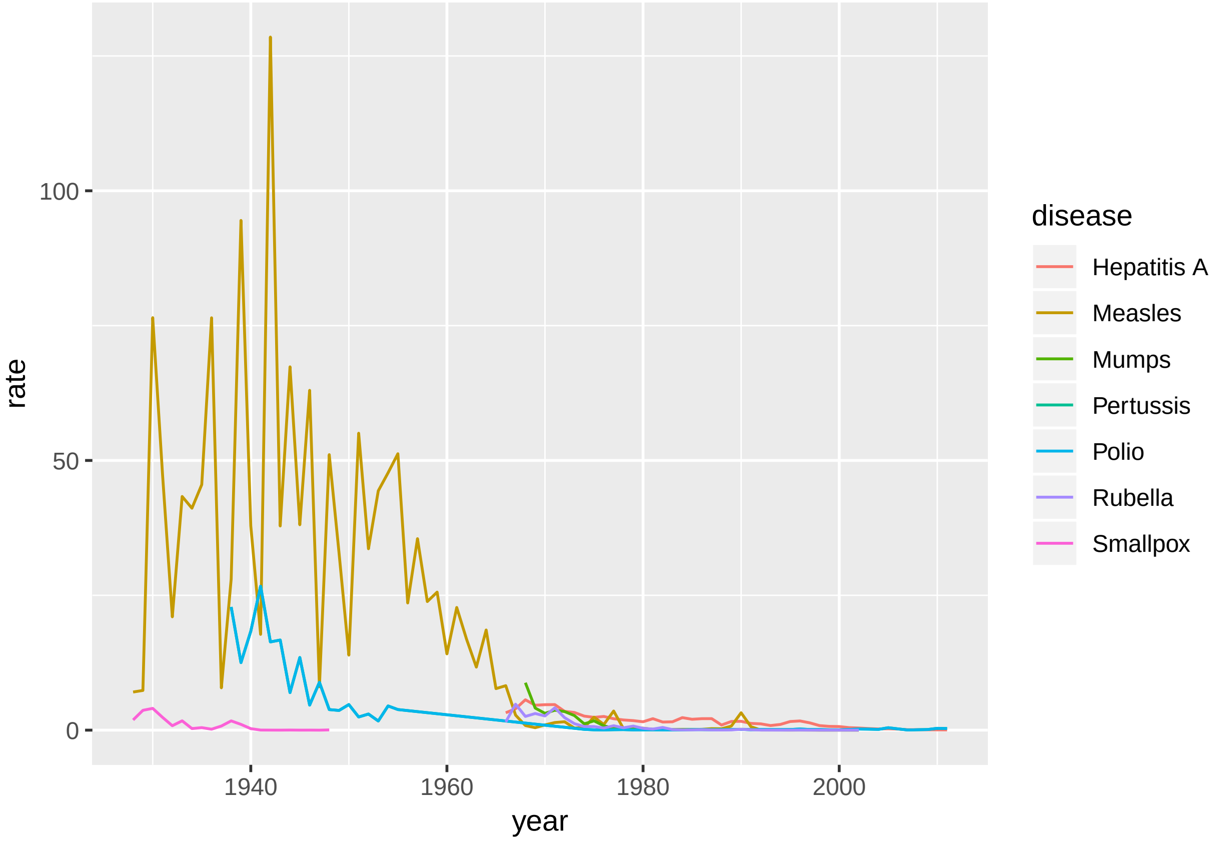
**Time series plot - all diseases in California Now we are going to look at the rates of all diseases in one state. Again, you will be modifying the sample code to produce the desired plot.

Instructions

- For the state of California, make a time series plot showing rates for all diseases.
- Include only years with 10 or more weeks reporting.
- Use a different color for each disease.

```
library(dplyr)
library(ggplot2)
library(dslabs)
library(RColorBrewer)
data(us_contagious_diseases)

us_contagious_diseases %>% filter(state=="California" & weeks_reporting>=10) %>%
  group_by(year, disease) %>%
  summarize(rate = sum(count)/sum(population)*10000) %>%
  ggplot(aes(year, rate, color=disease)) +
  geom_line()
```



Time series plot - all diseases in the United States

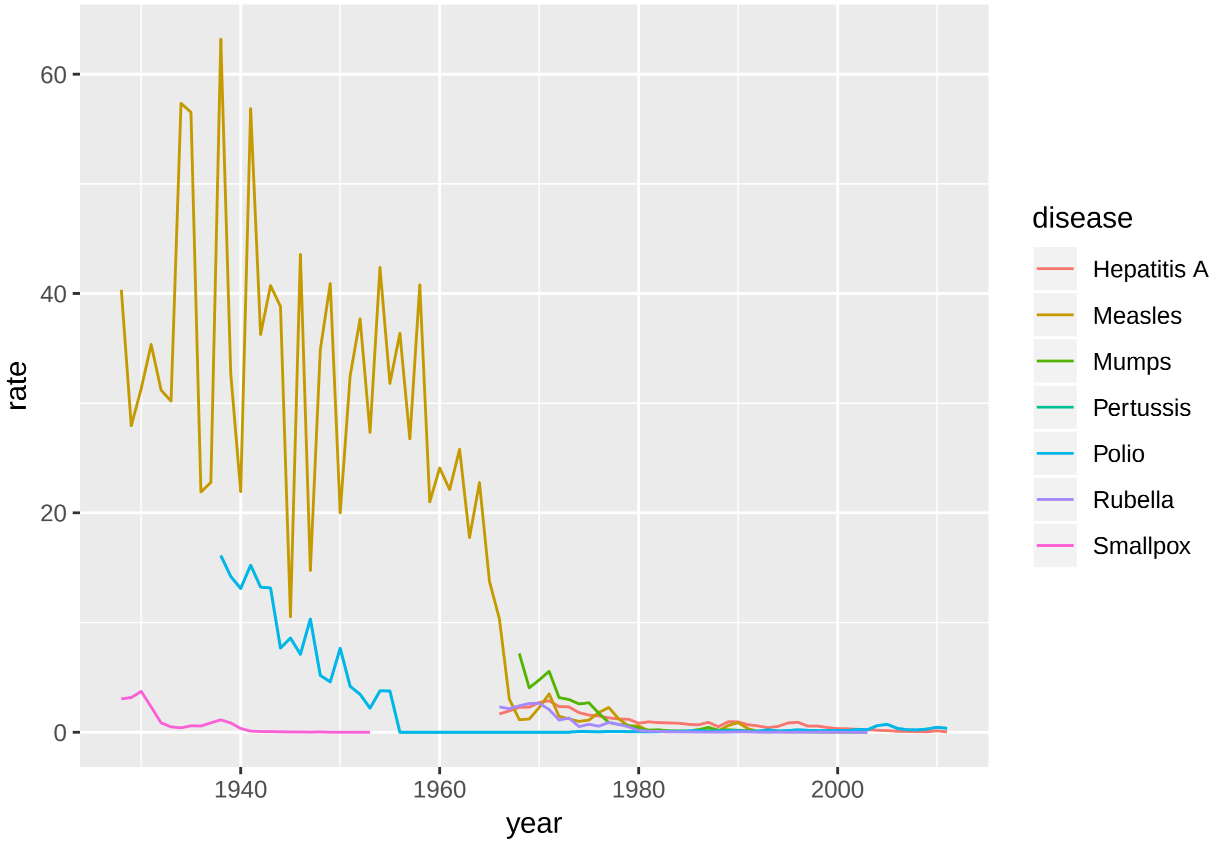
Now we are going to make a time series plot for the rates of all diseases in the United States. For this exercise, we have provided less sample code - you can take a look at the previous exercise to get you started.

Instructions

- Compute the US rate by using `summarize` to sum over states.
- The US rate for each disease will be the total number of cases divided by the total population.
- Remember to convert to cases per 10,000.
- You will need to filter for `!is.na(population)` to get all the data.
- Plot each disease in a different color.

```
library(dplyr)
library(ggplot2)
library(dslabs)
library(RColorBrewer)
data(us_contagious_diseases)

us_contagious_diseases %>% filter(!is.na(population)) %>%
  group_by(year, disease) %>%
  summarize(rate = sum(count)/sum(population)*10000) %>%
  ggplot(aes(year, rate, color=disease)) +
  geom_line()
```


rate
year
disease
Hepatitis A
Measles
Mumps
Pertussis
Polio
Rubella
Smallpox