Exploratory Data Analysis Notes

Coursera Course by John Hopkins University

INSTRUCTORS: Dr. Jeff Leek, Dr. Roger D. Peng, Dr. Brian Caffo

Contents

Intro

- **Slides and data for this course may be found at github**

Instructor's Note

This course covers the essential exploratory techniques for summarizing data. These techniques are typically applied before formal modeling commences and can help inform the development of more complex statistical models. Exploratory techniques are also important for eliminating or sharpening potential hypotheses about the world that can be addressed by the data. We will cover in detail the plotting systems in R as well as some of the basic principles of constructing data graphics.
We will also cover some of the common multivariate statistical techniques used to visualize high-dimensional data.

All the best,

Roger Peng

Introduction

- EDA allows you to develop a rough idea of what your data look like and what kinds of questions might be answered by them.
- EDA is often the "fun part" of data analysis, where you get to play around with the data and explore.

- These techniques for summarizing data are typically applied before formal modeling commences and can help inform the development of more complex statistical models.

Exploratory Data Analysis with R Book

- **Exploratory Data Analysis with R**

The Art of Data Science

- **The Art of Data Science eBook**
- **The Art of Data Science printed version**

Installing R on…

Windows

- Just go to **the cran site** and install the Windows version.
  + For an optimal experience, back up all of data onto a usb, then install your prefered version of Linux (I use Fedora) and install the Linux version instead.

Mac

- Just go to **the cran site** and install the Mac version.
  + If you don't have enough money to buy a Mac install Linux instead, it's open-source, meaning it's free!

Installing R Studio (Mac)

- Install from **the RStudio website** *after* you have `R` installed.

Setting Your Working Directory on…

Windows, Mac & Linux

- Your working directory is where R will look for all the files it reads and where all the files it writes will go
- `getwd()` will display your current working directory
- `dir()` will display all files in your wd
- `setwd(param)` will set your working directory to the character string that is represented by `param`

- `source("myFunction.R")` will load in `myFunction` script from wd and any functions that are within it.

Lesson 1: Graphs

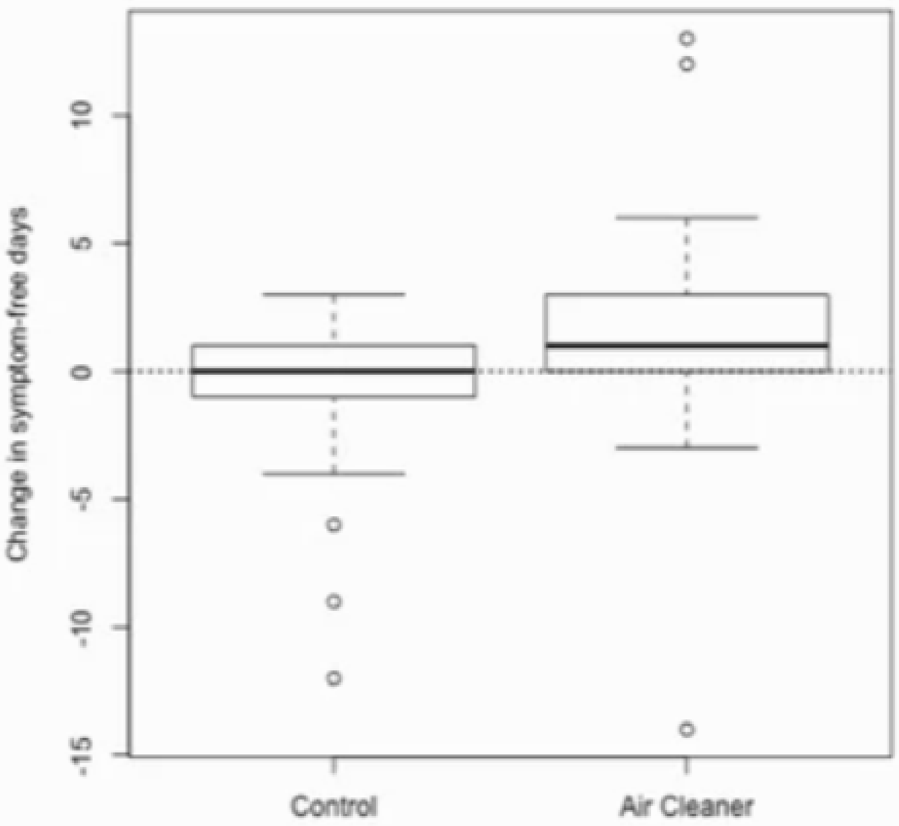
Principles of Analytic Graphics

- Some general rules to follow when building analytic graphics from data to tell the story the data hold.
- Principles:

1) Show Comparisons

- Evidence for a hypothesis is always *relative* to another competing hypothesis

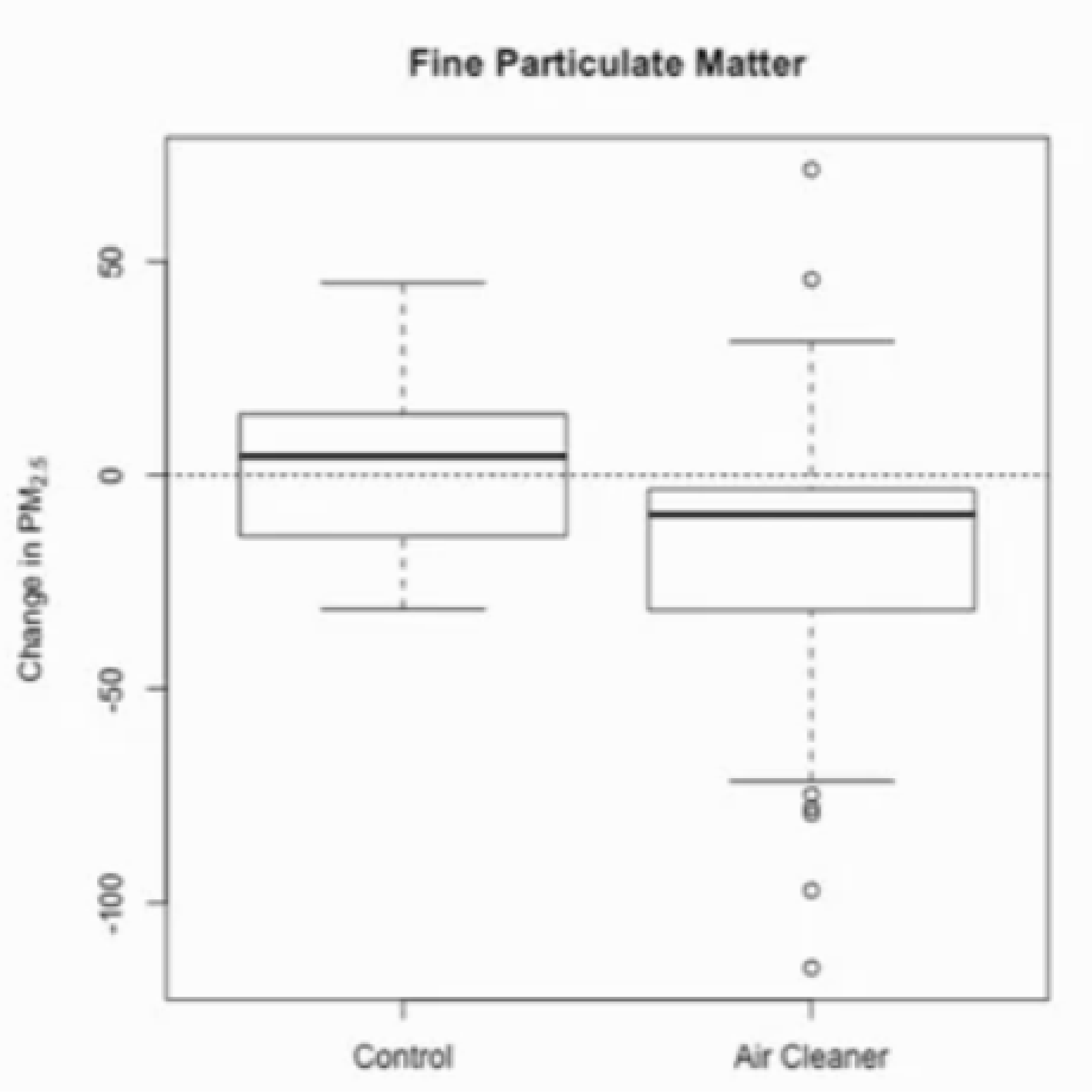
- Always ask “Commpared to What?”
- For example a box plot of **Change in symptom-free days** in children with asthma when an **Air Cleaner** is installed in their quarters should be shown in *comparison* to a control group



Reference: Butz AM, *et al.*, *JAMA Pediatrics*, 2011.

2) Show Causality, Mechanism, or Explanation

- To show what is going on/how you belive the system is operating and what is the cause for the result you are showing.
- What is your causal framework for thinking about a question
- This only shows a suggestion and indicates where futher investigation could go
- In the asthma example you would also want to show the **Change in PM** (Particulate Matter) in the childs home between the **Control** and **Air Cleaner**

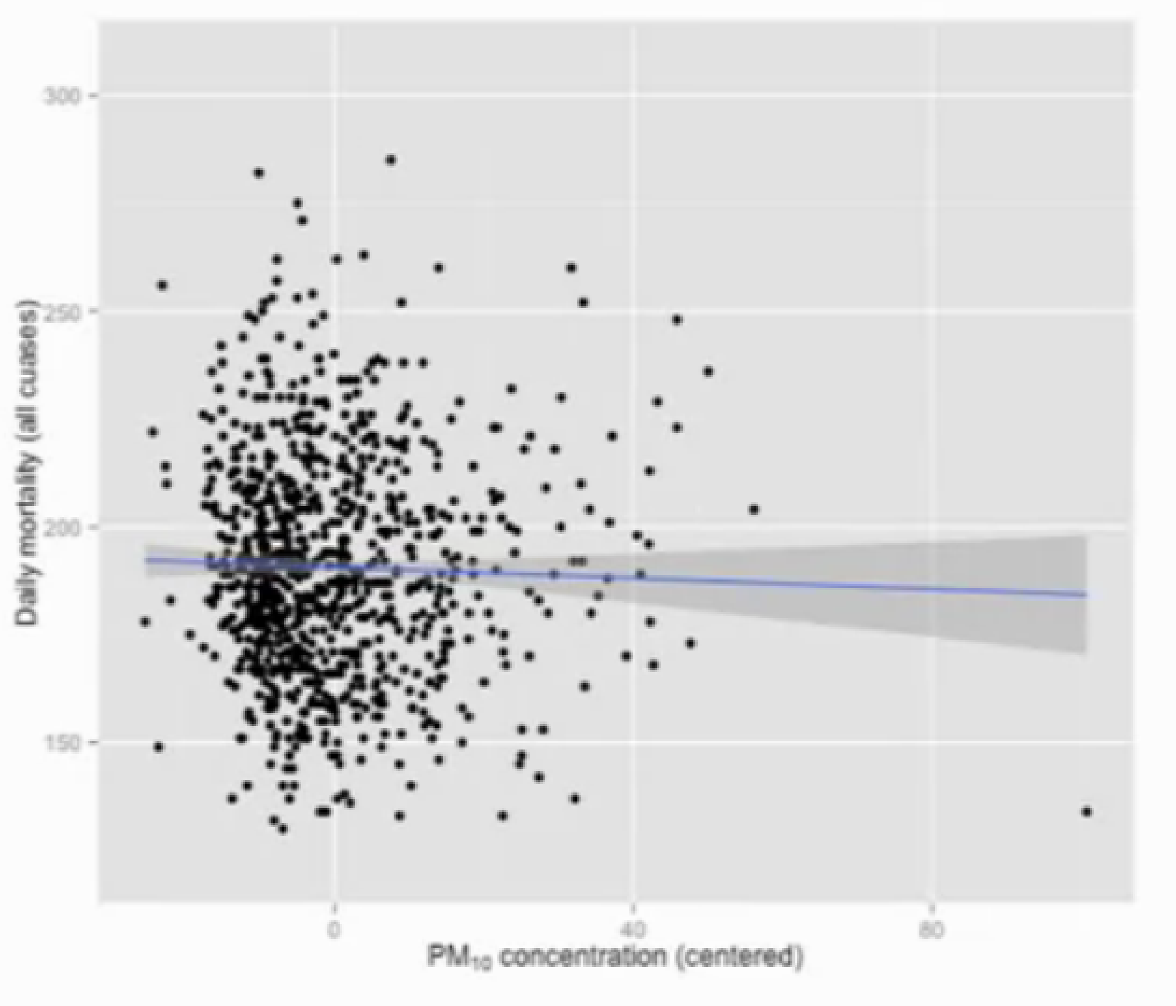


3) Show Multivariate Data

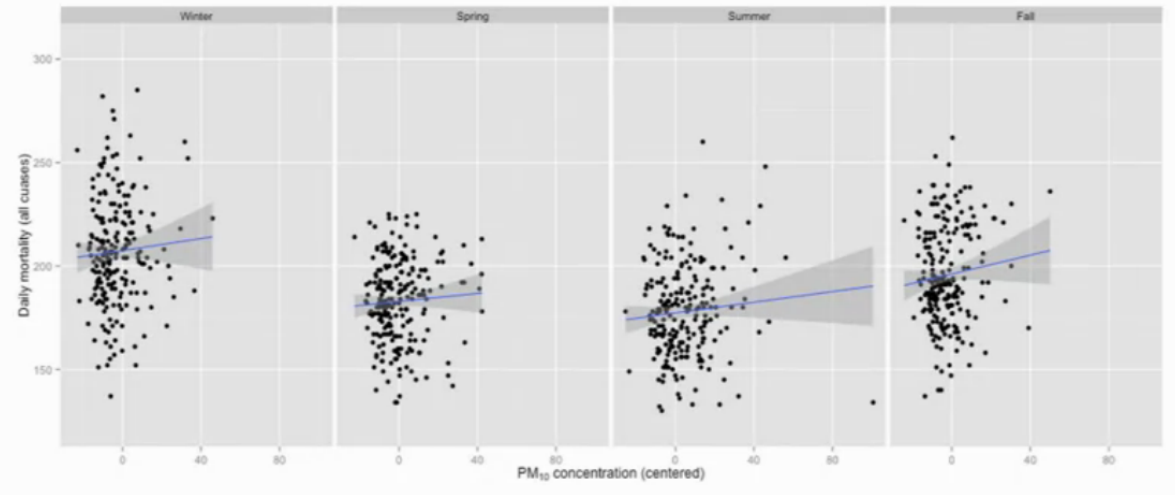
- Show as much data on a plot as you reasonably can, it tells a richer story

- The real world is multivariate, so your plots should reflect that

- Need to “escape flatland”
- For example, below is a 2-D plot of **Daily mortality** versus **PM concentration** in NYC, and it shows a slight decrease in mortality as PM increases.



- However, if we plot this across four plots for each season we can see an increase in mortality within each season



4) Integration of Evidence

- Don't let the tool drive the analysis

- Completely integrate words, numbers, images, and diagrams

- Data graphics should make use of many modes of data presentation

- Put lot of information on the plots rather than diffrent places where it may be hard to track down
- The following example shows a plot that has a column for probability that the hospitalizations are diffrent than 0, (the left side is labeling the rows), and the bottom is describing how the experiment was performed

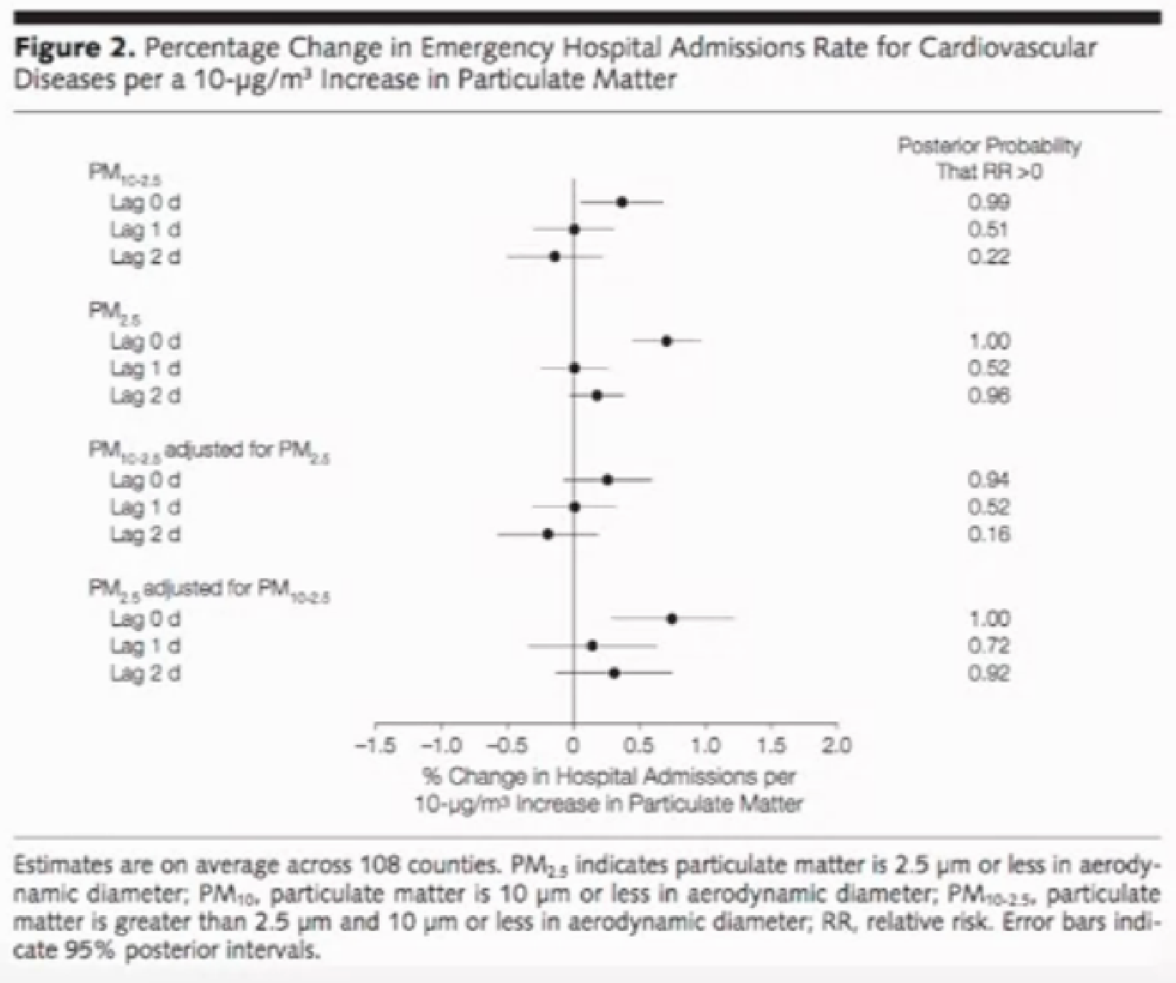


Figure 2. Percentage Change in Emergency Hospital Admissions Rate for Cardiovascular Diseases per a 10-μg/m³ Increase in Particulate Matter

Estimates are on average across 108 counties. $PM_{2.5}$ indicates particulate matter is 2.5 μm or less in aerodynamic diameter; PM_{10}, particulate matter is 10 μm or less in aerodynamic diameter; $PM_{10-2.5}$, particulate matter is greater than 2.5 μm and 10 μm or less in aerodynamic diameter; RR, relative risk. Error bars indicate 95% posterior intervals.

5) Describe and Document the Evidence

- Use appropriate labels, scales, sources, etc.

- A data graphic should tell a complete story that is also credible

6) Content is King

- If there isn't an intresting story to tell no amount of presentation will make it intresting

- Analytical presentations ultimately stand or fall depending on the quality, relevance, and integrity of their content
- Further Reading - **Edward Tufte's Beautiful Evidence ($32)**

Lesson with `swirl()`: Principles of Analytic Graphs

- This lesson runs through the 5 principles that were discussed in the above lecture.
- The multivate plot was an example of **Simpson's paradox, or the Yule-Simpson effect**
- With R, you want to preserve any code you use to generate your data and graphics so that the research can be replicated if necessary.
  - This allows for easy verfication or finding bugs in your analysis

Exploratory Graphs

- These are graphs that are made for yourself to look at and explore the data sets you're looking at

- Why do we use graphs in data analysis?
  - To undestand data properties

  - To find patterns in data

  - To suggest modeling strategies

  - To "debug" analyses

  - To communicate results

  - Exploratory graphs are for the first four of these reasons
- Characteristics of exploratory graphs
  - They are made quickly ("on the fly")

  - A large number are made
    - Looking through a lot of the varaibles and diffrent aspects of the data

  - The goal is for personal understanding
    - What are the properties, problems, and issues that need followed up

  - Axes/legends are generally cleaned up later

  - Color/size are primarily used for information, rather than presentation

- The following examples will be using data about ambient air pollution in the United States for particle pollution (PM2.5), the "annual mean, averged over 3 years" cannot exceed 12 micrograms/cubic meter

```
pollution <- read.csv("./data/avgpm25.csv",
                      colClasses = c("numeric", "character", "factor", "numeric", "numeric"))
head(pollution)
```

```
##        pm25  fips region longitude latitude
## 1  9.771185 01003   east -87.74826 30.59278
## 2  9.993817 01027   east -85.84286 33.26581
## 3 10.688618 01033   east -87.72596 34.73148
## 4 11.337424 01049   east -85.79892 34.45913
## 5 12.119764 01055   east -86.03212 34.01860
## 6 10.827805 01069   east -85.35039 31.18973
```

- The question we are looking into is: *Do any counties exceed the standard of* $12\mu g/m^3$*?*
  - You always want to have an underlying question in mind, even if it's kind of vague

- Simple summaries of Data
  - Five-number summary
  - Boxplots
  - Histograms
  - Density plot
  - Barplot

Five-number Summary

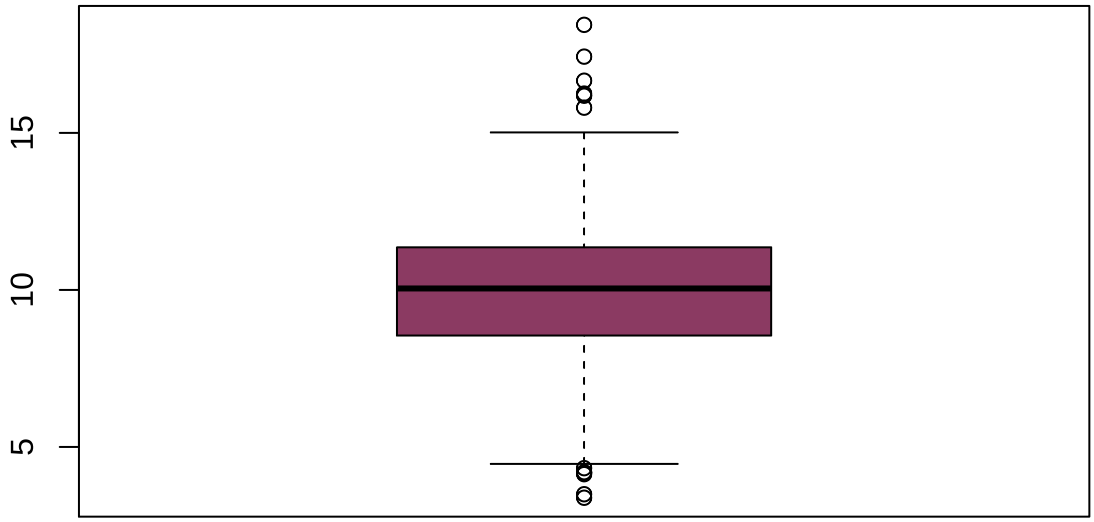
- The `summary` function in R gives the 5-number summary as well as the mean

```
summary(pollution$pm25)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.383   8.549  10.047   9.836  11.356  18.441
```

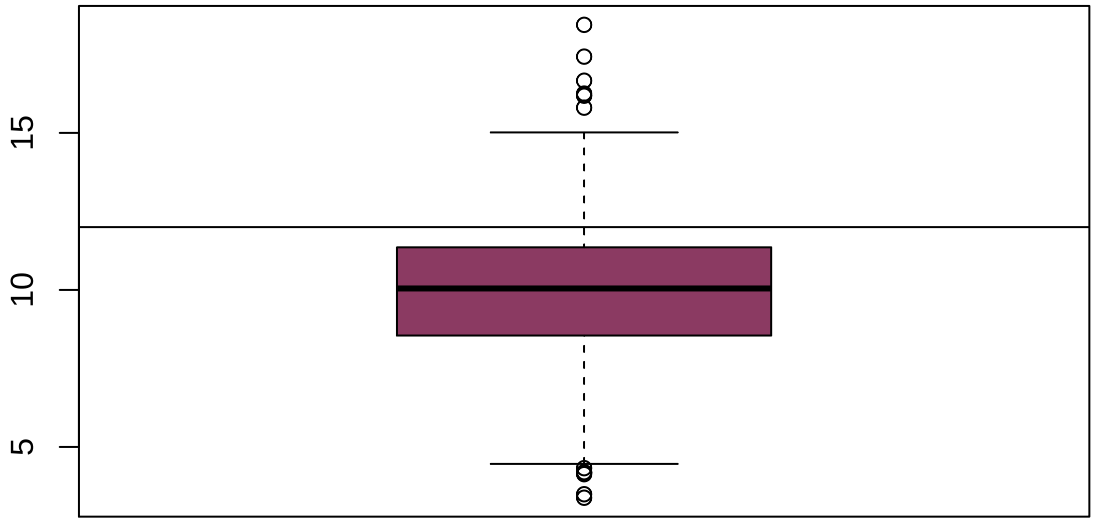
Boxplots

```
boxplot(pollution$pm25, col = "hotpink4") #Lecture used blue but I dislike that display
```



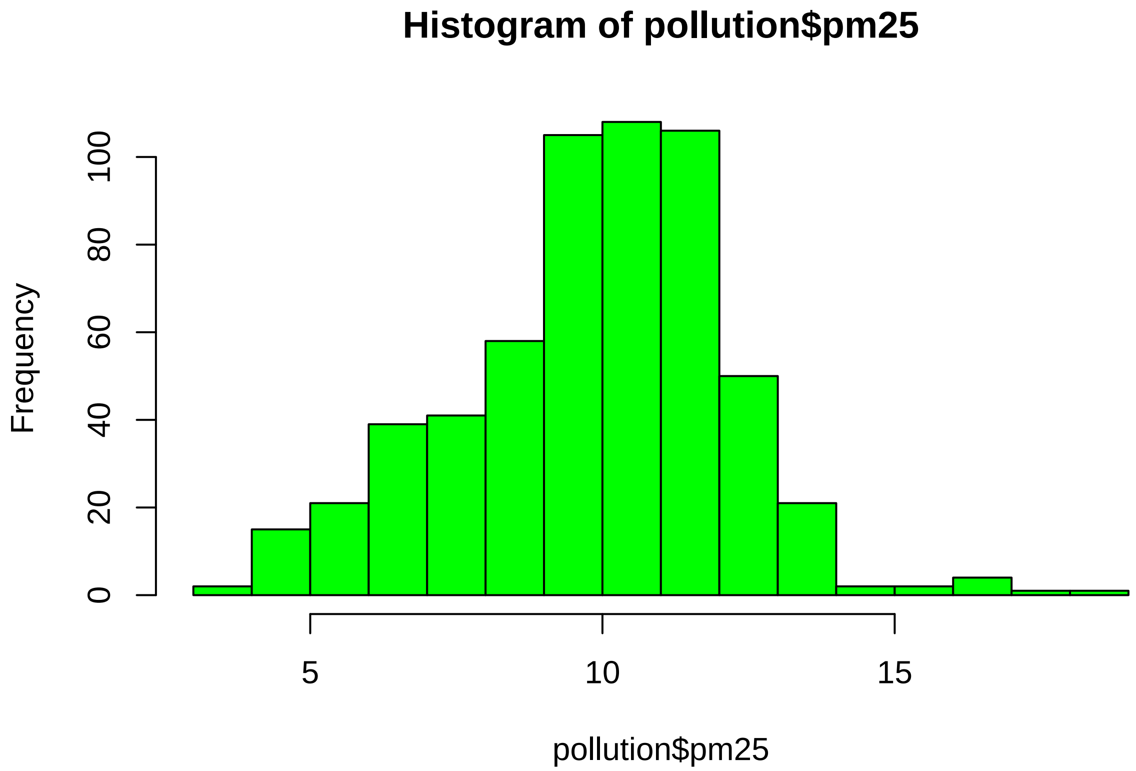
- Overlaying a horizontal line to help investigate our question

```
boxplot(pollution$pm25, col = "hotpink4")
abline(h = 12)
```



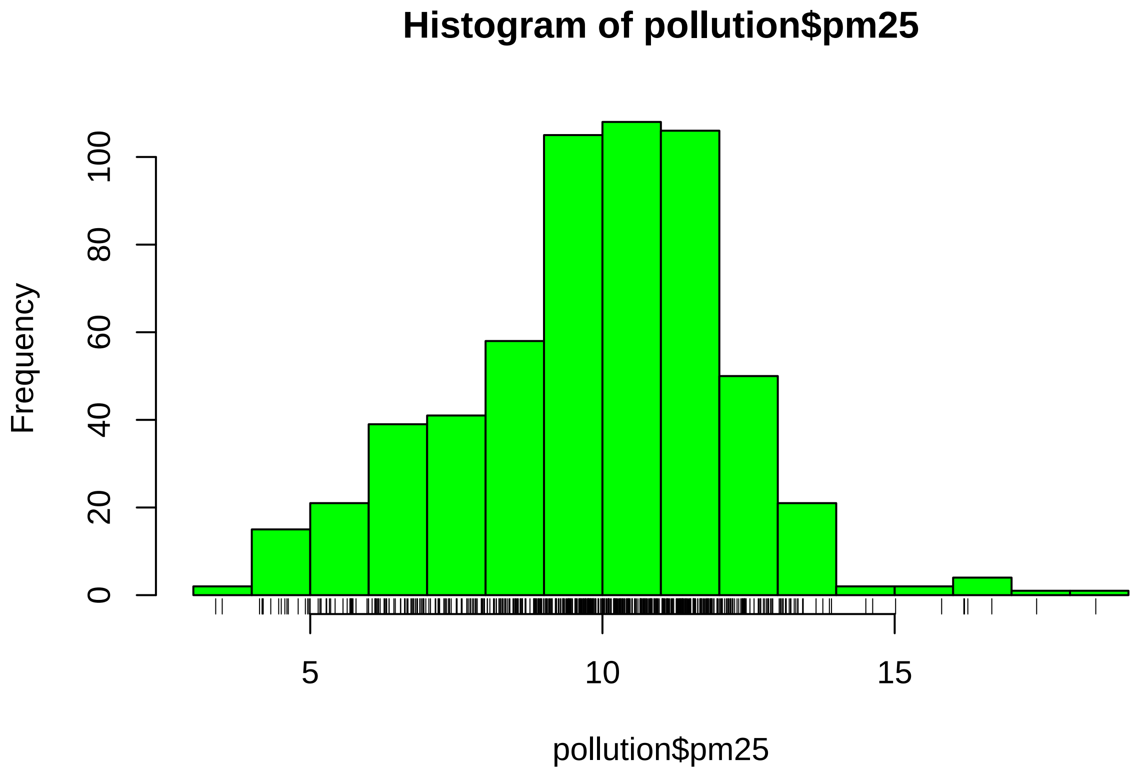
Histograms

```
hist(pollution$pm25, col = "green")
```



- Including a rug will show detail of the points that causing the plot

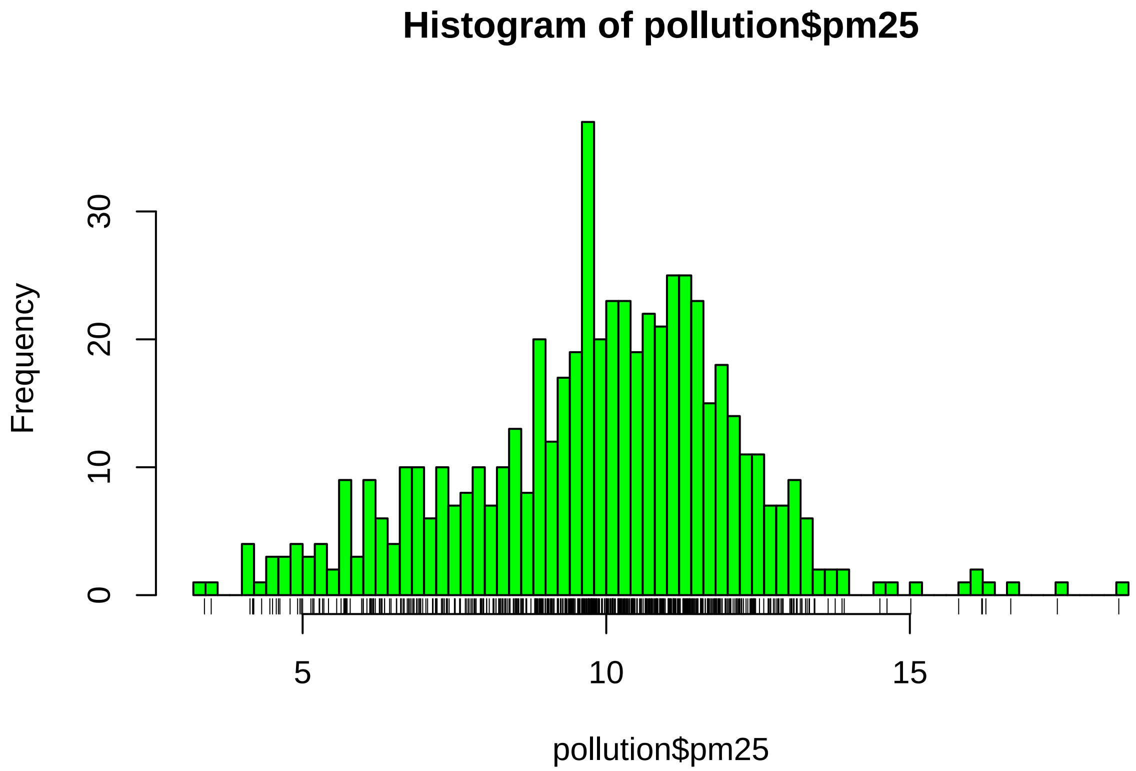
```r
hist(pollution$pm25, col = "green")
rug(pollution$pm25)
```



- One can also state the number of breaks that are to be in the histogram
  - too big of a number will make too much noise within the histogram

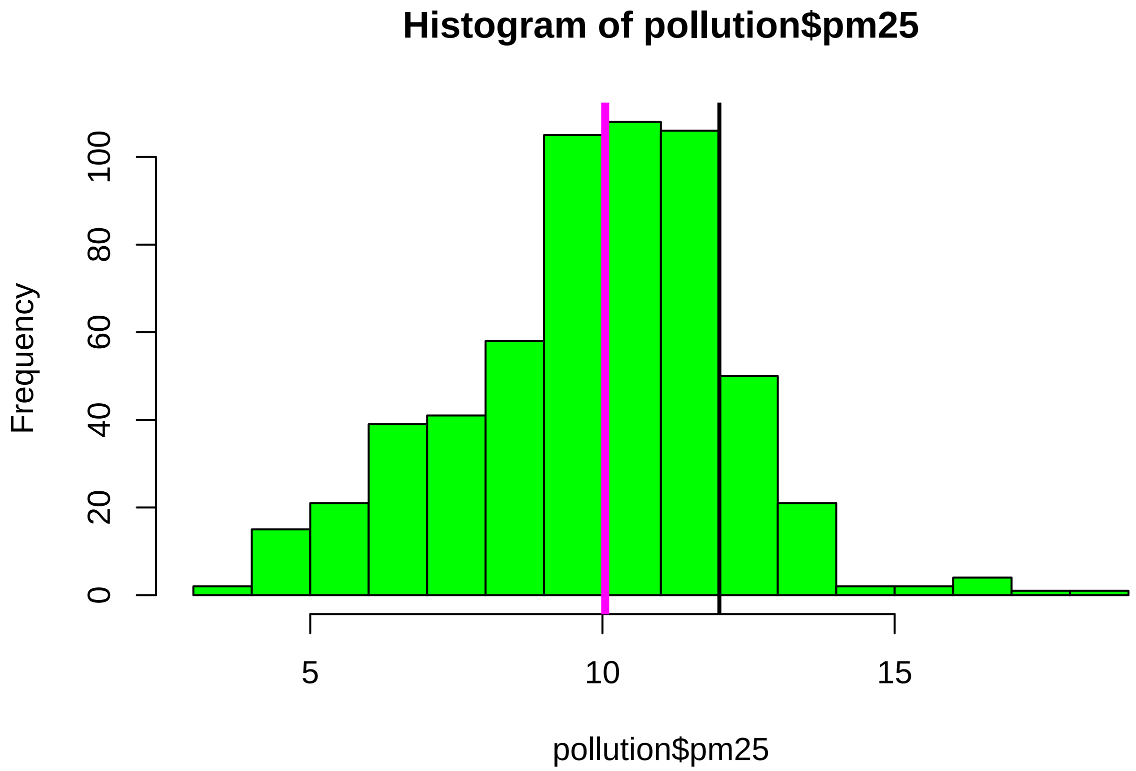
  - too small of a number won't show the shape of the distribution

```
hist(pollution$pm25, col = "green", breaks = 100)
rug(pollution$pm25)
```



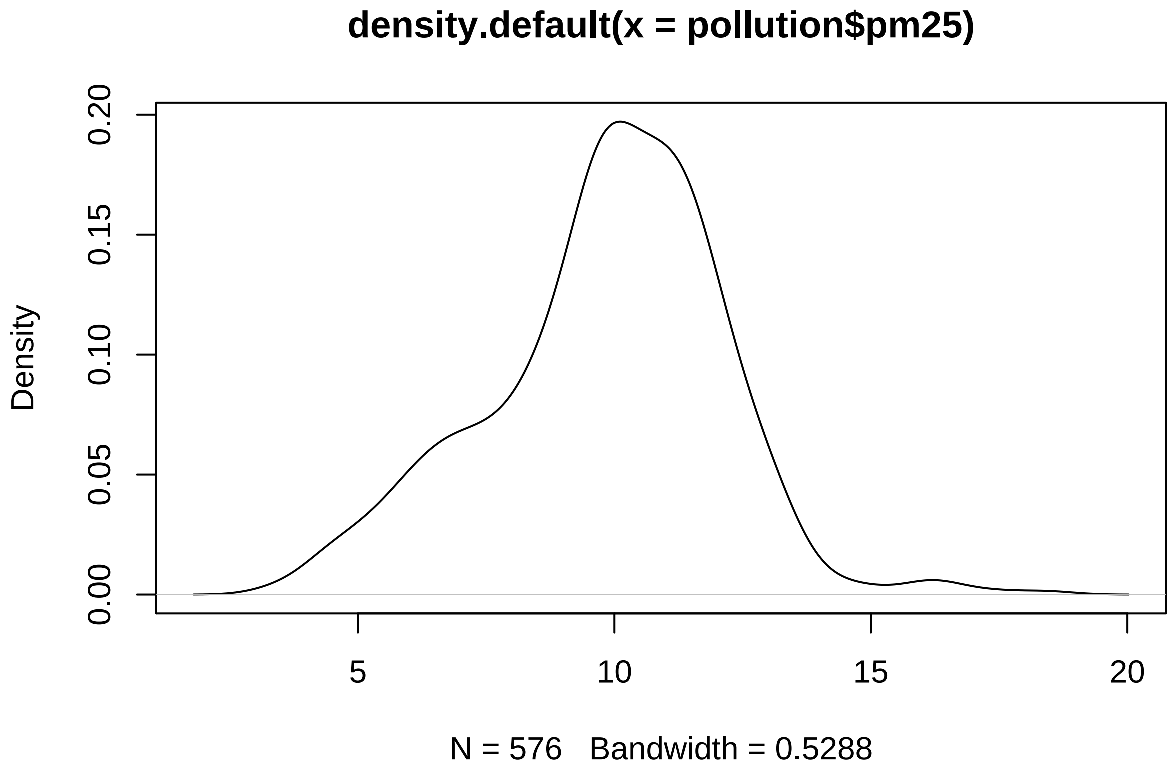
- Adding a vertical line and the median to the histogram

```
hist(pollution$pm25, col = "green")
abline(v = 12, lwd = 2) #lwd sets the width of the line
abline( v = median(pollution$pm25), col = "magenta", lwd = 4)
```



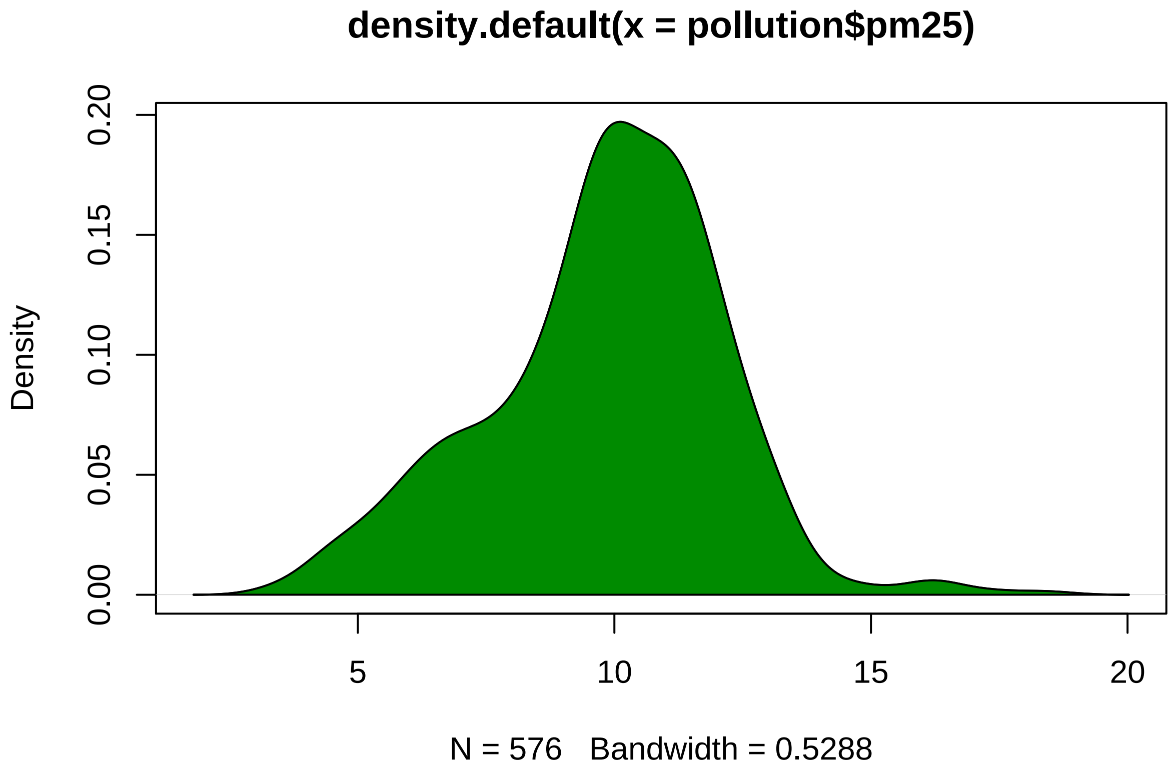
Density plot

```
plot(density(pollution$pm25))
```



- Adding a polygon to fill the area

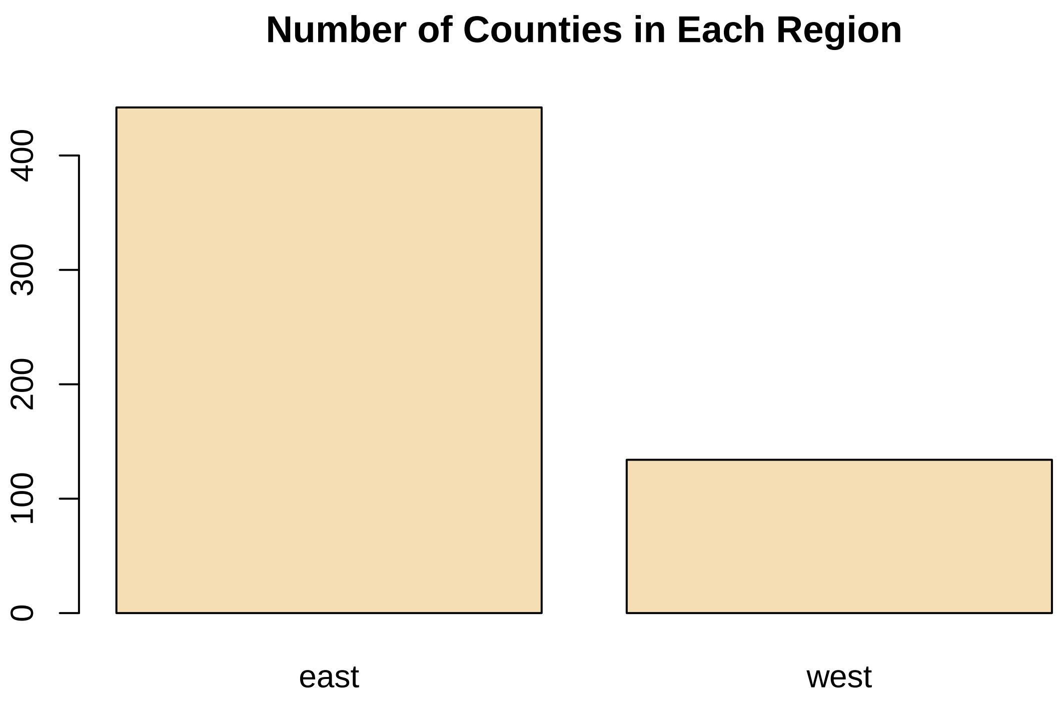
```
plot(density(pollution$pm25))
polygon(density(pollution$pm25), col = "green4")
```



Barplot

- used for comparing catagorical variables

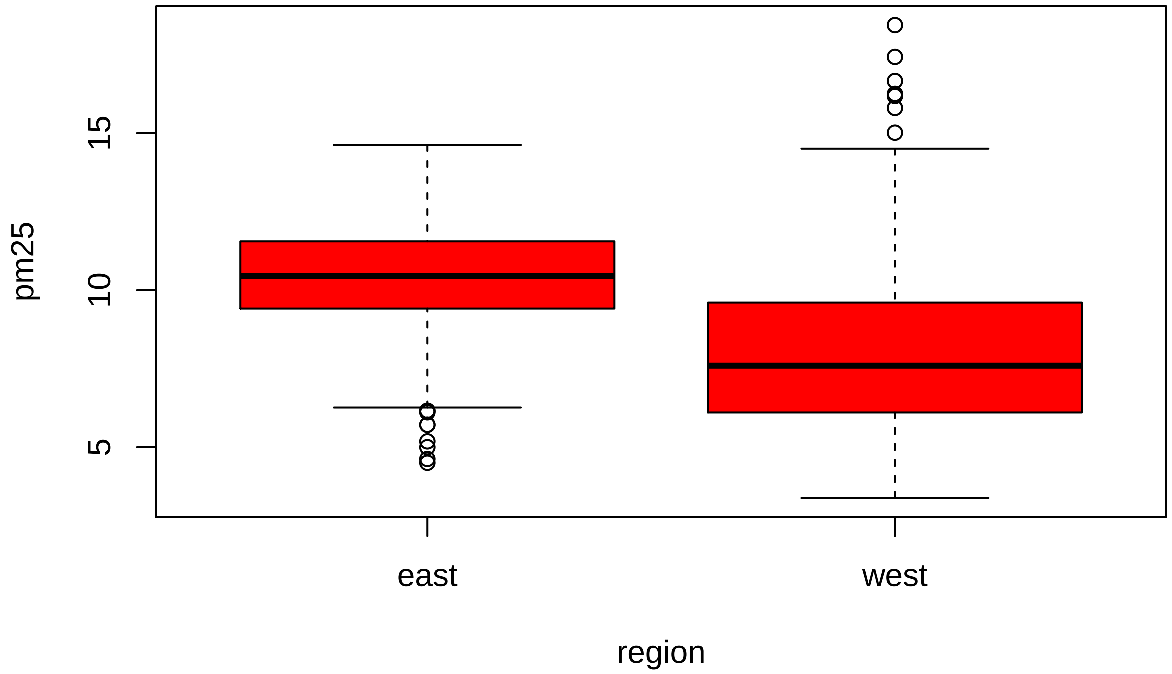
```
barplot(table(pollution$region), col = "wheat", main = "Number of Counties in Each Region")
```



- Simple Summaries of Data
    - Two dimensions
        - Multiple/overlayed 1-D plots (Lattice/ggplot2)
        - Scatterplots
        - Smooth scatterplots
    - Greater than 2 dimensions
        - Overlayed/multiple 2-D plots; coplots
        - Use color, size, shape to add dimensions
        - Spinning plots
        - Actual 3-D plots (not that useful)

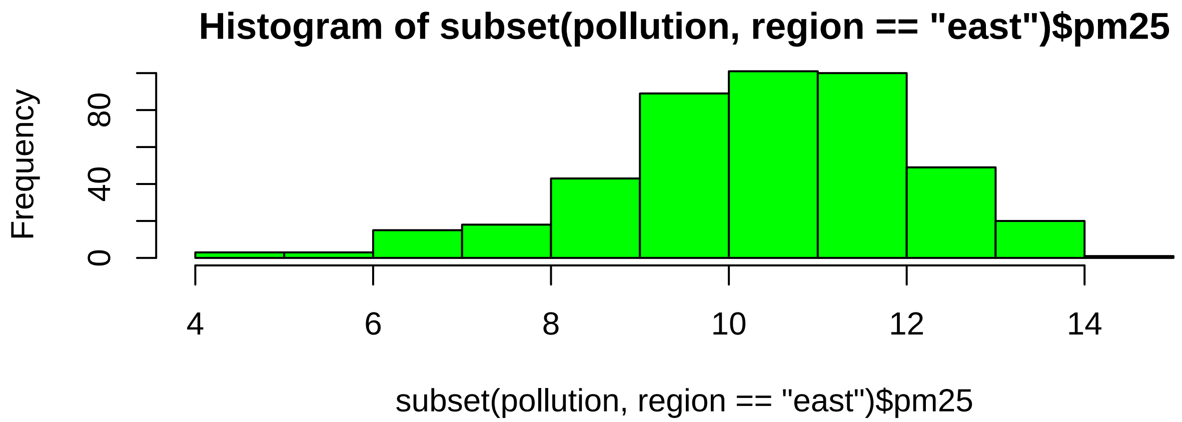
Multiple Boxplots

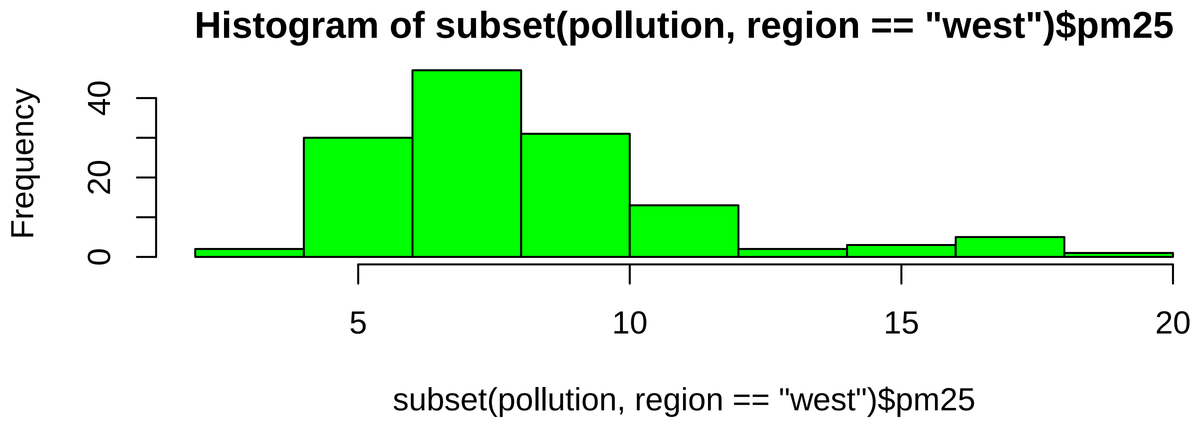
```r
#Look at pm25 ~(separated by) region
boxplot(pm25 ~ region, data = pollution, col = "red")
```



Multiple Histograms

```r
#mfrow determines the number of: c(row, col)
#mar is the size of the margins on the c(bottom, left, top, right)
par(mfrow = c(2, 1), mar = c(4, 4, 2, 1))
hist(subset(pollution, region == "east")$pm25, col = "green")
hist(subset(pollution, region == "west")$pm25, col = "green")
```

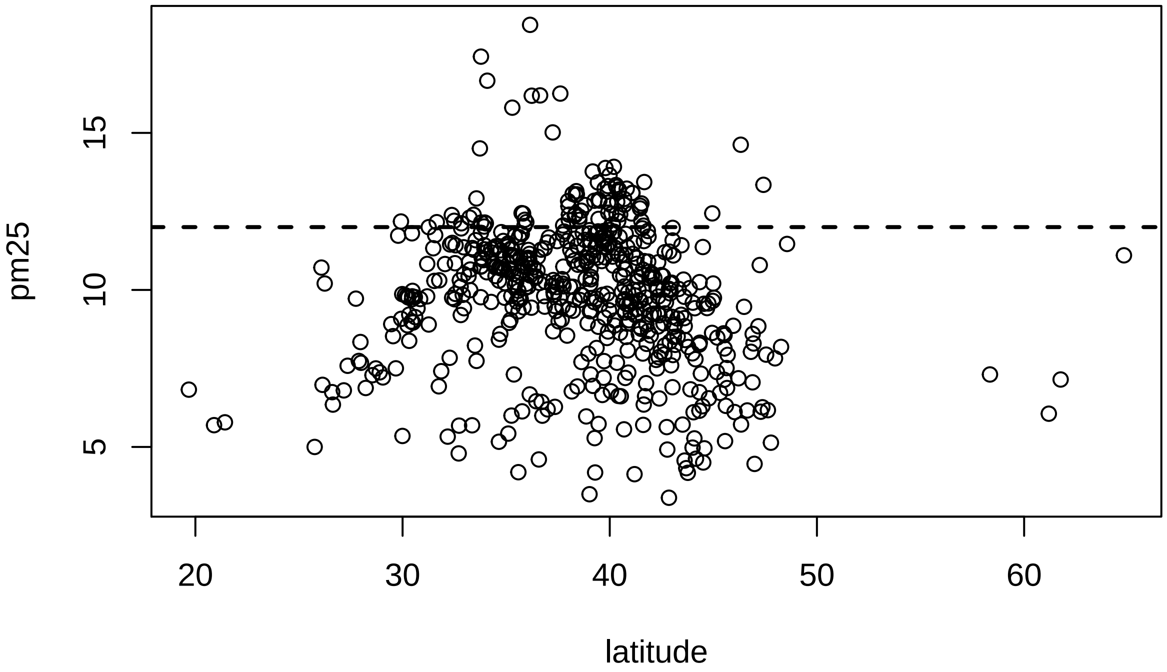




Scatterplot

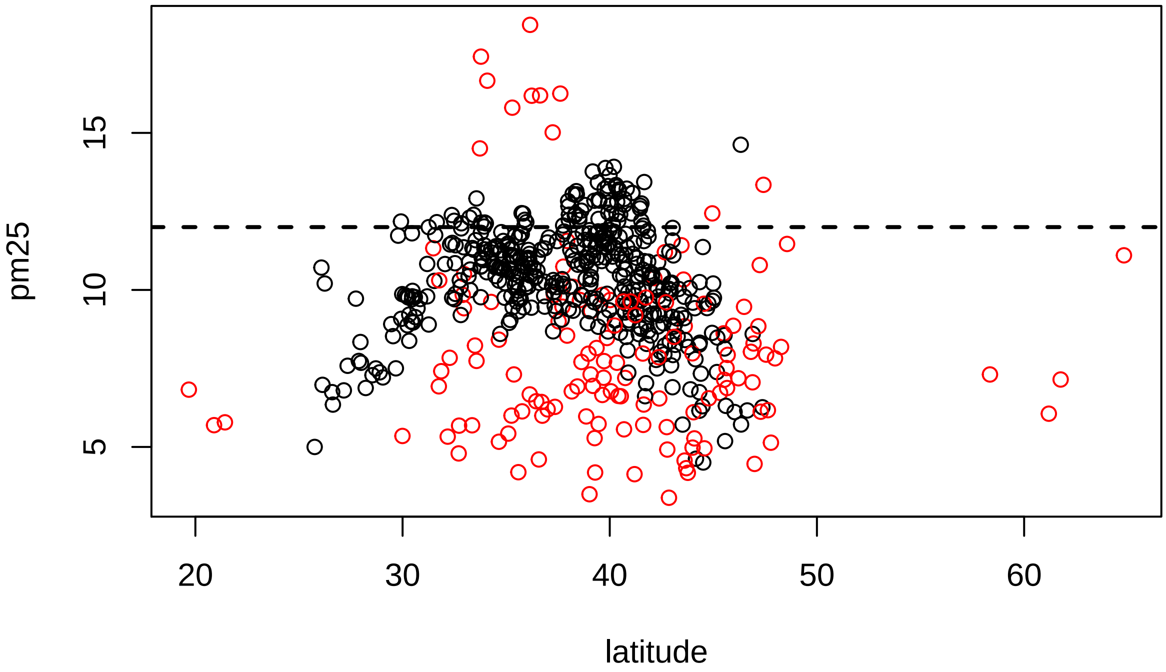
```
with(pollution, plot(latitude, pm25))

#lty = line type
abline(h = 12, lwd = 2, lty = 2)
```



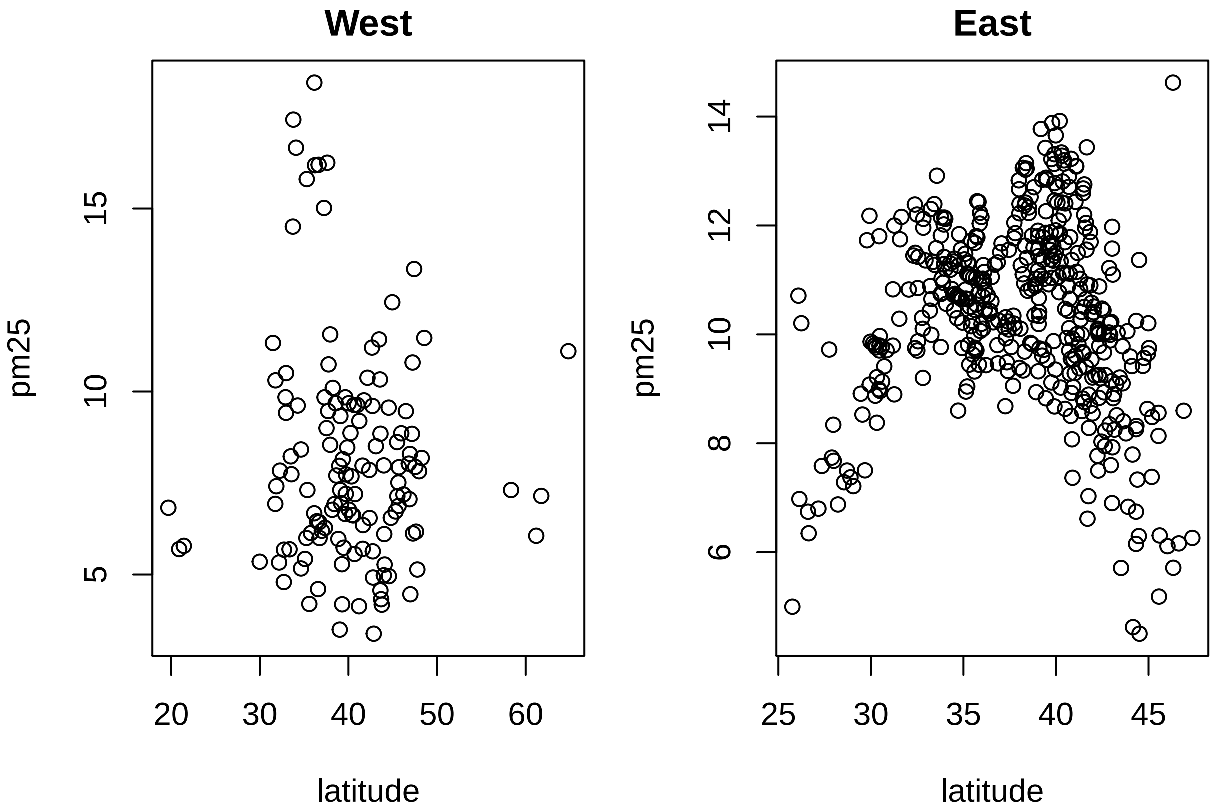
- Using Color

```
with(pollution, plot(latitude, pm25, col = region))
abline(h = 12, lwd = 2, lty = 2)
```



- Multiple Scatterplots

```
par(mfrow = c(1,2), mar = c(5, 4, 2, 1))
with(subset(pollution, region == "west"), plot(latitude, pm25, main = "West"))
with(subset(pollution, region == "east"), plot(latitude, pm25, main = "East"))
```



- Further Reading
  - **R Graph Gallery**

  - **R Bloggers**

Lesson with `swirl()`: Exploratory Graphs

- Since out brains are very good at seeing patterns, graphs give us a compact way to present data and find or display any pattern that may be present

- We ***don't*** use exploratory graphs to communicate results

- Exploratory graphs are the "quick and dirty" tool used to point the data scientist in a fruitful direction

- Plot details such as axes, legends, color, and size are cleaned up later to convey more information in an aesthetically pleasing way.

Reminder to commit to GitHub (Delete this line AFTER the commit)

Lesson 2: Plotting

Plotting Systems in R

Base Plotting System

Base Plotting Demonstration

Lesson with `swirl()`: Plotting Systems

Lesson with `swirl()`: Base Plotting System

Reminder to commit to GitHub (Delete this line AFTER the commit)

Lesson 3: Graphics Devices

Graphics Devices in R

Lesson with `swirl()`: Graphics Devices in R

Quiz 1 Scribbles

Reminder to commit to GitHub (Delete this line AFTER the commit)

Course Project 1

Reminder to commit to GitHub (Delete this line AFTER the commit)

Lesson 4: Lattice Plotting

Lattice Plotting System (part 1)

Lattice Plotting System (part 2)

Lesson with `swirl()`: Lattice Plotting System

Lesson with `swirl()`: Working with Colors

Reminder to commit to GitHub (Delete this line AFTER the commit)

Lesson 5: ggplot2 <3

Part 1

Part 2

Lesson with `swirl()`: GGPlot2 Part 1

Part 3

Part 4

Lesson with `swirl()`: GGPlot2 Part 2

Part 5

Lesson with `swirl()`: GGPlot2 Extras

Quiz 2 Scribbles

Reminder to commit to GitHub (Delete this line AFTER the commit)

Lesson 6: Hierarchical Clustering

Part 1

Part 2

Part 3

Lesson with `swirl()`: Hierarchical Clustering

Reminder to commit to GitHub (Delete this line AFTER the commit)

Lesson 7: K-Means Clustering & Dimension Reduction

K-Means Clustering (Part 1)

K-Means Clustering (Part 2)

Lesson with `swirl()`: K Means Clustering

Dimension Reduction (Part 1)

Dimension Reduction (Part 2)

Dimension Reduction (Part 3)

Lesson with `swirl()`: Dimension Reduction

Lesson with `swirl()`: Clustering Example

Reminder to commit to GitHub (Delete this line AFTER the commit)

Lesson 8: Working with Color in R Plots

Part 1

Part 2

Part 3

Part 4

Quiz 3 Scribbles

Reminder to commit to GitHub (Delete this line AFTER the commit)

Case Studies

Clustering Case Study

Air Pollution Case Study

Lesson with `swirl()`: CaseStudy

Quiz 4 Scribbles

Reminder to commit to GitHub (Delete this line AFTER the commit)

Course Project 2

Reminder to commit to GitHub (Delete this line BEFORE the commit)