smog

christine giang

4/28/2019

LIBRARIES

```
library(stringr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(latex2exp)
library(caret)
```

```
## Loading required package: lattice
```

```
library(class)
library(mclust)
```

```
## Package 'mclust' version 5.4.3
## Type 'citation("mclust")' for citing this R package in publications.
```

```
library(rworldmap)
```

```
## Loading required package: sp

## ### Welcome to rworldmap ###

## For a short introduction type :   vignette('rworldmap')
```

```
library(ggmap)
```

```
## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.

## Please cite ggmap if you use it! See citation("ggmap") for details.
```

```
library(rgdal)
```

```
## rgdal: version: 1.4-3, (SVN revision 828)
##  Geospatial Data Abstraction Library extensions to R successfully loaded
##  Loaded GDAL runtime: GDAL 2.1.3, released 2017/20/01
##  Path to GDAL shared files: /Library/Frameworks/R.framework/Versions/3.5/Resources/library/rgdal/gdal
##  GDAL binary built with GEOS: FALSE
##  Loaded PROJ.4 runtime: Rel. 4.9.3, 15 August 2016, [PJ_VERSION: 493]
##  Path to PROJ.4 shared files: /Library/Frameworks/R.framework/Versions/3.5/Resources/library/rgdal/pr
##  Linking to sp version: 1.3-1
```

```
library(raster)
```

```
##
## Attaching package: 'raster'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(sp)
library(GISTools)
```

```
## Loading required package: maptools

## Checking rgeos availability: TRUE

## Loading required package: RColorBrewer

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following objects are masked from 'package:raster':
##
##     area, select

## The following object is masked from 'package:dplyr':
##
##     select

## Loading required package: rgeos

## rgeos version: 0.4-3, (SVN revision 595)
##  GEOS runtime version: 3.6.1-CAPI-1.10.1
##  Linking to sp version: 1.3-1
##  Polygon checking: TRUE
```

```r
#install.packages("mclust")
```

```r
prov <- read.csv("~/Documents/caL/2019/cyplan101/projects/assignment3/Archive/chinese_provinces.csv")

ids <- c(11, 12, 13, 14, 15, 21, 22, 23, 31, 32, 34, 35, 36, 41, 42, 43, 44, 45, 46, 50, 51, 52, 53, 54

names <- c("beijing", "tianjin", "hebei", "shanxi", "inner mongolia", "liaoning (SHENYANG)", "jilin", "

name_frame <- data.frame(
  ids = ids,
  names = names,
  highlight = rep(0, 31)
)

name_frame[c(1, 6, 9, 18, 22),3] <- 1

smog_names <- c("beijing", "liaoning (SHENYANG)", "shanghai", "guangdong (GUANGZHOU)", "sichuan (CHENGDU

all_ids <- table(prov$prov_id)
all_id_int <- as.integer(names(all_ids))

full_name_vector <- NULL

indices <- NULL
summed <- 0

for (i in 1:nrow(prov)){
  if (prov[i,5] %in% ids){
    indices[i] <- i
    summed <- summed + 1
  } else{
    indices[i] <- 0
  }
}

for (i in 1:length(indices)){
  if (indices[i] == 0){
    full_name_vector[i] <- "unlabeled"
  } else{
    full_name_vector[i] <- names[name_frame$ids == prov[i,5]]
  }
}



prov$names <- full_name_vector

prov <- prov[, c(2,3,5,7)]

indicator <- NULL

for (i in 1:nrow(prov)){
  if (prov$names[i] %in% smog_names){
```

```
    indicator[i] <- 1
  } else{
    indicator[i] <- 0
  }
}
prov$smog_area <- indicator

#write.csv(provinces, "province_names.csv")
```

```
events <- read.csv("~/Documents/caL/2019/cyplan101/projects/assignment3/kaggle/events.csv")

# take out users with fewer than 5 entries
counts <- sort(table(events$device_id), decreasing = TRUE)


counts <- counts[counts > 5]

higher <- counts[1:26990]



filtered <- events[events$device_id %in% names(higher), ]

counted_frame <- data.frame(
  device_id = counts
)

merged <- merge(filtered, counted_frame, by.x = "device_id", by.y = "device_id.Var1")

events <- merged
```

```
beijing <- read.csv("~/Documents/caL/2019/cyplan101/projects/assignment3/smog/2016/beijing_2016.csv")

chengdu <- read.csv("~/Documents/caL/2019/cyplan101/projects/assignment3/smog/2016/chengdu_2016.csv")

guangzhou <- read.csv("~/Documents/caL/2019/cyplan101/projects/assignment3/smog/2016/guangzhou_2016.csv

shanghai <- read.csv("~/Documents/caL/2019/cyplan101/projects/assignment3/smog/2016/shanghai_2016.csv")

shenyang <- read.csv("~/Documents/caL/2019/cyplan101/projects/assignment3/smog/2016/shenyang_2016.csv")

beijing$Date <- str_extract(beijing$Date, pattern = "[0-9]+/[0-9]+/[0-9][0-9]")

chengdu$Date <- str_extract(chengdu$Date, pattern = "[0-9]+/[0-9]+/[0-9][0-9]")

guangzhou$Date <- str_extract(guangzhou$Date, pattern = "[0-9]+/[0-9]+/[0-9][0-9]")

shanghai$Date <- str_extract(shanghai$Date, pattern = "[0-9]+/[0-9]+/[0-9][0-9]")

shenyang$Date <- str_extract(shenyang$Date, pattern = "[0-9]+/[0-9]+/[0-9][0-9]")


# take out negative values
```

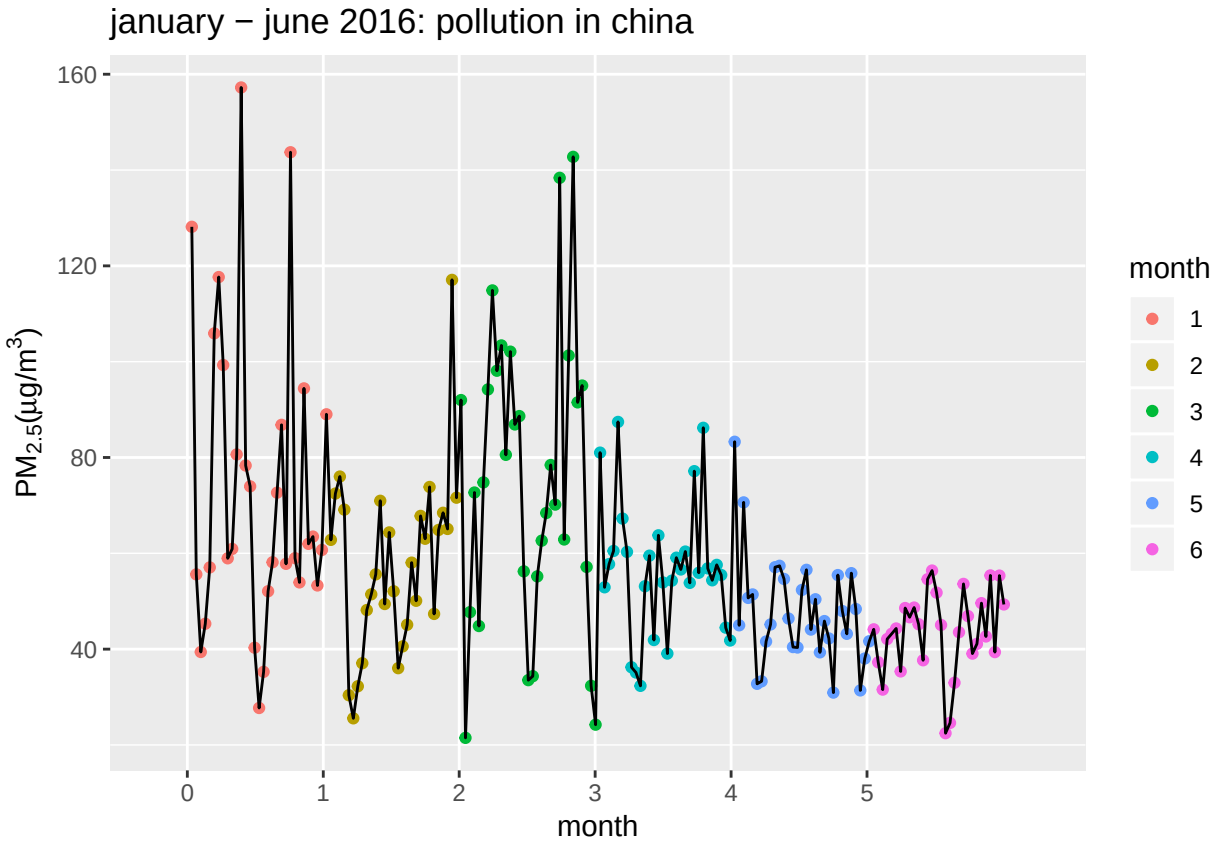
```r
beijing <- beijing[beijing$Value > 0, ]
chengdu <- chengdu[chengdu$Value > 0, ]
guangzhou <- guangzhou[guangzhou$Value > 0, ]
shanghai <- shanghai[shanghai$Value > 0, ]
shenyang <- shenyang[shenyang$Value > 0, ]
# aggregated by date, >> mean value of AQI of each day.

beijing_ag <- summarise(
  group_by(beijing, Date),
  mean_aqi = mean(Value),
  moe = sd(Value)
)

chengdu_ag <- summarise(
  group_by(chengdu, Date),
  mean_aqi = mean(Value),
  moe = sd(Value)
)

guangzhou_ag <- summarise(
  group_by(guangzhou, Date),
  mean_aqi = mean(Value),
  moe = sd(Value)
)

shanghai_ag <- summarise(
  group_by(shanghai, Date),
  mean_aqi = mean(Value),
  moe = sd(Value)
)

shenyang_ag <- summarise(
  group_by(shenyang, Date),
  mean_aqi = mean(Value),
  moe = sd(Value)
)


unhealthy = c(sum(beijing$Value > 150 & beijing$Value <= 300), sum(chengdu$Value > 150 & chengdu$Value

percent_unhealthy = c(sum(beijing$Value > 150 & beijing$Value <= 300)/nrow(beijing), sum(chengdu$Value

hazardous = c(sum(beijing$Value > 300), sum(chengdu$Value > 300), sum(guangzhou$Value > 300), sum(shangh

percent_hazardous = c(sum(beijing$Value > 300)/nrow(beijing), sum(chengdu$Value > 300)/nrow(chengdu), su


hazard_table <- data.frame(
  city = c("beijing", "chengdu", "guangzhou", "shanghai", "shenyang"),
  unhealthy = unhealthy,
  prop_unhealthy = (unhealthy/sum(unhealthy)),
  hazardous = hazardous,
  prop_hazardous = (hazardous/sum(hazardous))
```

```
)

#write.csv(hazard_table, "hazard_table.csv")

chengdu <- chengdu[,-2]

# site, month, day, value
```

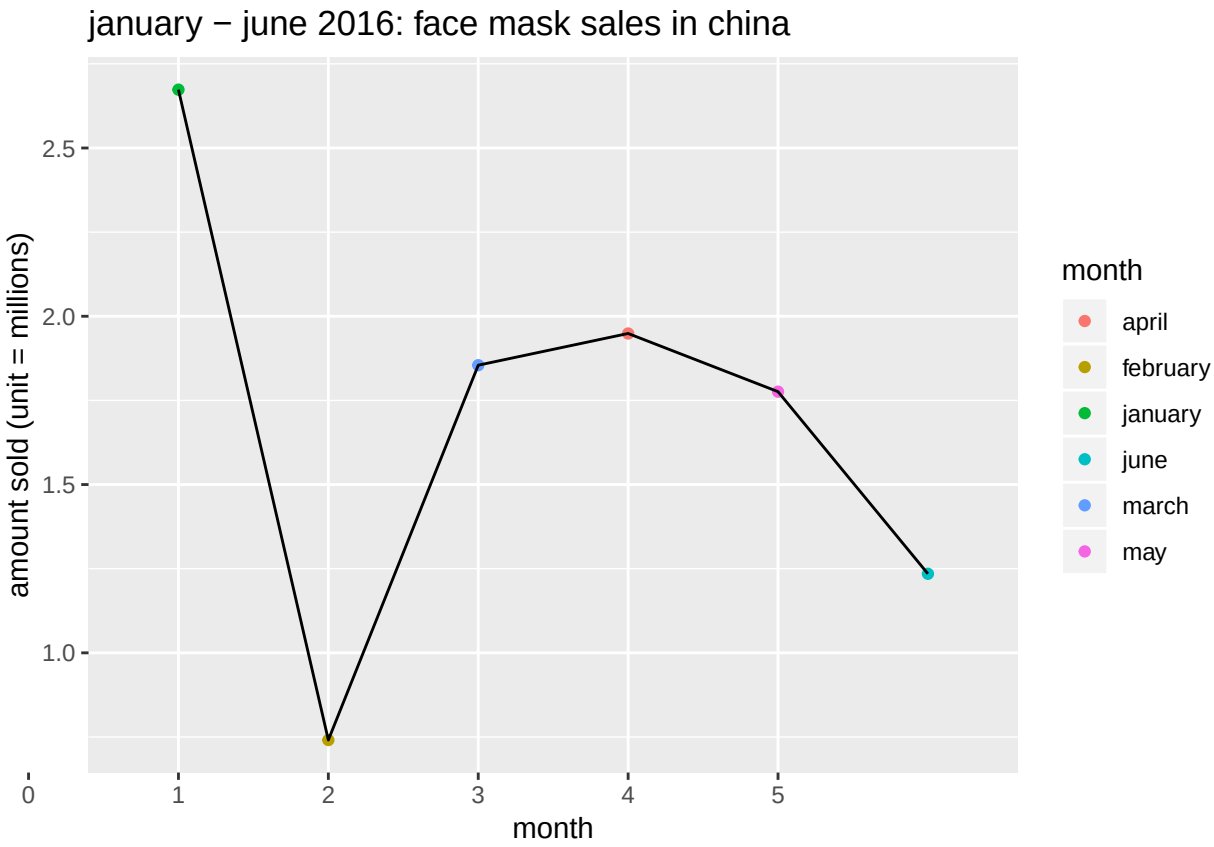
MASKS

```
masks <- read.csv("~/Documents/caL/2019/cyplan101/projects/assignment3/data/masks.csv")

five_table <- rbind(beijing, chengdu, guangzhou, shanghai, shenyang)


#month_only <- str_extract(five_table$Date, pattern = "[0-9]+/")
#month_only <- str_sub(month_only, end = -2L)

#five_table$month <- month_only

monthly_summary <- summarise(
  group_by(five_table, Month),
  monthly_mean_val = mean(Value),
  sd = sd(Value)
)


six_months <- monthly_summary[c(1:6), ]


mask_compare <- data.frame(
  month = six_months$Month,
  masks = masks$volume,
  mean_aqi = six_months$monthly_mean_val
)

#write.csv(mask_compare, 'mask_compare.csv')

five_hourly <- rbind(beijing, chengdu, guangzhou, shanghai, shenyang)

month_day <- str_c(five_hourly$Month, "/", five_hourly$Day, sep = "")

five_hourly$month_day <- month_day

five_summary <- summarise(
  group_by(five_hourly, month_day),
  mean_pm25 = mean(Value),
  sd = sd(Value)
)
five_summary$month <- str_extract(five_summary$month_day, pattern = '[0-9]+')

five_summary <- arrange(five_summary, month)
```

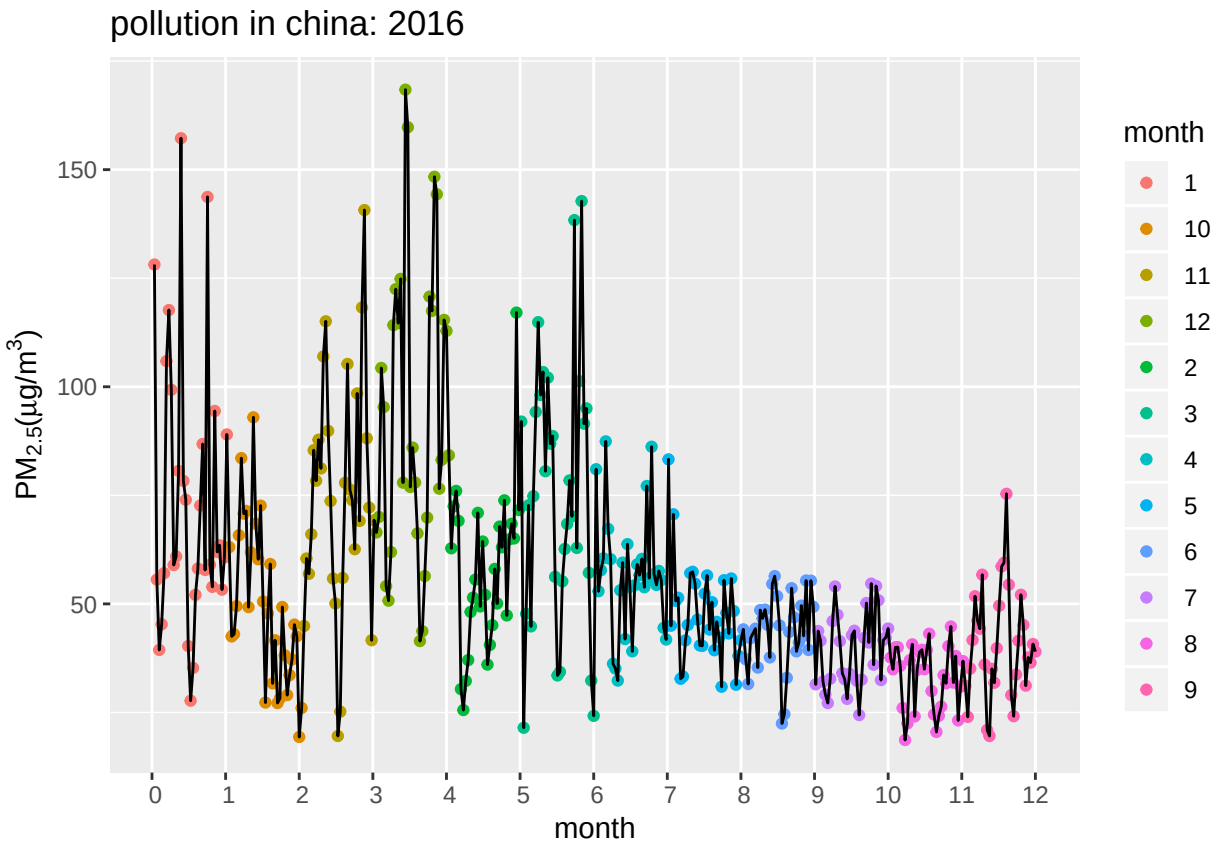
```
five_summary$month <- str_extract(five_summary$month_day, pattern = "[0-9]+/")
five_summary$month <- str_sub(five_summary$month, end = -2L)


first_six <- five_summary[five_summary$month %in% c(1:6), ]
month <- c("january", "february", "march", "april", "may", "june")
all_months <- c(0,1,2,3,4,5,6,7,8,9,10,11,12)

ggplot() + geom_point(aes(x = c(1:nrow(first_six))/30.3, y = first_six$mean_pm25, col = first_six$month
```



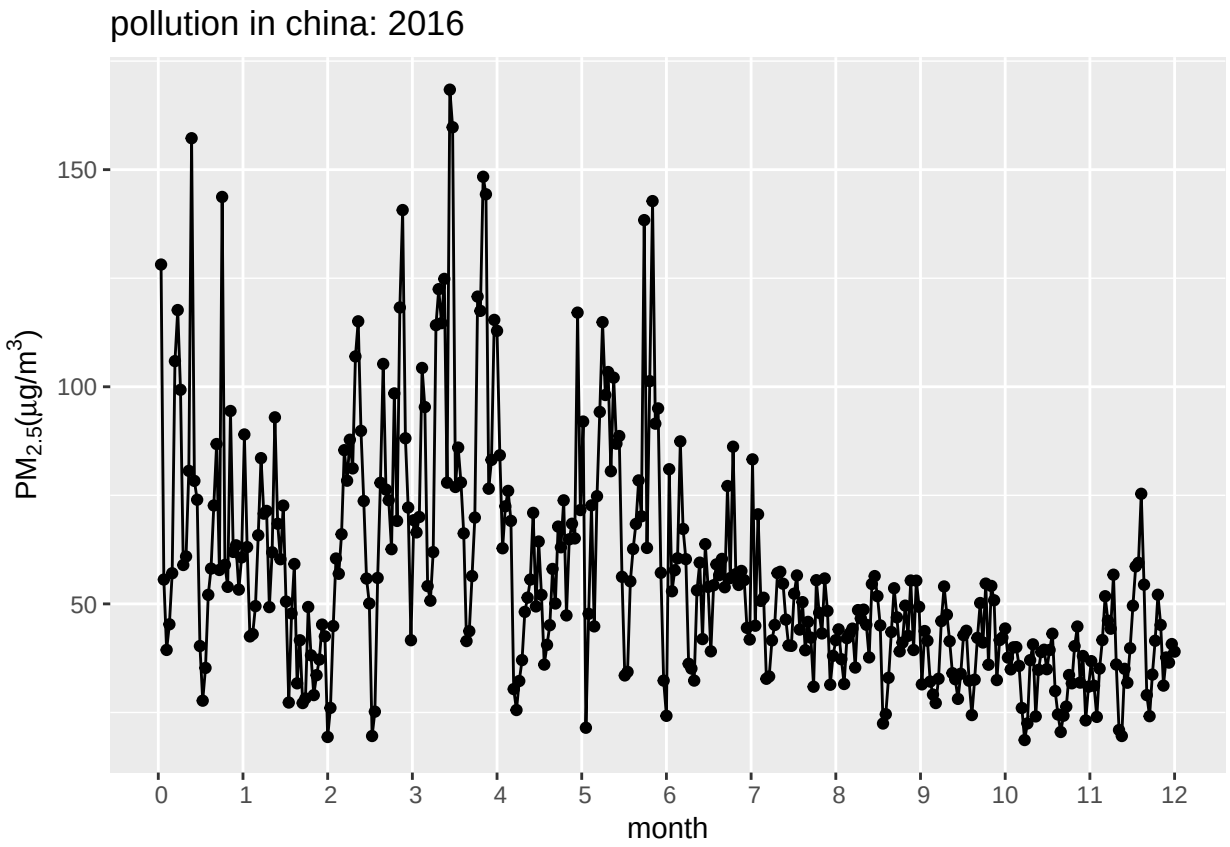

```
ggplot() + geom_point(aes(x =c(1:6) , y = masks$volume/1000000, col = month)) + geom_line(aes(x = c(1:6
```




```
ggplot() + geom_point(aes(x = c(1:nrow(five_summary))/30.5, y = five_summary$mean_pm25, col = five_summa
```



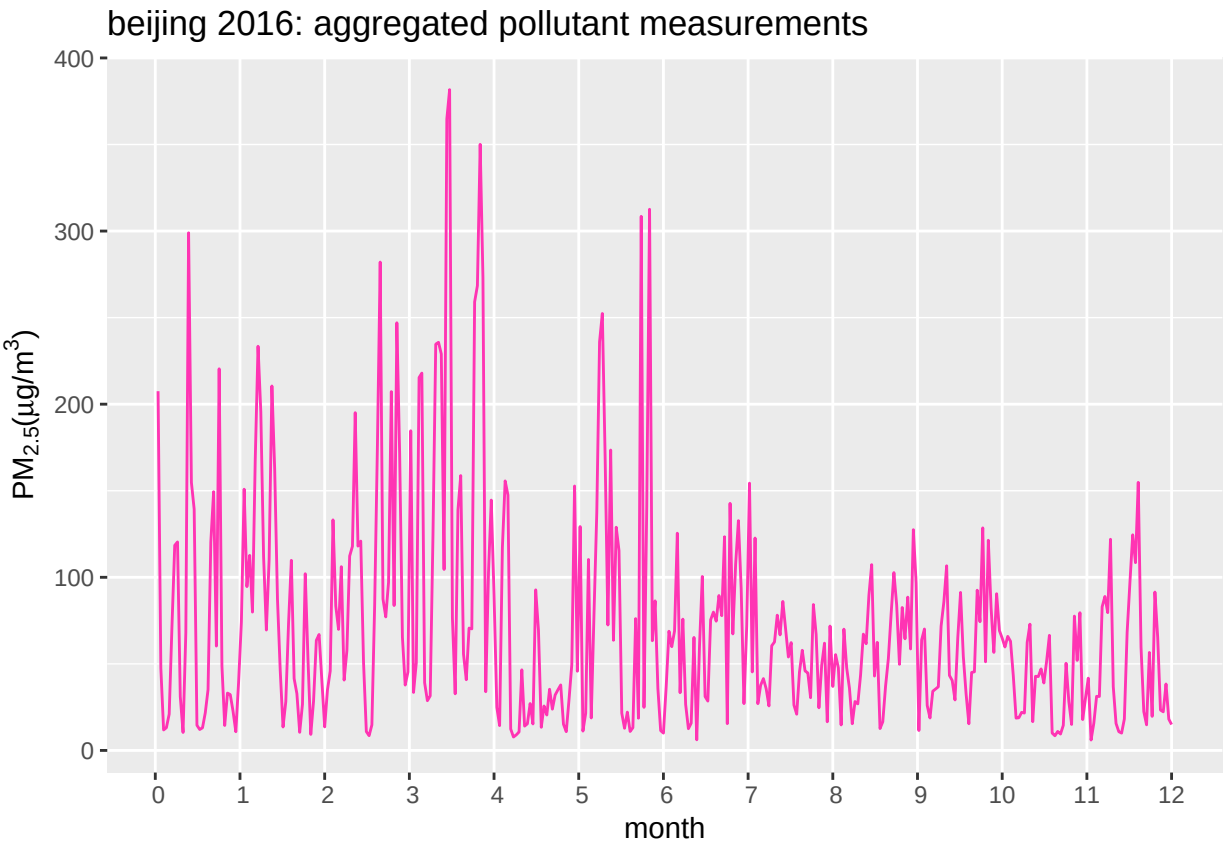

```
ggplot() + geom_point(aes(x = c(1:nrow(five_summary))/30.5, y = five_summary$mean_pm25)) + geom_line(aes
```



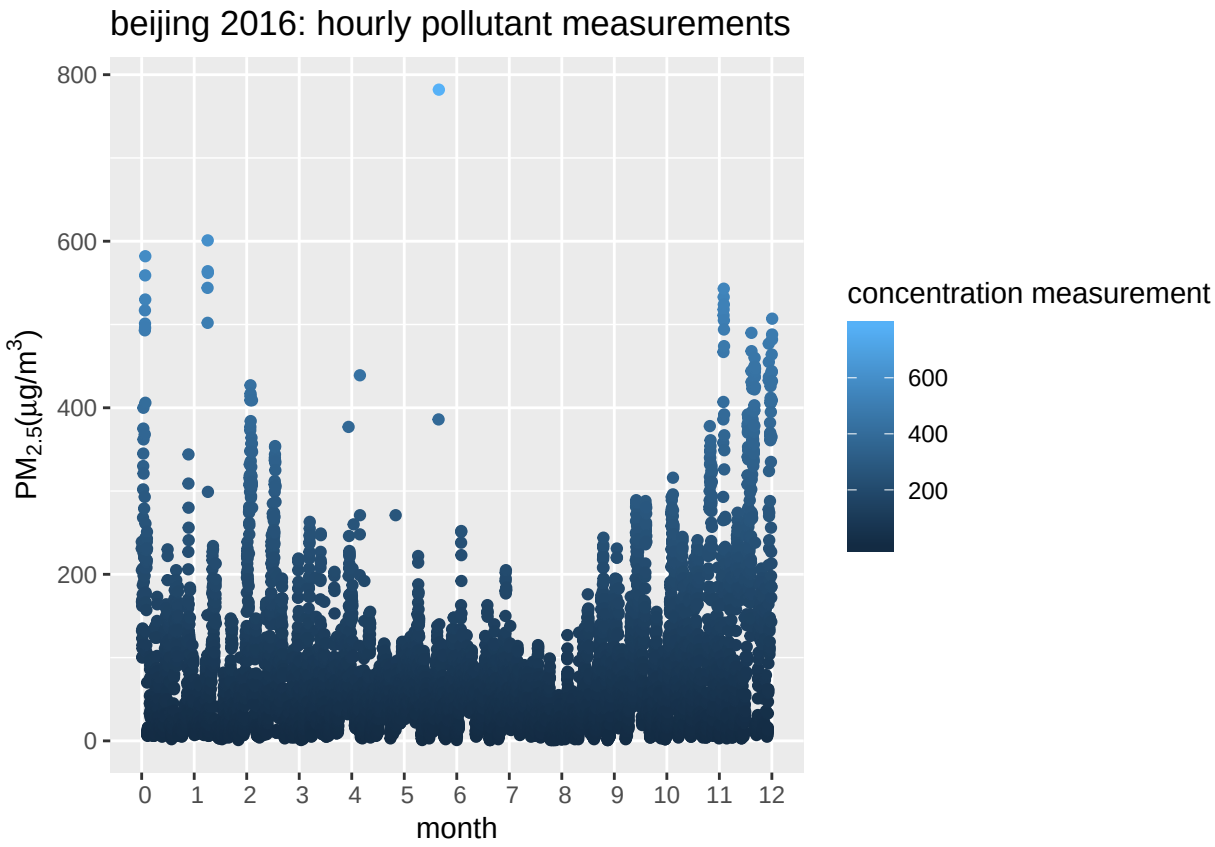

```
#write.csv(monthly_summary, "monthly_summary.csv")

#write.csv(five_summary, "five_summary.csv")
```

plots

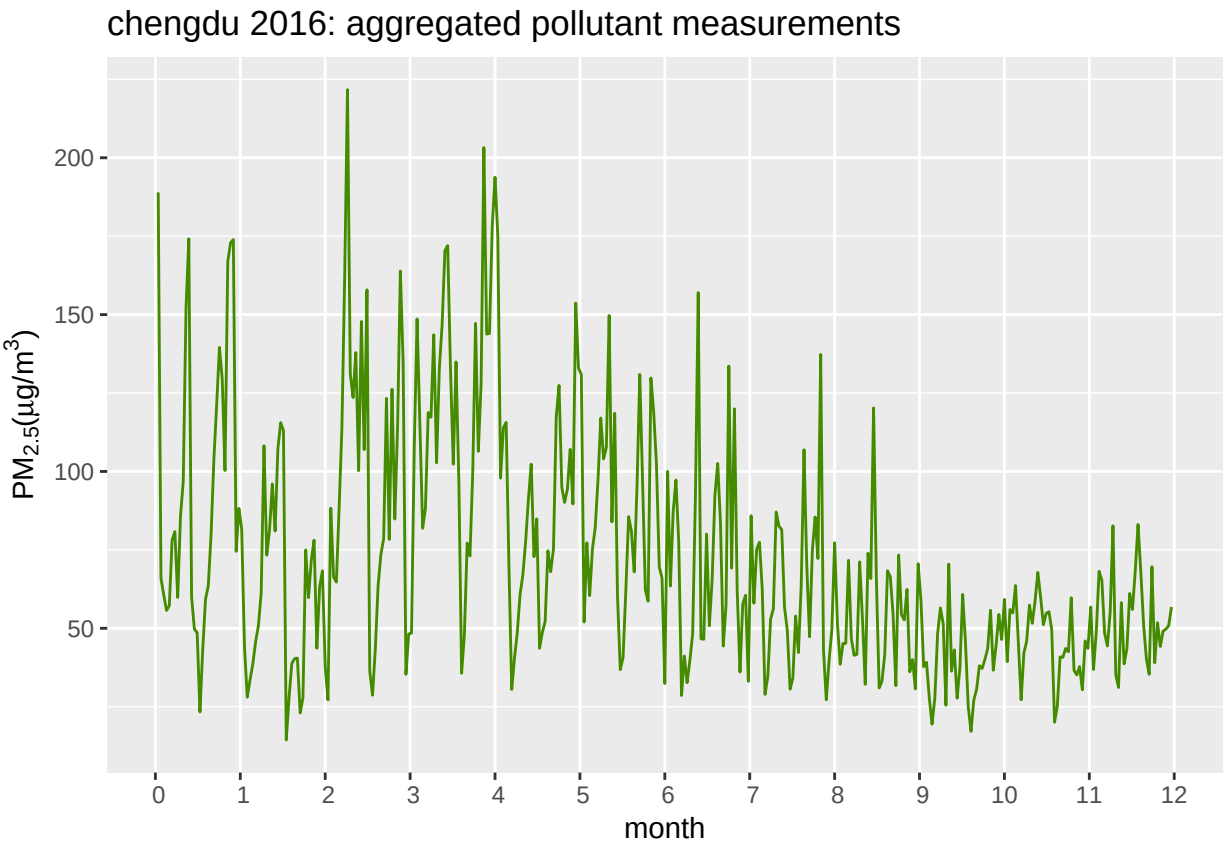
```
all_months <- c(0,1,2,3,4,5,6,7,8,9,10,11,12)
months <- c(0,1,2,3,4,5,6,7,8,9,10,11,12)
#png('cars_plot.png')

# BEIJING
ggplot() + geom_line(aes(x = c(1:nrow(beijing_ag))/30.5, y = beijing_ag$mean_aqi), col = "maroon1") + la
```



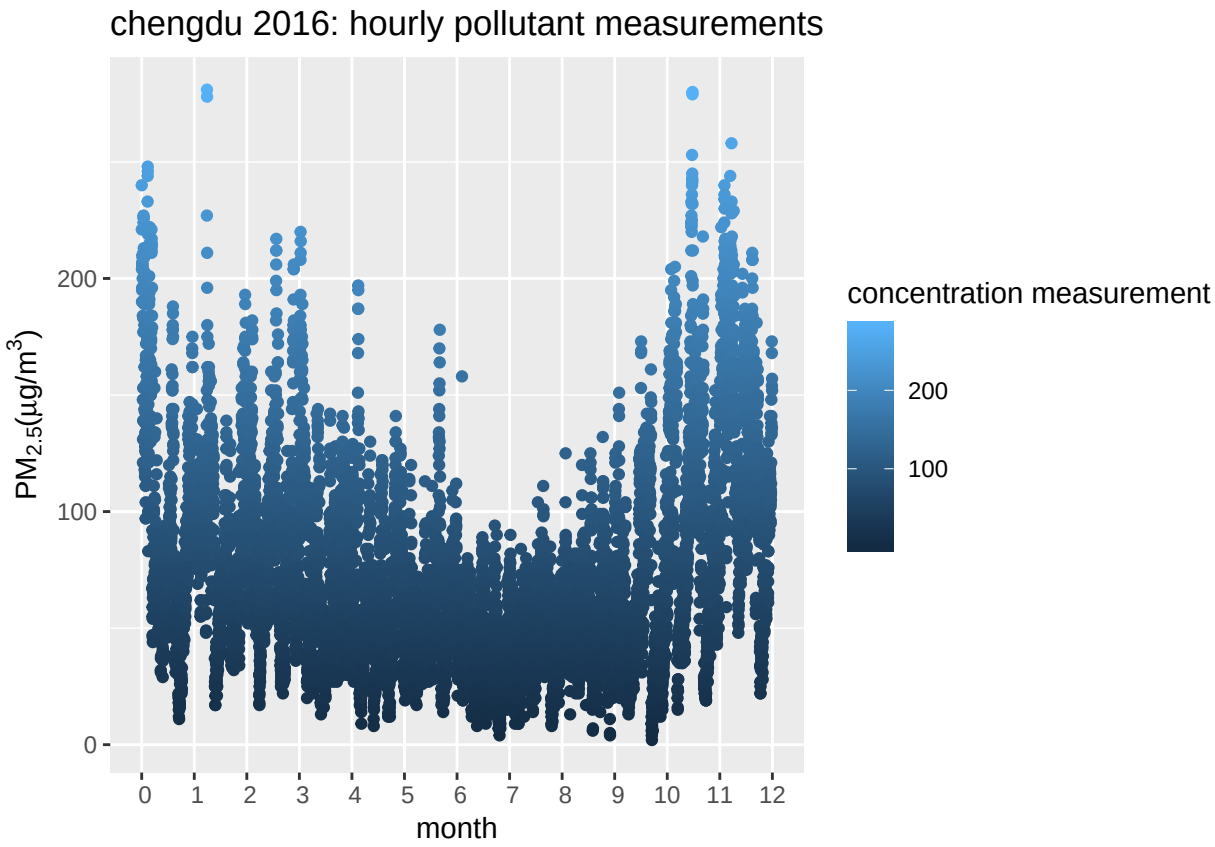

```
ggplot() + geom_point(aes(x = c(1:nrow(beijing))/727, y = beijing$Value, col = beijing$Value)) + labs(t
```



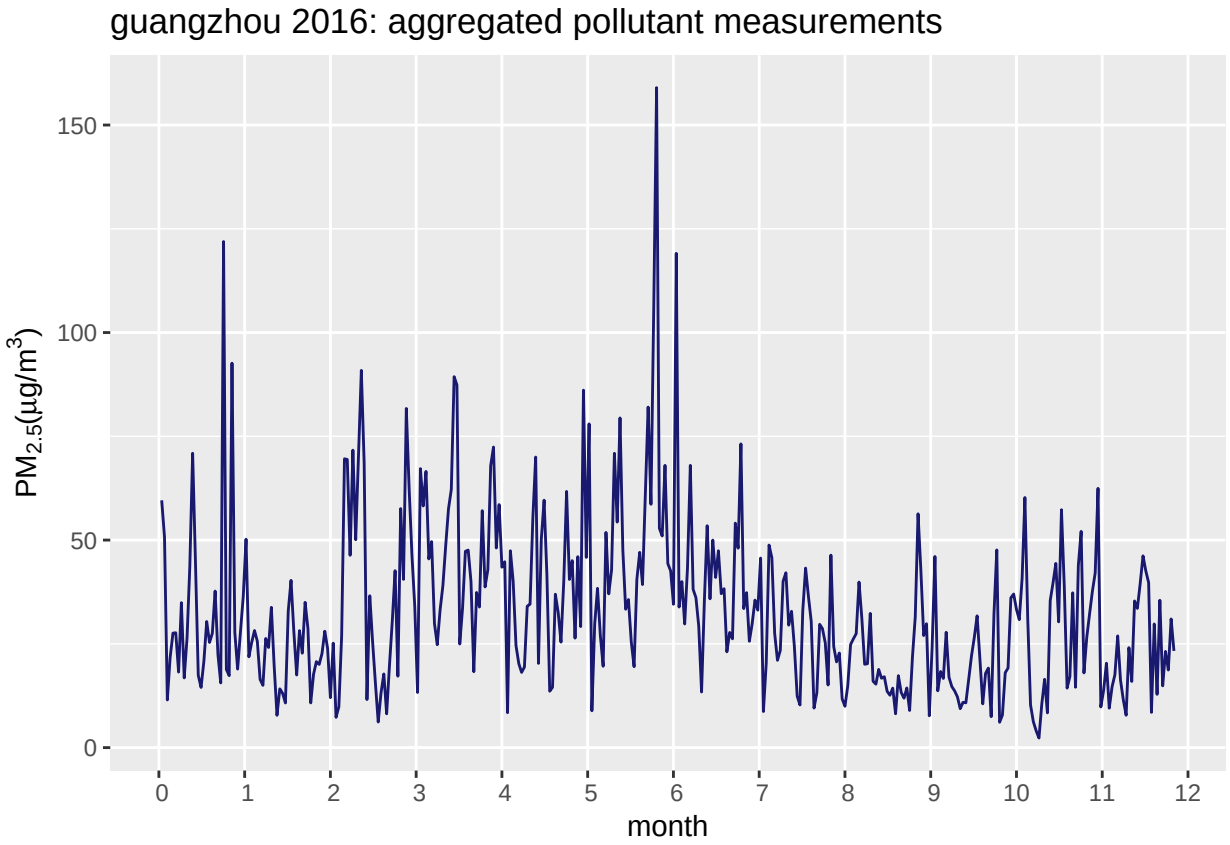

```
# CHENGDU
ggplot() + geom_line(aes(x = c(1:nrow(chengdu_ag))/30.5, y = chengdu_ag$mean_aqi), col = "chartreuse4")
```



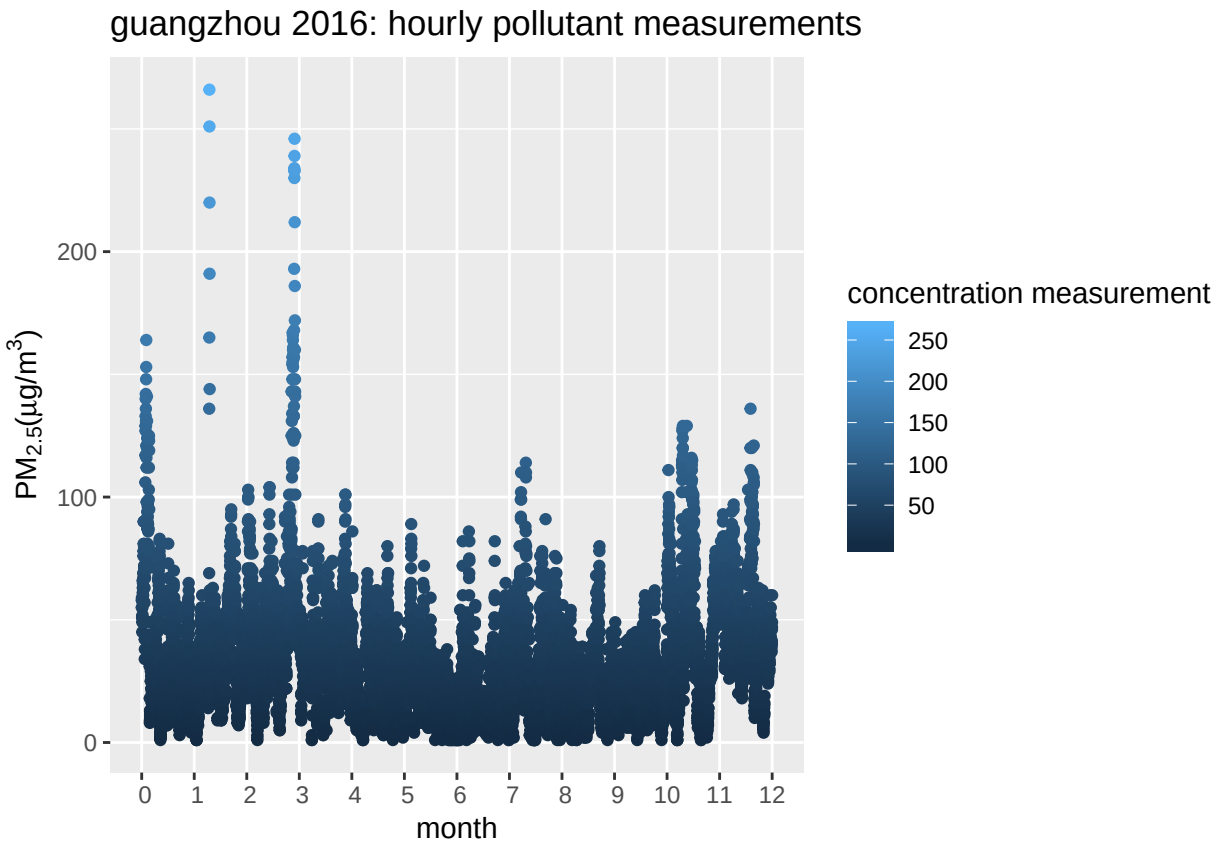

```
ggplot() + geom_point(aes(x = c(1:nrow(chengdu))/722, y = chengdu$Value, col = chengdu$Value)) + labs(t
```



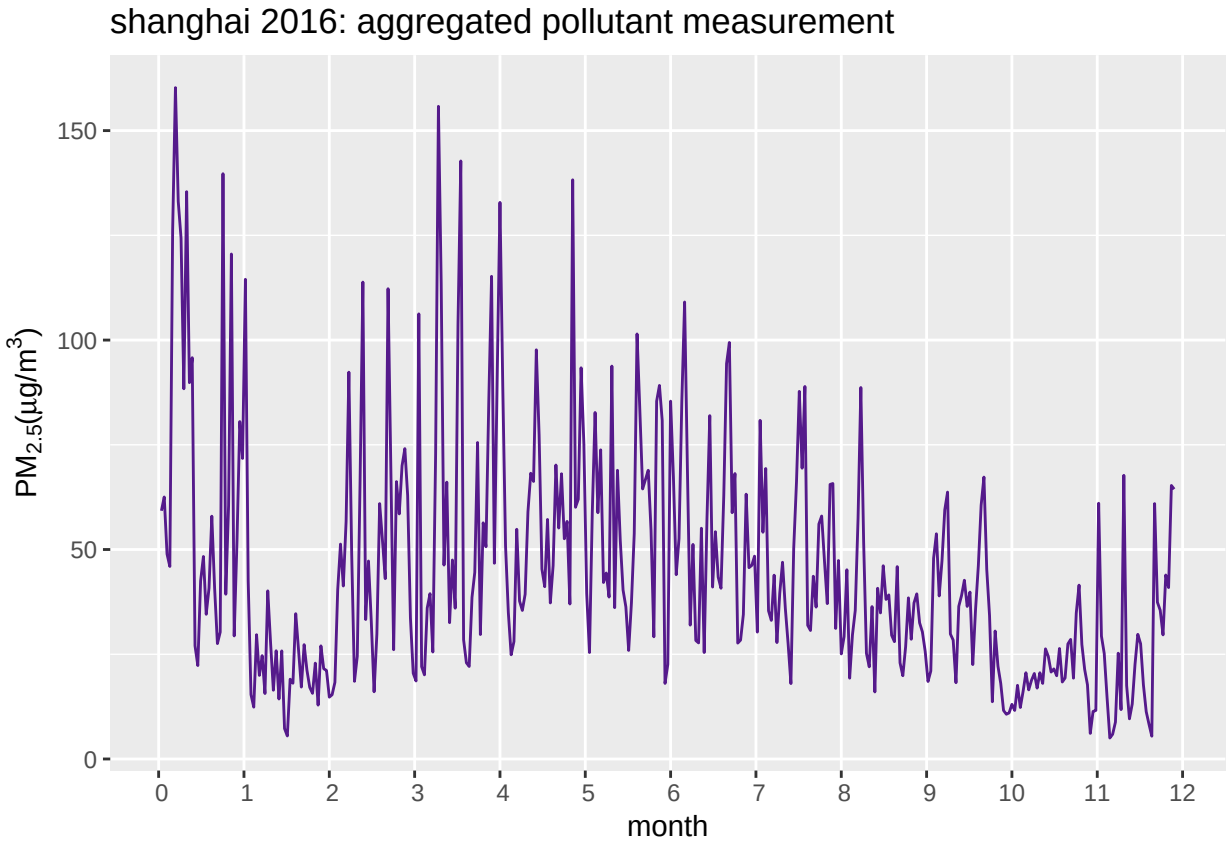

```
# GUANGZHOU

ggplot() + geom_line(aes(x = c(1:nrow(guangzhou_ag))/30.5, y = guangzhou_ag$mean_aqi), col = "midnight
```



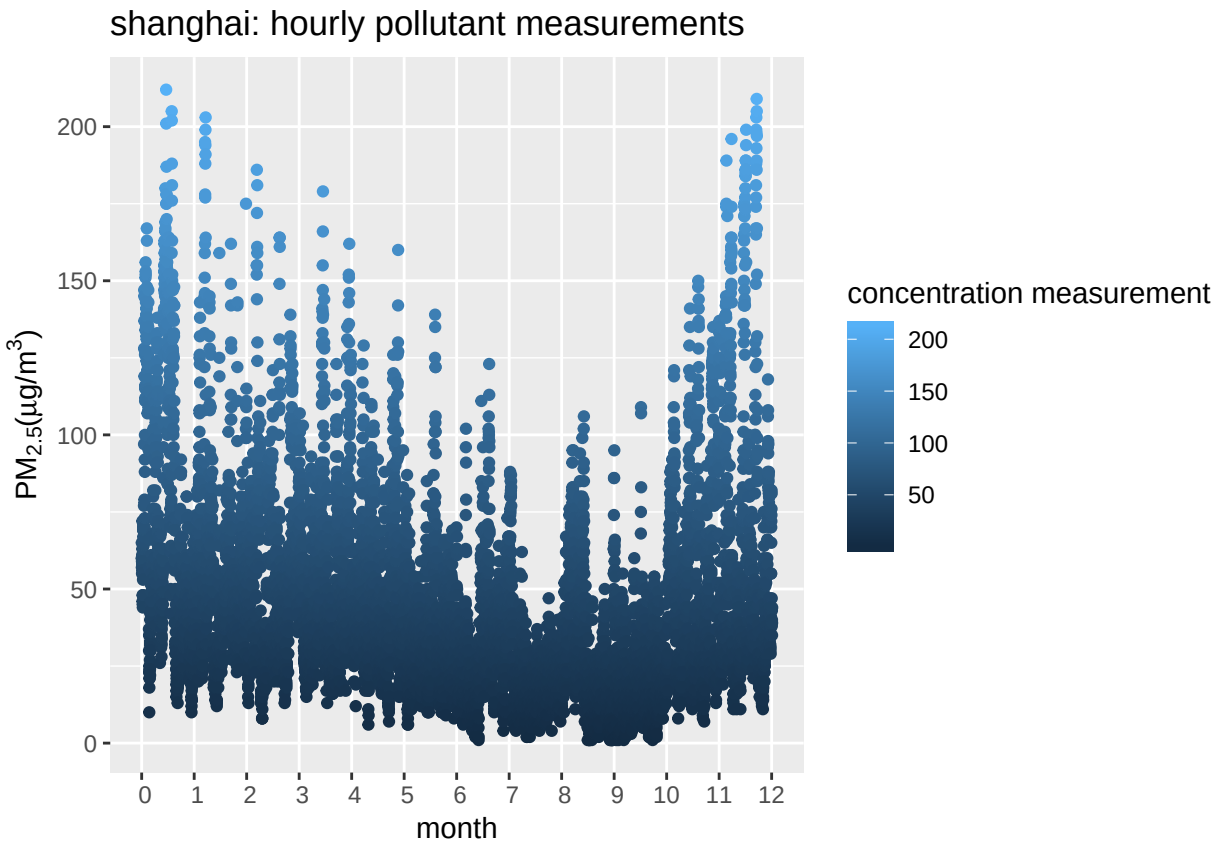

```
ggplot() + geom_point(aes(x = c(1:nrow(guangzhou))/676, y = guangzhou$Value, col = guangzhou$Value)) + l
```



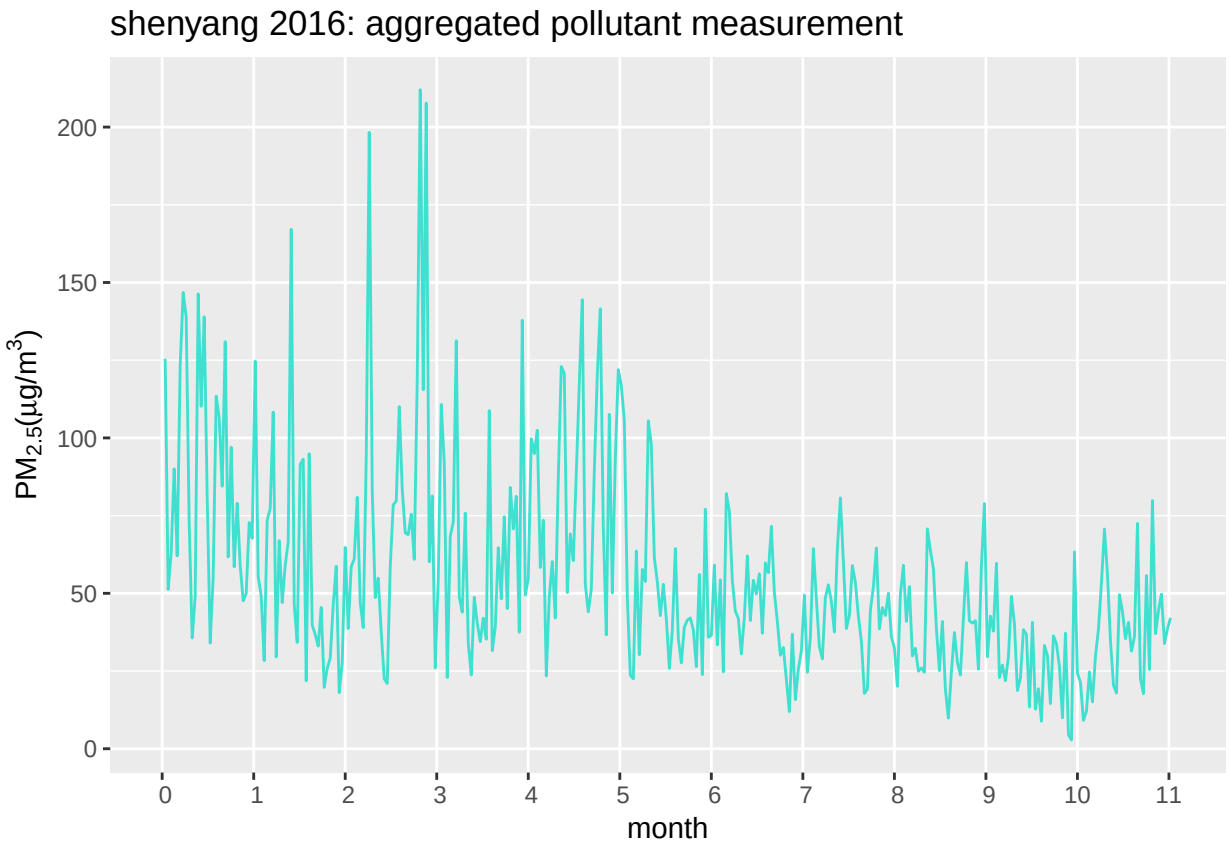

```
# SHANGHAI

ggplot() + geom_line(aes(x = c(1:nrow(shanghai_ag))/30.5, y = shanghai_ag$mean_aqi), col = "purple4") +
```



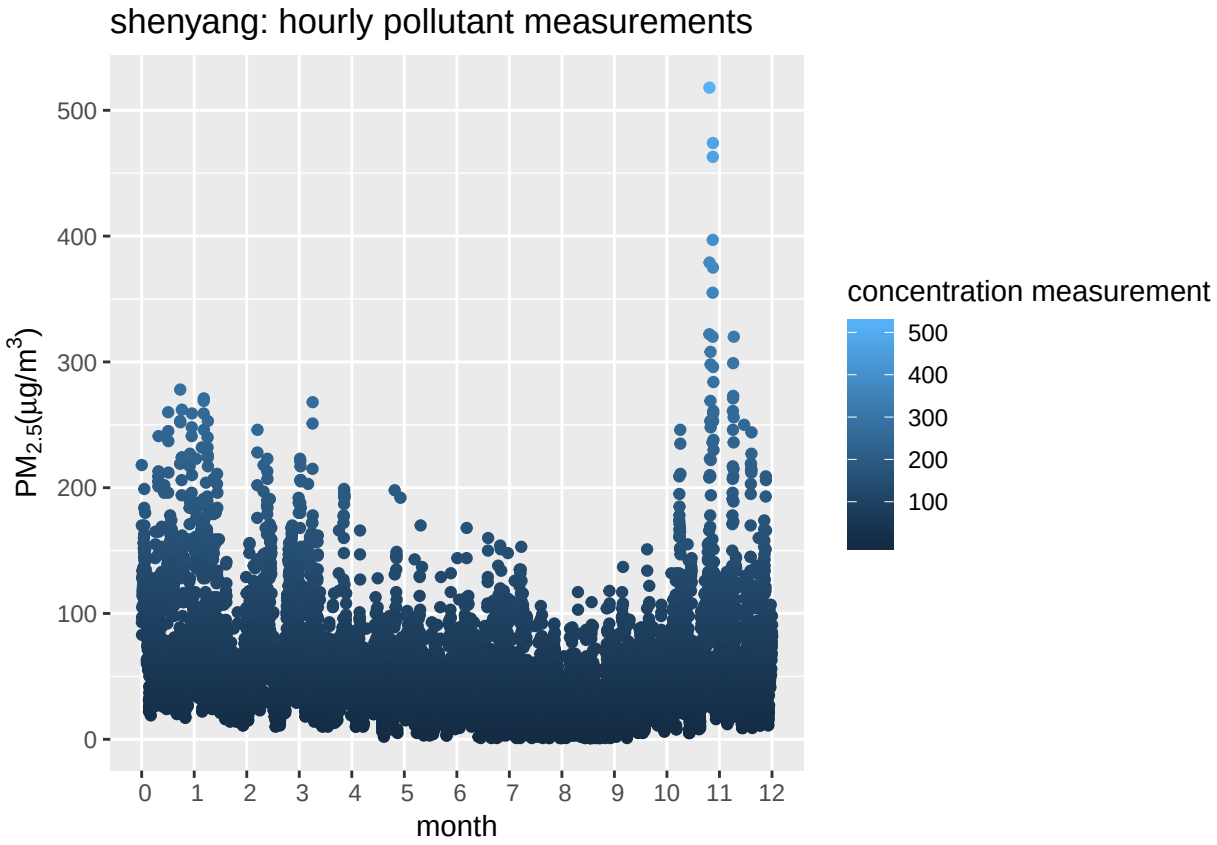

```
ggplot() + geom_point(aes(x = c(1:nrow(shanghai))/706, y = shanghai$Value, col = shanghai$Value)) + labs
```




```
# SHENYANG

ggplot()  + geom_line(aes(x = c(1:nrow(shenyang_ag))/30.5, y = shenyang_ag$mean_aqi), col = "turquoise"
```




```
ggplot() + geom_point(aes(x = c(1:nrow(shenyang))/634, y = shenyang$Value, col = shenyang$Value)) + labs
```




```
ggplot() +
  geom_line(aes(x = c(1:nrow(beijing_ag))/30.5, y = beijing_ag$mean_aqi), col = "red") +
  geom_line(aes(x = c(1:nrow(chengdu_ag))/30.5, y = chengdu_ag$mean_aqi), col = "blue") +
  geom_line(aes(x = c(1:nrow(guangzhou_ag))/30.5, y = guangzhou_ag$mean_aqi), col = "purple") +
  geom_line(aes(x = c(1:nrow(shanghai_ag))/30.5, y = shanghai_ag$mean_aqi), col = "green") +
  geom_line(aes(x = c(1:nrow(shenyang_ag))/30.5, y = shenyang_ag$mean_aqi), col = "pink") + labs(title =
```