EDA on Daily Air Quality Data

Gustavo Garcia-Franceschini

2023-12-05

UHF Air Quality

```
df_air = read_csv(here::here("data/cleaned_data/uhf_airquality.csv"))
```

```
## Rows: 6300 Columns: 16
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): name, measure, measure_info, geo_type_name, geo_place_name, time_pe...
## dbl (9): unique_id, indicator_id, geo_join_id, data_value, id, x_center, y_c...
##
## 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.
```

```
zip_shapes = read_sf(dsn = here::here("raw_shapes/"),
                     layer = 'tl_2019_us_zcta510') %>%
  rename(zip = ZCTA5CE10)

air_shapes = read_sf(dsn = here::here("raw_shapes/"), layer = "UHF42") %>%
  filter(id != "0") %>%
  st_transform(crs= st_crs(zip_shapes)) %>%
  rename(geo_join_id = id) %>%
  mutate(geo_join_id = as.numeric(geo_join_id))

df_air = df_air %>%
  left_join(air_shapes) %>%
   mutate(x_center = map_dbl(geometry, ~st_point_on_surface(.x)[[1]]),
         y_center = map_dbl(geometry, ~st_point_on_surface(.x)[[2]]),
         intplon10 = x_center, intplat10 = y_center)
```

```
## Joining with `by = join_by(geo_join_id)`
```

Alt Air Quality

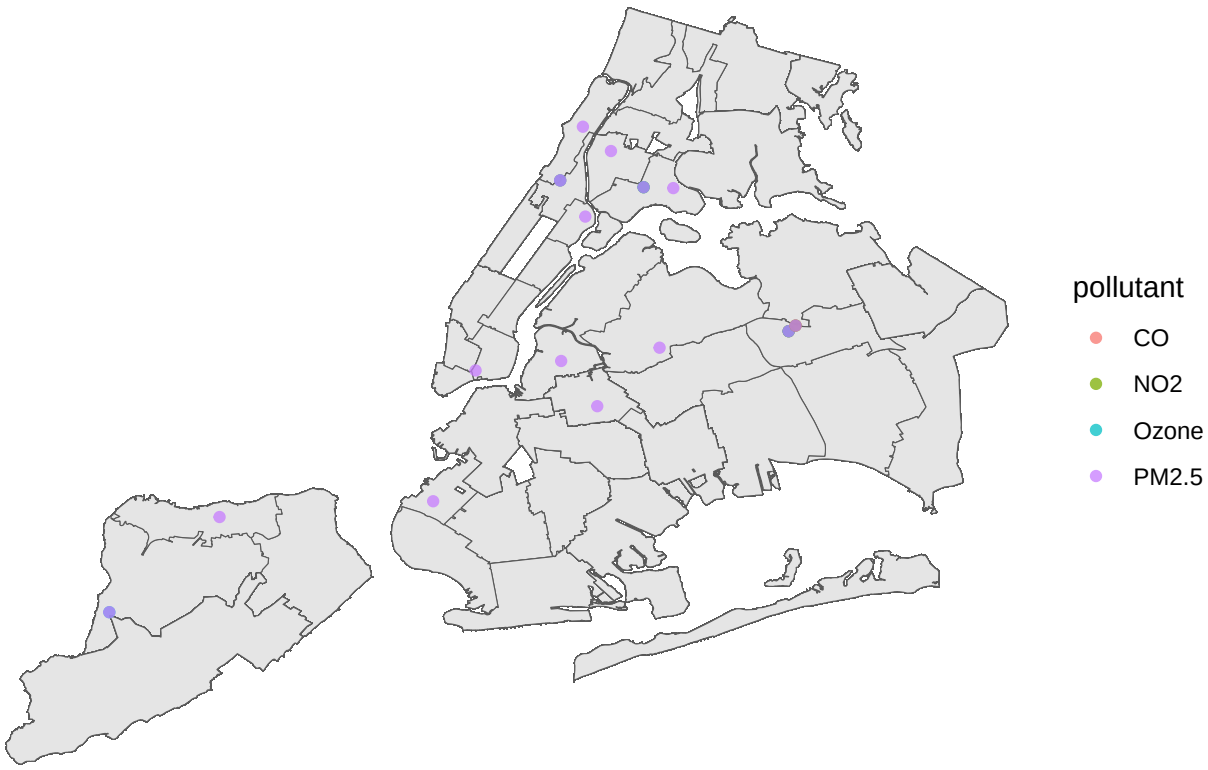
```
daily_air = read_csv(here::here("data/cleaned_data/alt_air_data.csv"))
```

```
## Rows: 23685 Columns: 26
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (11): date, source, units, site_name, aqs_parameter_desc, cbsa_name, sta...
## dbl (15): site_id, poc, daily_aqi_value, daily_obs_count, percent_complete, ...
##
## 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.
```

Visualization

```
daily_air %>%
  group_by(site_name, pollutant) %>%
  summarize(x = mean(x), y = mean(y))  %>%
  ggplot() + geom_sf(data = df_air, aes(geometry = geometry)) +
  geom_point(aes(x = x, y = y, color = pollutant), alpha = 0.75) + theme_void()
```

```
## `summarise()` has grouped output by 'site_name'. You can override using the
## `.groups` argument.
```



Points in each neighborhood

```
daily_air = daily_air %>%
  st_as_sf(coords = c("x", "y"), crs = st_crs(zip_shapes))
```

```
geo_joins_sites = st_join(daily_air, air_shapes, join = st_within)  %>%
  select(zip_code, site_name, geo_join_id, pollutant) %>%
  group_by(zip_code, site_name, pollutant) %>%
  summarize(geo_join_id = mean(geo_join_id))
```

```
## `summarise()` has grouped output by 'zip_code', 'site_name'. You can override
## using the `.groups` argument.
```

```
geo_join_neighborhoods = df_air %>%
  group_by(geo_place_name) %>%
  summarize(geo_join_id = mean(geo_join_id))
```

```
left_join(geo_joins_sites, geo_join_neighborhoods) %>%
  select(zip_code, site_name, pollutant, geo_place_name) %>%
  write_csv(here::here("data/cleaned_data/pollutant_per_neighborhood.csv"))
```

```
## Joining with `by = join_by(geo_join_id)`
```