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
```

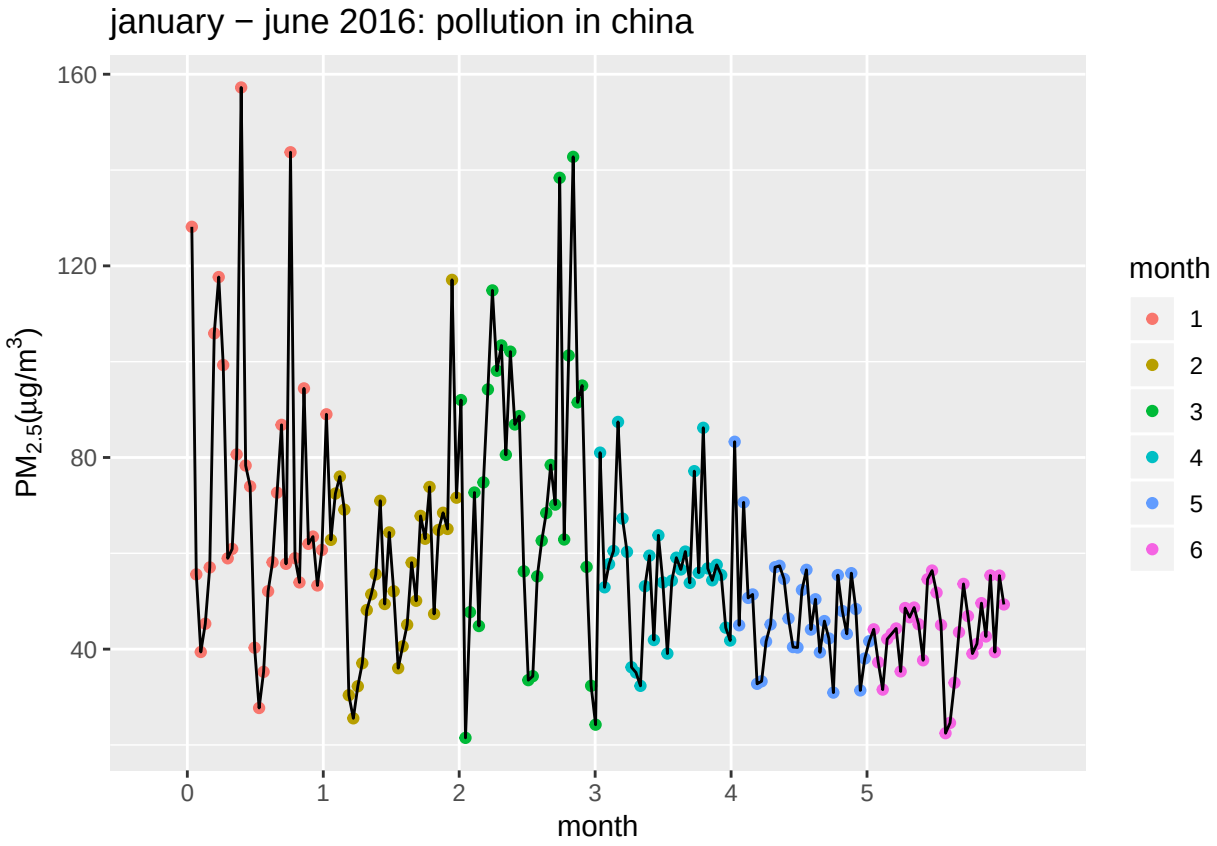
```r
beijing <- beijing[beijing$Value > 0, ]
chengdu <- chengdu[chengdu$Value > 0, ]
guangzhou <- guangzhou[guangzhou$Value > 0, ]
shanghai <- shanghai[shanghai$Value > 0, ]
shenyang <- shenyang[shenyang$Value > 0, ]
# aggregated by date, >> mean value of AQI of each day.

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
```

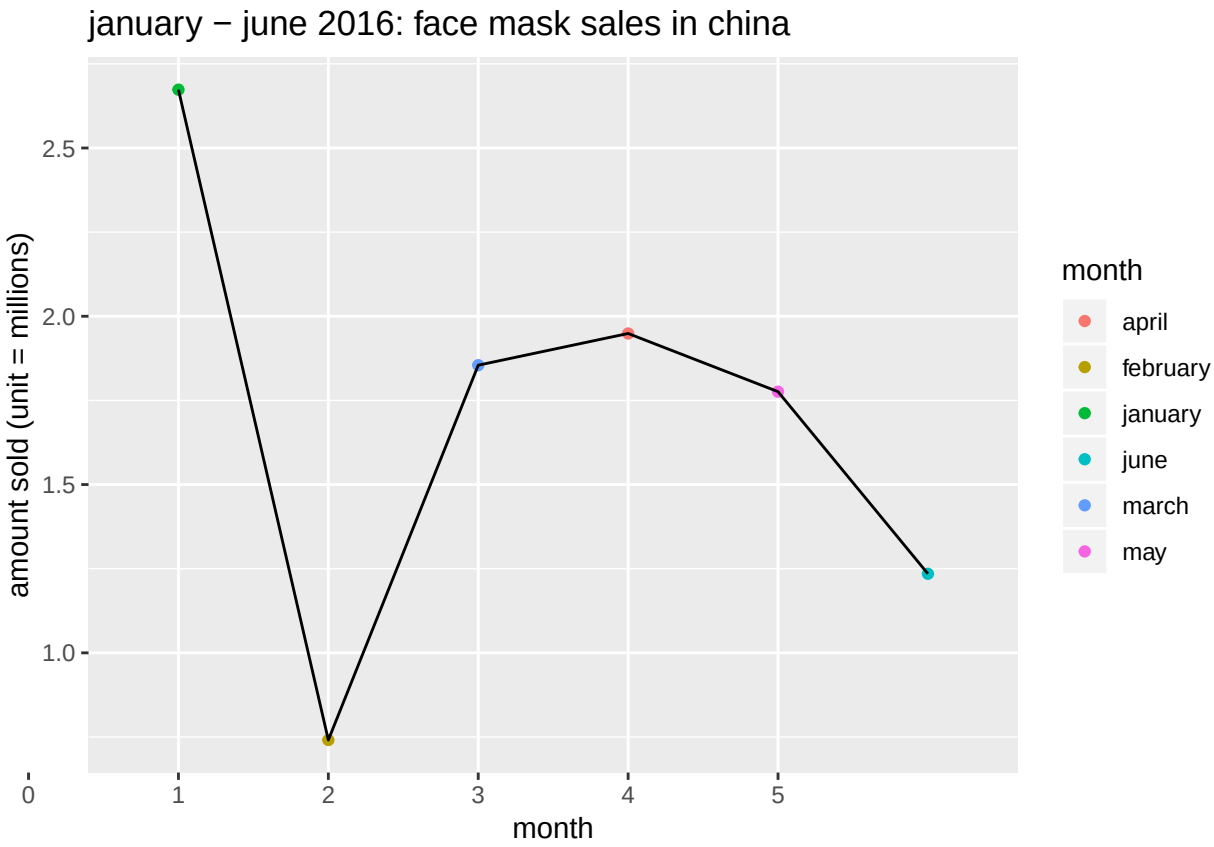
MASKS

```
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)
```

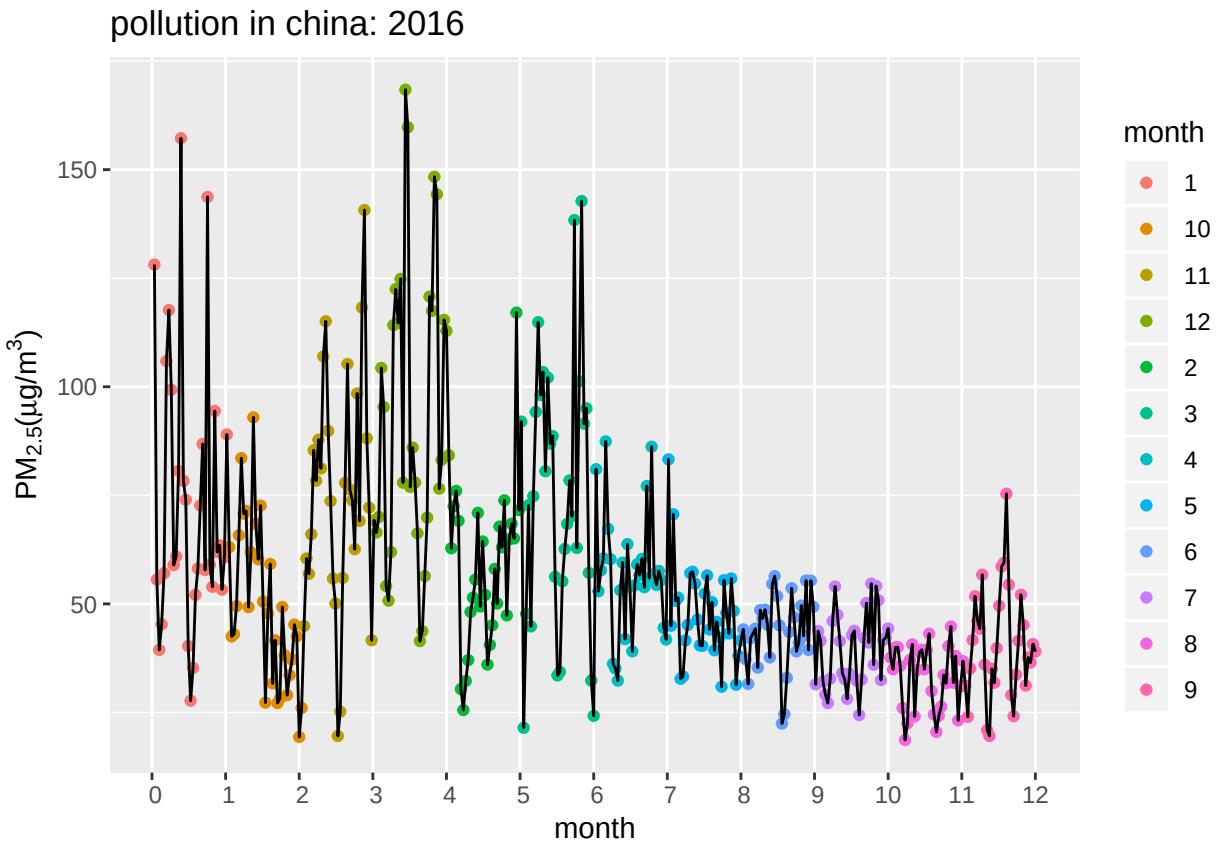
```
five_summary$month <- str_extract(five_summary$month_day, pattern = "[0-9]+/")
five_summary$month <- str_sub(five_summary$month, end = -2L)


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
```



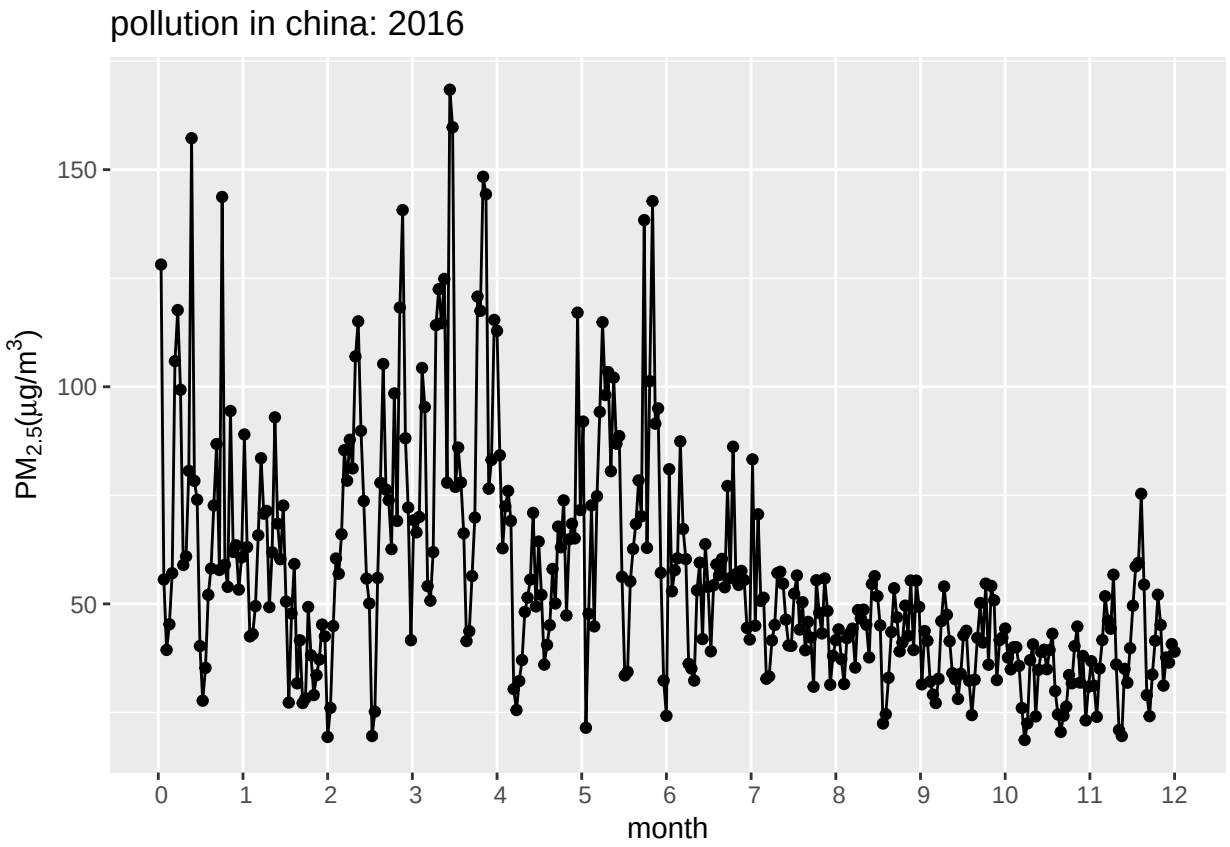

```
ggplot() + geom_point(aes(x =c(1:6) , y = masks$volume/1000000, col = month)) + geom_line(aes(x = c(1:6
```




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



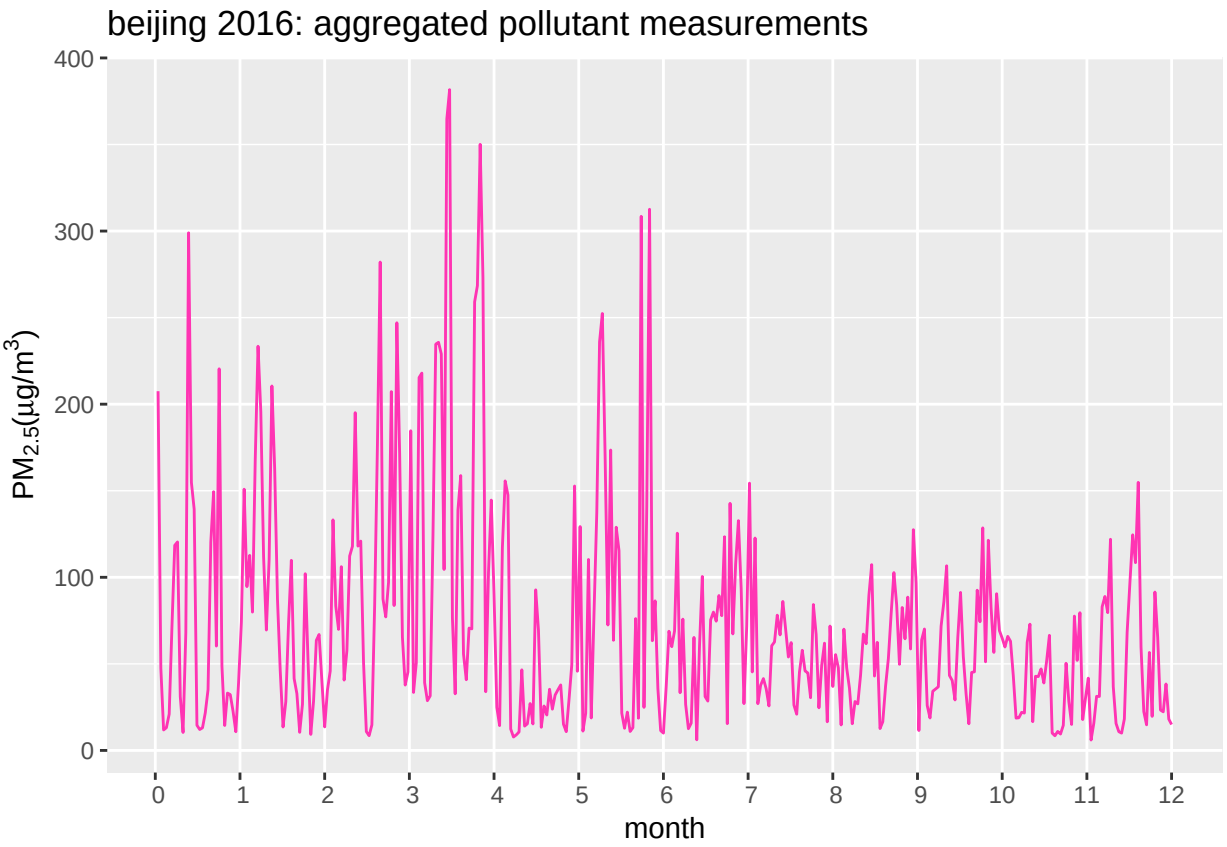

```
ggplot() + geom_point(aes(x = c(1:nrow(five_summary))/30.5, y = five_summary$mean_pm25)) + geom_line(aes
```



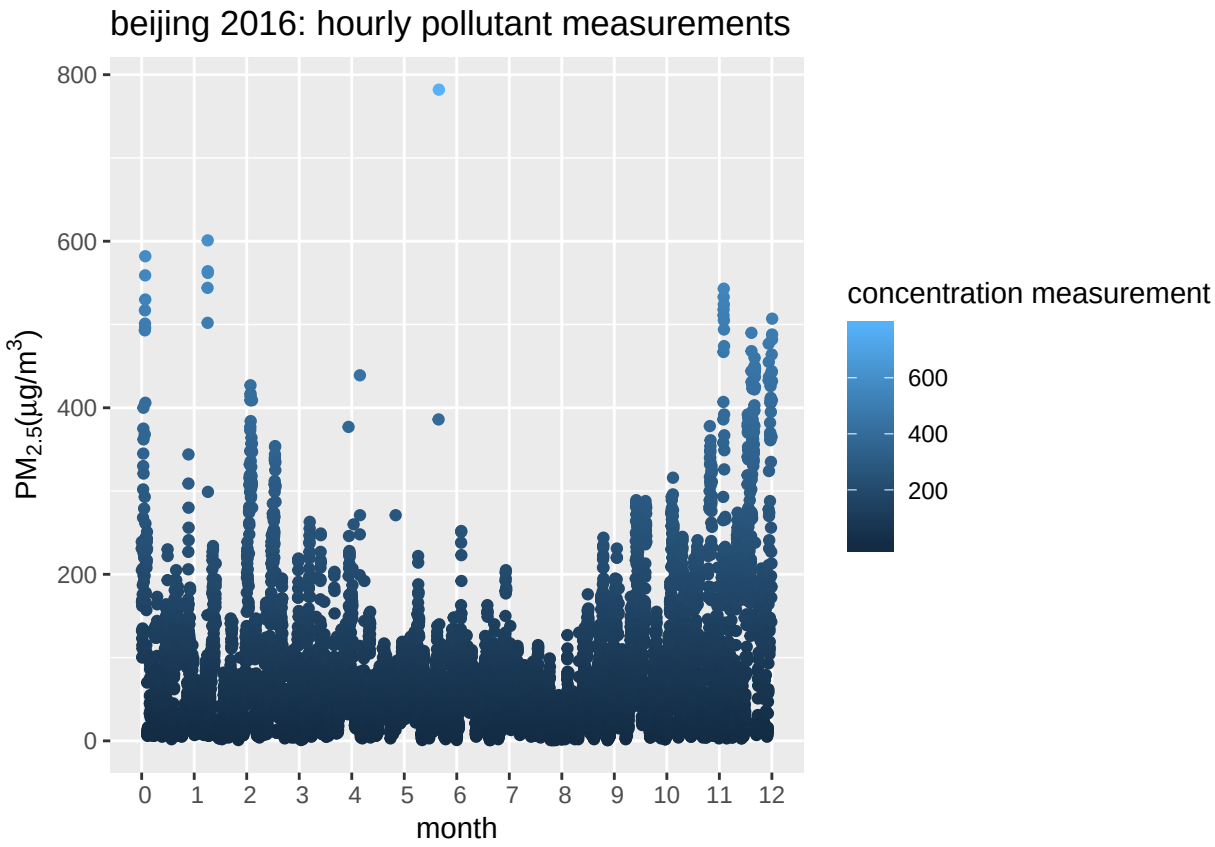

```
#write.csv(monthly_summary, "monthly_summary.csv")

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

plots

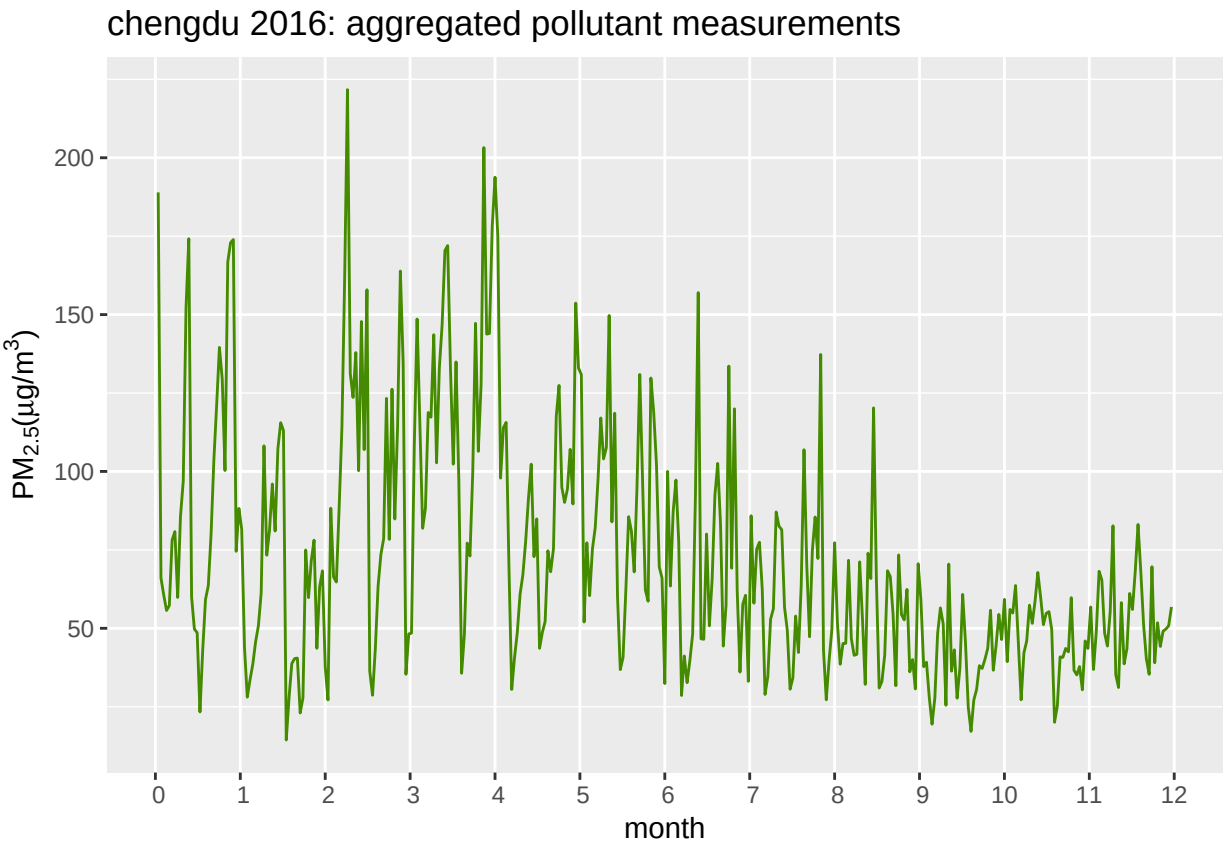
```
all_months <- c(0,1,2,3,4,5,6,7,8,9,10,11,12)
months <- c(0,1,2,3,4,5,6,7,8,9,10,11,12)
#png('cars_plot.png')

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



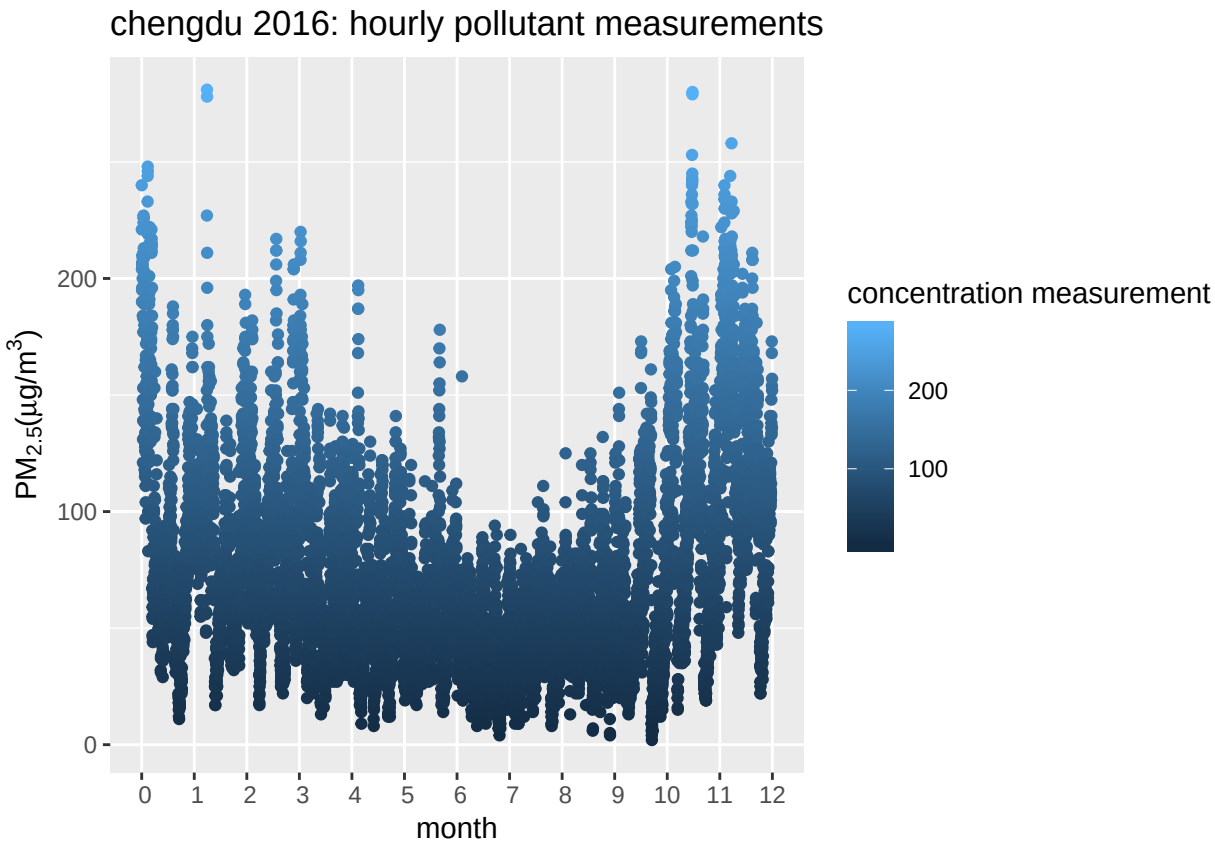

```
ggplot() + geom_point(aes(x = c(1:nrow(beijing))/727, y = beijing$Value, col = beijing$Value)) + labs(t
```



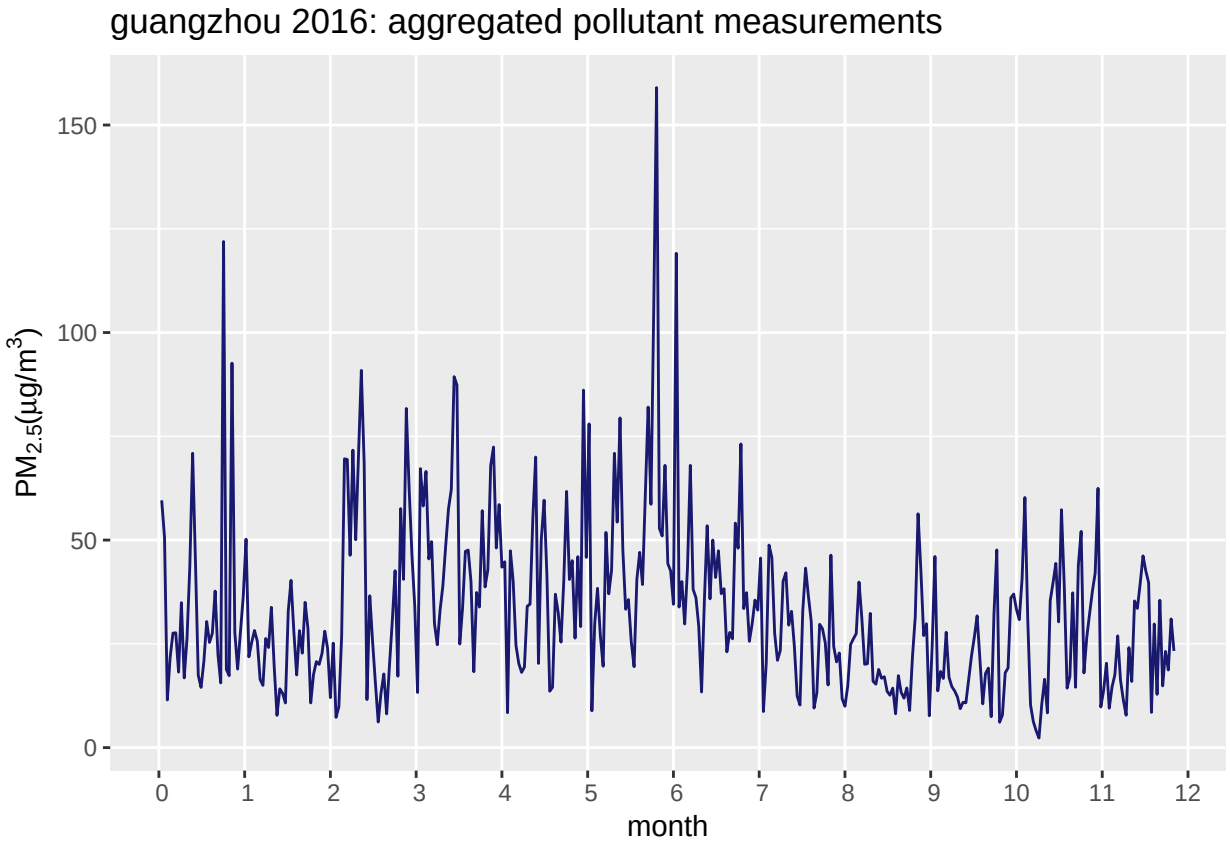

```
# CHENGDU
ggplot() + geom_line(aes(x = c(1:nrow(chengdu_ag))/30.5, y = chengdu_ag$mean_aqi), col = "chartreuse4")
```



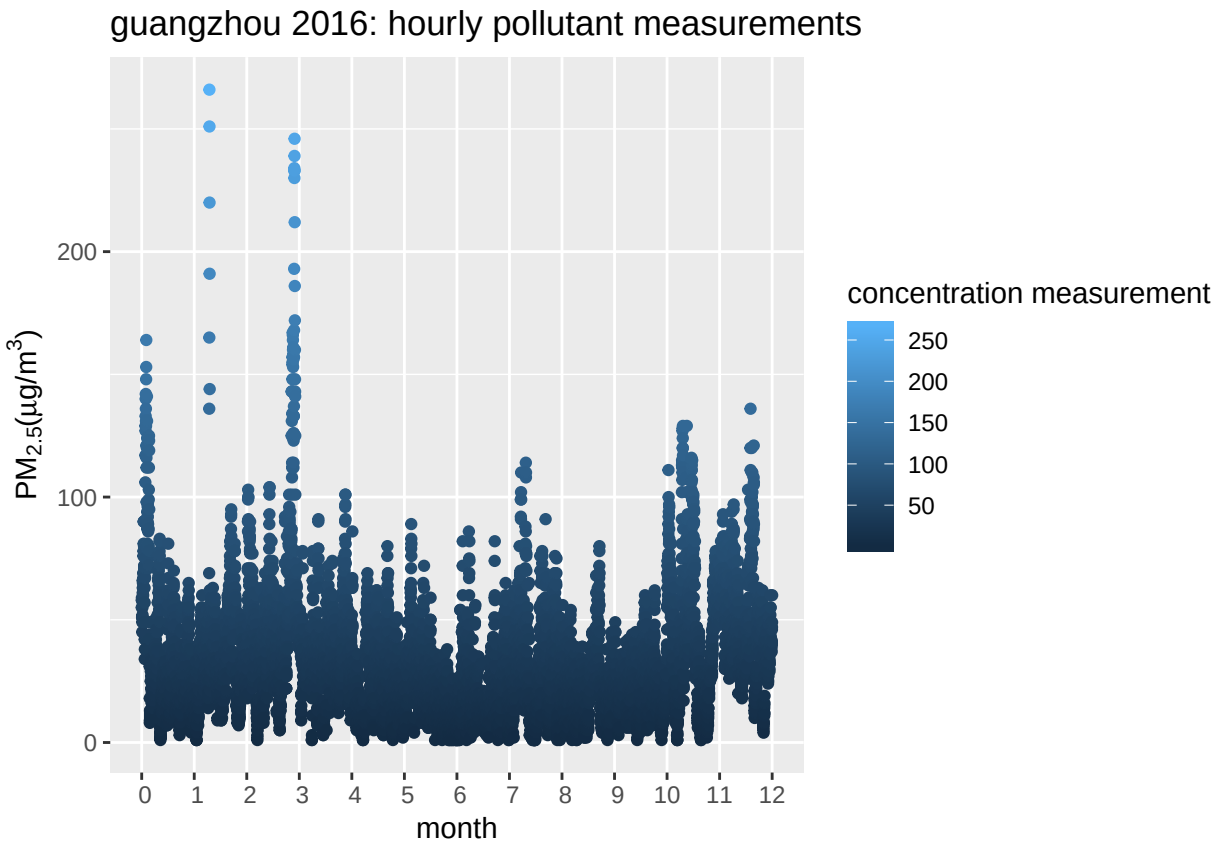

```
ggplot() + geom_point(aes(x = c(1:nrow(chengdu))/722, y = chengdu$Value, col = chengdu$Value)) + labs(t
```



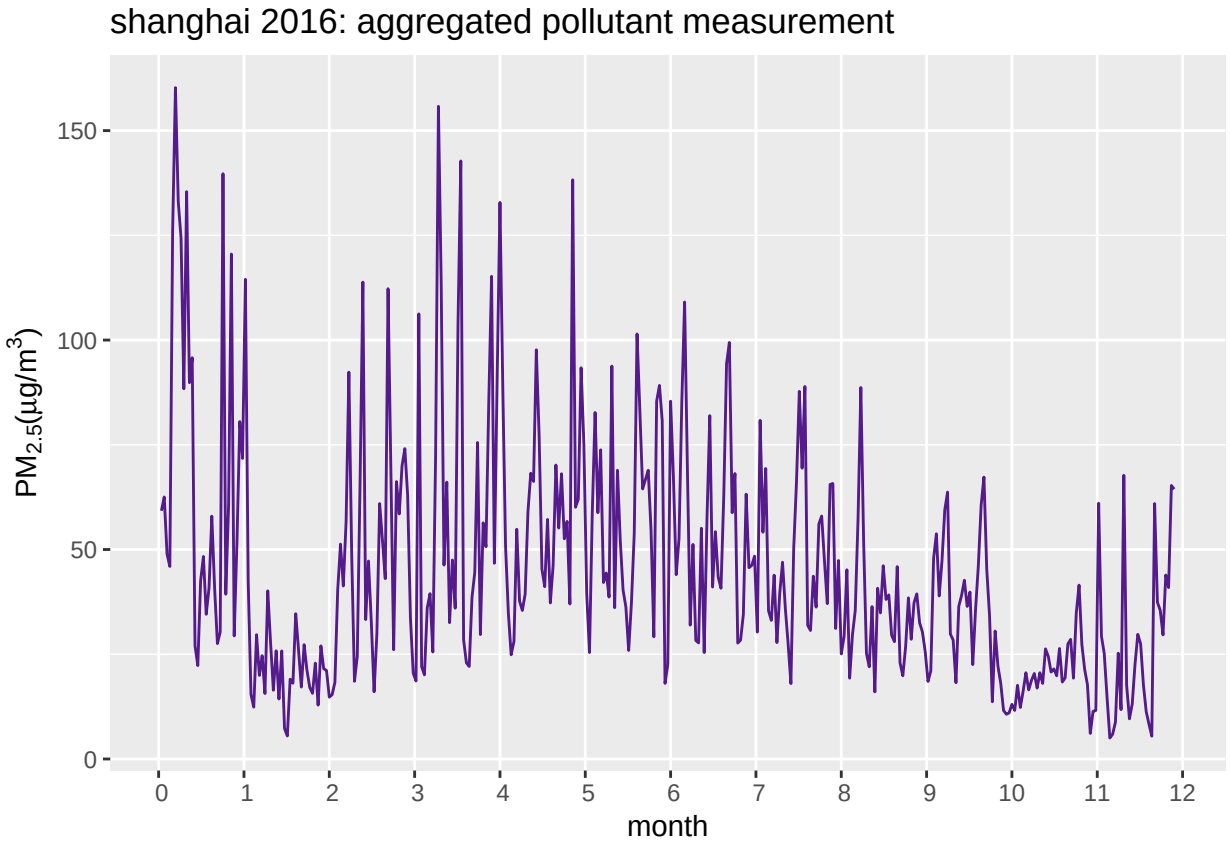

```
# GUANGZHOU

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



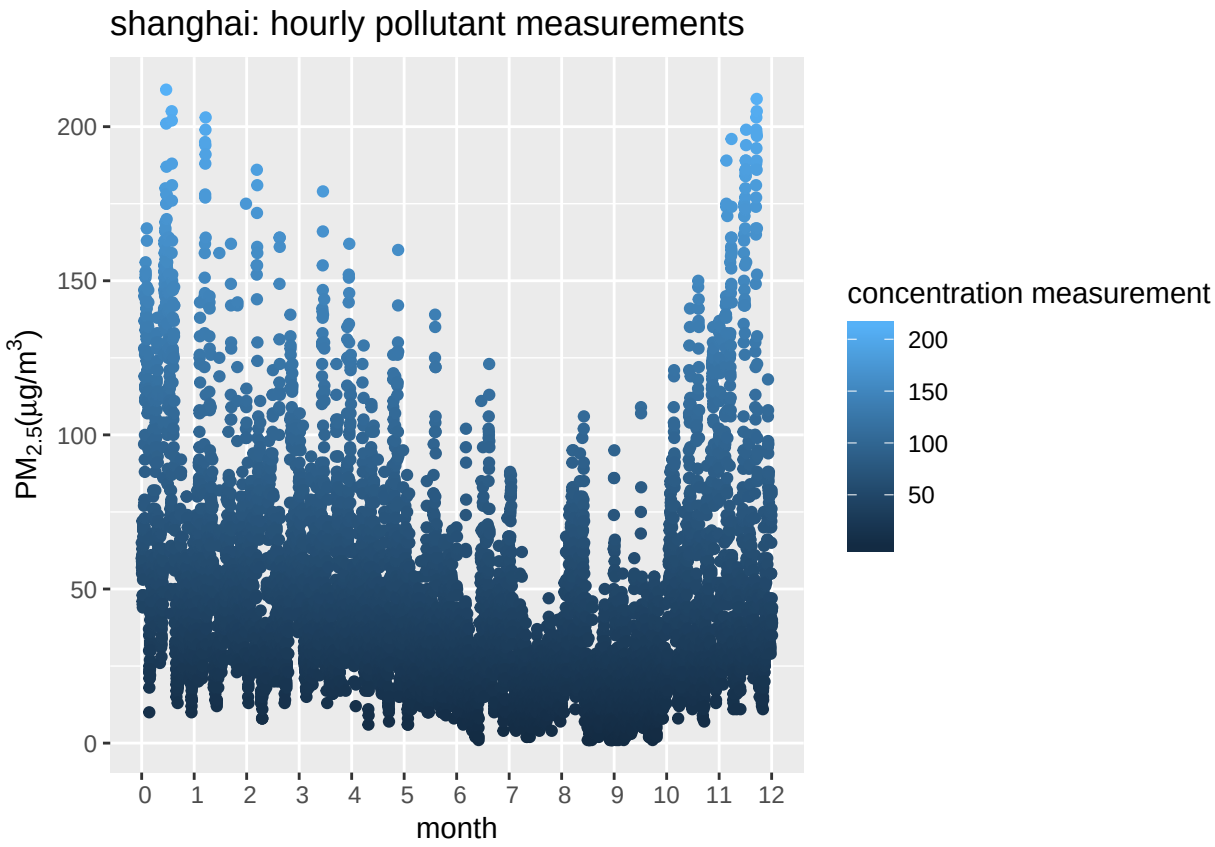

```
ggplot() + geom_point(aes(x = c(1:nrow(guangzhou))/676, y = guangzhou$Value, col = guangzhou$Value)) + l
```



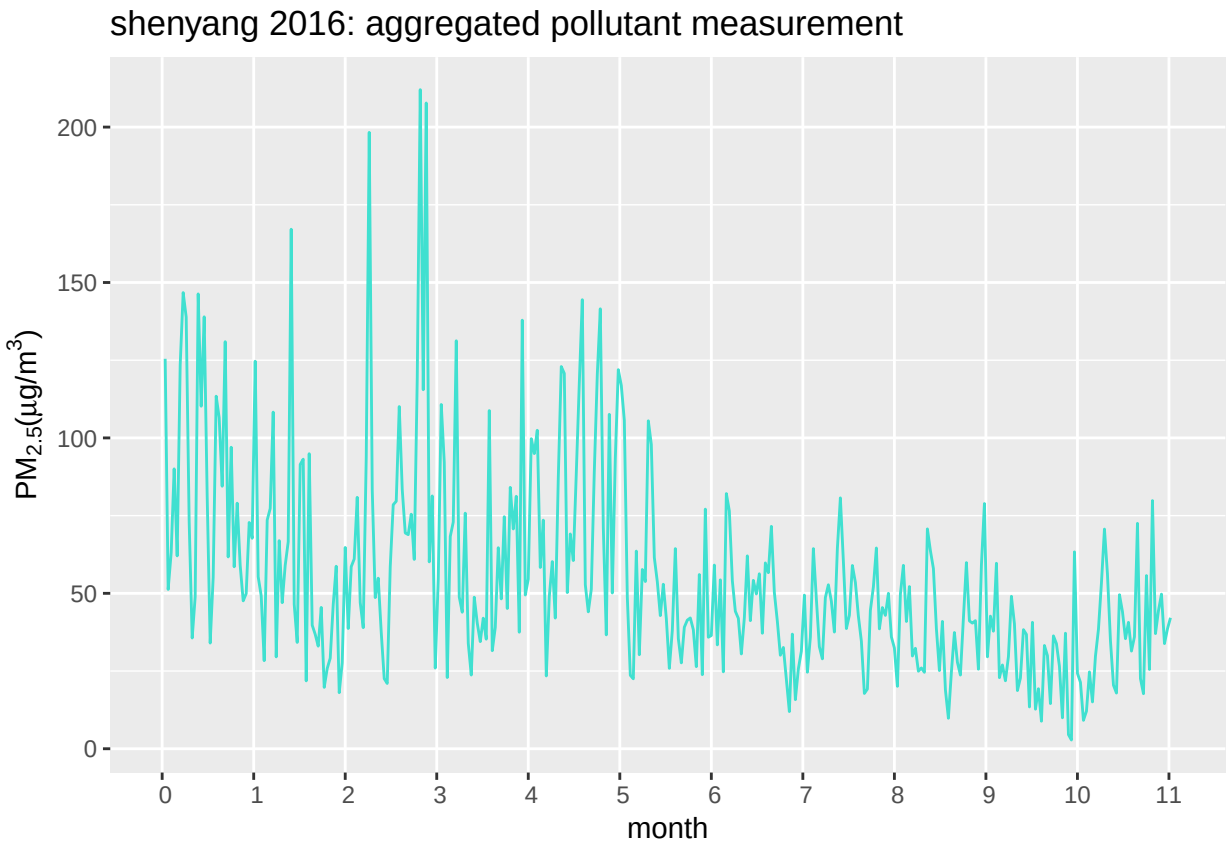

```
# SHANGHAI

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



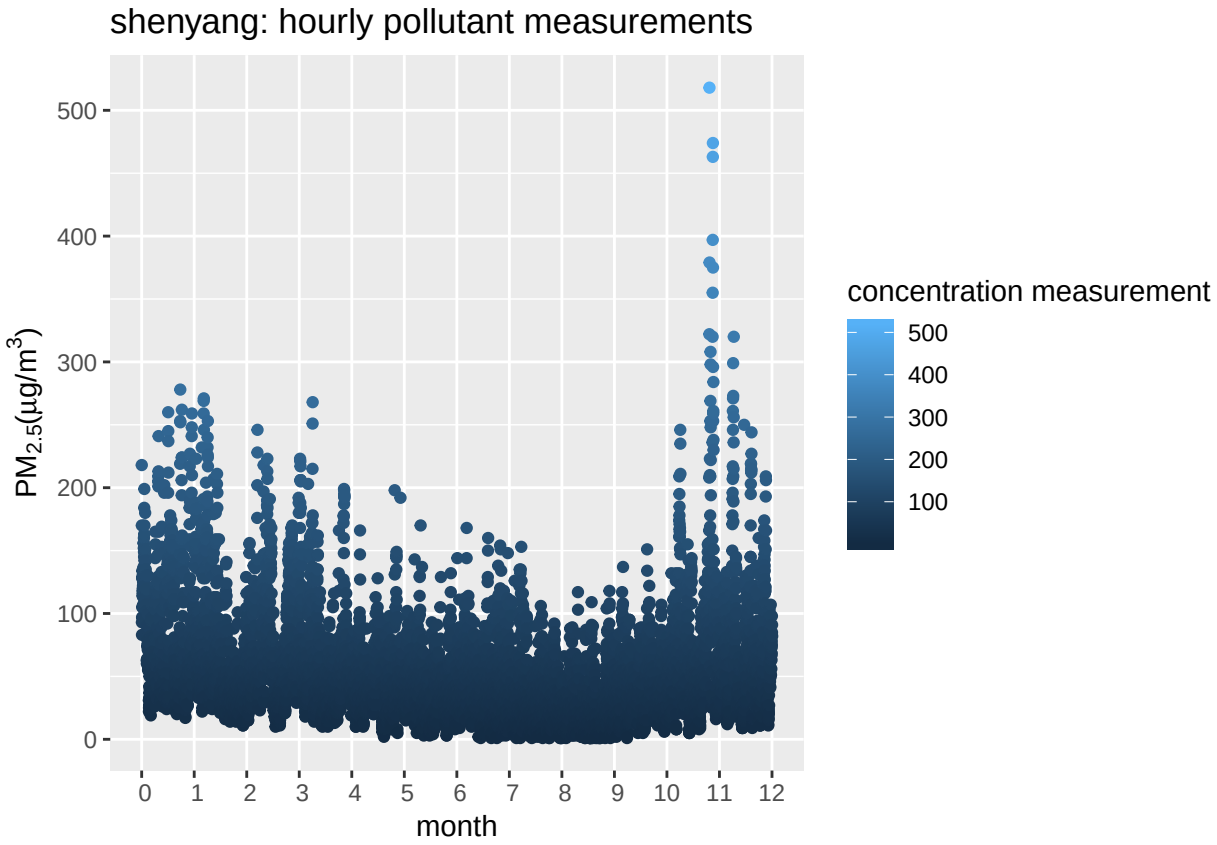

```
ggplot() + geom_point(aes(x = c(1:nrow(shanghai))/706, y = shanghai$Value, col = shanghai$Value)) + labs
```



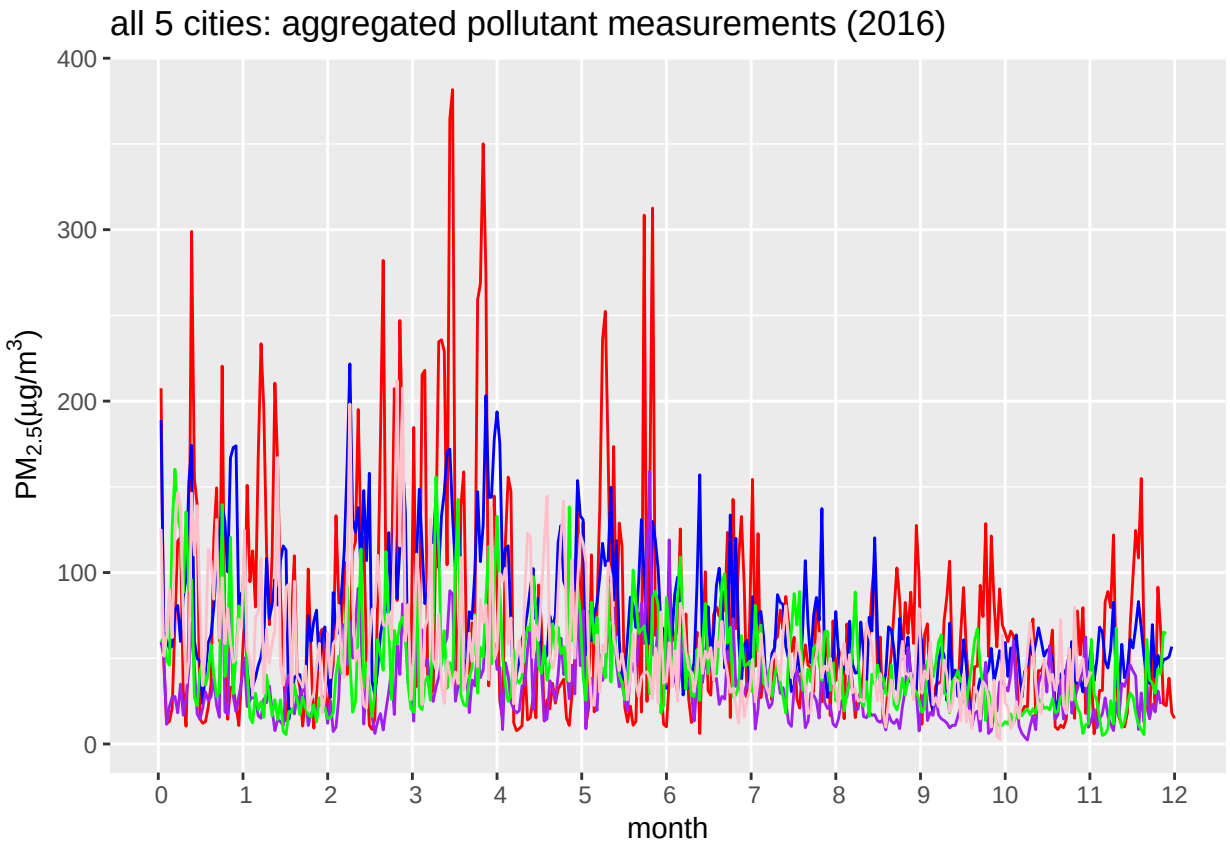

```
# SHENYANG

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
```



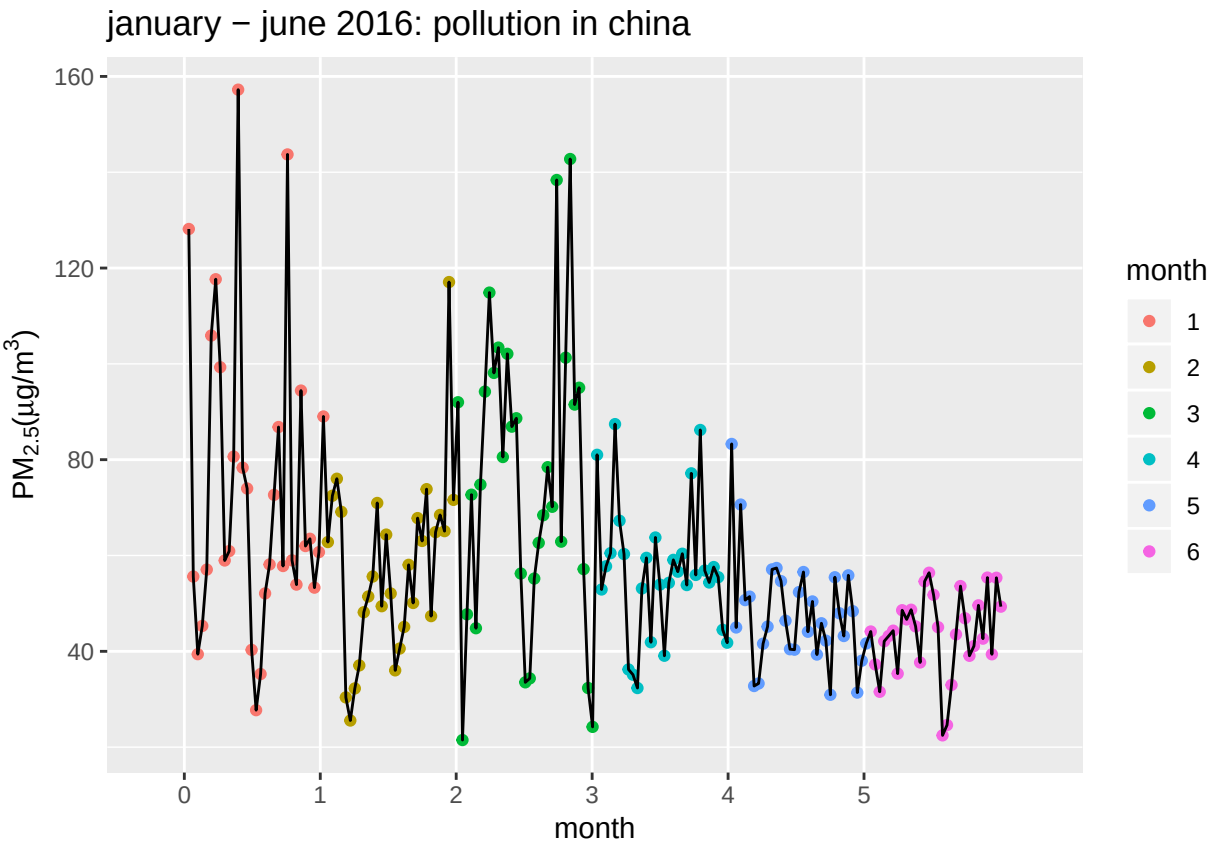

```
ggplot() +
  geom_line(aes(x = c(1:nrow(beijing_ag))/30.5, y = beijing_ag$mean_aqi), col = "red") +
  geom_line(aes(x = c(1:nrow(chengdu_ag))/30.5, y = chengdu_ag$mean_aqi), col = "blue") +
  geom_line(aes(x = c(1:nrow(guangzhou_ag))/30.5, y = guangzhou_ag$mean_aqi), col = "purple") +
  geom_line(aes(x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



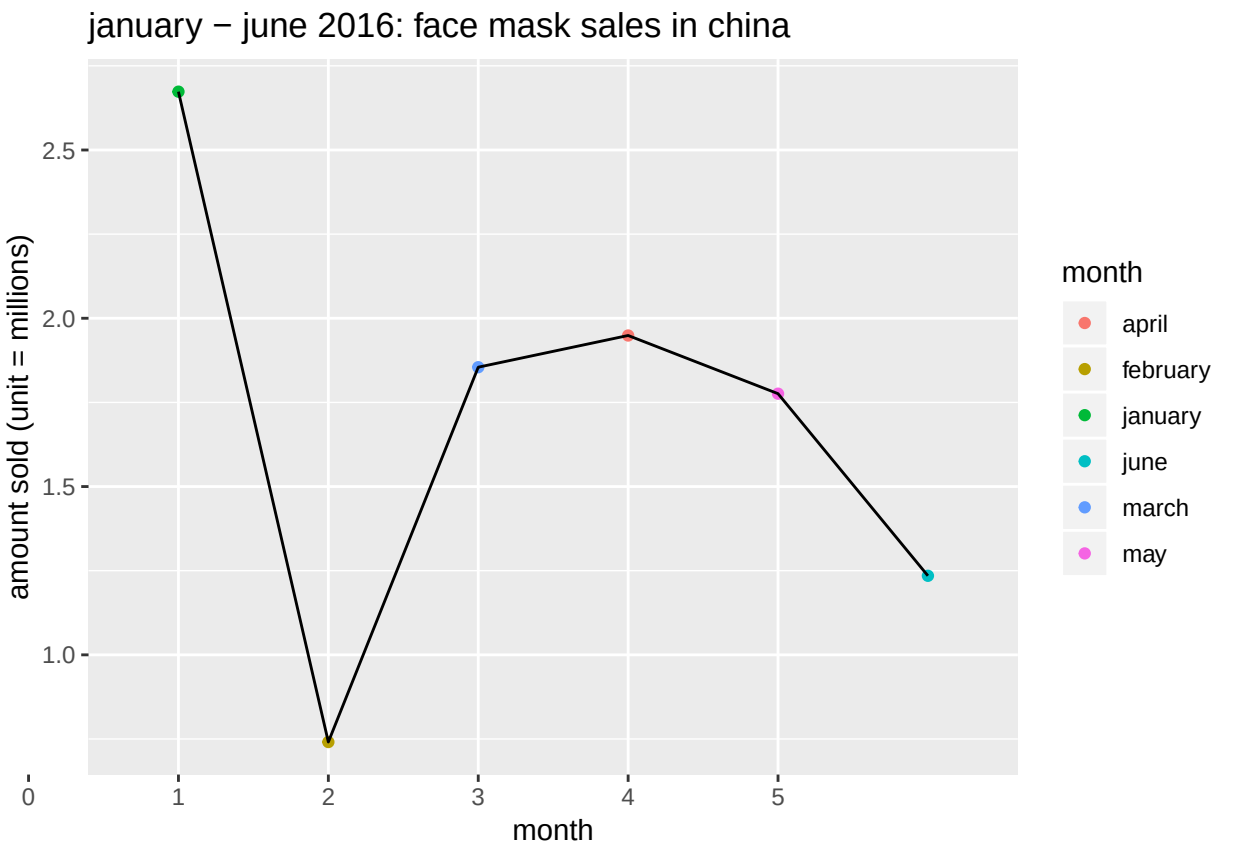

```
five_table <- rbind(beijing_ag, chengdu_ag, guangzhou_ag, shanghai_ag, shenyang_ag)

five_table$city <- rep(0, nrow(five_table))

length(chengdu_ag$Date)
```

```
## [1] 365
```

```
five_table$city[1:365] <- "beijing"
five_table$city[366:731] <- "chengdu"
five_table$city[732:1092] <- "guangzhou"
five_table$city[1093:1455] <- "shanghai"
five_table$city[1456:nrow(five_table)] <- "shenyang"

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)
```




```
?merge
```

```
## Help on topic 'merge' was found in the following packages:
##
##   Package               Library
##   data.table            /Library/Frameworks/R.framework/Versions/3.5/Resources/library
##   raster                /Library/Frameworks/R.framework/Versions/3.5/Resources/library
##   sp                    /Library/Frameworks/R.framework/Versions/3.5/Resources/library
##   base                  /Library/Frameworks/R.framework/Resources/library
##   lava                  /Library/Frameworks/R.framework/Versions/3.5/Resources/library
##
##
## Using the first match ...
```

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

```
## # A tibble: 366 x 3
##    Date    mean_aqi   moe
##    <chr>      <dbl> <dbl>
##  1 1/1/16     208.  88.8
##  2 1/10/16     47.5 22.0
##  3 1/11/16     11.8  2.91
##  4 1/12/16     13.1  5.81
##  5 1/13/16     21.0 13.3
```

```
##  6 1/14/16     71.6 33.8
##  7 1/15/16    118.  49.1
##  8 1/16/16    120.  24.8
##  9 1/17/16     31.1 34.3
## 10 1/18/16     10.4  5.66
## # ... with 356 more rows
```

```
cities <- merge(beijing_ag, chengdu_ag, by.x = "Date", by.y = "Date")
cities <- merge(cities, guangzhou_ag, by.x = "Date", by.y = "Date")

cities <- data.frame(
  Date = cities$Date,
  beijing = cities$mean_aqi.x,
  chengdu = cities$mean_aqi.y,
  guangzhou_ag = cities$mean_aqi
)

cities <- merge(cities, shanghai_ag, by.x = "Date", by.y = "Date")
cities <- merge(cities, shenyang_ag, by.x = "Date", by.y = "Date")

cities[1:10, -9]
```

```
##       Date   beijing   chengdu guangzhou_ag mean_aqi.x     moe.x
## 1   1/1/16 207.50000 188.95833     59.58333   59.25000  8.935761
## 2  1/10/16  47.50000  66.00000     50.75000   62.54167 10.496290
## 3  1/11/16  11.83333  60.54167     11.47826   48.91667 12.870312
## 4  1/12/16  13.12500  55.66667     21.70833   45.95833  9.493610
## 5  1/13/16  20.95833  57.25000     27.58333  126.09524 30.541619
## 6  1/14/16  71.58333  78.12500     27.66667  160.25000 19.460663
## 7  1/15/16 118.45833  80.81250     18.20000  133.08333 30.371492
## 8  1/16/16 120.41667  59.84615     34.91667  124.37500 13.454634
## 9  1/17/16  31.08333  85.58333     16.79167   88.37500 42.661267
## 10 1/18/16  10.37500  96.87500     26.29167  135.41667 25.286045
##    mean_aqi.y    moe.y
## 1   125.45833 26.97661
## 2    51.25000 24.49534
## 3    63.08333 41.03118
## 4    90.04167 51.72038
## 5    62.04167 36.66592
## 6   123.45833 65.50073
## 7   146.75000 24.10890
## 8   138.87500 16.18725
## 9    72.95833 33.26505
## 10   35.62500 11.26677
```

```
cities <- data.frame(
  Date = cities$Date,
  beijing = cities$beijing,
  chengdu = cities$chengdu,
  guangzhou = cities$guangzhou_ag,
  shanghai = cities$mean_aqi.x,
  shenyang = cities$mean_aqi.y
)
```

```r
#write.csv(cities, "cities.csv")


week <- c()
mult_7s <- c(0:46)

for ( i in 1: 46 ){
  week <- append(week, rep(mult_7s[i], 7))
}
week <- append(week, rep(47, 5))

cities$week <- week

bj <- summarise(
  group_by(cities[,c(1,2,7)], week),
  beijing = mean(beijing)
)

cd <- summarise(
  group_by(cities[,c(1,3,7)], week),
  chengdu = mean(chengdu)
)

gz <- summarise(
  group_by(cities[,c(1,4,7)], week),
  guangzhou = mean(guangzhou)
)

sha <- summarise(
  group_by(cities[,c(1,5,7)], week),
  shanghai = mean(shanghai)
)

she <- summarise(
  group_by(cities[,c(1,6,7)], week),
  shenyang = mean(shenyang)
)

cities_week <- data.frame(
  week = she$week,
  beijing = bj$beijing,
  chengdu = cd$chengdu,
  guangzhou = gz$guangzhou,
  shanghai = sha$shanghai,
  shenyang = she$shenyang
)

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

cities_week$week <- cities_week$week/3.9


#write.csv(cities_week, "cities_week_mo.csv")
```

```
event_days <- str_extract(events$timestamp, pattern = "2016-[0-9][0-9]-[0-9][0-9]")

unique_dates <- names(table(event_days))
unique_dates <- unique_dates[2:8]


event_dates <- c("4/30/16", "5/1/16", "5/2/16", "5/3/16", "5/4/16", "5/5/16", "5/6/16", "5/7/16", "5/8/
event_dates <- event_dates[2:8]

date_indices_b <- NULL
date_indices_c <- NULL
date_indices_g <- NULL
date_indices_sha <- NULL
date_indices_she <- NULL

for (i in 1:nrow(beijing_ag)){
  if (beijing_ag$Date[i] %in% event_dates){
    date_indices_b[i] <- i
  } else{
    date_indices_b[i] <- 0
  }
}

for (i in 1:nrow(chengdu_ag)){
  if (chengdu_ag$Date[i] %in% event_dates){
    date_indices_c[i] <- i
  } else{
    date_indices_c[i] <- 0
  }
}

for (i in 1:nrow(guangzhou_ag)){
  if (guangzhou_ag$Date[i] %in% event_dates){
    date_indices_g[i] <- i
  } else{
    date_indices_g[i] <- 0
  }
}

for (i in 1:nrow(shanghai_ag)){
  if (shanghai_ag$Date[i] %in% event_dates){
    date_indices_sha[i] <- i
  } else{
    date_indices_sha[i] <- 0
  }
}

for (i in 1:nrow(shenyang_ag)){
  if (shenyang_ag$Date[i] %in% event_dates){
    date_indices_she[i] <- i
  } else{
    date_indices_she[i] <- 0
  }
```

```
}
extract_b <- date_indices_b[date_indices_b != 0]
events_beijing <- beijing_ag[extract_b, ]
events_beijing$day <- unique_dates
events_beijing <- events_beijing[c(2:8), ]


extract_c <- date_indices_c[date_indices_c != 0]
events_chengdu <- chengdu_ag[extract_c, ]
events_chengdu$day <- unique_dates
events_chengdu <- events_chengdu[c(2:8), ]

extract_g <- date_indices_g[date_indices_g != 0]
events_guangzhou <- guangzhou_ag[extract_g, ]
events_guangzhou$day <- unique_dates
events_guangzhou <- events_guangzhou[c(2:8), ]

extract_sha <- date_indices_sha[date_indices_sha != 0]
events_shanghai <- shanghai_ag[extract_sha, ]
events_shanghai$day <- unique_dates
events_shanghai <- events_shanghai[c(2:8), ]

extract_she <- date_indices_sha[date_indices_sha != 0]
events_shenyang <- shenyang_ag[extract_she, ]
events_shenyang$day <- unique_dates
events_shenyang <- events_shenyang[c(2:8), ]


events$day <- event_days

time1 <- Sys.time()



# uhhhhh
for (i in 1:100){
  if (events$day[i] %in% events_beijing$day){
    events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day == events$day[i]]
  } else{
    events$beijing[i] <- 0
  }
}
```

```
## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
```

```
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
```

```
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
```

```
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
```

```
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
```

```
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length

## Warning in events$beijing[i] <- events_beijing$mean_aqi[events_beijing$day
## == : number of items to replace is not a multiple of replacement length
```

```
Sys.time() - time1
```

```
## Time difference of 8.463037 secs
```

```
# do this differently, assign all to zero
# index out matching date values
# only go thru those


events_beijing$mean_aqi[events_beijing$day == events$day[1]]
```

```
## [1] 61.95833       NA
```

```
t <- c(1:nrow(events))[events$day == events_beijing$day[2]]
unique_dates
```

```
## [1] "2016-05-01" "2016-05-02" "2016-05-03" "2016-05-04" "2016-05-05"
## [6] "2016-05-06" "2016-05-07"
```

```
events$date <- str_extract(events$timestamp, pattern = "[0-9]+-[0-9]+-[0-9]+")

event1 <- events[1:500000,]

event2 <- events[500001:1000000,]

event3 <- events[1000001:1500000,]

event4 <- events[1500001:2000000,]

event5 <- events[2000001:2500000,]

event5 <- events[2500001:3032372,]


dates_aqis <- data.frame(
  day = unique_dates,
  aqi_beijing = events_beijing$mean_aqi,
  aqi_chengdu = events_chengdu$mean_aqi,
  aqi_guangzhou = events_guangzhou$mean_aqi,
  aqi_shanghai = events_shanghai$mean_aqi,
  aqi_shenyang = events_shenyang$mean_aqi
)

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


randomized_index <- sample(1:nrow(events), 100000, replace = FALSE)

randomized_events <- events[randomized_index, ]

unique_dates
```

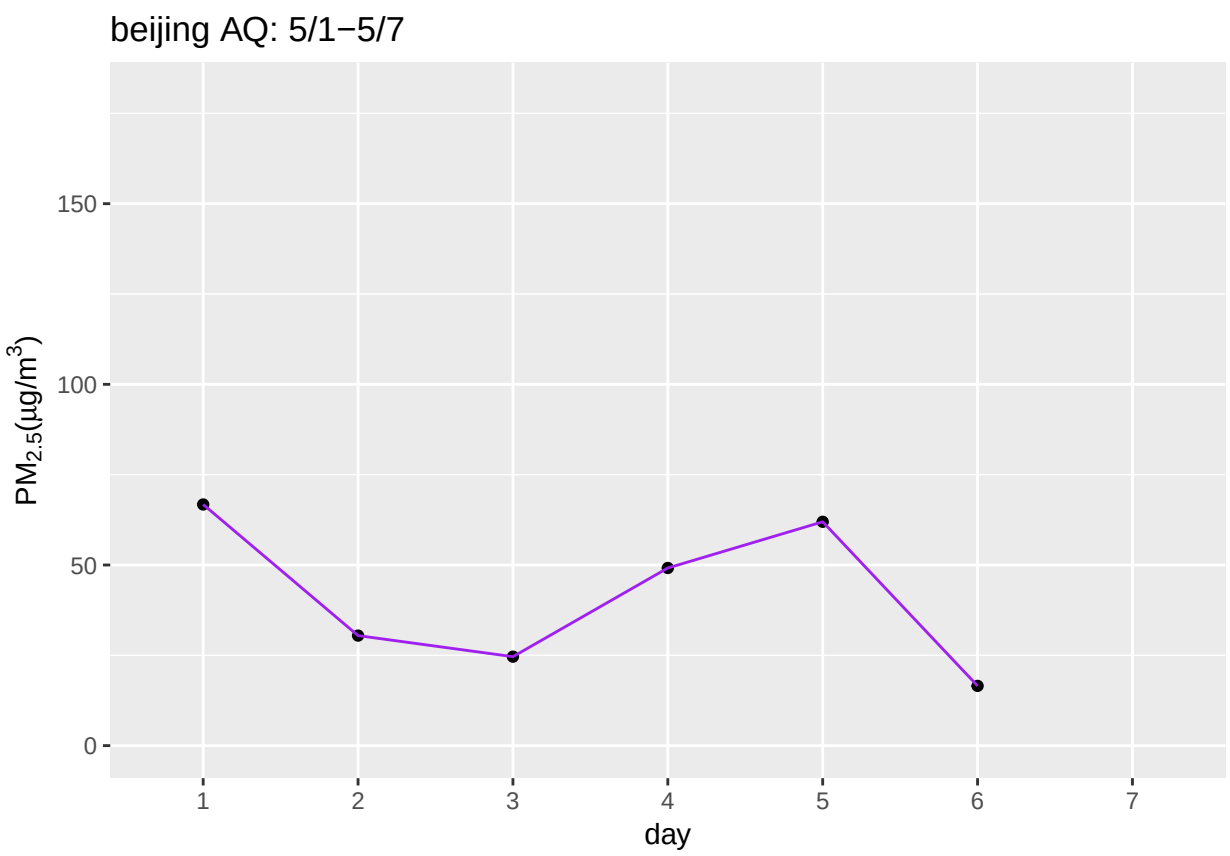
```
## [1] "2016-05-01" "2016-05-02" "2016-05-03" "2016-05-04" "2016-05-05"
## [6] "2016-05-06" "2016-05-07"
```

```
events_short <- events[ ,c(1, 3,4,5,9)]

apr_30 <- events_short[events$day == unique_dates[1], ]
may_1 <- events_short[events$day == unique_dates[2], ]
may_2 <- events_short[events$day == unique_dates[3], ]
may_3 <- events_short[events$day == unique_dates[4], ]
may_4 <- events_short[events$day == unique_dates[5], ]
may_5 <- events_short[events$day == unique_dates[6], ]
may_6 <- events_short[events$day == unique_dates[7], ]
may_7 <- events_short[events$day == unique_dates[8], ]
```

```r
may_8 <- events_short[events$day == unique_dates[9], ]



#write.csv(may_1, "may_1.csv")
#write.csv(may_2, "may_2.csv")
#write.csv(may_3, "may_3.csv")
#write.csv(may_4, "may_4.csv")
#write.csv(may_5, "may_5.csv")
#write.csv(may_6, "may_6.csv")
#write.csv(may_7, "may_7.csv")


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



#write.csv(randomized_events, "randomized_events.csv")
ggplot() + geom_point(aes(x = c(1:nrow(events_beijing)), y = events_beijing$mean_aqi)) + geom_line(aes(
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

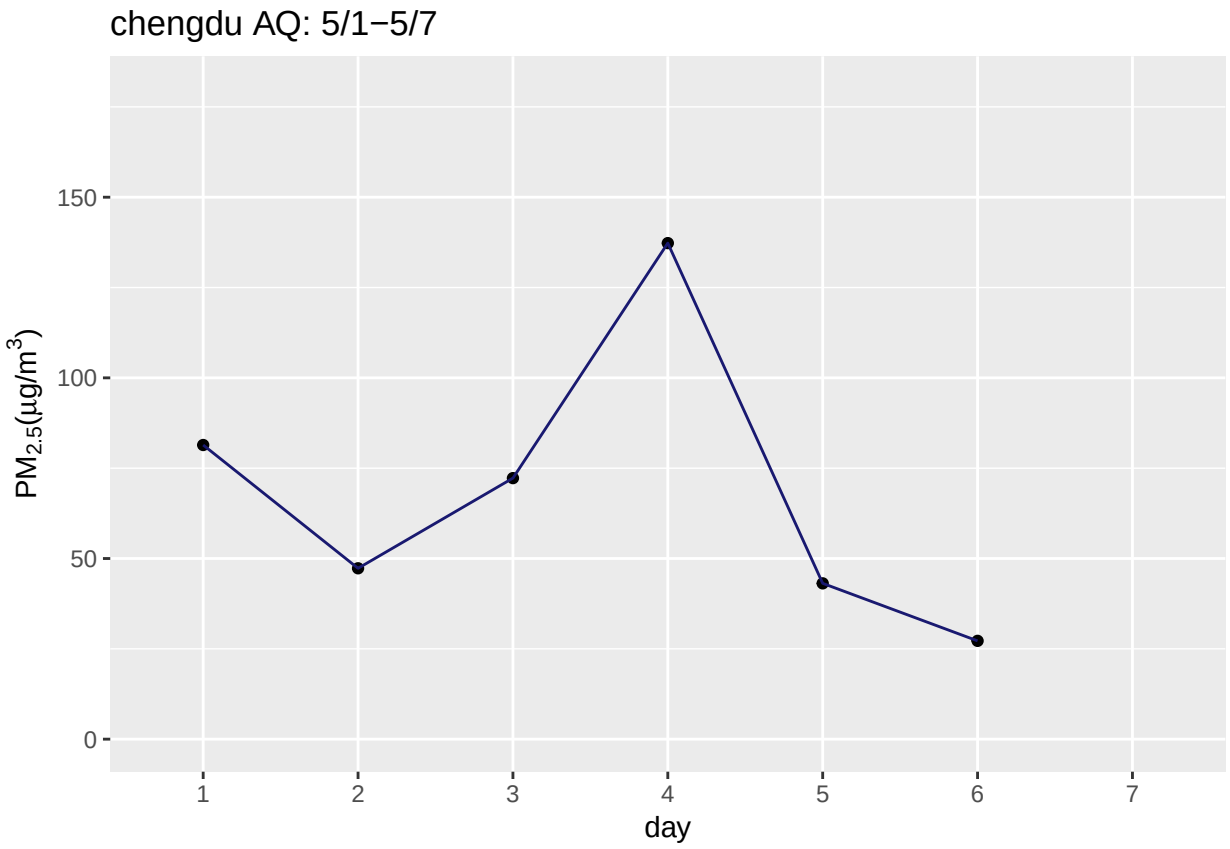
```
## Warning: Removed 1 rows containing missing values (geom_path).
```



```
ggplot() + geom_point(aes(x = c(1:nrow(events_chengdu)), y = events_chengdu$mean_aqi)) + geom_line(aes(
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

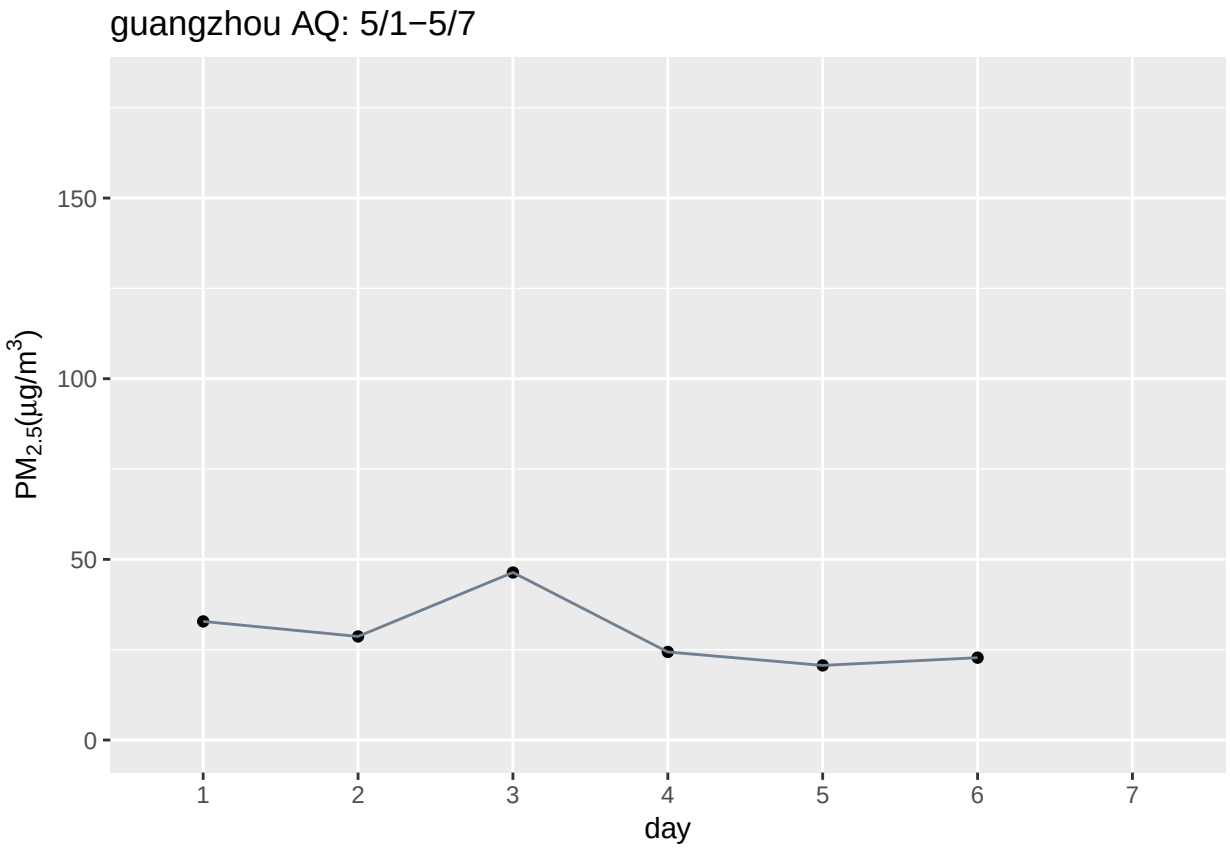
```
## Warning: Removed 1 rows containing missing values (geom_path).
```




```
ggplot() + geom_point(aes(x = c(1:nrow(events_guangzhou)), y = events_guangzhou$mean_aqi)) + geom_line(a
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

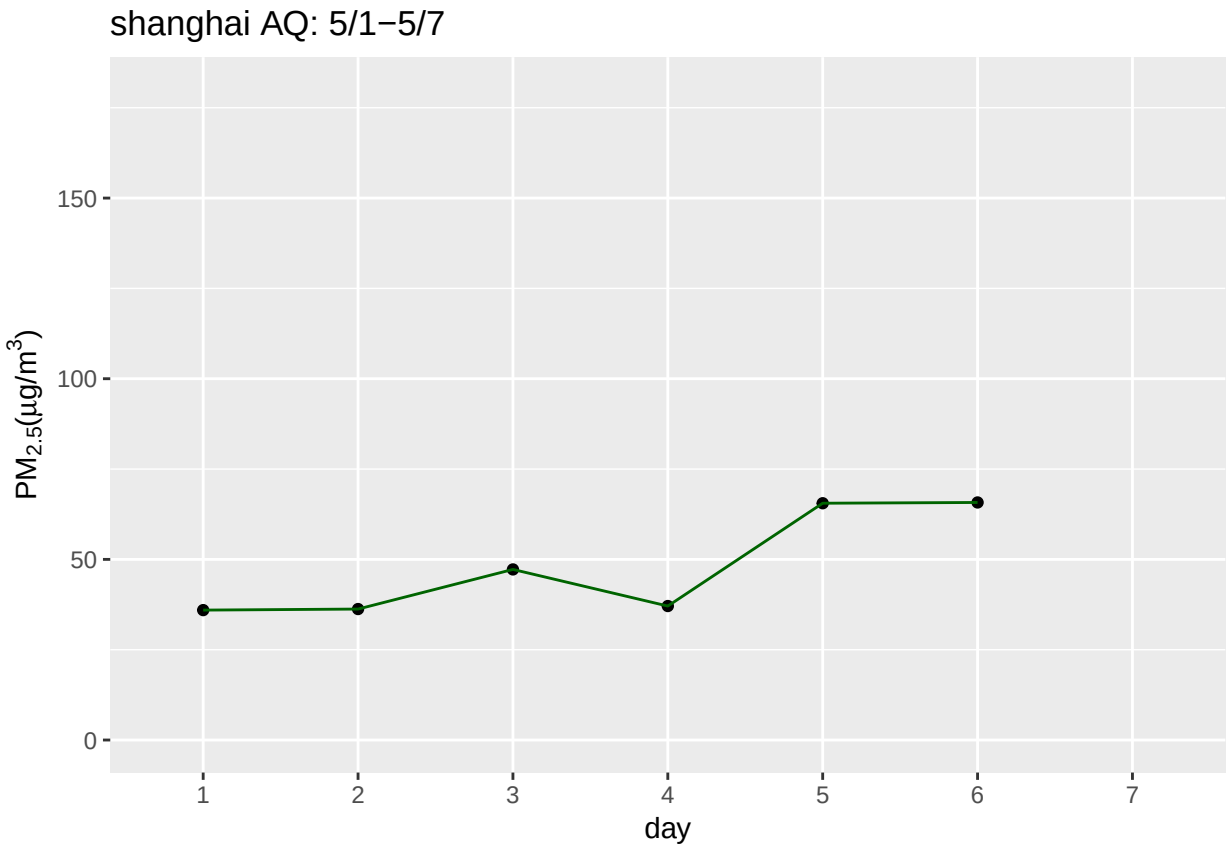
```
## Warning: Removed 1 rows containing missing values (geom_path).
```




```
ggplot() + geom_point(aes(x = c(1:nrow(events_shanghai)), y = events_shanghai$mean_aqi)) + geom_line(aes
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

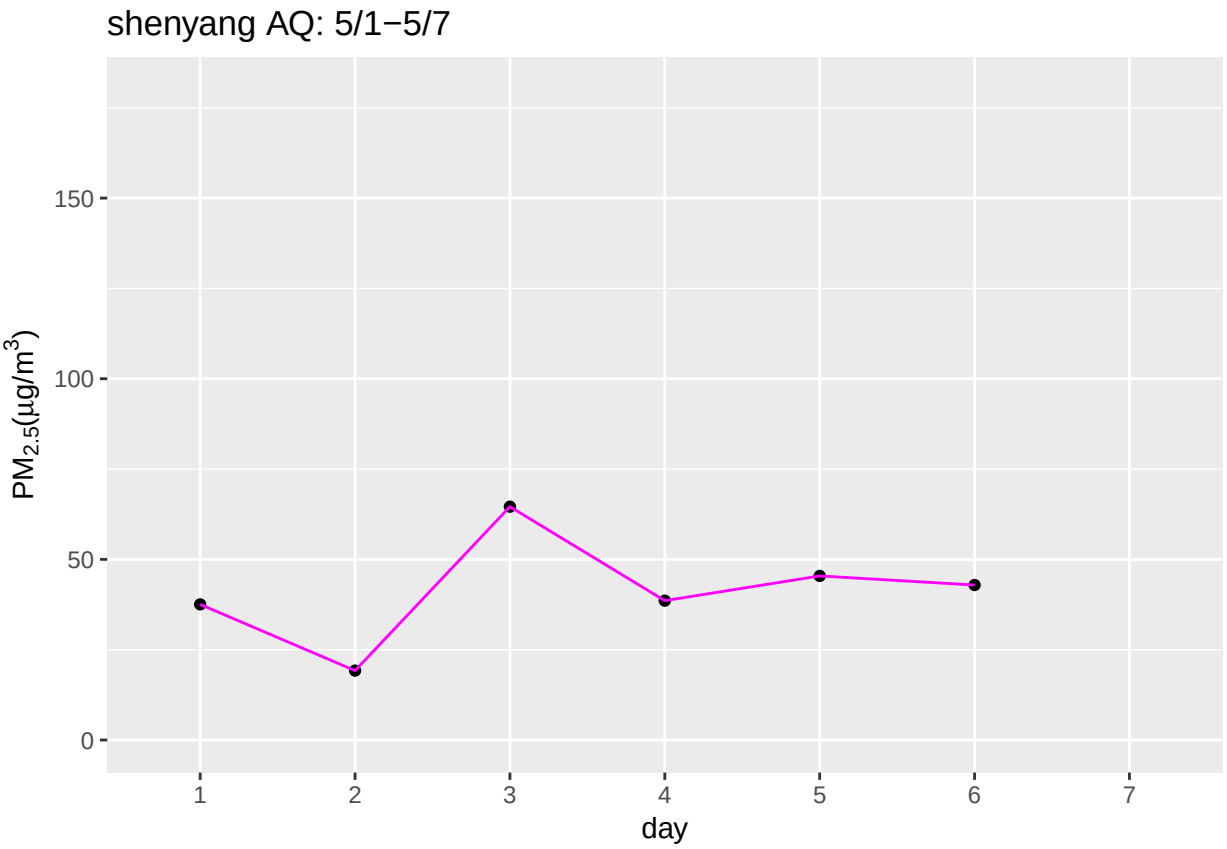
```
## Warning: Removed 1 rows containing missing values (geom_path).
```




```
ggplot() + geom_point(aes(x = c(1:nrow(events_shenyang)), y = events_shenyang$mean_aqi)) + geom_line(aes
```

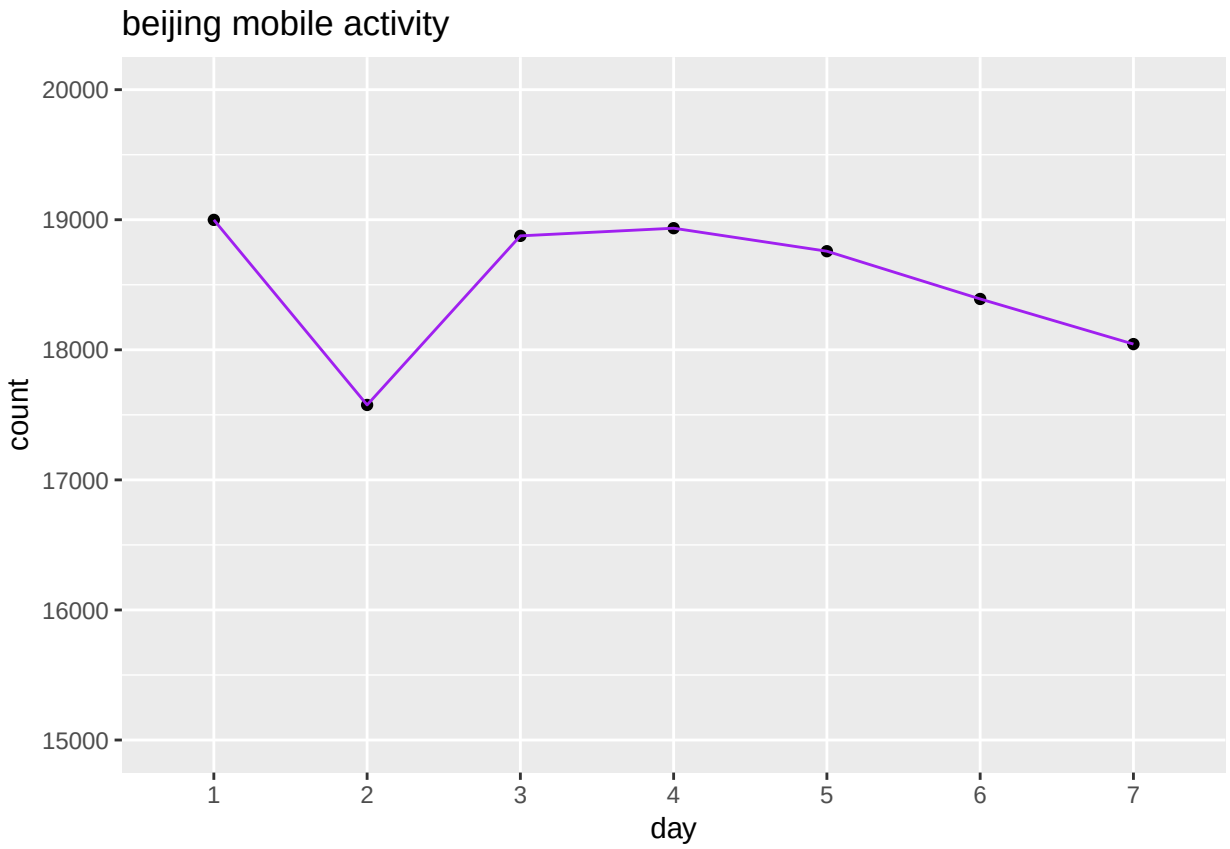
```
## Warning: Removed 1 rows containing missing values (geom_point).
```

```
## Warning: Removed 1 rows containing missing values (geom_path).
```

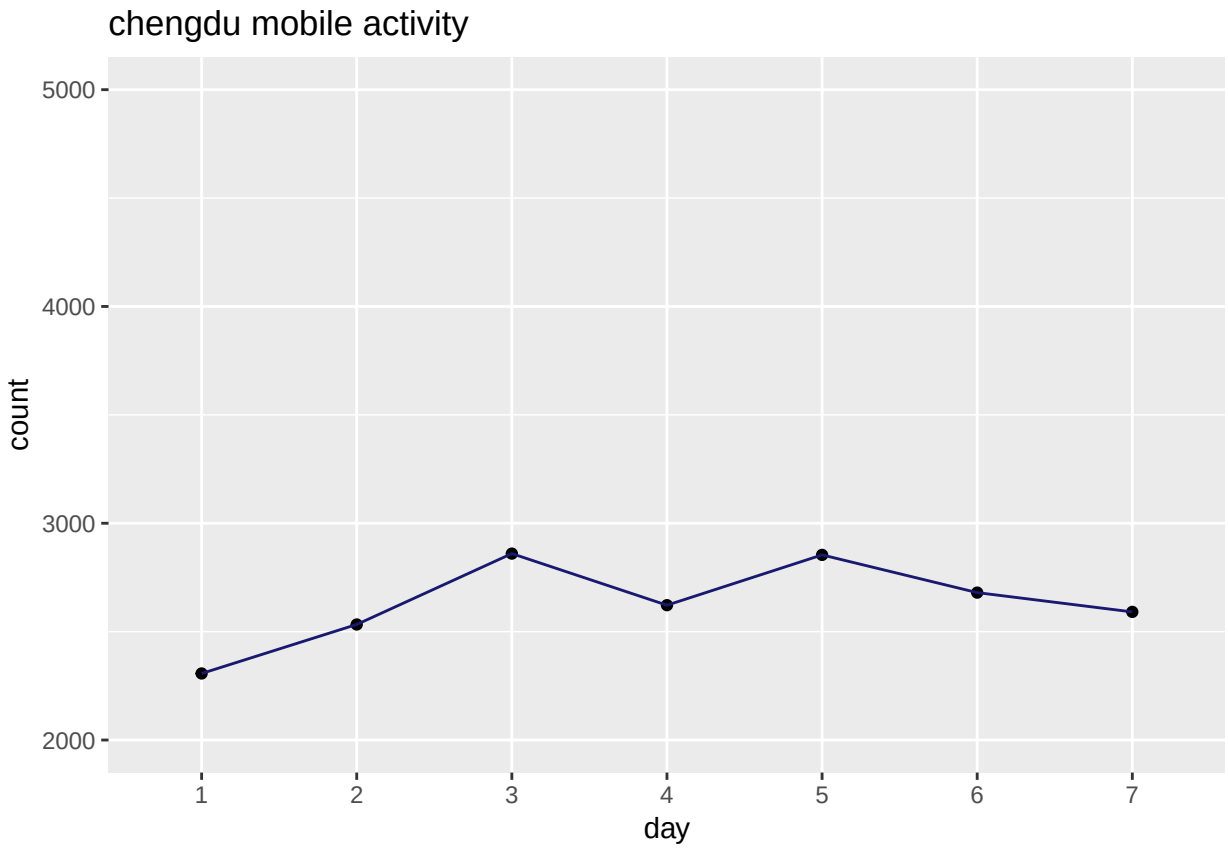


ACTIVITY PLOTS BY CITY

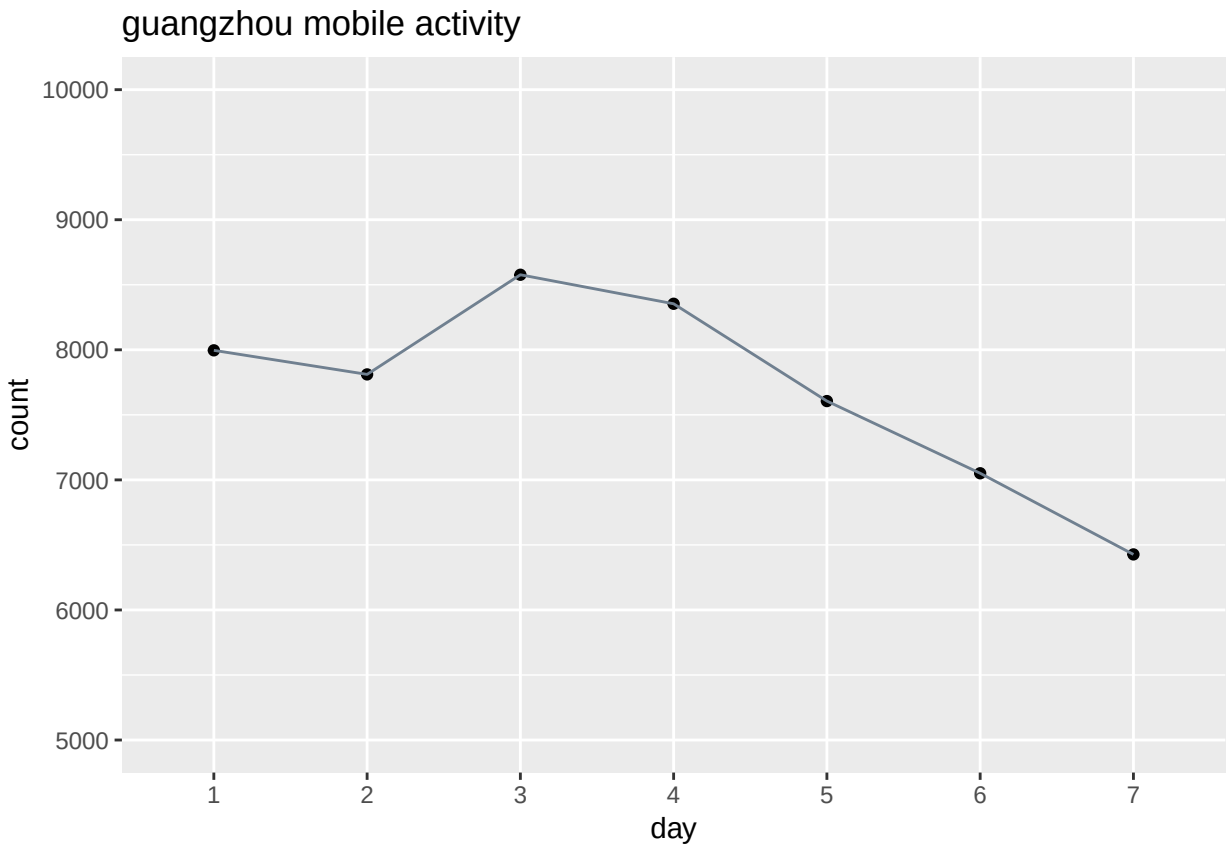
```
activity <- read.csv("~/Documents/caL/2019/cyplan101/projects/assignment3/activity_counts.csv")

ggplot() + geom_point(aes(x = c(1:7), y = activity$beijing)) + geom_line(aes(x = c(1:7), y = activity$be
```



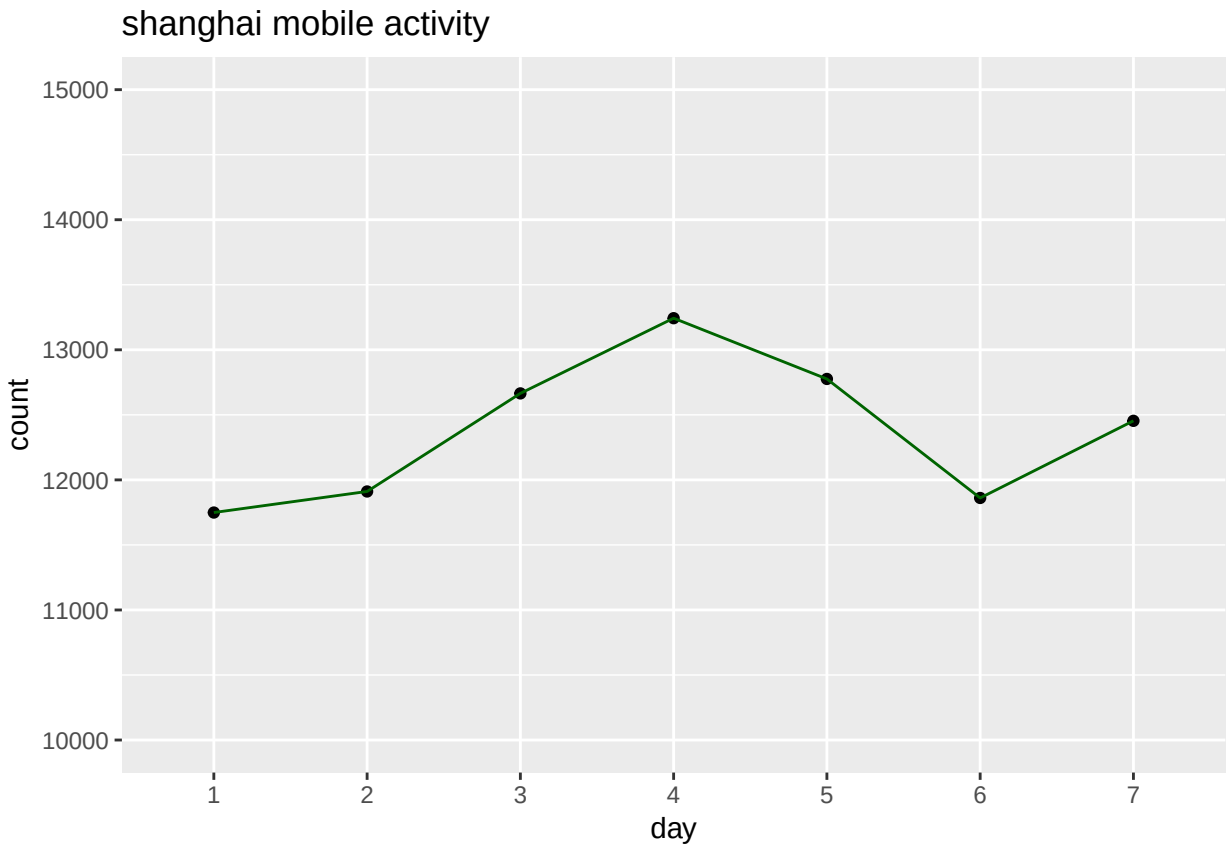

```
ggplot() + geom_point(aes(x = c(1:7), y = activity$chengdu)) + geom_line(aes(x = c(1:7), y = activity$cl
```



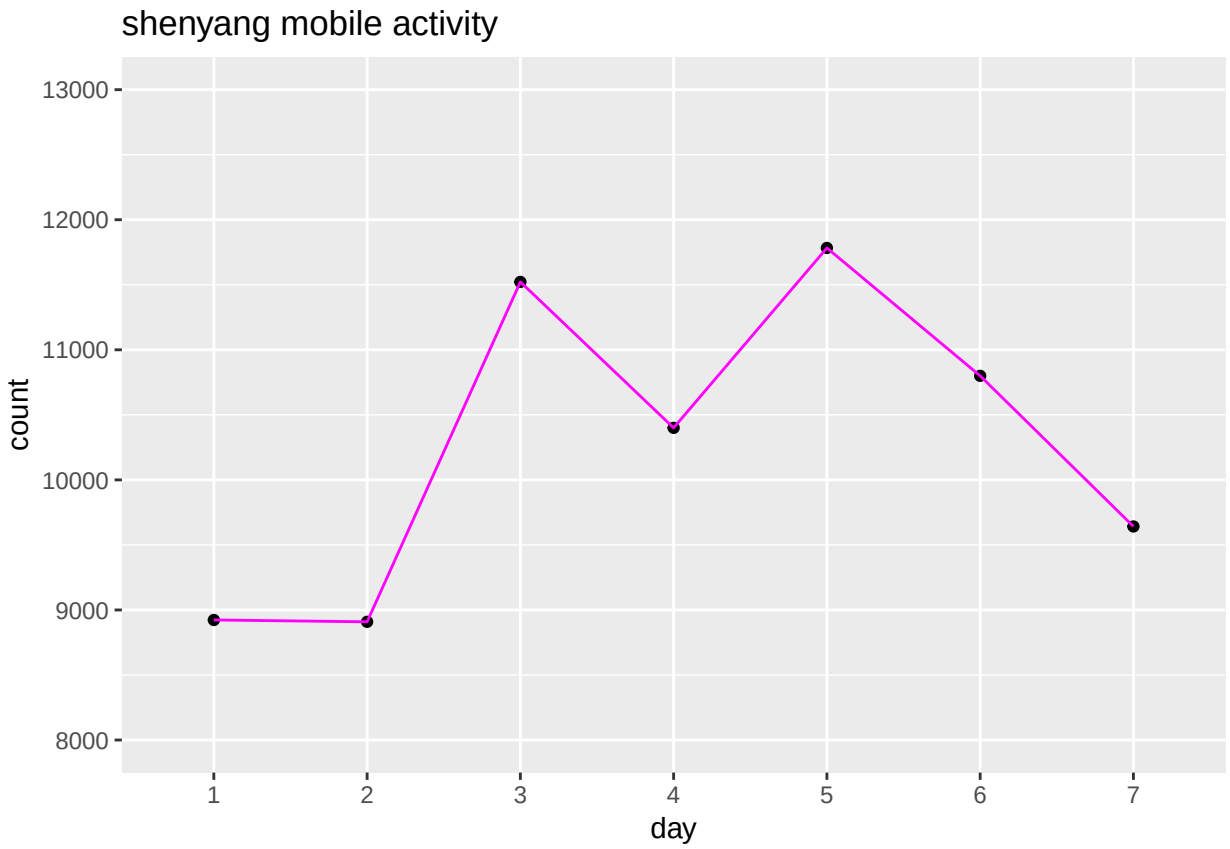

```
ggplot() + geom_point(aes(x = c(1:7), y = activity$guangzhou)) + geom_line(aes(x = c(1:7), y = activity$
```




```
ggplot() + geom_point(aes(x = c(1:7), y = activity$shanghai)) + geom_line(aes(x = c(1:7), y = activity$s
```




```
ggplot() + geom_point(aes(x = c(1:7), y = activity$shenyang)) + geom_line(aes(x = c(1:7), y = activity$
```



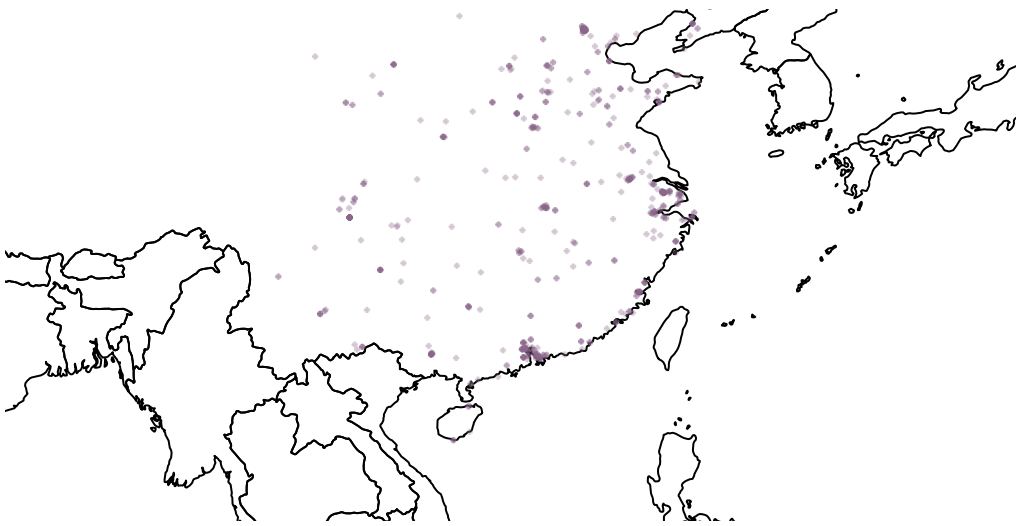
ACTIVITY ANALYSIS

```
activity
```

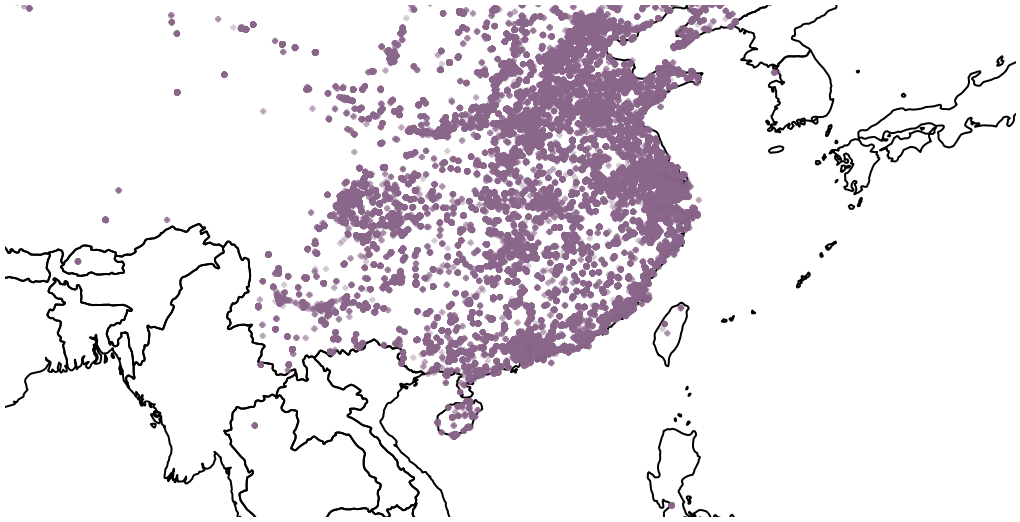
```
##    date beijing shenyang shanghai guangzhou chengdu
## 1 1-May   18999     8923    11749      7996    2307
## 2 2-May   17576     8909    11911      7811    2533
## 3 3-May   18876    11522    12665      8577    2860
## 4 4-May   18935    10400    13243      8354    2622
## 5 5-May   18758    11783    12776      7606    2854
## 6 6-May   18391    10800    11861      7051    2680
## 7 7-May   18044     9642    12454      6427    2591
```

PLOTS BY DAY

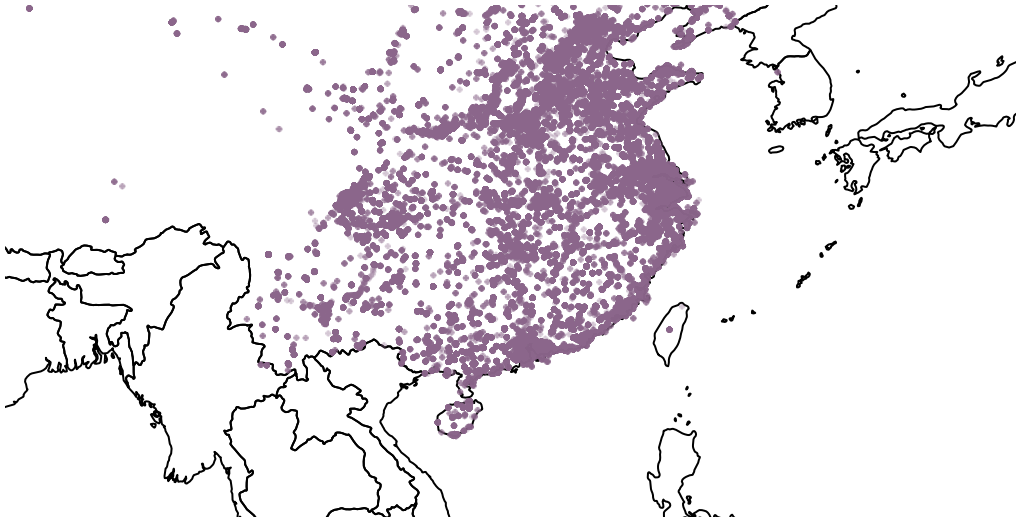
```
# APRIL 30
newmap <- getMap(resolution = "low")
plot(newmap, xlim = c(105, 120), ylim = c(15, 40), asp = 1, main = "mobile activity: april 30")
points(apr_30$longitude, apr_30$latitude, col = alpha("plum4", 0.2), cex = 0.3, pch = 10)
```



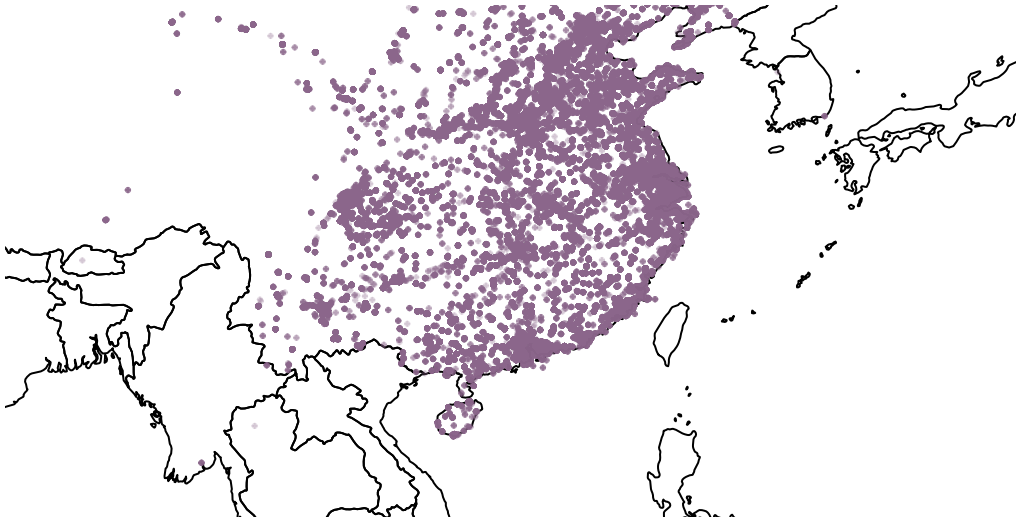

```
# MAY 1
plot(newmap, xlim = c(105, 120), ylim = c(15, 40), asp = 1, main = "mobile activity: may 1")
points(may_1$longitude, may_1$latitude, col = alpha("plum4", 0.2), cex = 0.3, pch = 10)
```



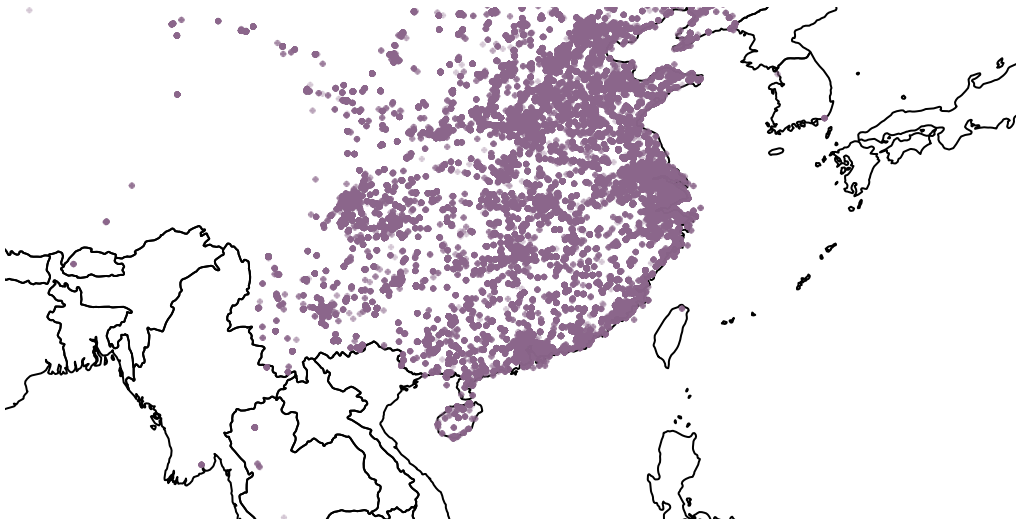

```
# MAY 2
plot(newmap, xlim = c(105, 120), ylim = c(15, 40), asp = 1, main = "mobile activity: may 2")
points(may_2$longitude, may_2$latitude, col = alpha("plum4", 0.2), cex = 0.3, pch = 10)
```



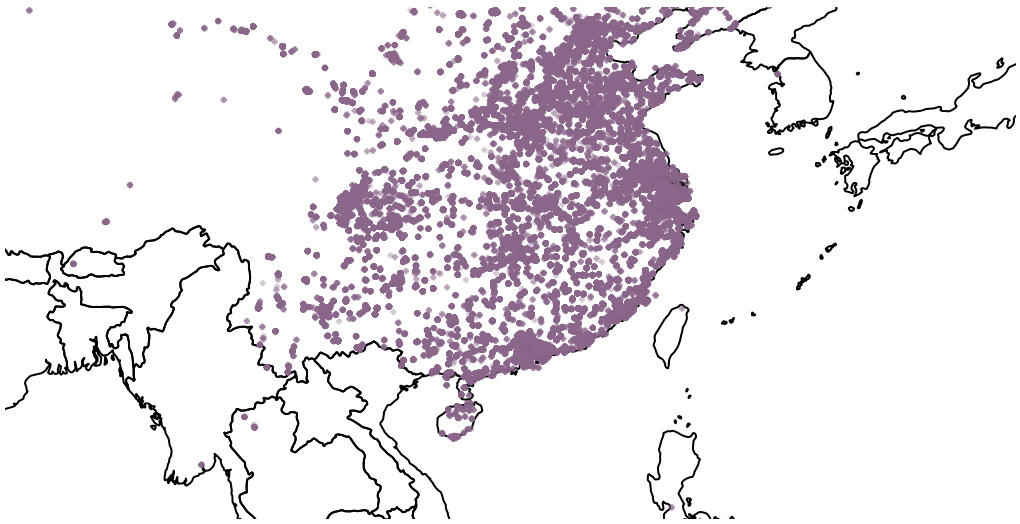

```
# MAY 3
plot(newmap, xlim = c(105, 120), ylim = c(15, 40), asp = 1, main = "mobile activity: may 3")
points(may_3$longitude, may_3$latitude, col = alpha("plum4", 0.2), cex = 0.3, pch = 10)
```



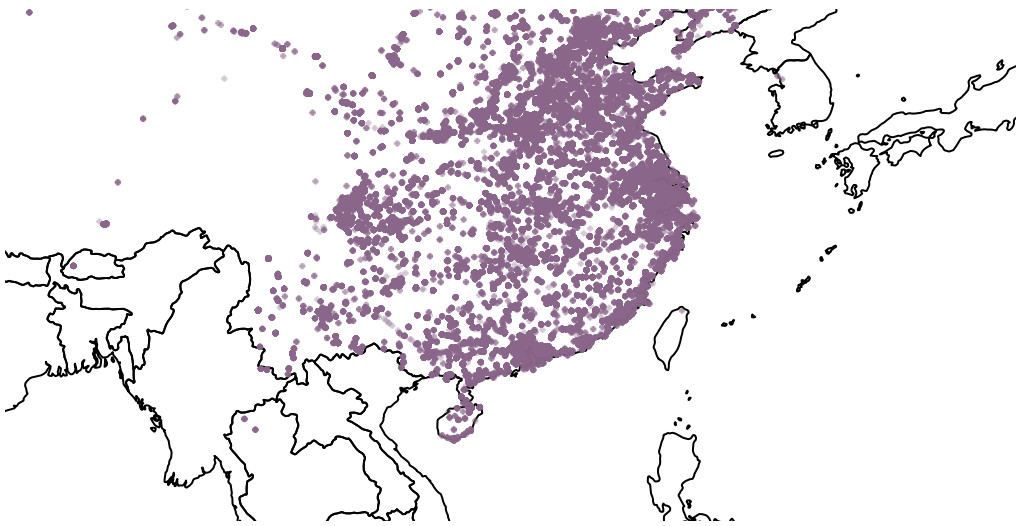

```
# MAY 4
plot(newmap, xlim = c(105, 120), ylim = c(15, 40), asp = 1, main = "mobile activity: may 4")
points(may_4$longitude, may_4$latitude, col = alpha("plum4", 0.2), cex = 0.3, pch = 10)
```



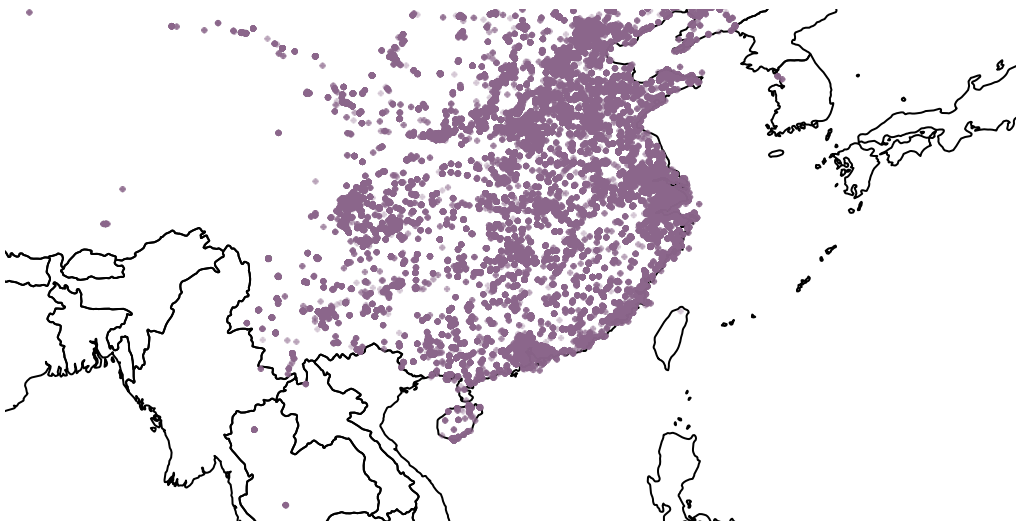

```
# MAY 5
plot(newmap, xlim = c(105, 120), ylim = c(15, 40), asp = 1, main = "mobile activity: may 5")
points(may_5$longitude, may_5$latitude, col = alpha("plum4", 0.2), cex = 0.3, pch = 10)
```




```
# MAY 6
plot(newmap, xlim = c(105, 120), ylim = c(15, 40), asp = 1, main = "mobile activity: may 6")
points(may_6$longitude, may_6$latitude, col = alpha("plum4", 0.2), cex = 0.3, pch = 10)
```




```
# MAY 7
plot(newmap, xlim = c(105, 120), ylim = c(15, 40), asp = 1, main = "mobile activity: may 7")
points(may_7$longitude, may_7$latitude, col = alpha("plum4", 0.2), cex = 0.3, pch = 10)
```




```
# MAY 8
plot(newmap, xlim = c(105, 120), ylim = c(15, 40), asp = 1, main = "mobile activity: may 8")
points(may_8$longitude, may_8$latitude, col = alpha("plum4", 0.2), cex = 0.3, pch = 10)
```

mobile activity: may 8

NEED GOOGLE API KEY FOR THIS MAP

```
register_google(key = 'AIzaSyAj98ZG3Tt8xFIdx78HRi3MMb1LRB_ZYfQ', "standard", client = "uc-berkeley-urba

has_google_key()
```

```
china <- ggmap(get_googlemap(center = c(lon = 112.363625, lat = 32),
                    zoom = 5, scale = 2,
                    maptype ='roadmap',
                    color = 'color'))
```

```
## Source : https://maps.googleapis.com/maps/api/staticmap?center=32,112.363625&zoom=5&size=640x640&scal
```

```
spatial_provinces <- readOGR("Archive/provinces")
```

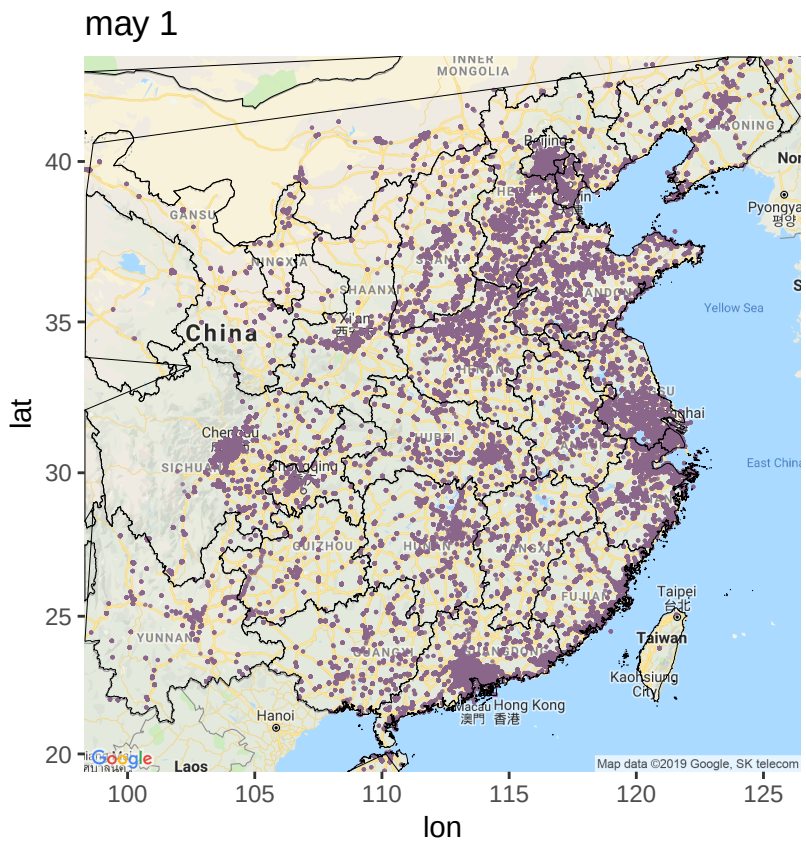
```
## OGR data source with driver: ESRI Shapefile
## Source: "/Users/pet/Documents/caL/2019/cyplan101/projects/assignment3/Archive/provinces", layer: "nyu
## with 925 features
## It has 6 fields
```

```
spatial_provinces <- spTransform(spatial_provinces, CRS("+proj=longlat +datum=WGS84"))

provinces <- fortify(spatial_provinces)
```

```
## Regions defined for each Polygons
```

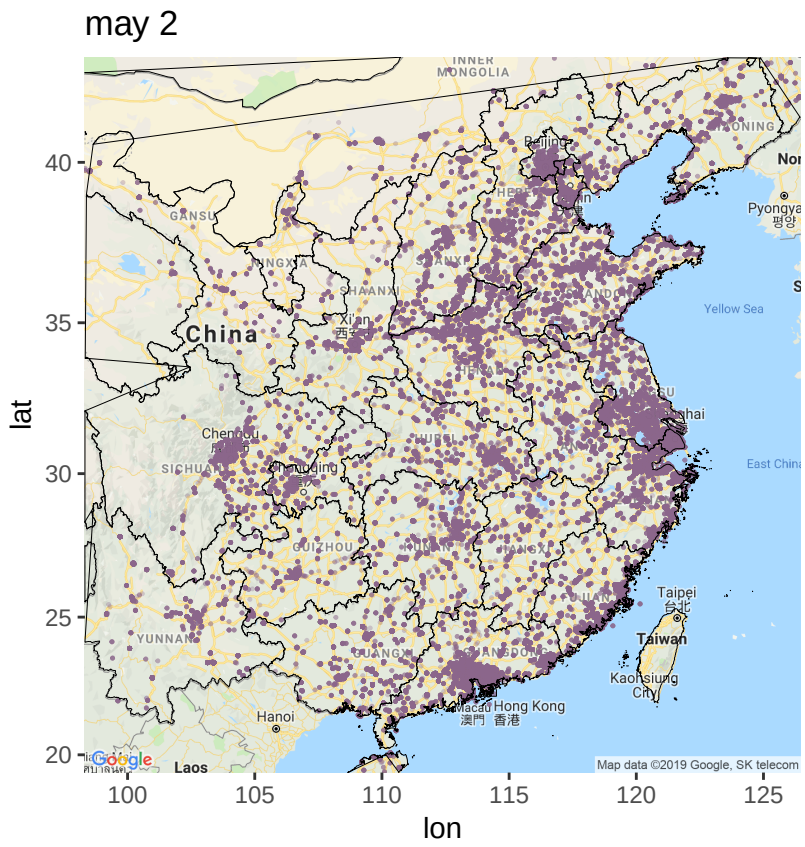
```
five_prov <- provinces[provinces$id %in% c(11, 31, 52, 45, 21), ]

china + geom_point(aes(x = longitude, y = latitude), data = may_1, size = 0.1, col = "plum4", alpha = 0
```

```
## Warning: Removed 133581 rows containing missing values (geom_point).
```



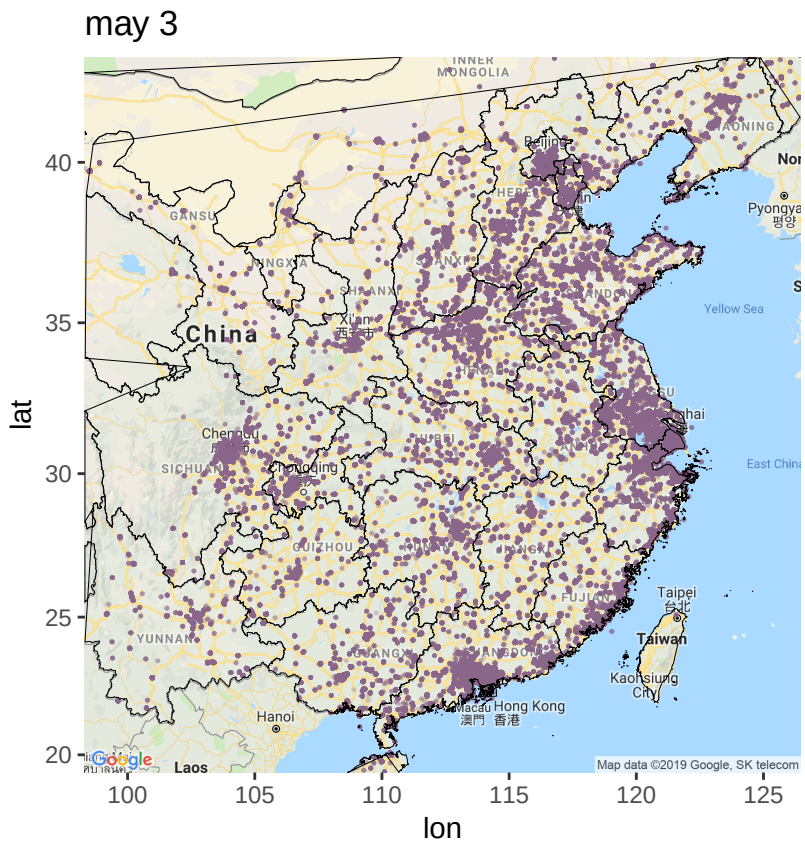

```
china + geom_point(aes(x = longitude, y = latitude), data = may_2, size = 0.2, col = "plum4", alpha = 0
```

```
## Warning: Removed 143134 rows containing missing values (geom_point).
```



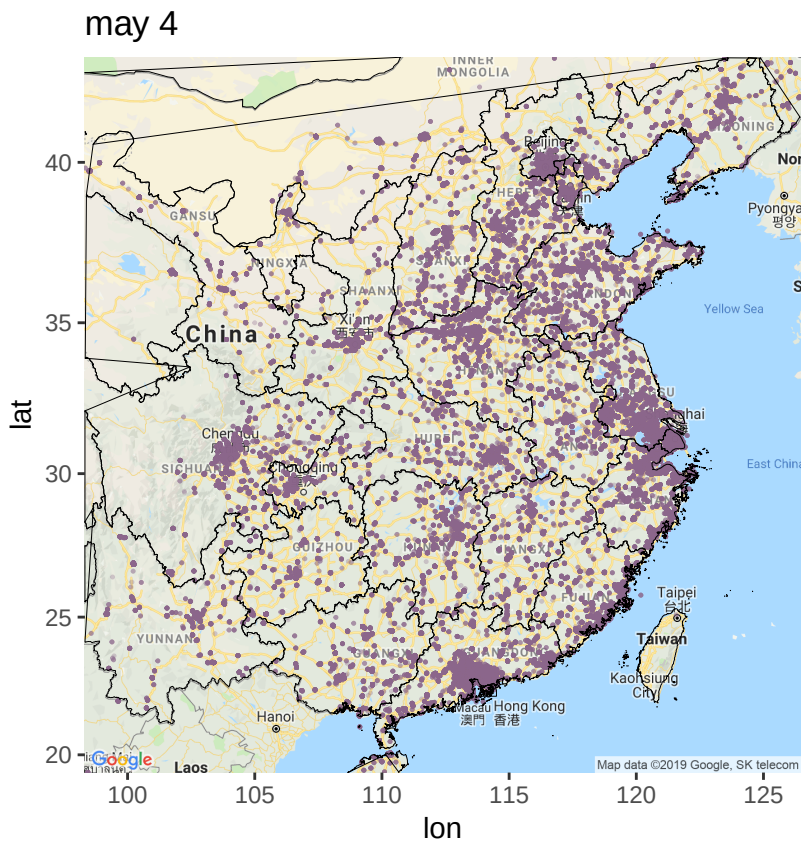

```
china + geom_point(aes(x = longitude, y = latitude), data = may_3, size = 0.2, col = "plum4", alpha = 0
```

```
## Warning: Removed 139702 rows containing missing values (geom_point).
```



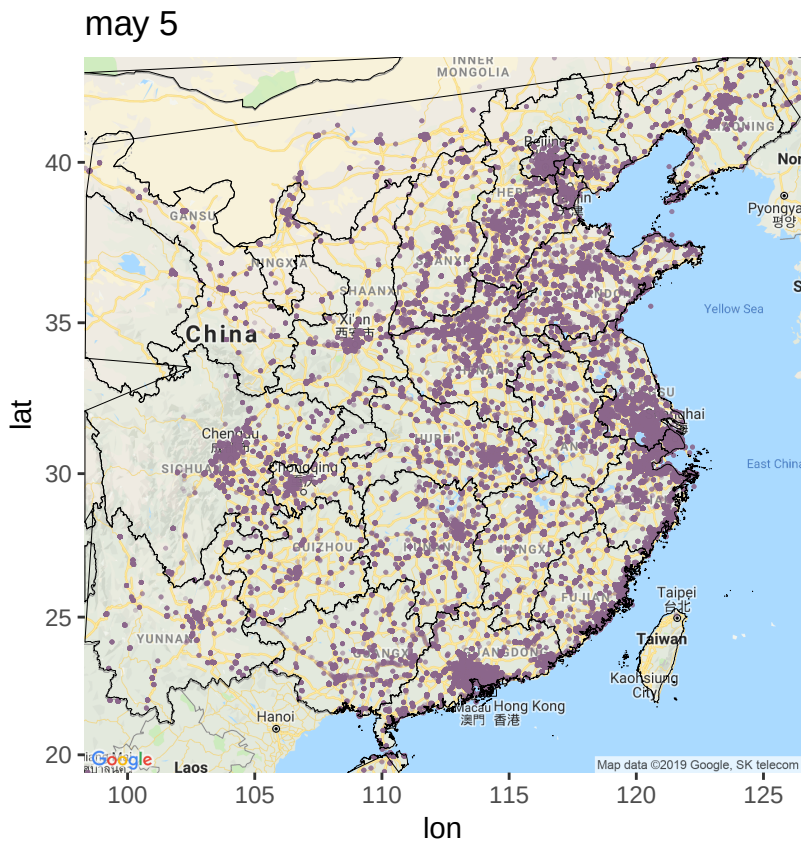

```
china + geom_point(aes(x = longitude, y = latitude), data = may_4, size = 0.2, col = "plum4", alpha = 0
```

```
## Warning: Removed 145330 rows containing missing values (geom_point).
```



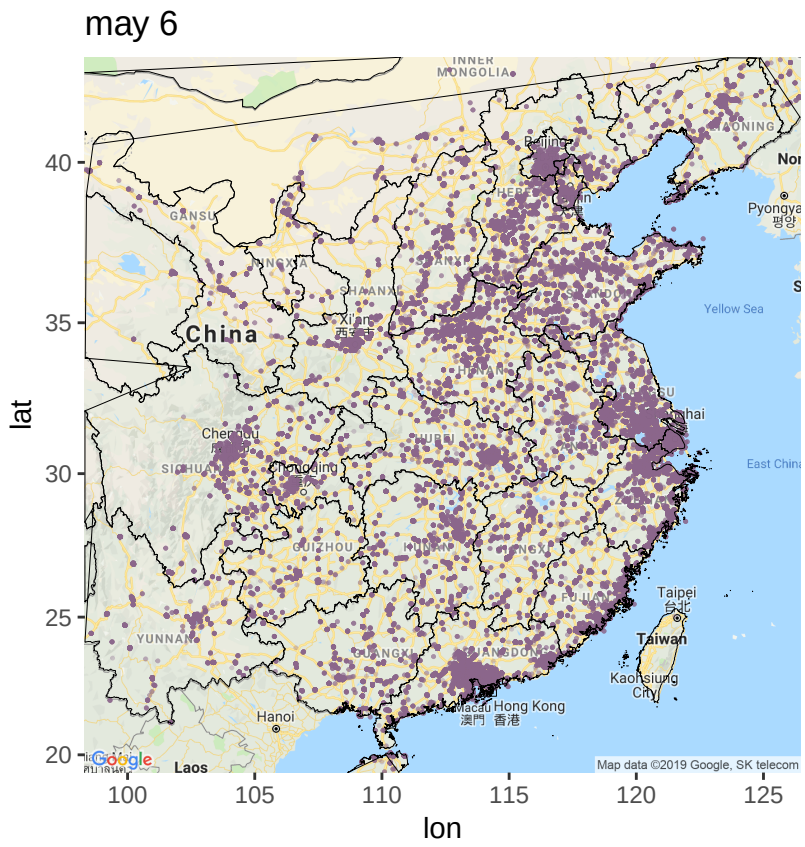

```
china + geom_point(aes(x = longitude, y = latitude), data = may_5, size = 0.2, col = "plum4", alpha = 0
```

```
## Warning: Removed 147263 rows containing missing values (geom_point).
```



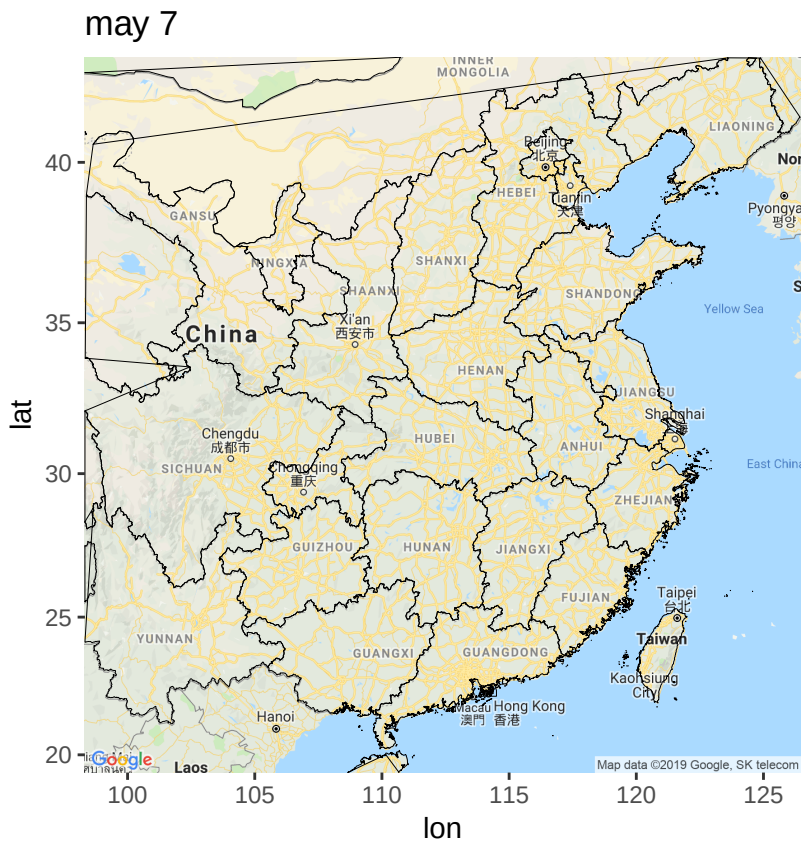

```
china + geom_point(aes(x = longitude, y = latitude), data = may_6, size = 0.2, col = "plum4", alpha = 0
```

```
## Warning: Removed 137721 rows containing missing values (geom_point).
```




```
china + geom_point(aes(x = longitude, y = latitude), data = may_7, size = 0.2, col = "plum4", alpha = 0
```

```
## Warning: Removed 3032372 rows containing missing values (geom_point).
```




```
may_1


p_may1 <- SpatialPointsDataFrame(may_1[,c(3,4)], may_1)
p_may2 <- SpatialPointsDataFrame(may_2[,c(3,4)], may_2)
p_may3 <- SpatialPointsDataFrame(may_3[,c(3,4)], may_3)
p_may4 <- SpatialPointsDataFrame(may_4[,c(3,4)], may_4)
p_may5 <- SpatialPointsDataFrame(may_5[,c(3,4)], may_5)
p_may6 <- SpatialPointsDataFrame(may_6[,c(3,4)], may_6)
p_may7 <- SpatialPointsDataFrame(may_7[,c(3,4)], may_7)


proj4string(spatial_provinces) <- CRS("+proj=longlat +ellps=WGS84 +datum=WGS84")

proj4string(p_may1) <- CRS("+proj=longlat +ellps=WGS84 +datum=WGS84")
proj4string(p_may2) <- CRS("+proj=longlat +ellps=WGS84 +datum=WGS84")
proj4string(p_may3) <- CRS("+proj=longlat +ellps=WGS84 +datum=WGS84")
proj4string(p_may4) <- CRS("+proj=longlat +ellps=WGS84 +datum=WGS84")
proj4string(p_may5) <- CRS("+proj=longlat +ellps=WGS84 +datum=WGS84")
proj4string(p_may6) <- CRS("+proj=longlat +ellps=WGS84 +datum=WGS84")
proj4string(p_may7) <- CRS("+proj=longlat +ellps=WGS84 +datum=WGS84")

res_51 <- over(p_may1, spatial_provinces)
res_52 <- over(p_may2, spatial_provinces)
res_53 <- over(p_may3, spatial_provinces)
res_54 <- over(p_may4, spatial_provinces)
```

```
res_55 <- over(p_may5, spatial_provinces)
res_56 <- over(p_may6, spatial_provinces)
res_57 <- over(p_may7, spatial_provinces)


sort(as.integer(names(table(res_51$prov_id))))


count1 <- table(res_51$prov_id)[names(table(res_51$prov_id)) %in% c(11,31,52,45,21)]
count2 <- table(res_52$prov_id)[names(table(res_52$prov_id)) %in% c(11,31,52,45,21)]
count3 <- table(res_53$prov_id)[names(table(res_53$prov_id)) %in% c(11,31,52,45,21)]
count4 <- table(res_54$prov_id)[names(table(res_54$prov_id)) %in% c(11,31,52,45,21)]
count5 <- table(res_55$prov_id)[names(table(res_55$prov_id)) %in% c(11,31,52,45,21)]
count6 <- table(res_56$prov_id)[names(table(res_56$prov_id)) %in% c(11,31,52,45,21)]
count7 <- table(res_57$prov_id)[names(table(res_57$prov_id)) %in% c(11,31,52,45,21)]

activity_counts <- data.frame(
  may_1 = count1,
  may_2 = count2,
  may_3 = count3,
  may_4 = count4,
  may_5 = count5,
  may_6 = count6,
  may_7 = count7
)


activity_counts <- activity_counts[,c(2, 4, 6, 8, 10, 12, 14)]


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

# beijing : 11, 31, 52, 45, 21


# beijing, shenyang, chengdu, guangzhou


#beijing, shenyang, shanghai, guangzhou, chengdu
```

% total checkins in each city

$ total checkins of all days in each city

CLUSTERING

```
knn_events <- kmeans(events[,c(4,5)], 20, nstart = 20)

kmeans_51 <- kmeans(may_1[,c(3,4)], 20, nstart = 20)
```

```
kmeans_52 <- kmeans(may_2[,c(3,4)], 20, nstart = 20)
kmeans_53 <- kmeans(may_3[,c(3,4)], 20, nstart = 20)
kmeans_54 <- kmeans(may_4[,c(3,4)], 20, nstart = 20)
kmeans_55 <- kmeans(may_5[,c(3,4)], 20, nstart = 20)
kmeans_56 <- kmeans(may_6[,c(3,4)], 20, nstart = 20)
kmeans_57 <- kmeans(may_7[,c(3,4)], 20, nstart = 20)
```

plots

```
newmap <- getMap(resolution = "low")
plot(newmap, xlim = c(105, 120), ylim = c(15, 40), asp = 1)
points(events$longitude, events$latitude, col = "rosybrown2", cex = 0.05)
points(knn_events$centers[,1], knn_events$centers[,2], col = "red", pch = 19)
#summary(events$longitude)
#summary(events$latitude)

centers <- data.frame(
  events_long = knn_events$centers[,1],
  events_lat = knn_events$centers[,2],
  long_51 = kmeans_51$centers[,1],
  lat_51 = kmeans_51$centers[,2],
  long_52 = kmeans_52$centers[,1],
  lat_52 = kmeans_52$centers[,2],
  long_53 = kmeans_53$centers[,1],
  lat_53 = kmeans_53$centers[,2],
  long_54 = kmeans_54$centers[,1],
  lat_54 = kmeans_54$centers[,2],
  long_55 = kmeans_55$centers[,1],
  lat_55 = kmeans_55$centers[,2],
  long_56 = kmeans_56$centers[,1],
  lat_56 = kmeans_56$centers[,2],
  long_57 = kmeans_57$centers[,1],
  lat_57 = kmeans_57$centers[,2]
)

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

```
plot(newmap, xlim = c(105, 120), ylim = c(15, 40), asp = 1, main = "may 1 clusters")
points(may_1$longitude, may_1$latitude, col = "rosybrown2", cex = 0.05)
points(kmeans_51$centers[,1], kmeans_51$centers[,2], col = "red", pch = 19)

plot(newmap, xlim = c(105, 120), ylim = c(15, 40), asp = 1, main = "may 2 clusters")
points(may_2$longitude, may_2$latitude, col = "rosybrown2", cex = 0.05)
points(kmeans_52$centers[,1], kmeans_52$centers[,2], col = "red", pch = 19)

plot(newmap, xlim = c(105, 120), ylim = c(15, 40), asp = 1, main = "may 3 clusters")
points(may_3$longitude, may_3$latitude, col = "rosybrown2", cex = 0.05)
points(kmeans_53$centers[,1], kmeans_53$centers[,2], col = "red", pch = 19)

plot(newmap, xlim = c(105, 120), ylim = c(15, 40), asp = 1, main = "may 4 clusters")
points(may_4$longitude, may_4$latitude, col = "rosybrown2", cex = 0.05)
points(kmeans_54$centers[,1], kmeans_54$centers[,2], col = "red", pch = 19)
```

```r
plot(newmap, xlim = c(105, 120), ylim = c(15, 40), asp = 1, main = "may 5 clusters")
points(may_5$longitude, may_5$latitude, col = "rosybrown2", cex = 0.05)
points(kmeans_55$centers[,1], kmeans_55$centers[,2], col = "red", pch = 19)

plot(newmap, xlim = c(105, 120), ylim = c(15, 40), asp = 1, main = "may 6 clusters")
points(may_6$longitude, may_6$latitude, col = "rosybrown2", cex = 0.05)
points(kmeans_56$centers[,1], kmeans_56$centers[,2], col = "red", pch = 19)

plot(newmap, xlim = c(105, 120), ylim = c(15, 40), asp = 1, main = "may 7 clusters")
points(may_7$longitude, may_7$latitude, col = "rosybrown2", cex = 0.05)
points(kmeans_57$centers[,1], kmeans_57$centers[,2], col = "red", pch = 19)
```