ds_final_data_cleaning

Chuyuan XU

2025-11-21

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()
```

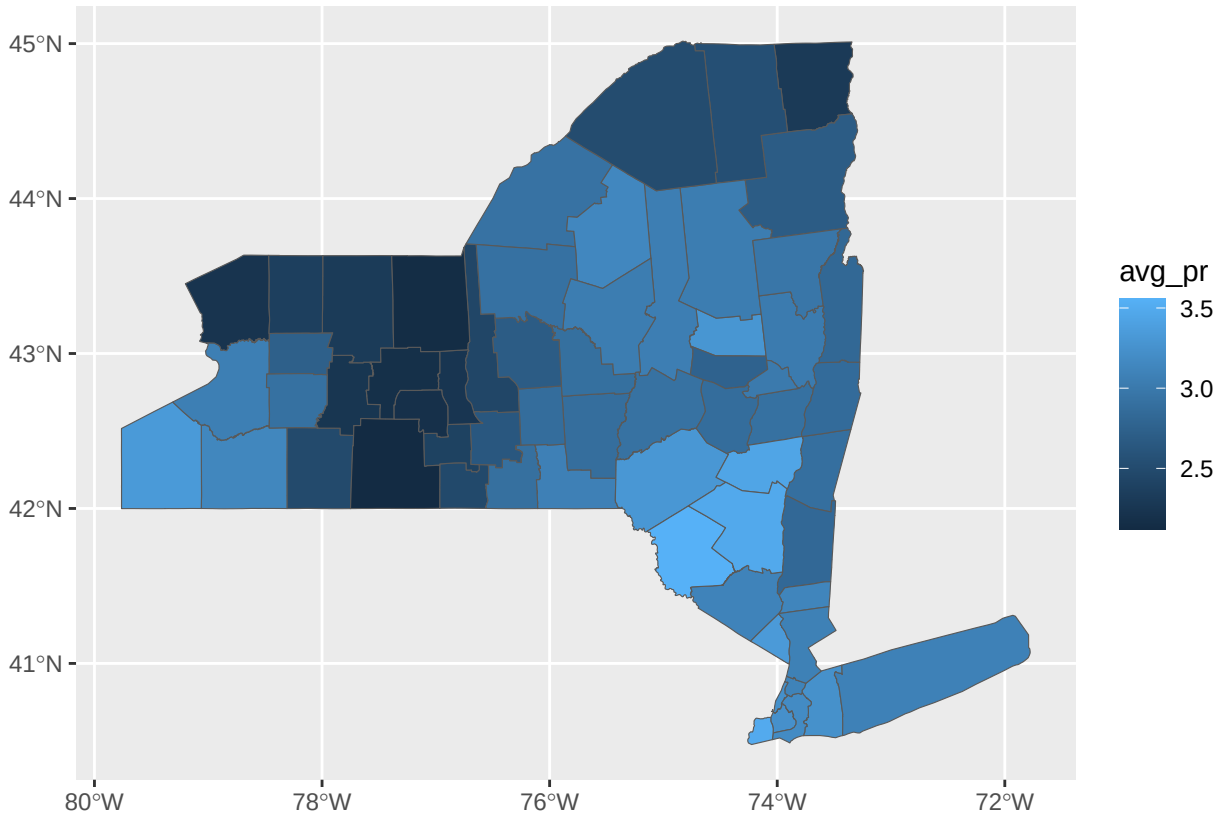
import all NY meteor data and filter county in NY state

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

meteor_1620 = 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_1620 = bind_rows(meteor_temp, meteor_1620)
}

meteor_1620 = meteor_1620 |>
  filter(county %in% fips_ny) |>
  mutate(
    fips = county
  )
```

```
meteor_1620_sf = meteor_1620 |>
  left_join(county_ny, by = "fips") |>
  st_as_sf()
```

try some visualization to test the dataset

```
meteor_1620_sf |>
  filter(year(date) == 2020) |>
  group_by(fips) |>
  summarise(
    avg_pr = mean(pr)
  ) |>

  ggplot() +
  geom_sf(aes(fill = avg_pr))
```



I think it works?

output tht data

```
dsn_meteor = str_c(output, "/meteor_2016-2020_NY state.csv")

write_csv(
  meteor_1620,
  dsn_meteor,
  na = "NA"
)
```

```
dsn_geo = str_c(output, "/NY county_fips")

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

Raw Dataset Variable Introduction

In the meteorology datasets from 2016-2020, each dataset includes the following variables:
`county`: Federal Information Processing Series (FIPS), the id of the counties;
`date`: date of the estimated meteorological variables;
`sph`: near-surface specific humidity (Mass fraction);
`vpd`: mean vapor pressure deficit (kPa);
`tmmn`: minimum near-surface air temperature (Kelvin);
`tmmx`: maximum near-surface air temperature (Kelvin);
`pr`: precipitation (mm, daily total);
`rmin`: minimum near-surface relative humidity (%);
`rmax`: maximum near-surface relative humidity (%);
`srad`: surface downwelling solar radiation (W/mˆ2);
`vs`: wind speed at 10m (m/s);
`th`: wind direction at 10m (degree).

Key variables in `tl_2024_county` are:
`GEOID`: Current county identifier; a concatenation of current state FIPS code and county FIPS code;
`STUSPS`: Current United States Postal Service (USPS) state abbreviation;
`NAME`: Current state name;
`NAMELSAD`: Current name and the translated LSAD code for county;
`geometry`: the geological boundaries in the geometry of multipolygon.

Key variables in `tl_2020_state` are:
`STATEFP`: State FIPS code;
`STUSPS`: Current United States Postal Service (USPS) state abbreviation;
`NAME`: Current county name.