

Chapter 4

EXPLORATORY GRAPHS

DATA ANALYSIS IN R

Outline

- Characteristics of exploratory graphs
- Use Case: Air Pollution in the United States
- Getting the Data
- Simple Summaries: One Dimension
- Five Number Summary
- Boxplot
- Histogram
- Overlaying Features
- Barplot
- Simple Summaries: Two Dimensions and Beyond
- Multiple Boxplots
- Multiple Histograms
- Scatterplots

Exploratory graphs

- To understand basic properties of the data
- To find simple patterns in data
- To suggest possible modeling strategies.
- In later stages of an analysis, graphics can be used:
 - To "debug" an analysis, if an unexpected (but not necessarily wrong) result occurs,
 - To communicate your findings to others.

Characteristics of exploratory graphs

- They are made quickly
- A large number is made
- The goal is for personal understanding
- Axes/legends are generally cleaned up (later)
- Color/size are primarily used for information

Use Case: Air Pollution in the US

- **Ambient air quality standards in the U.S. concerns the long-term average level of fine particle pollution, also referred to as PM2.5.**

The standard says that the "annual mean averaged over 3 years" cannot exceed 12 micrograms per cubic meter.

- **Data on daily PM2.5 are available from the U.S. EPA web site, or specifically, the EPA Air Quality System website.**

<https://ags.epa.gov/agsweb/airdata/download files.html>

https://github.com/jtleek/modules/blob/master/04_ExploratoryAnalysis/exploratoryGraphs/data/avgpm25.csv

- **One key question we are interested is:**

Are there any counties in the U.S. that exceed the national standard for fine particle pollution?

- counties that are found to be in violation of the national standards can face serious legal consequences
- In particular, states that have counties in violation of the standards are required to create a State Implementation Plan (SIP) that shows how those counties will come within the national standards within a given period of time

Getting the data

Read the data into R with `read.csv()`.

This dataset contains the annual mean PM2.5 averaged over the period 2008 through 2010

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

Do any counties exceed the national standard for fine particle pollution (12 micrograms per cubic meter.)?

Each row contains the 5-digit code indicating the county (fips), the region of the country in which the county resides, the longitude and latitude of the centroid for that county, and the average PM2.5 level.

Getting the data

Check more information on the dataset as given by str().

```
> str(pollution)
'data.frame':  576 obs. of  5 variables:
 $ pm25      : num  9.77 9.99 10.69 11.34 12.12 ...
 $ fips      : chr  "01003" "01027" "01033" "01049" ...
 $ region    : Factor w/ 2 levels "east","west": 1 1 1 1 1 1 1 1 1 1 ...
 $ longitude: num  -87.7 -85.8 -87.7 -85.8 -86 ...
 $ latitude : num  30.6 33.3 34.7 34.5 34 ...
```

How can we see if any counties exceed the standard of 12 micrograms per cubic meter?

Simple Summaries: One Dimension

- For one dimensional summarize, there are number of options in R.
- Five-number summary
- Boxplots
- Barplot
- Histograms:
- Density plot:
density() function computes a non-parametric estimate of the distribution of a variable

Five Number Summary

- This gives the minimum, 25th percentile, median, 75th percentile, maximum of the data and is quick check on the distribution of the data

A five-number summary can be computed with the `fivenum()` function, which takes a vector of numbers as input.

- Five-number summary of the PM2.5 data in the pollution dataset.

```
> fivenum(pollution$pm25)
[1] 3.382626 8.547590 10.046697 11.356829 18.440731
```

We can see that the median across all the counties in the dataset is about 10 micrograms per cubic meter.

- Let's check the `summary()` function.

```
> summary(pollution$pm25)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 3.383   8.549  10.047   9.836  11.356  18.441
```

In addition to the five-number summary the `summary()` function also adds the mean of the data, which can be compared to the median to identify any skewness in the data. (No dramatic amount of skewness in the distribution of PM2.5 in this dataset)

Boxplot

- **Boxplots are a visual representation of the five-number summary plus a bit more information.**

Boxplots commonly plot outliers that go beyond the bulk of the data.

Implemented via the `boxplot()` function

- **Boxplot of the PM2.5 data.**

In a boxplot, the "whiskers" that stick out above and below the box have a length of 1.5 times the *inter-quartile range*, or **IQR**, which is simply the distance from the bottom of the box to the top of the box.

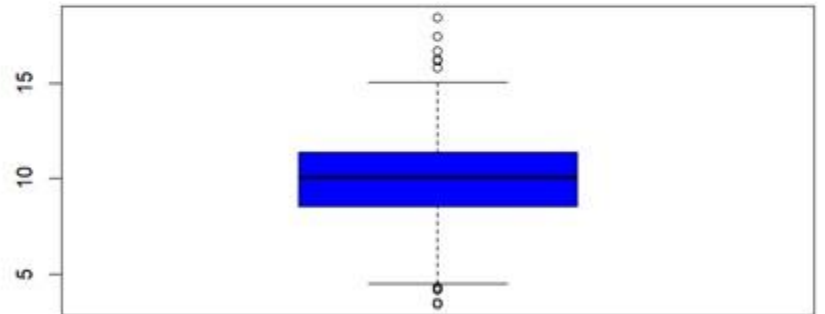
Anything beyond the whiskers is marked as an "outlier" and is plotted separately as an individual point.

```
> boxplot(pollution$pm25, col = "blue")
```

From the boxplot:

1

- few points on both the high and the low end that appear to be outliers according to the `boxplot()` algorithm.
- the high points are all above the level of 15, so we can take a look at those data points directly.



Analysis of PM2.5 data

Check the high points above the level of 15

```
library(dplyr)
filter(pollution, pm25>15)
```

	pm25	fips	region	longitude	latitude
1	16.19452	06019	west	-119.9035	36.63837
2	15.80378	06029	west	-118.6833	35.29602
3	18.44073	06031	west	-119.8113	36.15514
4	16.66180	06037	west	-118.2342	34.08851
5	15.01573	06047	west	-120.6741	37.24578
6	17.42905	06065	west	-116.8036	33.78331
7	16.25190	06099	west	-120.9588	37.61380
8	16.18358	06107	west	-119.1661	36.23465

- These counties are all in the western U.S. (region==west) specifically all in California because the first two digits of the fips code are 06.
- We can make a quick map of these counties to get a sense of where they are in California.

Analysis of PM2.5 data

Check the high points above the level of 15 on map

```
install.packages("maps")  
library(maps)  
map("county", "california")  
with(filter(pollution, pm25>15), points(longitude,  
latitude))
```



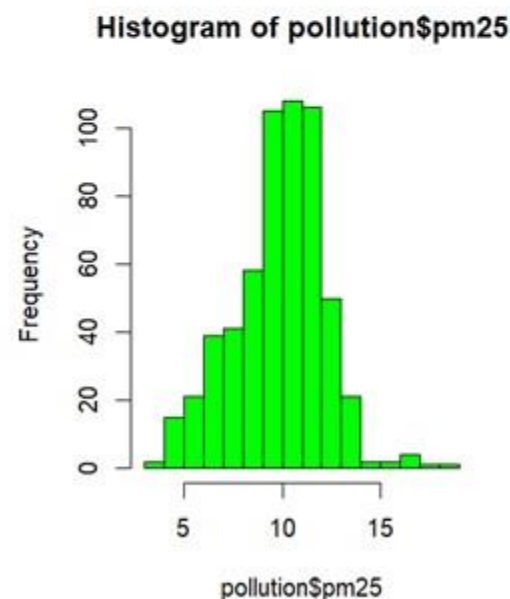
At this point, you might decide to follow up on these counties that are potentially in violation of the standards.
or ignore them if you are interested in other features.

Histogram

- A histogram is useful to look at when we want to see more detail on the full distribution of the data.
- Histograms show the complete empirical distribution of the data, beyond the five data points shown by the boxplots.
- You can easily check skewness of the data, symmetry, multi-modality, and other features.
- The `hist()` function makes a histogram, and a handy function to go with it sometimes is the `rug()` function

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

This distribution is interesting because there appears to be a high concentration of counties in the neighborhood of 9 to 12 micrograms per cubic meter.



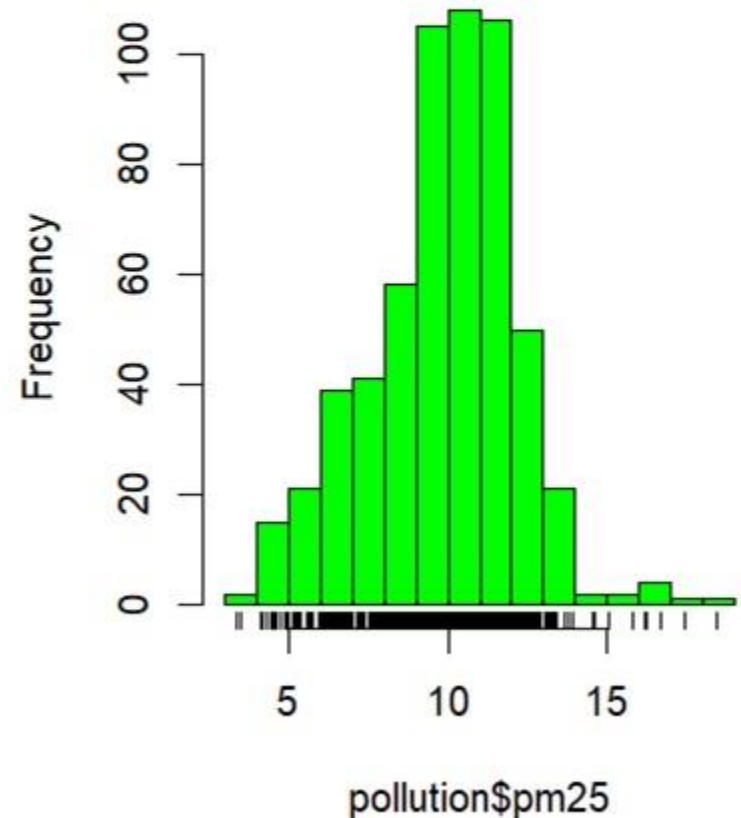
Histogram

To get more detail we use the `rug()` function to show us the actual data points.

```
rug(pollution$pm25)
```

- The large cluster of data points is in the 9 to 12 Range
- There are still quite a few counties above the level of 12, which may be worth investigating.
- The `hist()` function has a default algorithm for determining the number of bars to use in the histogram based on the density of the data.
- You can override the default option by setting the **breaks** argument to something else.

Histogram of pollution\$pm25

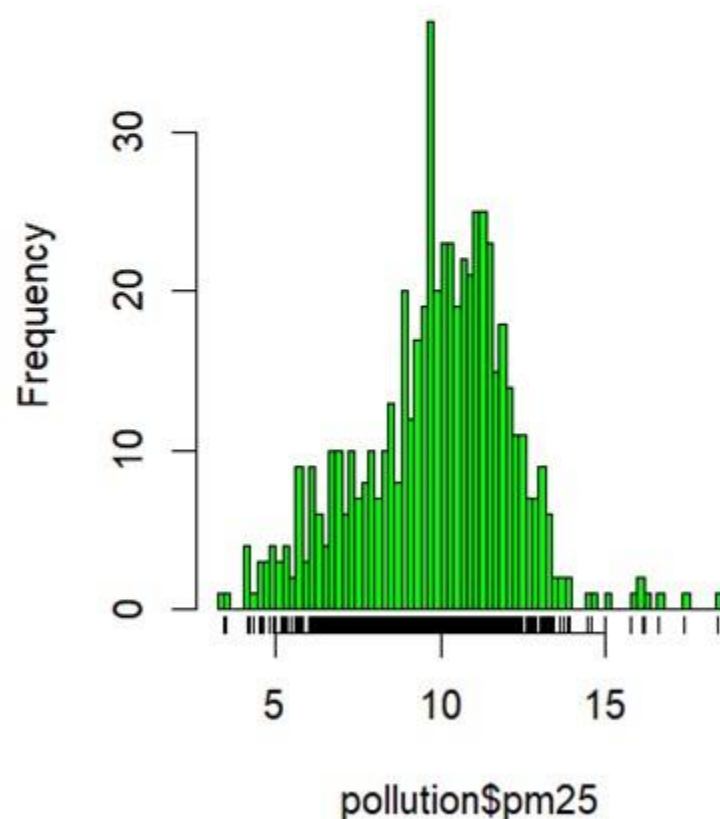


Histogram: breaks

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

- Now we see that there's a rather large spike 9 micrograms per cubic meter. It's not immediately clear why, but again, it might be worth following up on.

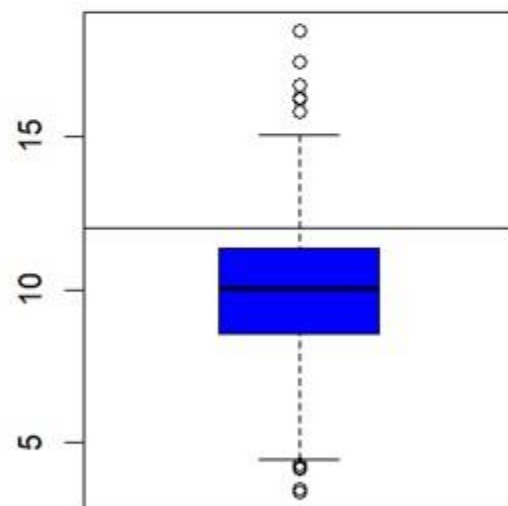
Histogram of pollution\$pm25



Overlaying Features

- Once we start seeing interesting features in our data, it's often useful to lay down annotations on our plots as reference points and markers.
- For example, in our boxplot: draw a horizontal line at 12 where the national standard is.

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



- A reasonable portion of the distribution as displayed by the boxplot is above the line (i.e. potentially in violation of the standard).

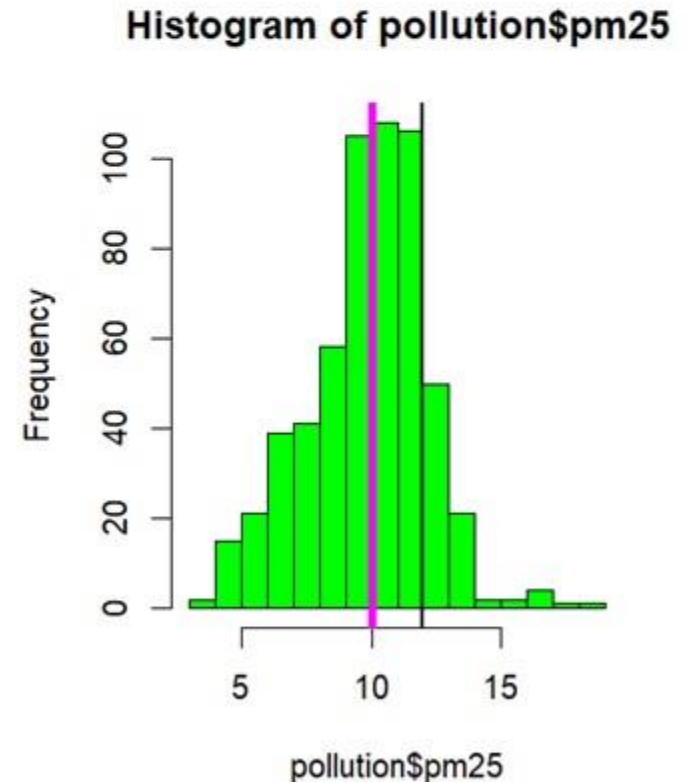
Overlaying Features

While the boxplot gives a sense, the histogram might be better suited to visualizing the data here.

In the plot below, we see the histogram and draw two lines, one at the median of the data and one at 12, the level of the standard.

```
hist(pollution$pm25, col = "green")  
abline(v = 12, lwd = 2)  
abline(v = median(pollution$pm25), col = "magenta", lwd = 4)
```

color (col) and the line width (lwd) indicate components of information.

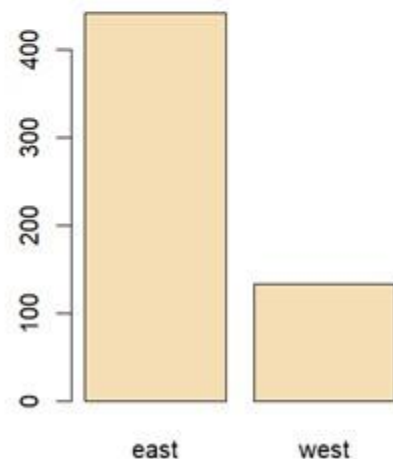


Barplot

- Useful for summarizing categorical data, with the number of entries for each category being proportional to the height of the bar
- The barplot can be made with the `barplot()` function.
- **Example: In PM2.5 data:**
 - one categorical variable the region in which a county resides (east or west).
 - Check how many western and eastern counties there are with `barplot()`.
 - Use the `table()` function to do the actual tabulation of how many counties there are in each region.

```
table(pollution$region) %>% barplot(col = "wheat")
```

We can see quite clearly that there are many more counties in the eastern U.S. in this dataset than in the western U.S.



Simple Summaries: Two Dimensions and Beyond

- For investigating data in two dimensions and beyond:
- Multiple or overlaid 1-D plots (Lattice/ggplot2): Using multiple boxplots or multiple histograms can be useful for seeing the relationship between two variables, especially when one is naturally categorical.
- Scatterplots: Scatterplots are the natural tool for visualizing two continuous variables. Transformations of the variables (e.g. log or square-root transformation) may be necessary for effective visualization.
- Smooth scatterplots: Similar in concept to scatterplots but rather plots a 2-D histogram of the data. Can be useful for scatterplots that may contain many many data points.

Simple Summaries: Two Dimensions and Beyond

- For visualizing data in more than 2 dimensions, we can often combine the tools:
- Overlaid or multiple 2-D plots; conditioning plots (coplots):

A conditioning plot, or coplot shows the relationship between two variables as a third (or more) variable changes.

For example, you might want to see how the relationship between air pollution levels and mortality changes with the season of the year. Here, air pollution and mortality are the two primary variables and season is the third variable varying in the background.
- Use color, size, shape to add dimensions:

Plotting points with different colors or shapes is useful for indicating a third dimension where different colors can indicate different categories or ranges of something. Plotting symbols with different sizes can also achieve the same effect when the third dimension is continuous.

Simple Summaries: Two Dimensions and Beyond

- For visualizing data in more than 2 dimensions, we can often combine the tools:
- **Spinning/interactive plots:**

Spinning plots can be used to simulate 3-D plots by allowing the user to essentially quickly cycle through many different 2-D projections so that the plot feels 3-D. These are sometimes helpful to capture unusual structure in the data but rarely used.
- **Actual 3-D plots (not that useful):** Actual 3-D plots (for example, requiring 3-D glasses) are relatively few and far between and are generally impractical for communicating to a large audience.

Multiple Boxplots

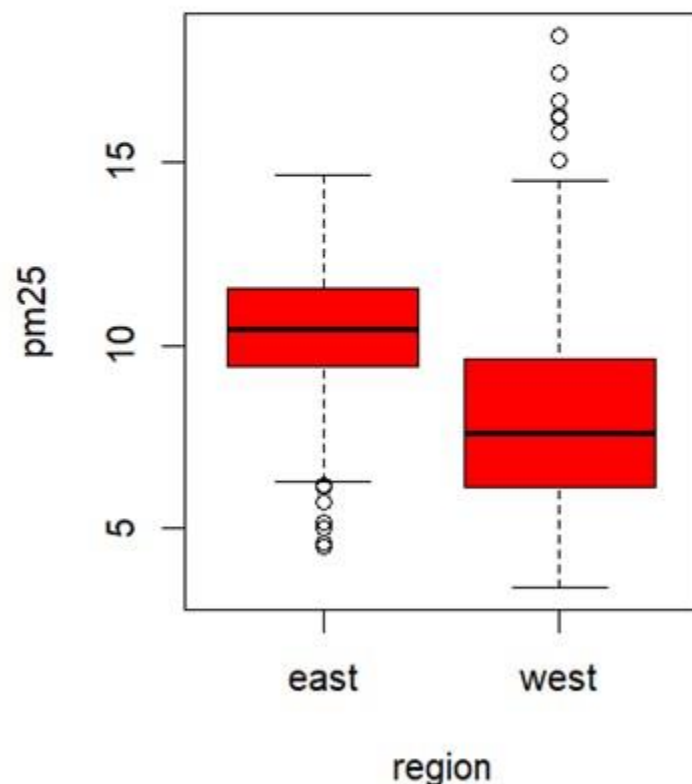
- One of the simplest ways to show the relationship between two variables (in this case, one categorical and one continuous) is to show side-by-side boxplots.

Using the pollution data, we can show the difference in PM2.5 levels between the eastern and western parts of the U.S. with the `boxplot()` function.

```
boxplot(pm25~region, data = pollution, col = "red")
```

The `boxplot()` function can take *a formula*, indicating the variable for which we want to create the boxplot (continuous) and the variable that stratifies the left hand side into categories. Since the region variable only has two categories we end up with two boxplots.

From the plot: the levels in eastern counties are on average higher than the levels in western counties

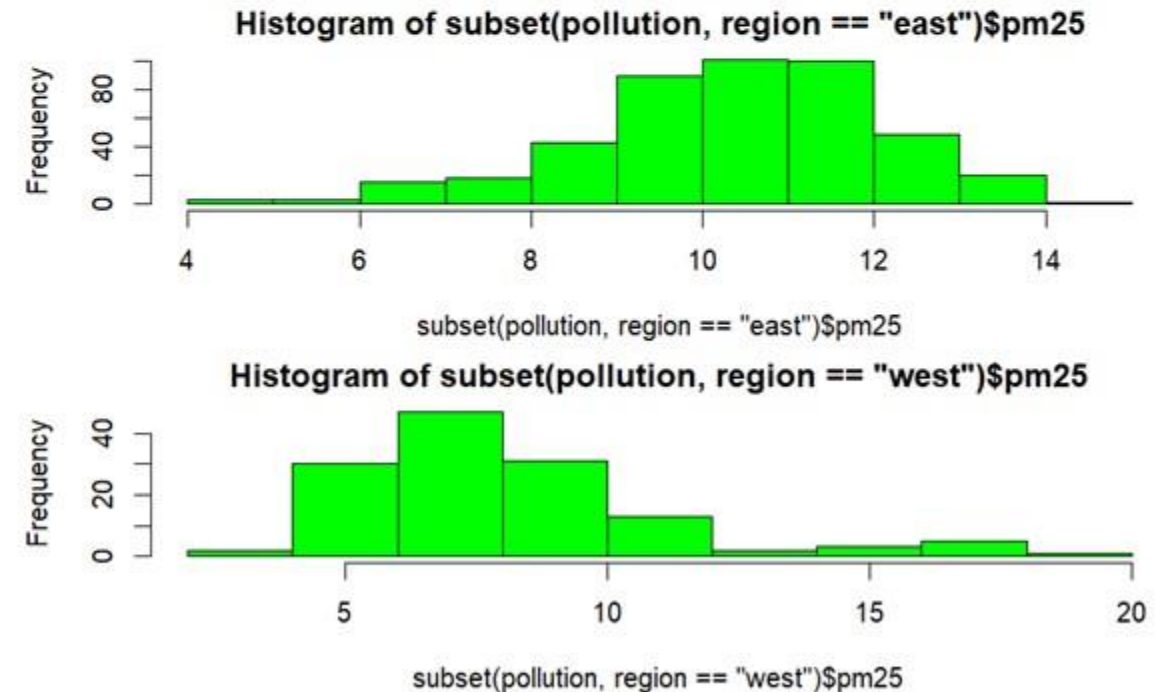


Multiple Histograms

- It can sometimes be useful to plot multiple histograms, much like with side-by-side boxplots, to see changes in the shape of the distribution of a variable across different categories.
- Example: the distribution of PM2.5 in the eastern and western regions of the U.S.

```
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")
```

You can see here that the PM2.5 in the western U.S. tends to be right-skewed with some outlying counties with very high levels. The PM2.5 in the east tends to be left skewed some counties having very low levels.

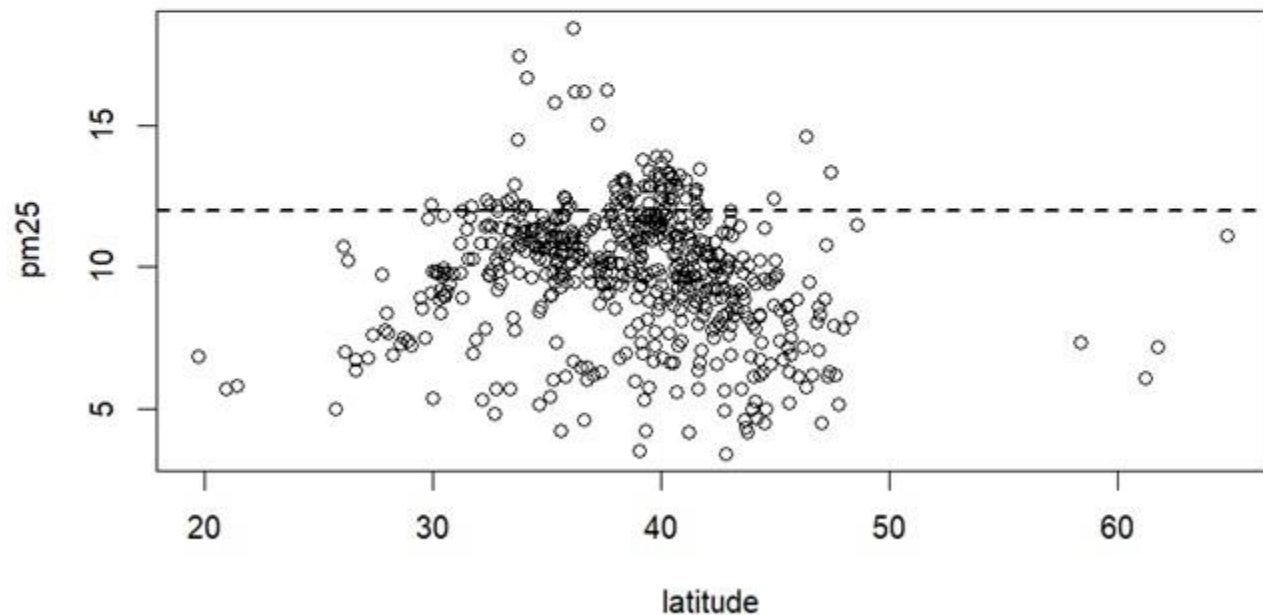


Scatterplots

- For continuous variables, the most common visualization technique is the scatterplot, which simply maps each variable to an x- or y-axis coordinate.
- Example: scatterplot of latitude and PM2.5, which can be made with the `plot()` function.

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

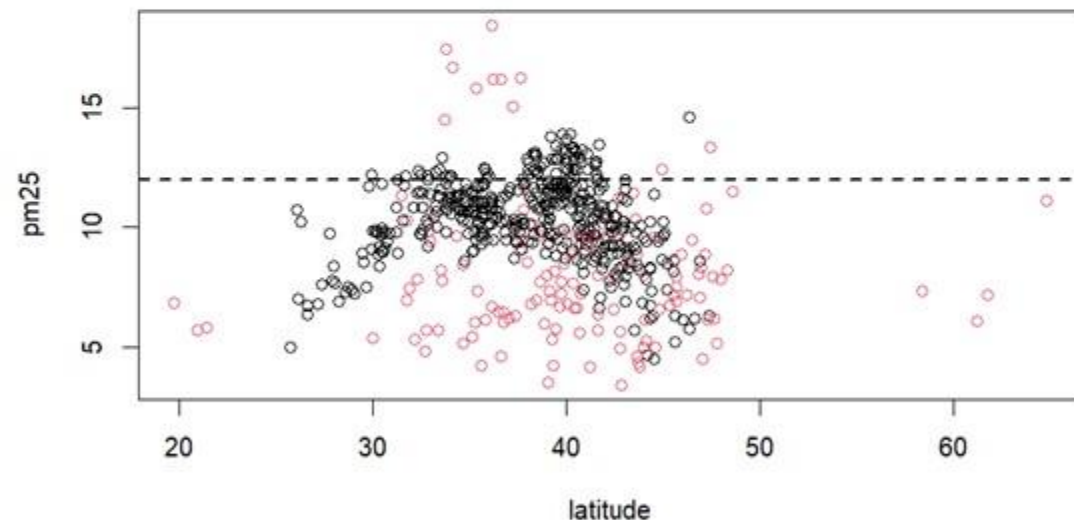
As you go from south to north in the U.S., we can see that the highest levels of PM2.5 tend to be in the middle region of the country.



Scatterplot - Using Color

Add a third dimension to the scatterplot, say the region variable indicating east and west we could use color to highlight that dimension. Here we color the circles in the plot to indicate east (black) or west (red).

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



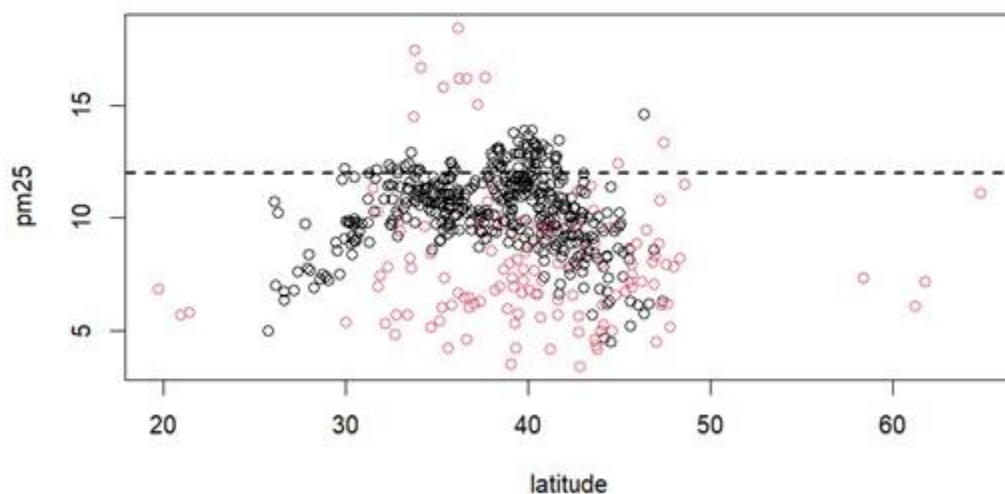
Which color gets mapped to which region?

Scatterplots

- It may be confusing at first to figure out which color gets mapped to which region. We can find out by looking directly at the levels of the region variable.

```
> levels(pollution$region)
[1] "east" "west"
```

- The first level is "east" and the second level is "west".
- So the color for "east" will get mapped to 1 and the color for "west" will get mapped to 2.
 - For plotting functions, col = 1 is black (the default color) and col = 2 is red.



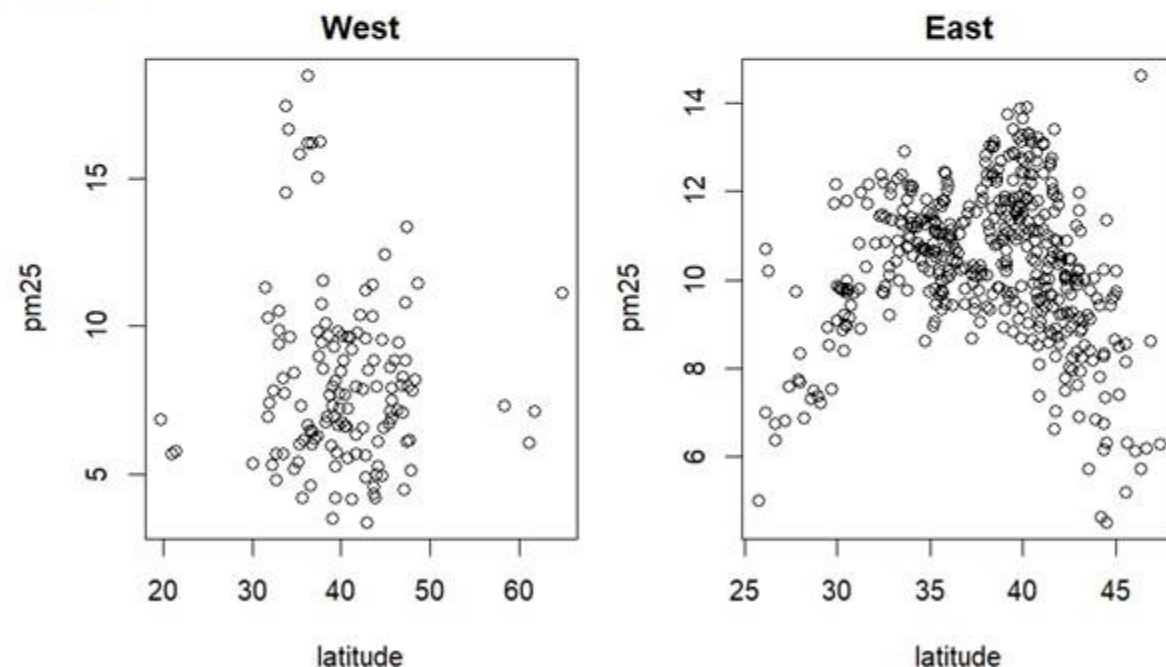
Multiple Scatterplots

- Using multiple scatterplots can be necessary when overlaying points with different colors or shapes is confusing (sometimes because of the volume of data).

Separating the plots out can sometimes make visualization easier.

```
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"))
```

These kinds of plots, sometimes called panel plots, are generally easier to make with either the lattice or ggplot2 system, which we will learn about in greater detail in later chapters.



Summary

- Exploratory plots are "quick and dirty" and their purpose is to let you summarize the data and highlight any broad features.
- They are also useful for exploring basic questions about the data and for judging the evidence for or against certain hypotheses.
- They may be useful for suggesting modeling strategies that can be employed in the "next step" of the data analysis process.

Questions