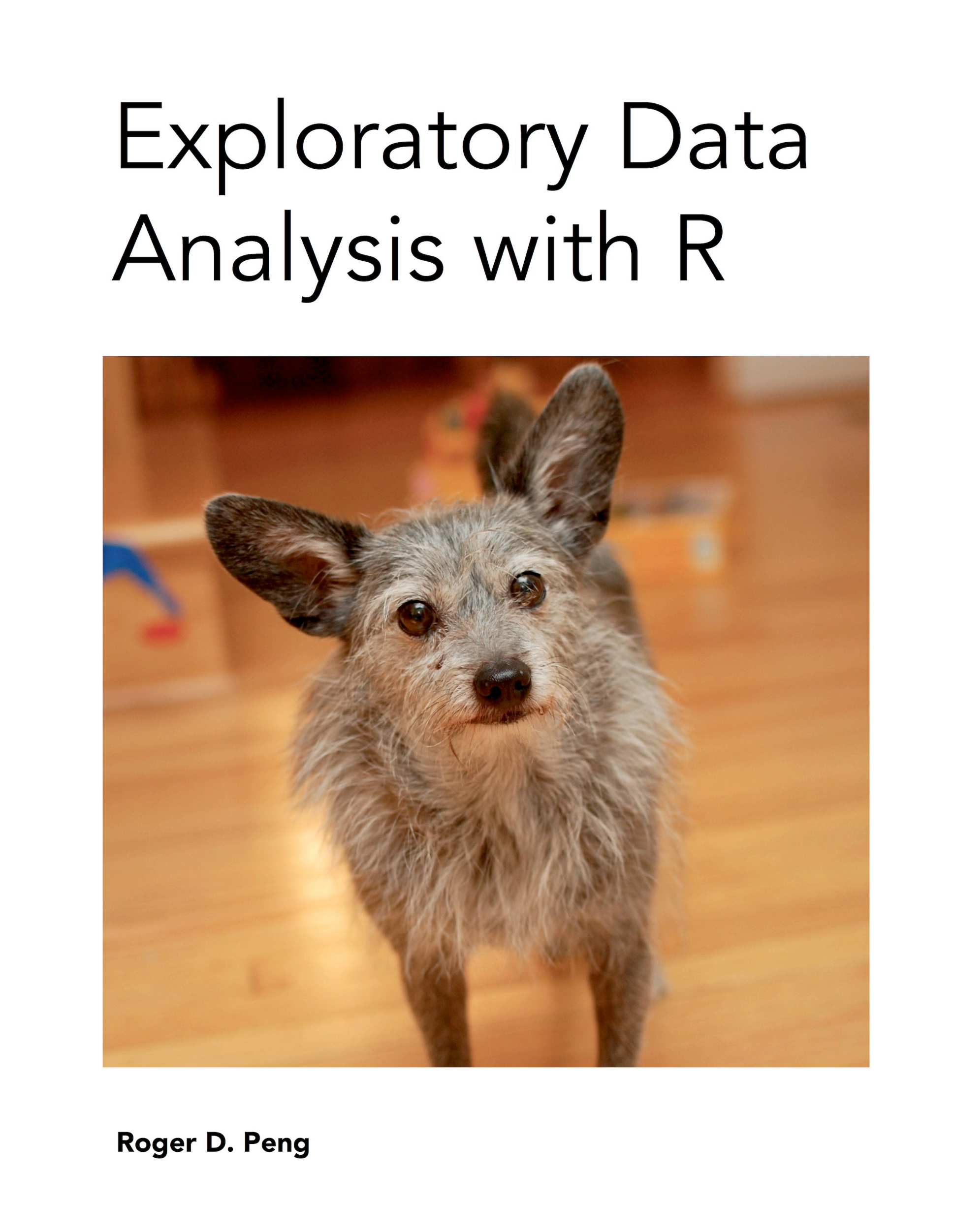
DATA VISUALIZATION

SOWK3136 L3



Exploratory Data Analysis with R

Roger D. Peng

Exploratory Data Analysis with R

Roger D. Peng

This book is for sale at http://leanpub.com/exdata
This version was published on 2015-11-12

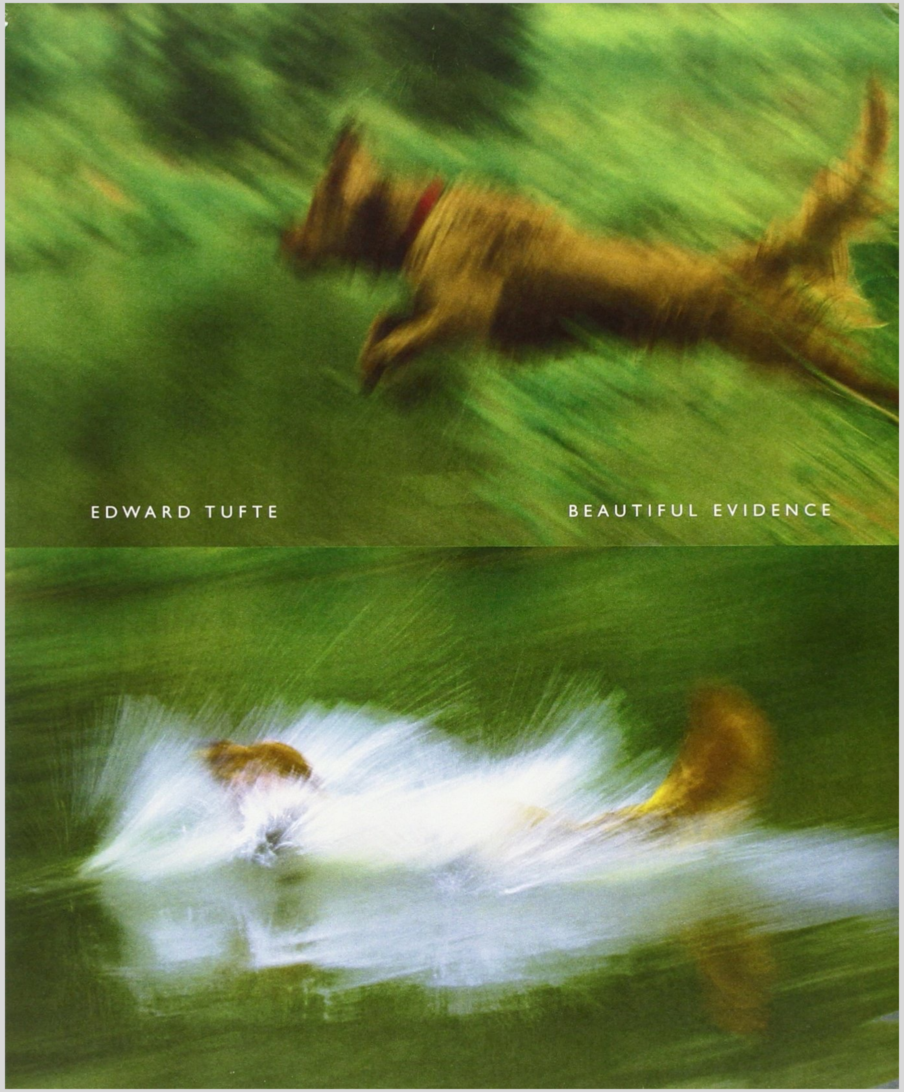
“In the editing room”

- The process of making the “rough cut” for a data analysis.
- The goals：

1. identifying relationships between variables that are particularly interesting or unexpected;
2. checking to see if there is any evidence for or against a stated hypothesis;
3. checking for problems with the collected data, such as missing data or measurement error;
4. identifying certain areas where more data need to be collected

Principles of Analytic Graphics

Edward Tufte <Beautiful Evidence>



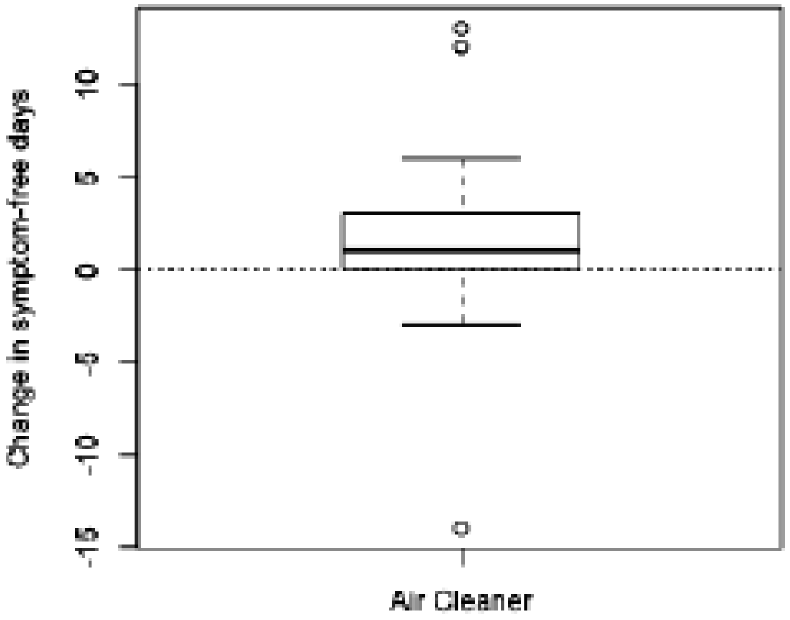
Show comparisons

- The basis of all good scientific investigation
- Evidence for a hypothesis is always *relative* to another competing hypothesis.
- “the evidence favors hypothesis A”
- “the evidence favors hypothesis A versus hypothesis B”
- A good scientist is always asking “Compared to What?” when confronted with a scientific claim or statement.
- Data graphics should generally follow this same principle. You should always be comparing at least two things.

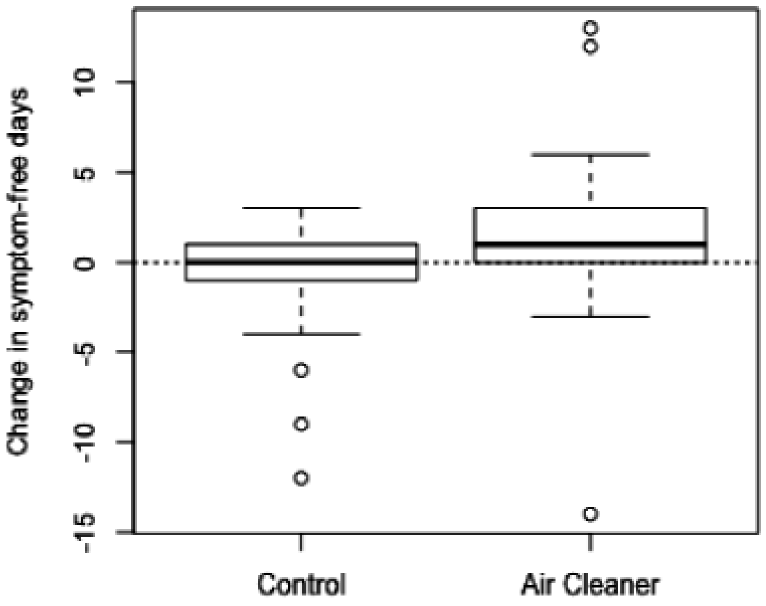
An air cleaner installed in a child’s home improves their asthma-related symptoms?

- This study was conducted at the Johns Hopkins University School of Medicine and was conducted in homes where a smoker was living for at least 4 days a week. Each child was assessed at baseline and then 6-months later at a second visit. The aim was to improve a child’s symptom-free days over the 6-month period. In this case, a higher number is better, indicating that they had *more* symptom-free days.

Huang, F., & Kim, J. S. (2012). A Randomized Trial of Air Cleaners and a Health Coach to Improve Indoor Air Quality for Inner-City Children With Asthma and Secondhand Smoke Exposure. *Pediatrics*, *130*(Supplement 1), S33-S34.



Change in symptom-free days with air cleaner



Change in symptom-free days by treatment group

Huang, F., & Kim, J. S. (2012). A Randomized Trial of Air Cleaners and a Health Coach to Improve Indoor Air Quality for Inner-City Children With Asthma and Secondhand Smoke Exposure. *Pediatrics*, *130*(Supplement 1), S33-S34.

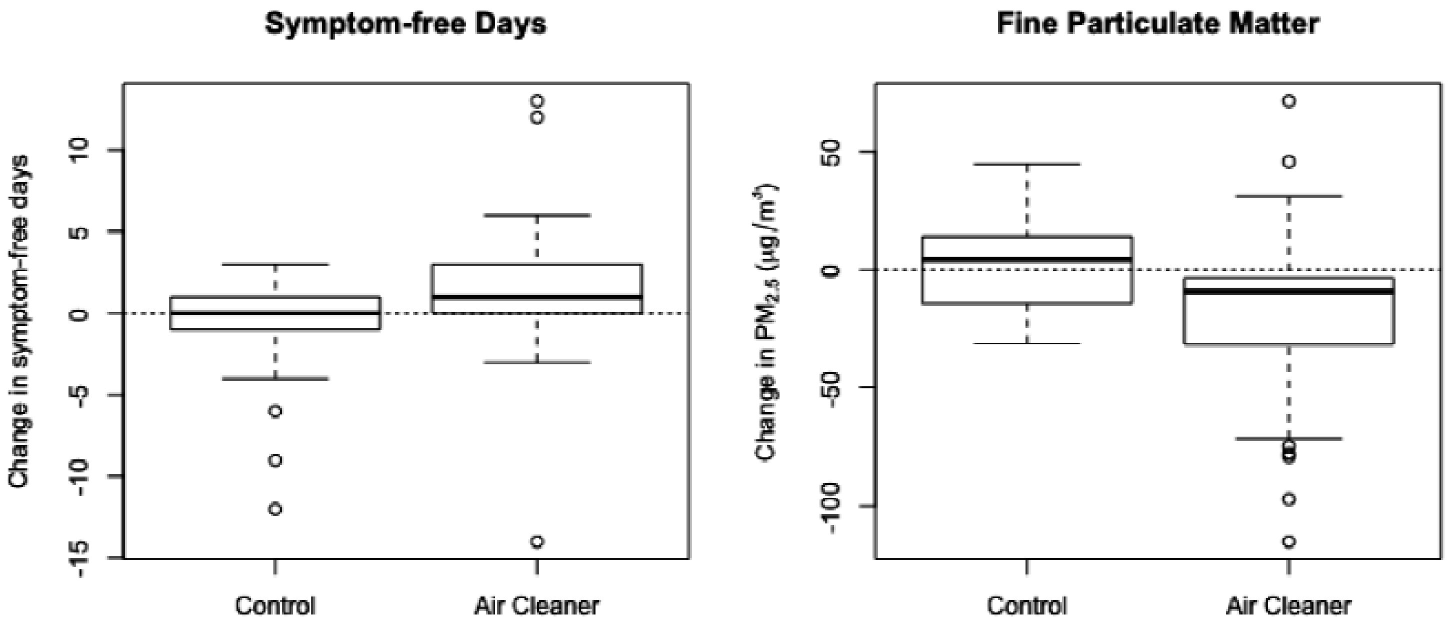
Show causality, mechanism, explanation, systematic structure

- If possible, it’s always useful to show your causal framework for thinking about a question.
- Generally, it’s difficult to prove that one thing causes another thing even with the most carefully collected data.
- But it’s still often useful for your data graphics to indicate what you are thinking about in terms of cause.
- Such a display may suggest hypotheses or refute them, but most importantly, they will raise new questions that can be followed up with new data or analyses.

“Why do the children with the air cleaner improve?”

- The hypothesis behind air cleaners improving asthma morbidity in children is that the air cleaners remove airborne particles from the air.

A decrease in airborne particles?



Change in symptom-free days and change in PM2.5 levels in-home

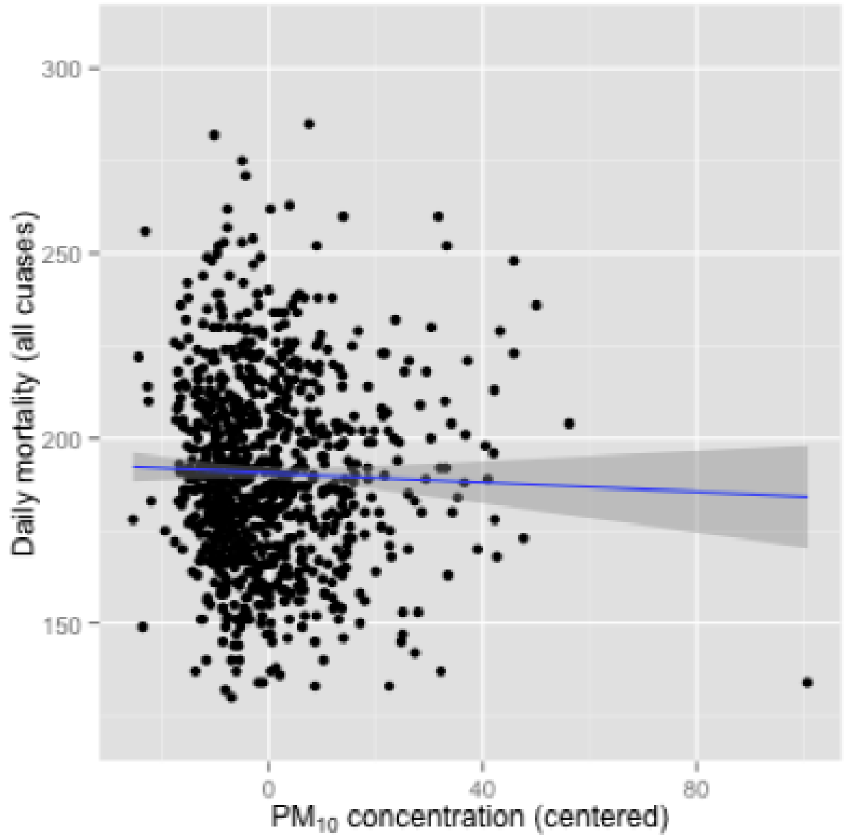
However, it is not conclusive proof of this idea because there may be other unmeasured confounding factors that can lower levels of PM2.5 and improve symptom-free days.

Show multivariate data

- The real world is multivariate.
- For anything that you might study, there are usually many attributes that you can measure.
- The point is that data graphics should attempt to show this information as much as possible, rather than reduce things down to one or two features that we can plot on a page.
- There are a variety of ways that you can show multivariate data.

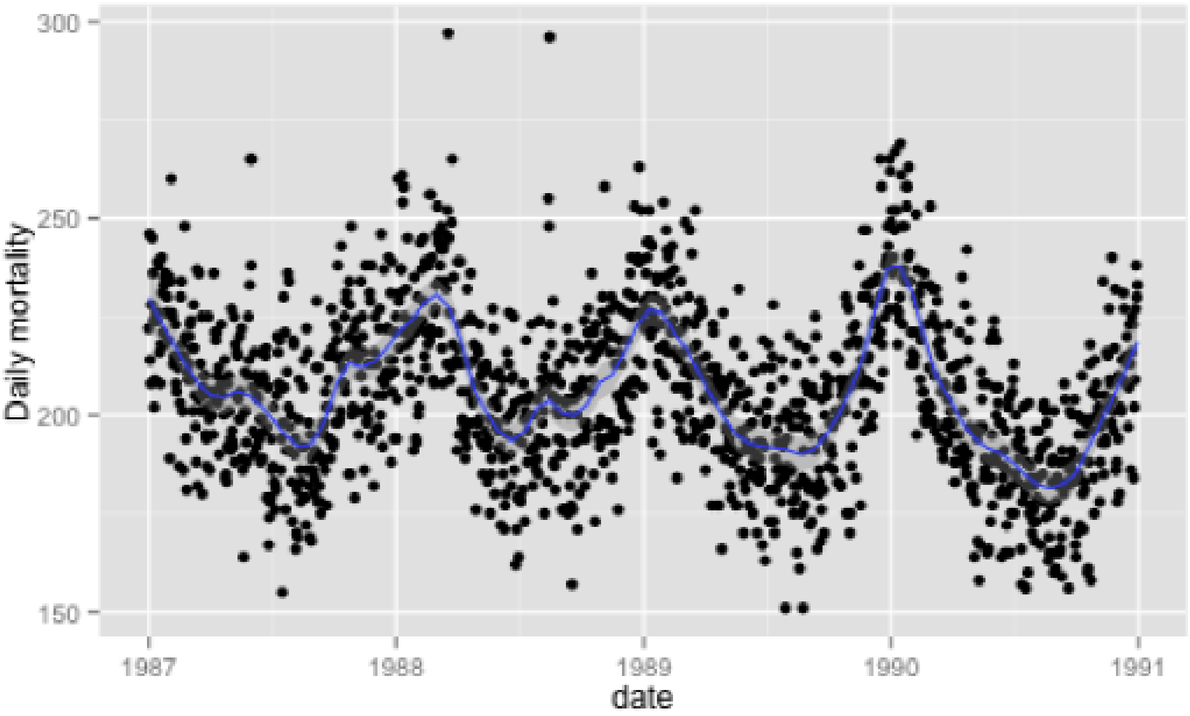
Daily airborne particulate matter (“PM10”) and mortality

- Daily PM10 in New York City and mortality from 1987 to 2000
- The PM10 data come from the U.S. Environmental Protection Agency
- The mortality data come from the U.S. National Center for Health Statistics.

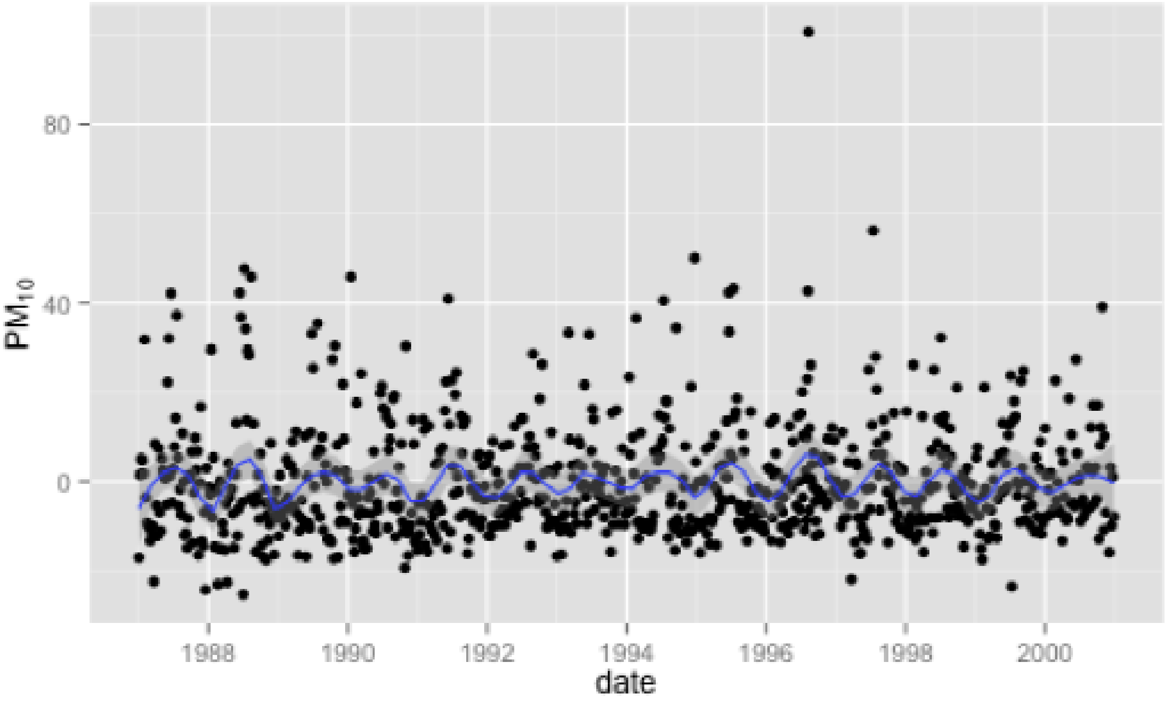


PM10 and mortality in New York City

Season?

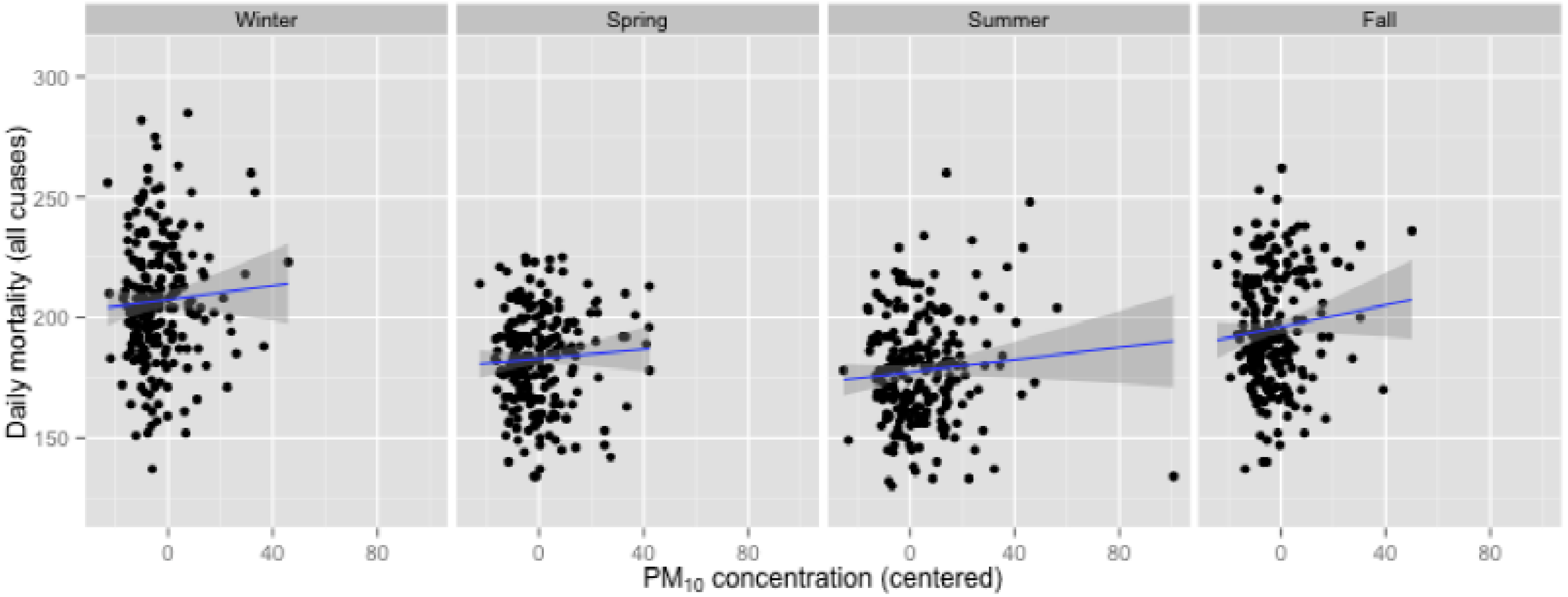


Daily mortality in New York City



Daily PM10 in New York City

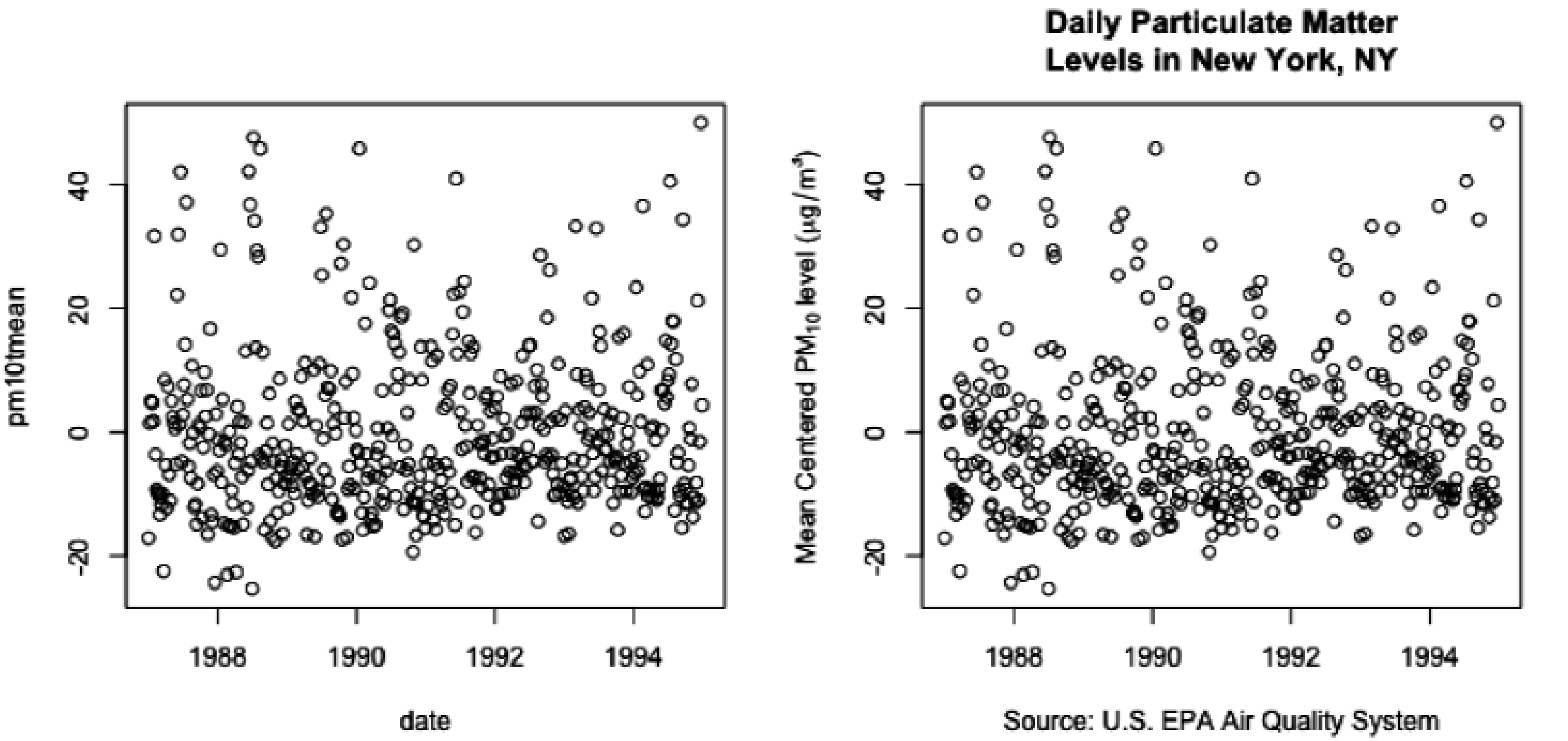
What happens if we plot the relationship between mortality and PM10 *by season*?



PM10 and mortality in New York City by season

Describe and document the evidence

- Data graphics should be appropriately documented with labels, scales, and sources.
- A general rule for me is that a data graphic should tell a complete story all by itself.



Daily Particulate Matter Levels in New York, NY

Source: U.S. EPA Air Quality System

Labelling and annotation of data graphics

Content, content, content!

- Analytical presentations ultimately stand or fall depending on the quality, relevance, and integrity of their content.
- This includes the question being asked and the evidence presented in favour of certain hypotheses.
- Starting with a good question, developing a sound approach, and only presenting information that is necessary for answering that question, is essential to every data graphic.

No amount of visualization magic or bells and whistles can make poor data, or more importantly, a poorly formed question, shine with clarity.

Exploratory graphs

- Why?
  - The understand the characteristics of the data/variables/features
  - Find patterns
  - Suggest modeling strategies
  - Identify potential problems (“debug”)
  - Communicate results, internally and externally

Simple data summaries

- One dimension summary
  - Five-number summary (min, Q1, Q2[median], Q3, max)
  - Ba chart
  - Boxplots
  - Histogram
  - Density plot

Try the following function

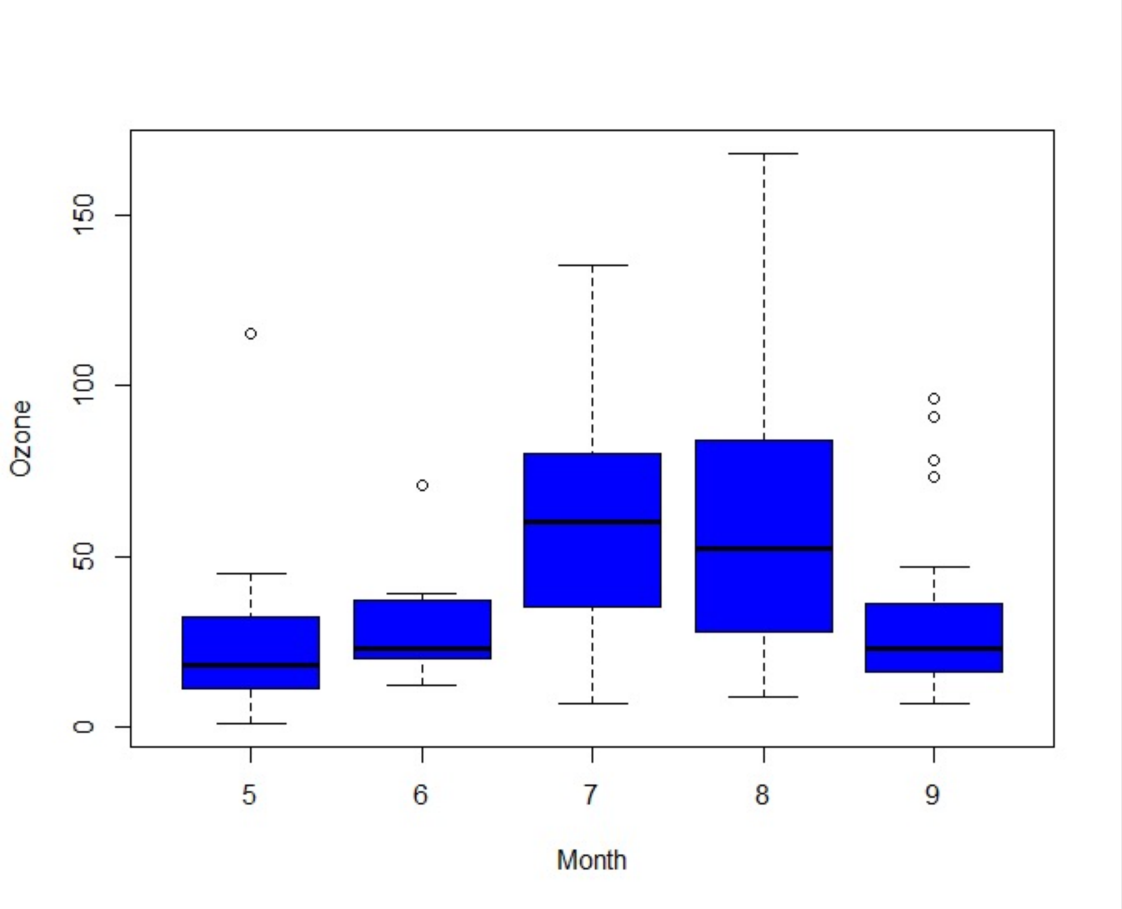
- summary( )
- boxplot(, col = “blue”)
- abline(h =)
- hist(, col = “green”, breaks =)
- rug( )
- abline(v = , lwd =)
- abline(v = median(), col = ””, lwd =)
- barplot(table(), col = , main = “”)

Simple data summaries

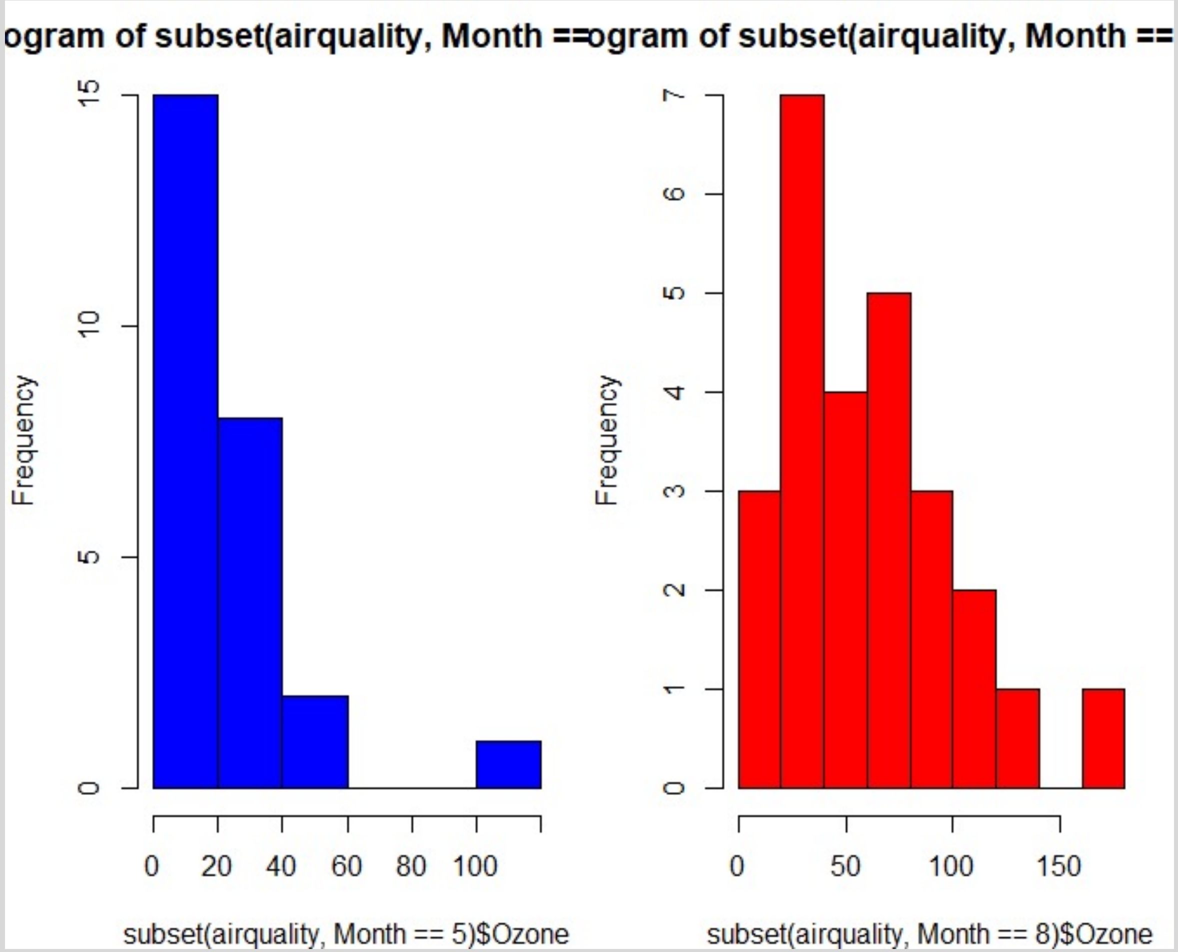
- Two dimensions
  - Scatterplots
  - Multiple (overlayed) 1-D plots
- Three or more
  - 3-D?
  - Multiple (overlayed) 2-D plots
  - Use of color and other parameters

Try the following functions on your data

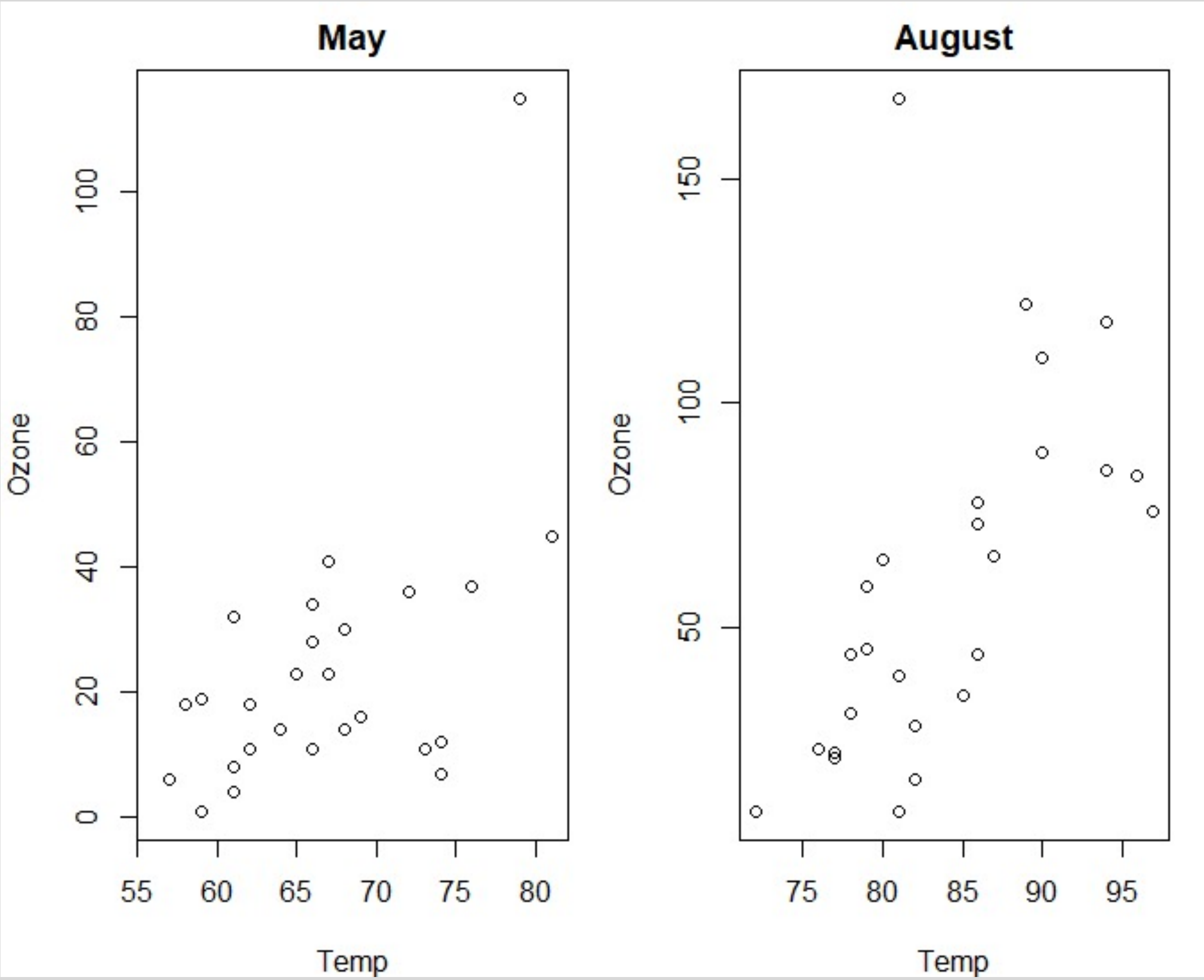
- `boxplot(Ozone ~ Month, data = airquality, col = "blue")`



```
par(mfrow = c(1, 2), mar = c(4, 4, 2, 1))
    hist(subset(airquality, Month == 5)$Ozone, col = "blue")
    hist(subset(airquality, Month == 8)$Ozone, col = "red")
```



```
par(mfrow = c(1, 2), mar = c(4, 4, 2, 1))
    with(subset(airquality, Month == 5), plot(Temp, Ozone, main = "May"))
    with(subset(airquality, Month == 8), plot(Temp, Ozone, main = "August"))
```



Plotting systems in R

- The Base plotting system
- The Lattice system
- The ggplot2 system

The Base plotting system

- Base function
- “Artist’s palette” model
- Start with blank canvas and plot function
- Add lines, points, text, etc. piece by piece – annotation functions
- The end product is a series of R commands

- `data(cars)`
- `with(cars, plot(speed, dist))`

The Lattice system

- Single function call `xyplot(), bwplot(),` etc
- Conditioning types of plot (multivariate plots/3 [or more] dimensional plot): explore the relationship between x and y, by z
- Margins, spacing, and detailed plot parameters will be set automatically at once
- Annotation is not intuitive and not flexible
- `library(lattice)`
- `state <- data.frame(state.x77, region = state.region)`
- `xyplot(Life.Exp ~ Income | region, data = state, layout = c(4,1))`

The ggplot2 system

- Base + lattice
- Automatically deals with spacing, text, title but also allow flexible annotations
- Default mode is usually “pretty” enough

- `library(ggplot2)`
- `data(mpg)`
- `qplot(displ, hwy, data = mpg)`

Plotting systems in R - summary

- The Base plotting system – “artist’s palette” model
- The Lattice system – one function for the entire plot
- The ggplot2 system – mixed elements of Base and Lattice

The Base plotting system

- Two phases to creating a base plot
  - Initializing a new plot: `plot(x, y)` `hist(x)` – they have many arguments
  - Annotating (adding to) an existing lot
- The base graphics system has many many parameters that you can set
- Check `?par` for details of these parameters

- `library(datasets)`
- `hist(airquality$Ozone)`
- `with(airquality, plot(Wind, Ozone))`
- `airquality <- transform(airquality, Month = factor(Month))`
- `boxplot(Ozone ~ Month, airquality, xlab = “Month”, ylab = “Ozone (ppb)”)`

Important parameters:

`pch`: plotting ‘character’, i.e., symbol to use.

`lty`: The line type

`lwd`: The line width, a positive number, defaulting to 1.

`col`: color code or name.

`xlab`: character string for the x-axis label

`ylab`: character string for the y-axis label

`xlim`

`ylim`

Global graphics parameters

- `par()` function is used to specify global graphics parameters that affect all plots
- `mfrow`: number of plots per row, column (row-wise fill)
- `mfcol`: number of plots per column (column-wise fill)
- `las`: the orientation of the axis labels
- `bg`: background color
- `mar`: margin size
- `oma`: outer margin size

Check the default values:

```
par("bg")
par("mar")
par("mfrow")
```

Base plotting functions

- plot(): Draw a scatter plot with decorations such as axes and titles in the active graphics window.
- lines(): A generic function taking coordinates given in various ways and joining the corresponding points with line segments.
- points(): A generic function to draw a sequence of points at the specified coordinates.
- text(): Add labels to a plot using x, y coordinates
- title(): Add annotations to x, y axis labels, title, subtitle, outer margin
- axis(): Adding axis ticks/labels

- `with(airquality, plot(Wind, Ozone))`
- `title(main = “Air quality in NYC”) # Add a title`
- `with(airquality, plot(Wind, Ozone, main = “Air quality in NYC”))`
- `with(subset(airquality, Month == 5), points(Wind, Ozone, col = “pink”))`
- `with(subset(airquality, Month != 5), points(Wind, Ozone, col = “blue”))`
- `legend(“topright”, pch = 1, col = c(“pink”, “blue”), legend = c(“May”, “Other”))`

Base plot with regression line

- `with(airquality, plot(Wind, Ozone, main = “Air quality in NYC”, pch = 20))`
- `model <- lm(Ozone ~ Wind, airquality)`
- `abline(model, lwd = 2)`

- `par(mfrow = c(1,3))`
- `with(airquality, {`
- `  plot(Ozone, Solar.R, main = "Ozone and Solar")`
- `  plot(Solar.R, Wind, main = "Solar and Wind")`
- `  plot(Ozone, Temp, main = "Ozone and Temp")`
- `  })`

- `par(mfrow = c(1,3), oma = c(0, 0, 2, 0))`
- `with(airquality, {`
- `   plot(Ozone, Solar.R, main = “Ozone and Solar”)`
- `   plot(Solar, Wind, main = “Solar and Wind”)`
- `   plot(Ozone, Temp, main = “Ozone and Temp”)`
- `   mtext(“Air quality in NYC”, outer = T)`
- `})`

Assignment

- Type in the commends we covered during the lecture
- Experiment different parameter settings

- The “covid.Rdata” file contains two dataframes: daily and weekly
- These datasets contain daily/weekly COVID cases, deaths, and government response variables of different countries
- See the official website of “COVID-19 GOVERNMENT RESPONSE TRACKER” for details
- https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker
- Explore!