ds_final_data_cleaning

Chuyuan XU

2025-12-03

import the county and state data

```
state_r = read_sf(str_c(import,"/tl_2020_us_state")) |>
  janitor::clean_names() |>
  as.data.frame() |>
  filter(stusps == "NY") |>
  select(statefp, stusps, name)

stsfp_ny = state_r |>
  distinct(statefp) |>
  pull()

county_ny = read_sf(str_c(import,"/tl_2024_us_county")) |>
  janitor::clean_names() |>
  filter(statefp %in% stsfp_ny) |>
  select(statefp, countyfp, fips = geoid, name, namelsad)

fips_ny = county_ny |>
  distinct(fips) |>
  pull()
```

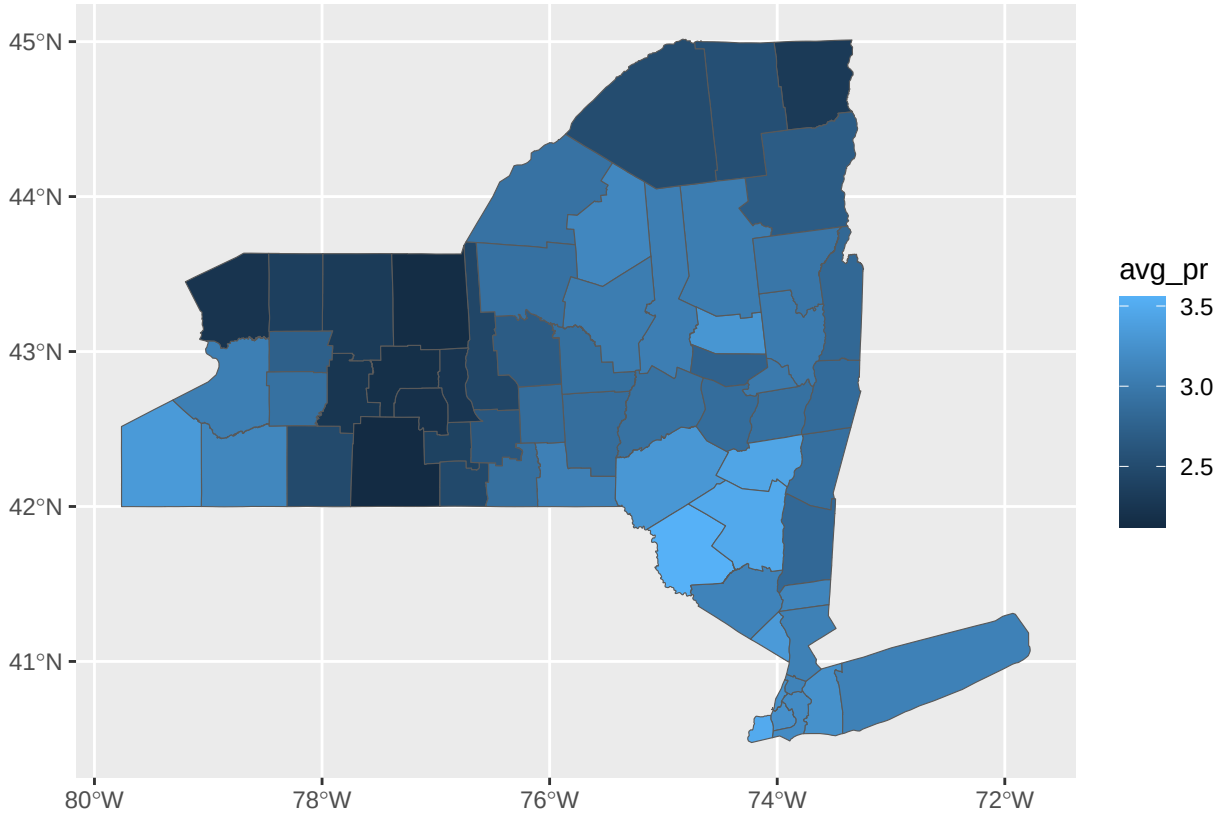
import all NY meteor data and filter county in NY state

```
meteor_path = str_c(import, "/meteorology")
meteor_files = list.files(meteor_path)

meteor_16to20 = read_parquet(str_c(meteor_path, '/', meteor_files[1])) |>
  janitor::clean_names()

for (i in 2:length(meteor_files)){
  meteor_temp = read_parquet(str_c(meteor_path, '/', meteor_files[i])) |>
    janitor::clean_names()

  meteor_16to20 = bind_rows(meteor_temp, meteor_16to20)
}

meteor_16to20 = meteor_16to20 |>
  filter(county %in% fips_ny) |>
  mutate(
    fips = county
  ) |>
  select(county, fips, date, everything()) |>
```

```
  arrange(county, fips, date)

# still too large (more than 50mb)
meteor_16to20_sf = meteor_16to20 |>
  left_join(county_ny, by = "fips") |>
  st_as_sf()
```

try some visualization to test the dataset

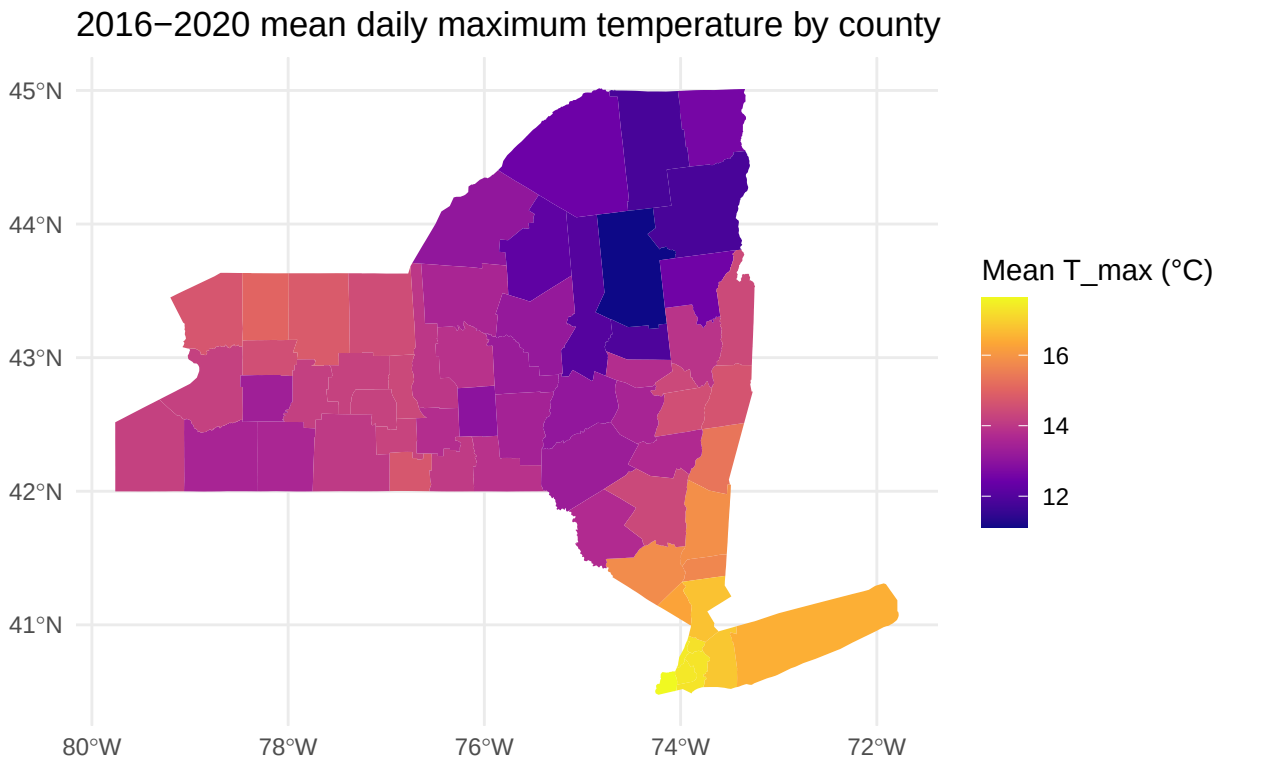
```
meteor_16to20_sf |>
  filter(year(date) == 2020) |>
  group_by(fips) |>
  summarise(
    avg_pr = mean(pr)
  ) |>
  ggplot() +
  geom_sf(aes(fill = avg_pr))
```



I think it works?

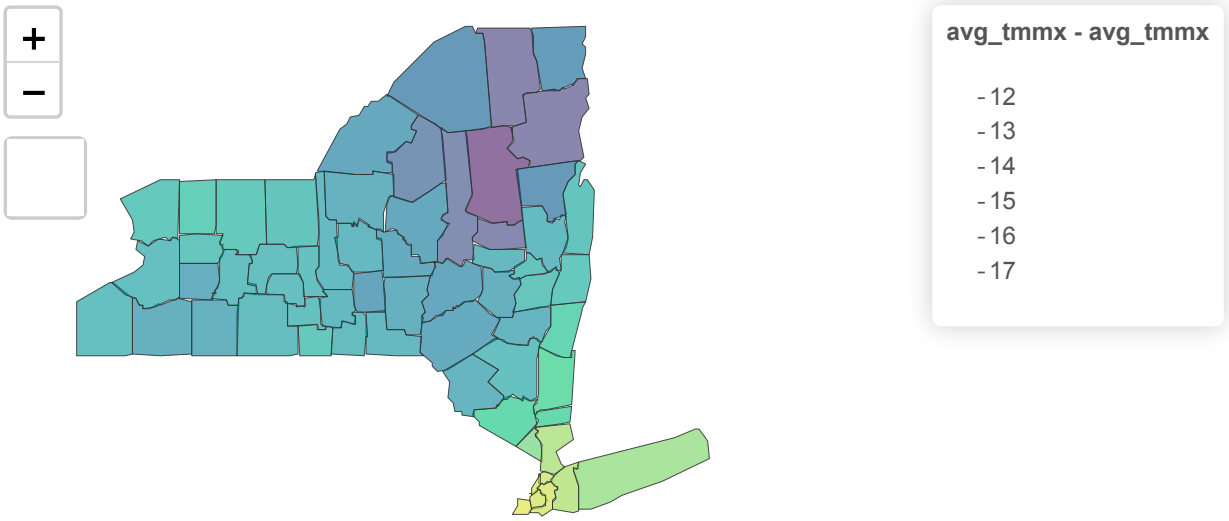
```
avg_tmmx =  meteor_16to20 |>
  group_by(fips) |>
  summarise(
    avg_tmmx = mean(tmmx) - 273.15
  ) |>
  mutate(
    avg_tmmx = round(avg_tmmx, 3)
  ) |>
```

```
  left_join(county_ny, join_by(fips)) |>
  st_as_sf()

  ggplot(avg_tmmx) +
  geom_sf(aes(fill = avg_tmmx), color = NA) +
  scale_fill_viridis_c(
    name = "Mean T_max (°C)",
    option = "plasma",
    na.value = "grey90"
  ) +
  labs(
    title = "2016-2020 mean daily maximum temperature by county",
    x = NULL,
    y = NULL
  ) +
  theme_minimal()
```




```
mapview(avg_tmmx, zcol = "avg_tmmx")
```

```
## file:////private/var/folders/hv/8w8v7z815jvbpzwwdfgpnfwr0000gn/T/RtmpsrVZus/file469c6b7ad9ca/widget4
```


avg_tmmx - avg_tmmx
- 12
- 13
- 14
- 15
- 16
- 17
100 km
100 mi
avg_tmmx - avg_tmmx
Leaflet | © OpenStreetMap contributors © CARTO

output tht data

```
dsn_meteor = str_c(output, "/meteor_2016to20_NY.csv")

write_csv(
  meteor_16to20,
  dsn_meteor,
  na = "NA",
  append = FALSE
)

dsn_geo = str_c(output, "/NYcounty_fips")

st_write(
  county_ny,
  dsn_geo,
  driver = "ESRI Shapefile",
  append = FALSE
)
```

Check

```
file_check = read_csv(
  '/Users/chuyuanxu/Desktop/Columbia/25_fall/BIST8105_DS I/final project/df_gen/data/meteor_2016to20_NY
```

```
## Rows: 20 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (12): county, fips, sph, vpd, tmmn, tmmx, pr, rmin, rmax, srad, vs, th
## dttm  (1): date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```