Week 4 DS Recitation: Data Visualization Part 1

SIBDS 2024 @ Columbia

20 June, 2024

Here is a useful link of ggplot cheatsheets: https://github.com/rstudio/cheatsheets/blob/main/data-visualization.pdf

Getting Started

1. Open the R project `week4_DS_recitation`
2. Create a `data` folder. Download the NOAA data and put it in the data folder.
3. Put the `week4_DS_recitation_data_viz_1.Rmd` file into the same folder of the R project.

Visualization

We are going to use NY NOAA data.

```
ny_noaa = read_csv("data/nynoaadat.csv")
```

Here, I did some basic cleaning for you:

```
noaa_df = ny_noaa %>%
  separate(date, into = c("year","month","day"),sep = "-") %>%
  mutate(
    year = as.factor(year),
    month = as.integer(month),
    month = (month.name[month]),
    day = as.factor(day),
    prcp = prcp / 10,
    tmax = as.numeric(tmax),
    tmin = as.numeric(tmin),
    tmax = tmax / 10,
    tmin = tmin / 10
  )

# Here is a table shows the most common observations of snowfall
noaa_df %>%
  count(snow, name = "n_of_observations") %>%
  drop_na(snow) %>%
  arrange(desc(n_of_observations)) %>%
  head(n = 5)
```

```
## # A tibble: 5 x 2
##    snow n_of_observations
##   <dbl>             <int>
## 1     0           2008508
## 2    25             31022
## 3    13             23095
## 4    51             18274
## 5    76             10173
```

Tasks:

Please complete the following tasks:

1. Make a two-panel plot showing the average max temperature in January and in July in each station across years. Is there any observable / interpretable structure? Any outliers?

```
# You may start with this processed data
noaa_df %>%
  #focus on Jan and Jul
  filter(month %in% c("January", "July")) %>%
  group_by(year, month, id) %>%
  summarise(mean_of_tmax = mean(tmax,na.rm = TRUE))
```

```
## # A tibble: 14,111 x 4
## # Groups:   year, month [60]
##    year  month   id          mean_of_tmax
##    <fct> <chr>   <chr>              <dbl>
##  1 1981  January USC00300023         -3.17
##  2 1981  January USC00300055         -4.28
##  3 1981  January USC00300063         -4.04
##  4 1981  January USC00300085         -3.95
##  5 1981  January USC00300093         -4.23
##  6 1981  January USC00300183         -3.00
##  7 1981  January USC00300220         -4.56
##  8 1981  January USC00300254        NaN
##  9 1981  January USC00300331         -4.07
## 10 1981  January USC00300343        NaN
## # i 14,101 more rows
```

2. Make a two-panel plot showing:

(i) `tmax` vs `tmin` for the full dataset (note that a scatterplot may not be the best option); and
(ii) make a plot showing the distribution of snowfall values greater than 0 and less than 100 separately by year.