hw2

Andi Liao

January 19, 2019

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.2.1     v forcats 0.3.0

## -- Conflicts ------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(ggplot2)

data = read_csv('environment_index_province.csv', skip = 1)
```

```
## Warning: Missing column names filled in: 'X1' [1]

## Parsed with column specification:
## cols(
##   .default = col_double(),
##   X1 = col_character(),
##   region = col_character()
## )

## See spec(...) for full column specifications.
```

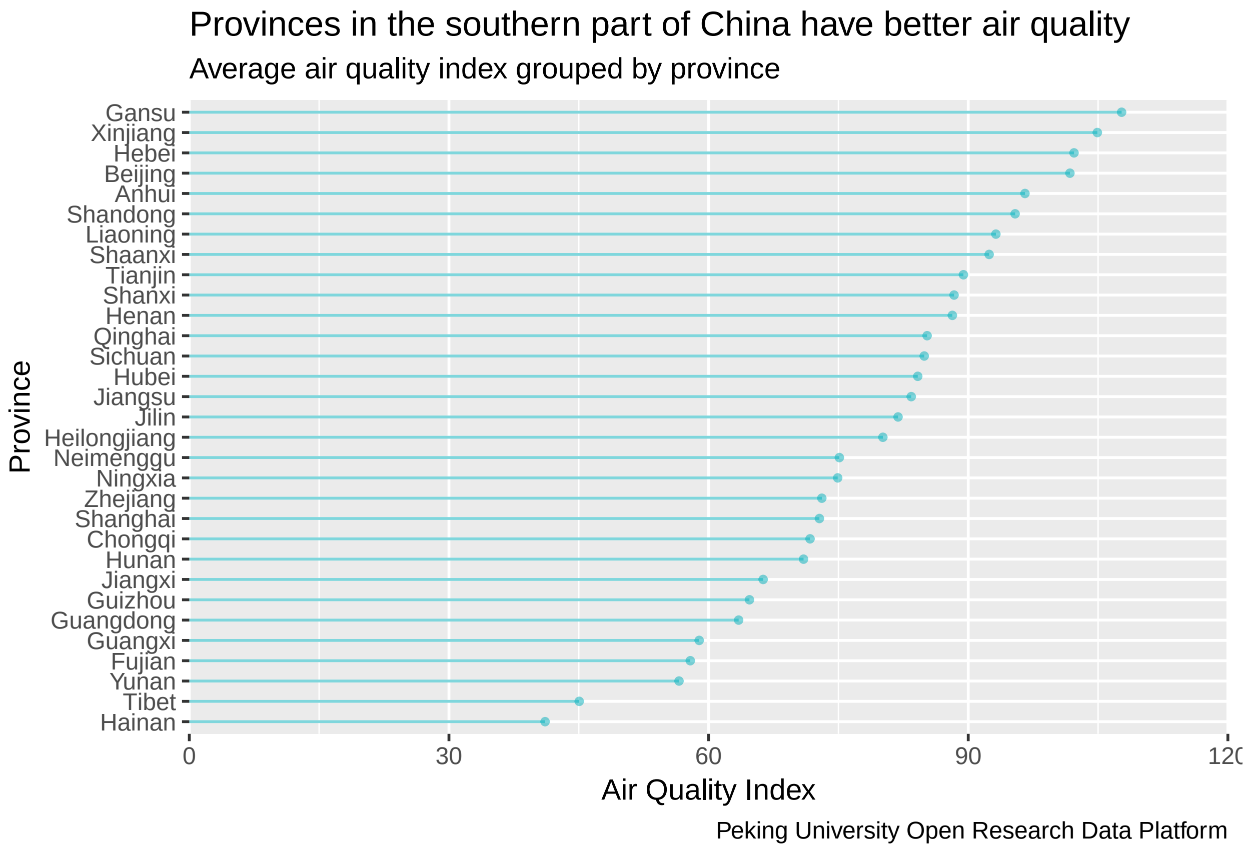
```r
data = data[, -1]

data_long = gather(data, key = "metrics", value = "value", water_2011:aqi_2014)
data_long = cbind(data_long, year = 0)
data_long = cbind(data_long, measurement = 0)

for(i in 1:dim(data_long)[1]){
  tmp = unlist(strsplit(as.character(data_long[i, "metrics"]),"_"))
  data_long[i, 5] = tmp[2]
  data_long[i, 6] = tmp[1]
}
```

```r
data_long %>%
  select(-("metrics")) %>%
  filter(measurement == "aqi") %>%
  spread(measurement, value) %>%
  group_by(region) %>%
  summarise(aqi = mean(aqi)) %>%
  rownames_to_column("province") %>%
  arrange(aqi) %>%
  mutate(region = factor(region, levels = .$region)) %>%

  ggplot(aes(x = aqi, y = region)) +
  geom_segment(aes(x = 0, xend = aqi, y = region, yend = region), color = "#00AFBB", alpha = 0.5) +
  geom_point(color = "#00AFBB", size = 1, alpha = 0.5 ) +
  scale_x_continuous(expand = c(0, 0), limits = c(0, 120)) +
  labs(title = "Provinces in the southern part of China have better air quality",
       subtitle = "Average air quality index grouped by province",
       caption = "Peking University Open Research Data Platform",
       x = "Air Quality Index",
       y = "Province")
```

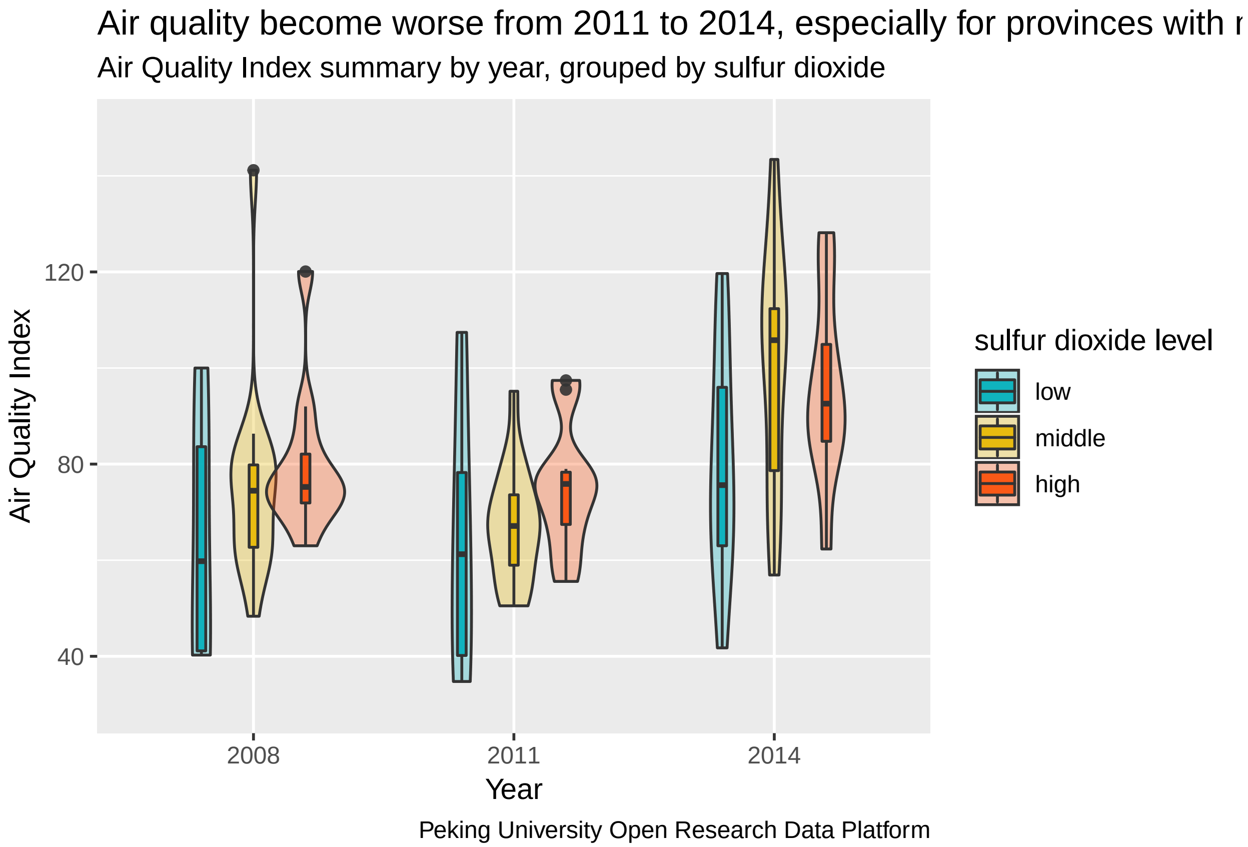

Provinces in the southern part of China have better air quality

Average air quality index grouped by province

Peking University Open Research Data Platform

```
data_long %>%
  select(-("metrics")) %>%
  filter(measurement == "aqi" | measurement == "sulfur") %>%
  spread(measurement, value) %>%
  mutate(category = cut(sulfur, breaks=c(0, 100, 200, Inf),labels = c("low","middle", "high"))) %>%
  group_by(category) %>%

  ggplot(aes(x = year, y = aqi, fill = category)) +
  geom_violin(position = position_dodge(width = 0.6), width = 0.9, alpha = 0.3) +
  geom_boxplot(position = position_dodge(width = 0.6), width = 0.1, alpha = 0.9) +
  scale_fill_manual(values = c("#00AFBB", "#E7B800", "#FC4E07")) +
  ylim(30, 150) +
  labs(
    title = "Air quality become worse from 2011 to 2014, especially for provinces with middle and high
    subtitle = "Air Quality Index summary by year, grouped by sulfur dioxide",
    caption = "Peking University Open Research Data Platform",
    x = "Year",
    y = "Air Quality Index",
    fill = "sulfur dioxide level"
  )
```



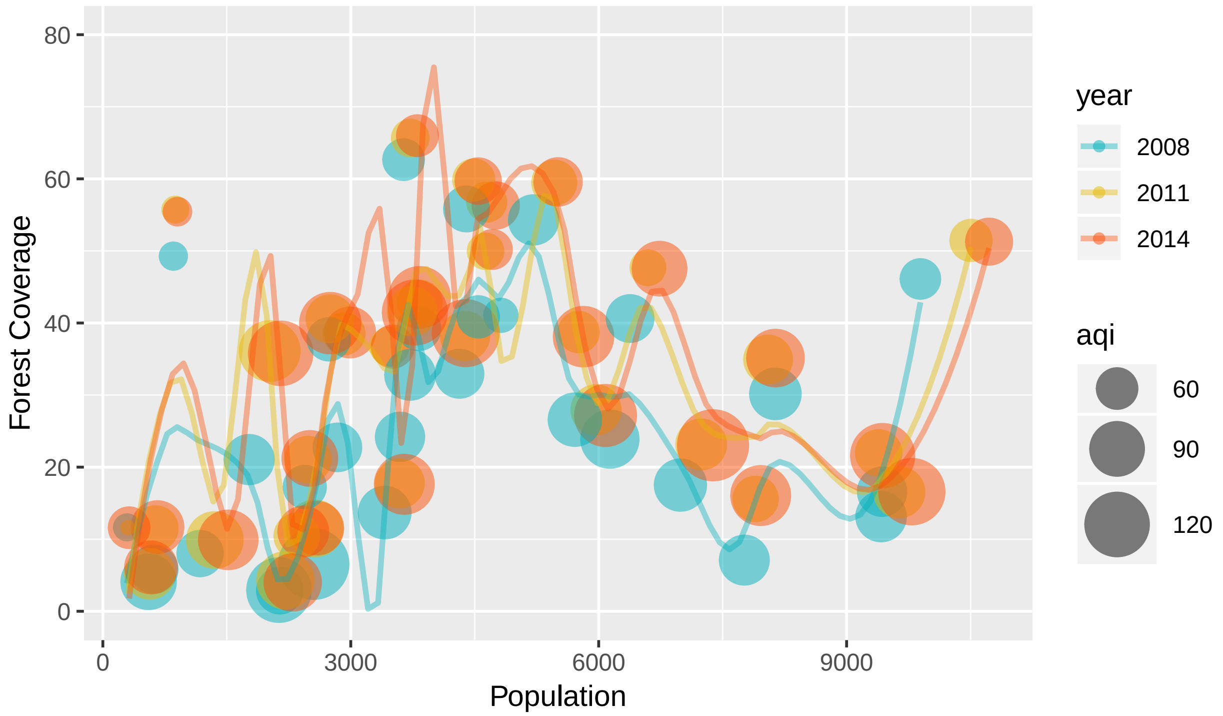
```
data_long %>%
  select(-("metrics")) %>%
  filter(measurement == "aqi" | measurement == "forest" | measurement == "population") %>%
  spread(measurement, value) %>%

  ggplot(aes(x = population, y = forest, size = aqi, color = year)) +
  geom_jitter(aes(color = year), width = 0.5, height = 0.5, alpha = 0.5) +
  stat_smooth (geom = "line", alpha = 0.4, size = 1, span = 0.2) +
  scale_size(range = c(2, 12)) +
  scale_color_manual(values = c("#00AFBB", "#E7B800", "#FC4E07")) +
  ylim(0, 80) +
  labs(
    title = "Provinces with less population and higher forest coverage have better air quality",
    subtitle = "Population by the end of Year vs. Average Forest Coverage, a circle represents a provin
    caption = "Peking University Open Research Data Platform",
    x = "Population",
    y = "Forest Coverage",
    color = "year"
  )
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

Provinces with less population and higher forest coverage have better air qu

Population by the end of Year vs. Average Forest Coverage, a circle represents a province



Peking University Open Research Data Platform