Chapter 4

Understanding and Comparing Distributions

D. Raffle
5/20/2015

Overview

In previous chapters, we looked at:

- The distributions of single quantitative variables
- The distributions of single categorical variables
- Compared distributions of two categorical variables

In this chapter, we will:

- Compare the distributions of numeric variables across groups
- Look at the distribution of numeric variables across time

The Data

For this chapter, we will use the Motor Trend Cars data set we looked at in Chapter 2.

```
##                    mpg cyl disp  hp    wt  qsec vs     am
## Mazda RX4         21.0   6  160 110 2.620 16.46  V   auto
## Mazda RX4 Wag     21.0   6  160 110 2.875 17.02  V   auto
## Datsun 710        22.8   4  108  93 2.320 18.61  S   auto
## Hornet 4 Drive    21.4   6  258 110 3.215 19.44  S manual
## Hornet Sportabout 18.7   8  360 175 3.440 17.02  V manual
## Valiant           18.1   6  225 105 3.460 20.22  S manual
```

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

The Variables

Variable	Description
mpg	Miles/(US) Gallon
cyl	Number of Cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
wt	Weight (lb/1000)
qsec	1/4 mile time
vs	V or Straight Configuration Engine
am	Transmission Type (auto/manual)

Comparing Groups

Generally, looking at the distribution of a single variable is less informative than see how things change across groups.

- Do men and women live as long?
- Are young people more dangerous behind the wheel than older people?
- Does being in a fraternity affect academic performance?
- Do manual cars have higher performance than automatics?

We will discuss statistical tests to do this later on, but we can get a good idea of where differences exist by graphing the distribution of a variable within each group and seeing if there's a difference.

Flashback: Comparing Two Categorical Variables

To compare two categorical variables, we look at the frequencies (or relative frequencies) of one variable within the levels of another.

We can, for example, look engine configuration vs. transmission type.

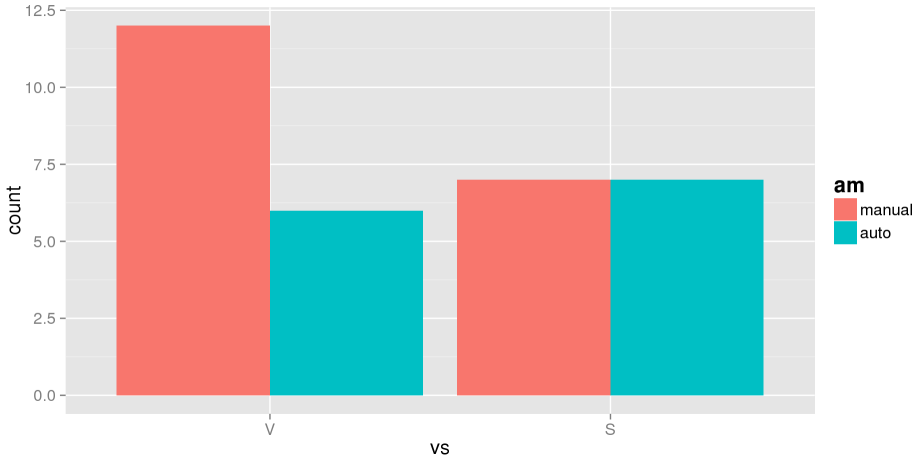
```
##    am manual auto
## vs
## V         12    6
## S          7    7
```

What can we conclude?

- It looks like straight cylinders were equally likely to appear in manuals and automatics
- The V configuration appeared twice as often in manuals than automatics

Flashback: Visualizing Two Categorical Variables

We can visualize this with a stacked barplot.



Comparing Groups: Histograms

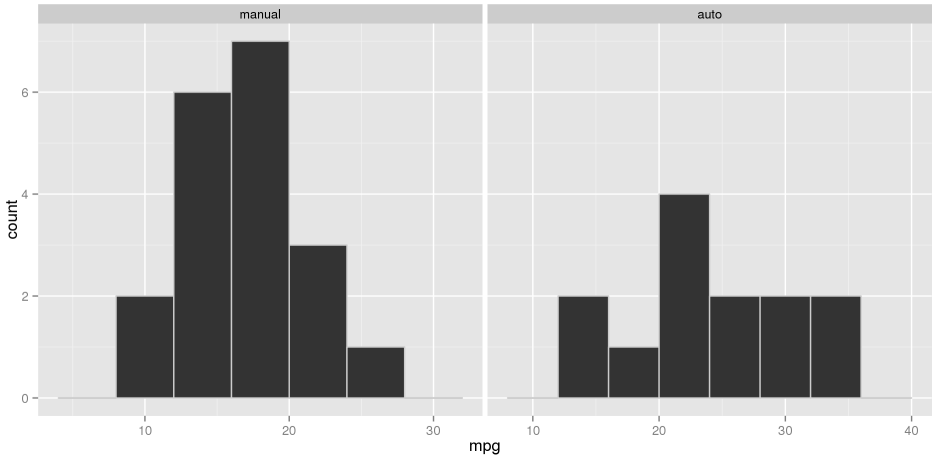
This simplest option is to create a histogram for each group. There are a few things to keep in mind:

- The axes need to have the same scale in every group, or the comparison is meaningless
- Differences in the centers tell us that the variable has different typical values between groups
- Differences in spread tell us that there is more variability in different groups
- Different shapes tell us that the distribution completely shifts between groups

First, let's compare the fuel efficiency for manuals and automatics.

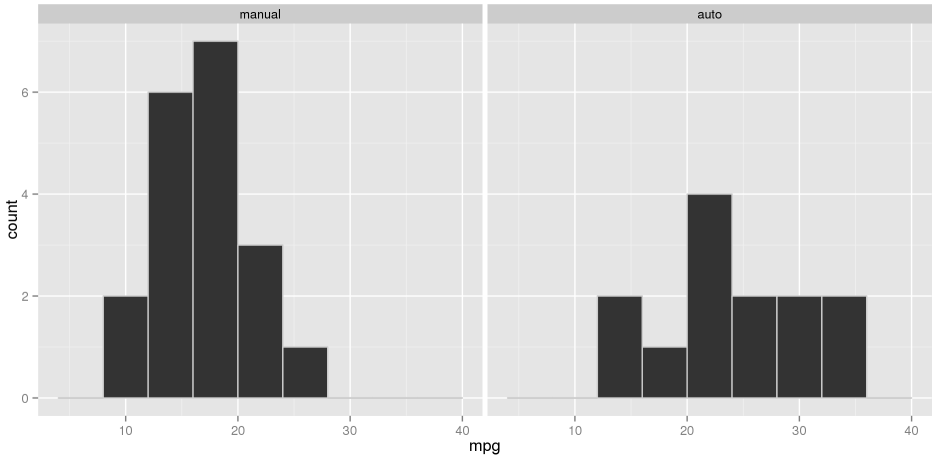
MPG vs. Transmission: Side-by-Side Histograms

First, the wrong way:



MPG vs. Transmission: Side-by-Side Histograms

Now the right way:



MPG vs. Transmission: Numerical Summaries

To further investigate what we saw in the histograms, we'll look at the five number summaries of MPG within each group.

```
##        Min. 1st Qu. Median 3rd Qu. Max.
## manual 10.4   14.95   17.3    19.2 24.4
## auto   15.0   21.00   22.8    30.4 33.9
```

Each quartile looks higher for the autos, but so is the IQR. Because the distributions are *mostly* symmetric, we can also look at the mean and standard deviation:

```
##        Mean     St.Dev
## manual 17.14737 3.833966
## auto   24.39231 6.166504
```

Note that not only is the mean higher in the auto, but they have almost twice as much variability.

Comparing Distributions: Boxplots

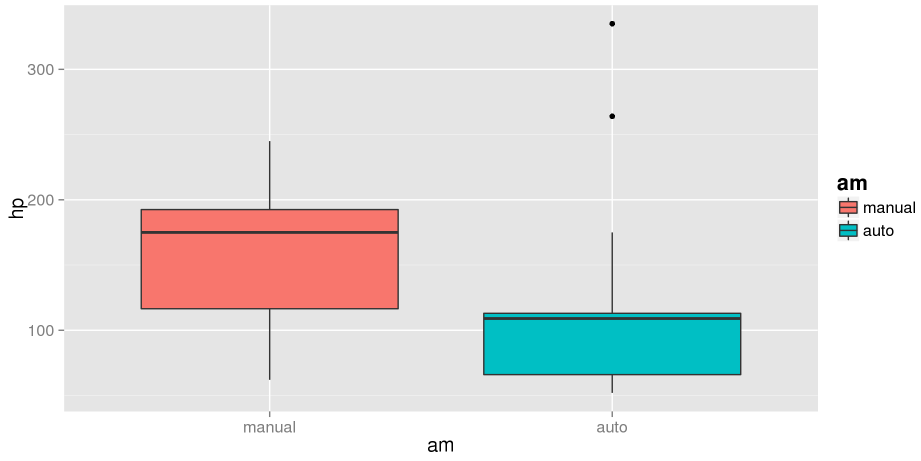
We previously used boxplots to examine the distribution of a single numeric variable. The main advantage of boxplots is that they also offer us a compact way to compare the distribution of a numeric variable across groups.

Recall that a boxplot is a visualization of the five number summary:

- The box height (or length) of the box is a representation of the IQR, with the edges representing Q1 and Q3
- The median is represented as a line in the box
- The whiskers either represent the range of the data *or* the cut-off for outliers (1.5 IQRs away from the median)
- If there are outliers, they're represented by points extending past the whiskers

In *side-by-side boxplots*, we simply stack them next to each other.

Transmission vs. Horsepower: Boxplots



Transmission vs. Horsepower: Numerical Summary

There certainly looks like a difference in the distribution of horsepower across transmission types. However, there are some automatics which have horsepowers greater than the max of the manuals.

```
##        Min. 1st Qu. Median 3rd Qu. Max.
## manual   62   116.5    175   192.5  245
## auto     52    66.0    109   113.0  335
```

We can also check which cars have the highest horsepower:

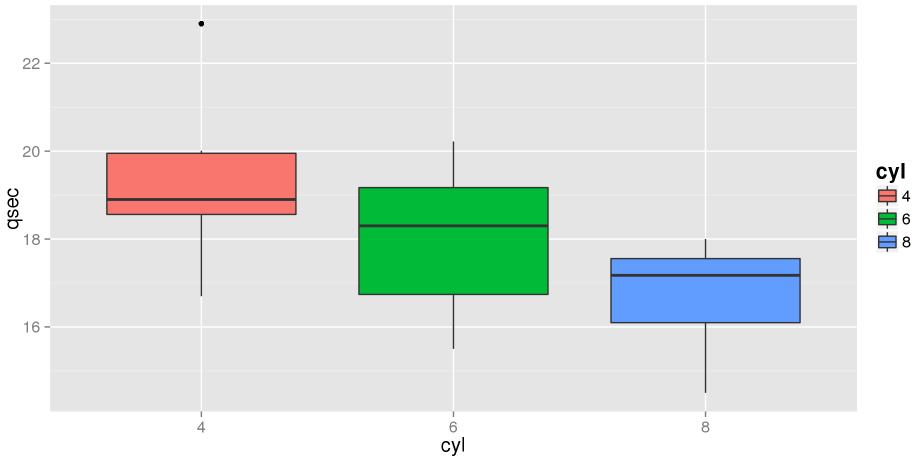
```
##                 mpg cyl disp  hp   wt  qsec vs     am
## Maserati Bora  15.0   8  301 335 3.57 14.60  V   auto
## Ford Pantera L 15.8   8  351 264 3.17 14.50  V   auto
## Duster 360     14.3   8  360 245 3.57 15.84  V manual
## Camaro Z28     13.3   8  350 245 3.84 15.41  V manual
```

Using Numeric Variables as Groups

Think about the variable cyl, which records the number of cylinders.

- Does it really make sense to view it as a numeric variable?
- Typically, we only see cars with 4, 6, or 8 cylinders
- A mean number of cylinders isn't really interpretable – a car can't really have 4.5 cylinders
- It may be more reasonable to treat it as a categorical variable

Cylinders vs. Quarter Mile Time



Cylinders vs. Quarter Mile Time

A couple nuances show up here that we haven't seen before:

- There isn't a huge differnce in time between 4 & 6 or 6 & 8 cylinder cars. There is substantial overlap in the boxes if you only compare groups to their neighbors
- There is, however, a fairly substantial difference in 4 & 8 cylinder cars
- The six cylinder cars have much more spread than the other types

```
##   Min. 1st Qu. Median 3rd Qu.  Max.
## 4 16.7   18.56  18.90   19.95 22.90
## 6 15.5   16.74  18.30   19.17 20.22
## 8 14.5   16.10  17.18   17.56 18.00
```

Dealing with Outliers

We've seen outliers in several of these plots, so dealing with them should be mentioned.

- Outliers can either be indicative of a problem in our data, or they may simply show us that our sample wasn't large enough to capture the true patterns that exist
- For example, they may be extreme in our sample, but they may not be that rare in a larger population or in the longterm
- In the case of controlled experiments, that particular test tube, petri dish, or other experimental unit may have been contaminated
- It could simply be that someone made a typo when entering the data

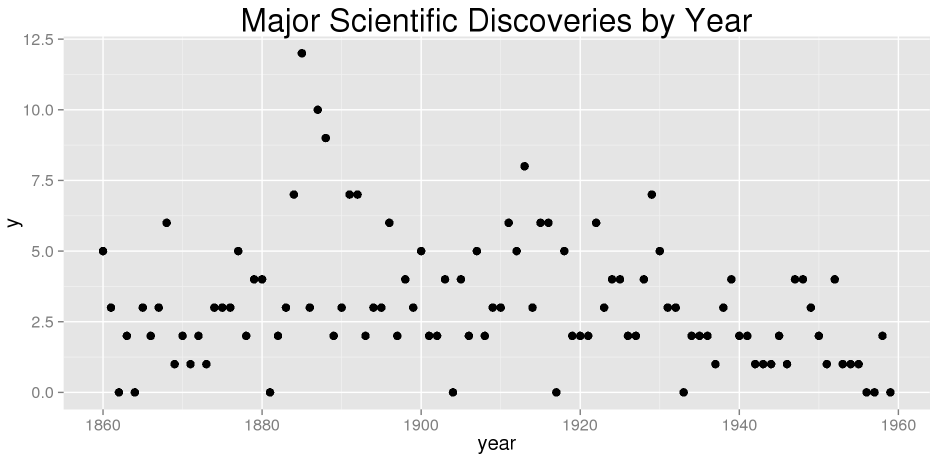
All outliers should be investigated, especially if you are collecting data yourself. However, you should **never** remove data unless you have a good reason to.

Graphing Data Across Time

Time based data needs special attention:

- Time does not work like a normal variable, it only ever varies in one direction
- Finding the mean day of the week or standard deviation of the month doesn't make intuitive sense
- Typically, we treat time data as its own variable type.
- The models and techniques for dealing with time can be very complex, so we can't get into them in this class
- For now, we will just look at plotting time data.

Timeplots



20/41

Timeplots

Timeplots or time series plots graph time, measured in hours, days, months, year, etc., against a numeric variable

- We usually put time on the x-axis and the variable on the y-axis
- Each observation is represented as a point
- We often draw some line to represent the overall pattern

Time data can be highly variable, so we often attempt to highlight the overall pattern with a *trace*.

- A *trace* is a line that represents the overall pattern

Traces

Traces are drawn through the points in an attempt to highlight the overall pattern. There are two extreme types of traces:

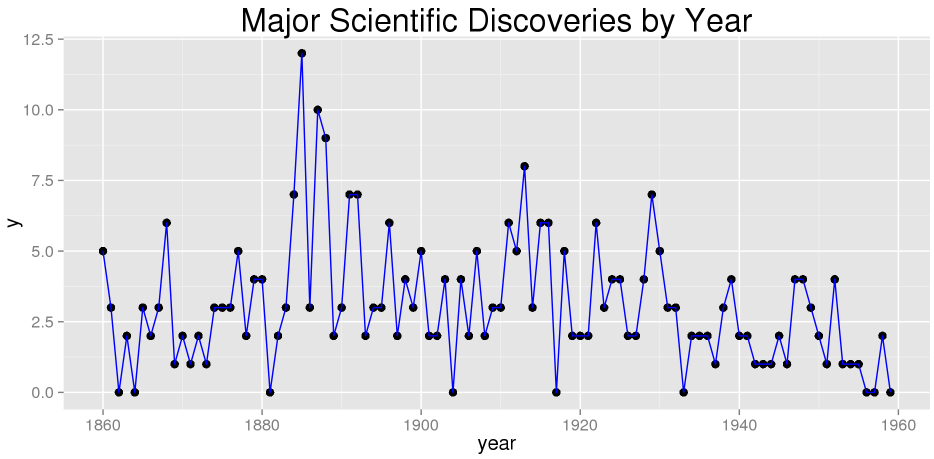
Connecting Every Point:

- This is the most variable (wiggly) version. Statisticians often call this an *interpolating fit*.
- These traces show the exact pattern, but it's often hard to see the true pattern
- We might see more *noise* than *signal*

Fitting a Straight Line:

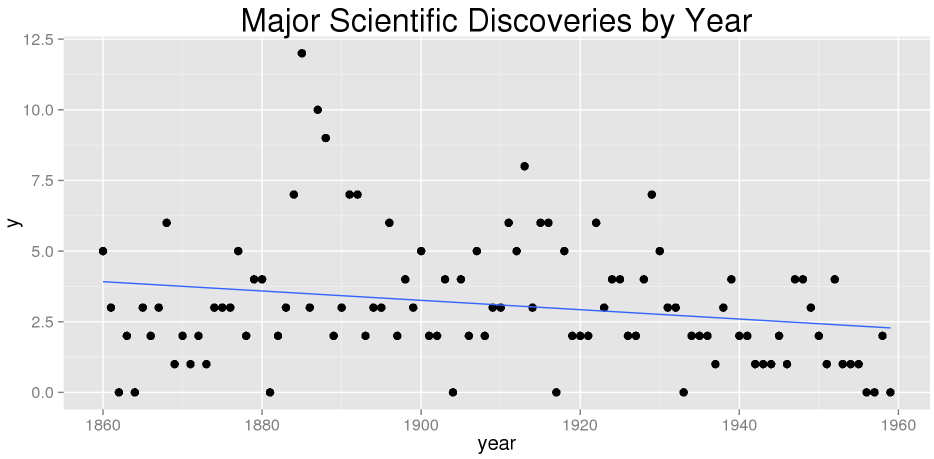
- This is the least variable (smoothest) version. We usually call this a *linear fit*
- Linear fits show us the broadest pattern possible: does the variable go up or down over time?
- They can fail to detect more subtle patterns

An Interpolating Trace



23/41

A Linear Trace



The Middle Ground: A Smoothing Trace

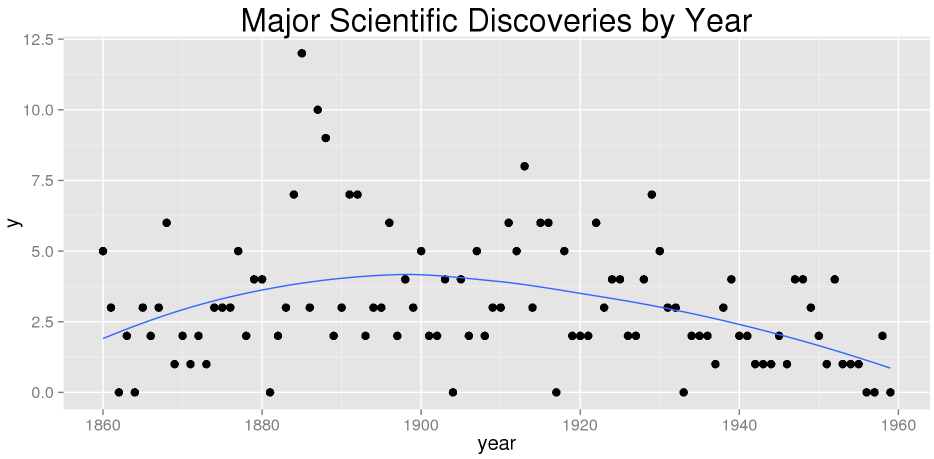
We probably want to be somewhere between the two extremes. This middle ground is called a *smoothing trace*

- We want to see peaks and valleys, but we don't want to connect every dot
- Ideally, we'll see a *smooth* curve

This is typically accomplished using a *moving average* and a *window*

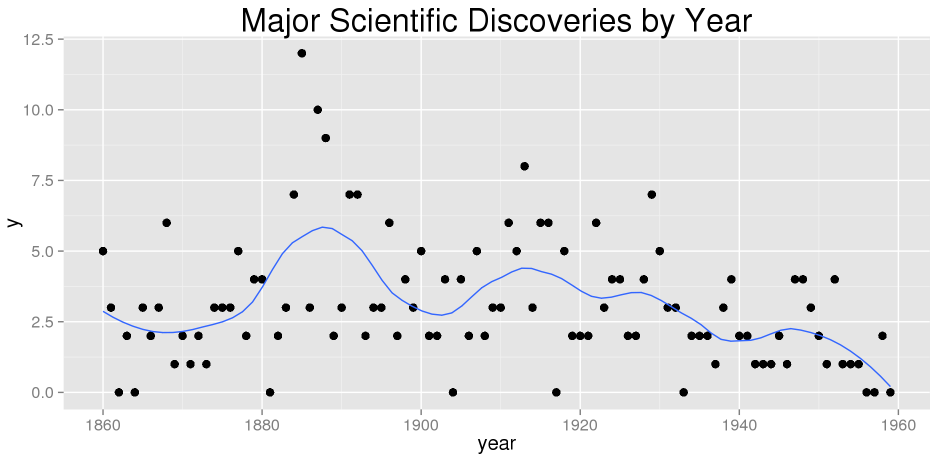
- For an individual point, we average only a few points in a *window* around it
- We do this for every point, then draw a smooth curve connected each average
- The larger the window, the smoother the curve
- Some types of windows only look in the past and give lower weights to points further back. We call this *exponential smoothers*.

Smoothing Trace: Large Window



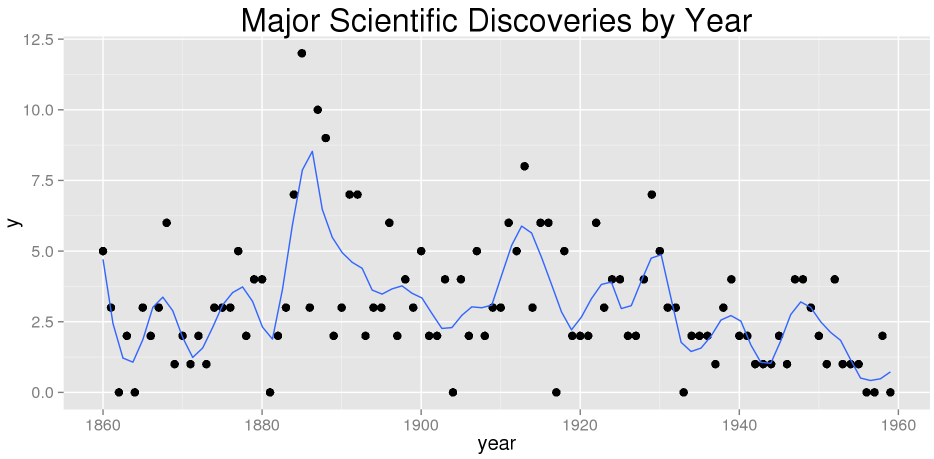
26/41

Smoothing Trace: Medium Window



27/41

Smoothing Trace: Small Window



28/41

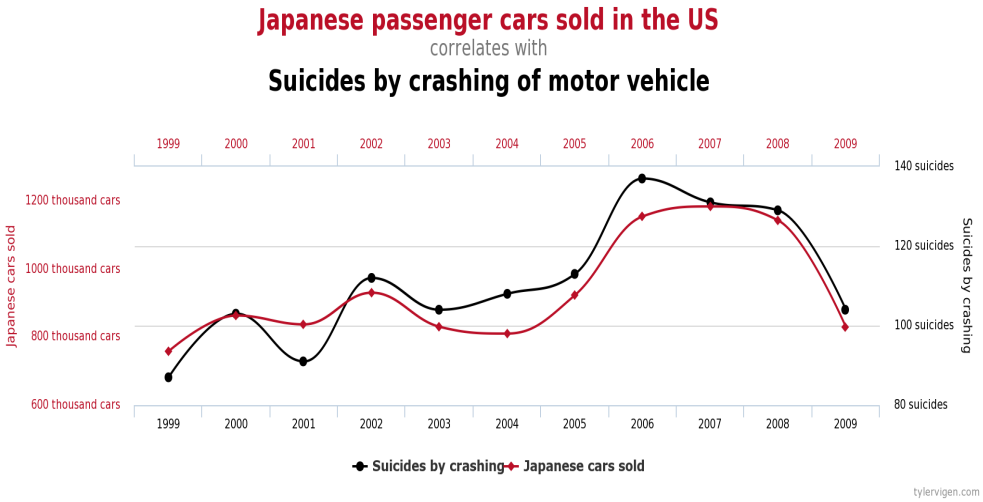
Timplots: Watch the Axis

Statistics is a tool for analyzing data. Unfortunately, some people will intentionally use bad statistics to prove a point or intentionally mislead an audience.

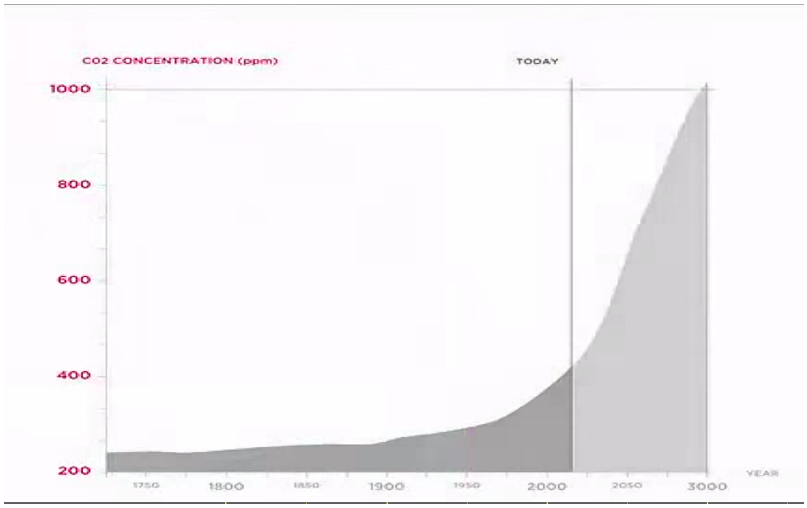
This is especially common in time plots, but holds for many visualizations. Some things to watch out for:

- Make sure both axes are consistent (start at zero, no breaks or gaps)
- Make sure each axis has only one scale
- Everything is properly labelled with units

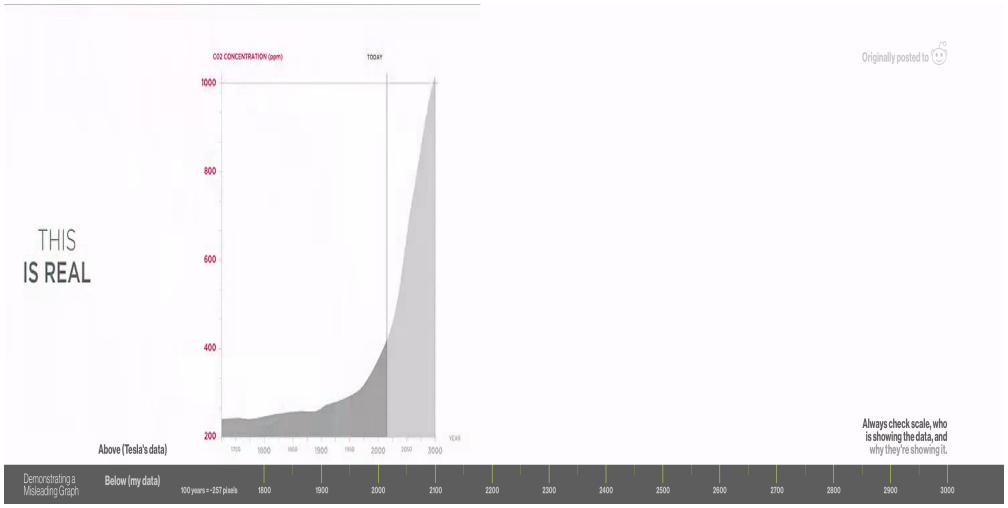
Double Scales



Broken Scale



Fixed Scale


CO2 CONCENTRATION (ppm)
TODAY
1000
800
600
400
200
YEAR
THIS
IS REAL
Originally posted to
Always check scale, who is showing the data, and why they're showing it.
Above (Tesla's data)
Demonstrating a Misleading Graph
Below (my data)
100 years = ~257 pixels
1800
1900
2000
2100
2200
2300
2400
2500
2600
2700
2800
2900
3000

Re-Expressing/Transforming Data

As we've seen, data is often skewed in one direction. In practice, right-skew is far more common:

- Certain Variables, like time, height, weight, or income are bounded on one side
- Values must be greater than zero, but can still be have extremely large

The problem is that the median and IQR, while good for summarizing data, can be difficult to use in analysis methods we'll discuss later. How can we fix this?

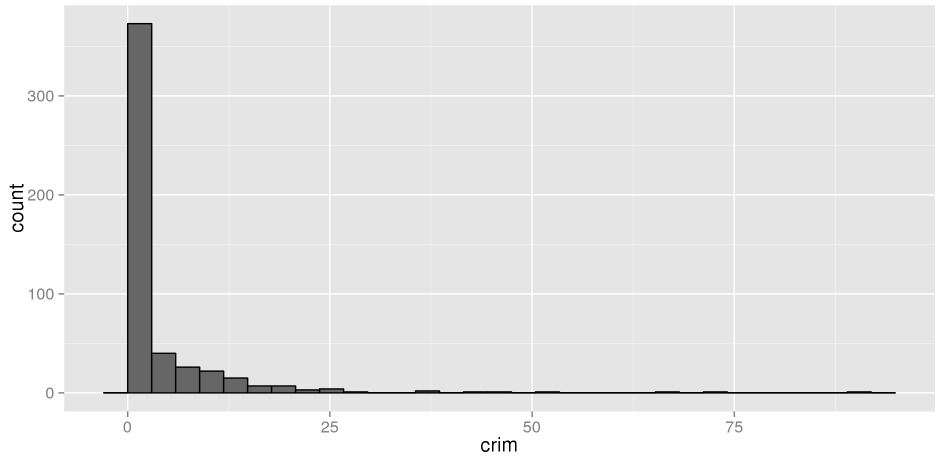
- Certain mathematical functions, like logarithms, square roots, or inversions (i.e., $x \to 1/x$) can be used to correct skew
- Additionally, they can let us better compare distributions with different spread
- Note: I don't expect you to use these, but it's commonly done and I want you to see why

Boston Housing Data

The Boston housing data set contains property value information for neighborhoods in Boston in 1978.

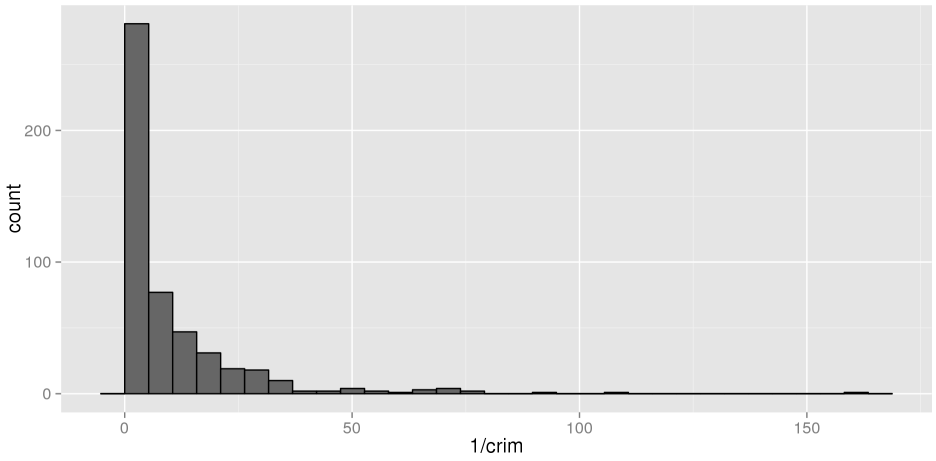
- Of particular interest is the `crim` variable, which lists the per capita crime rate by neighborhood
- This variable has extreme right-skew
- Additionally, we will compare the neighborhood crime rate to the accessibility of major highways, which has very different ranges of criminality across its values
- We will see how re-expressing `crim` helps to correct these issues

Criminality Histogram

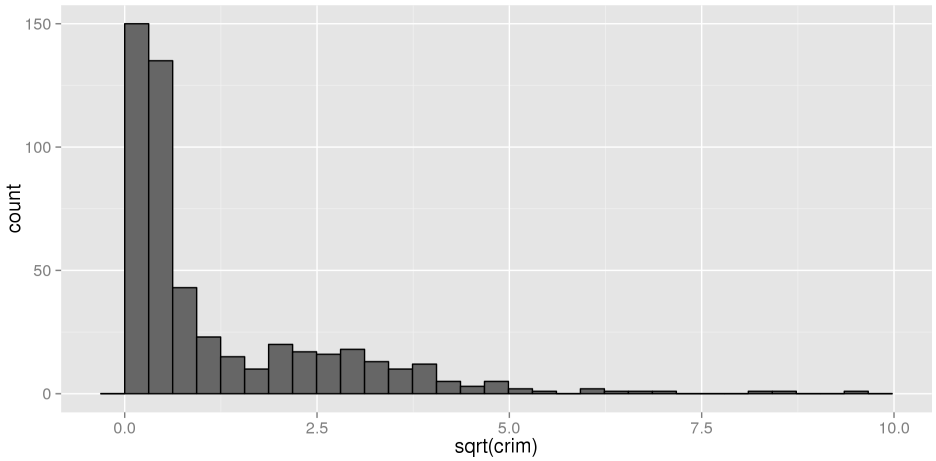


35/41

$\frac{1}{\text{Criminality}}$ Histogram

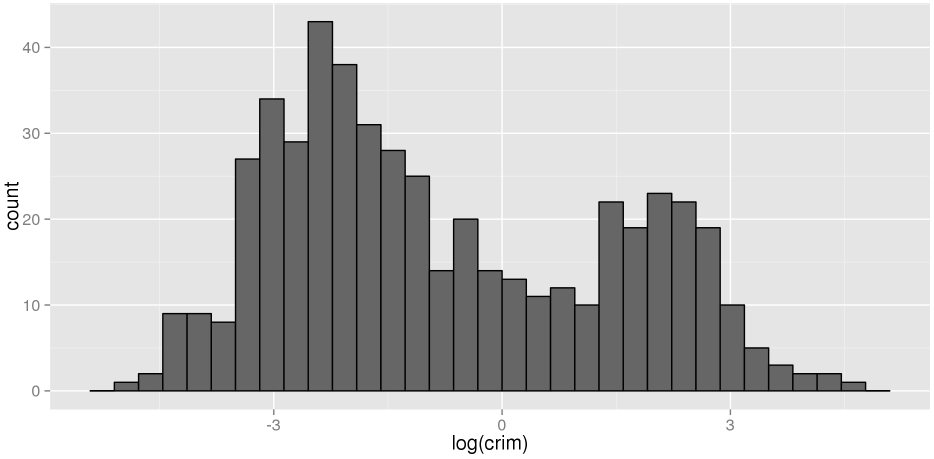


$\sqrt{\text{Criminality}}$ Histogram



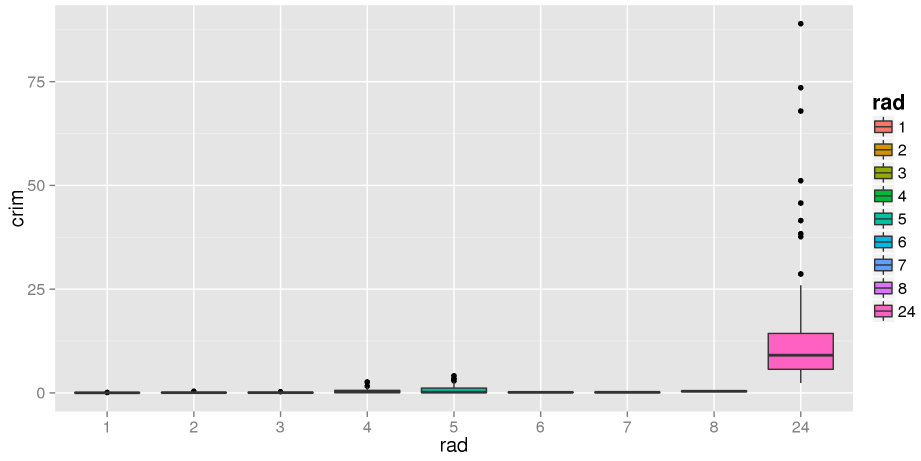
37/41

Log(Criminality) Histogram



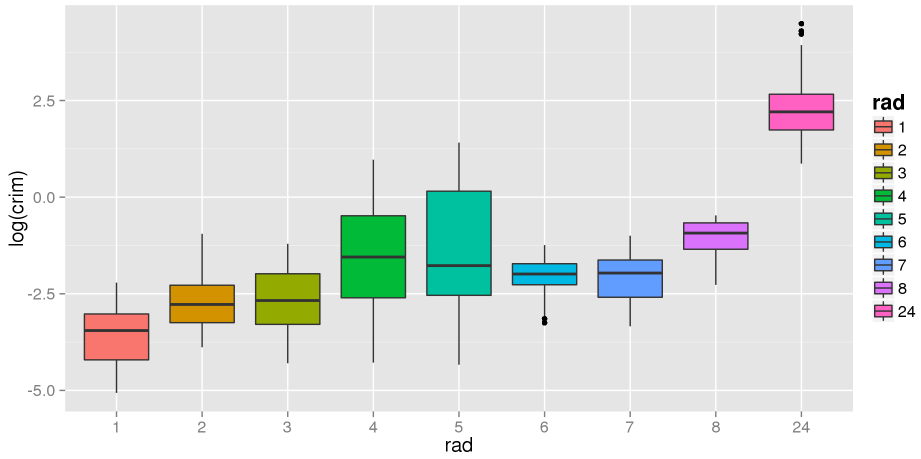
38/41

Criminality by Highway Access



39/41

Log(Criminality) by Highway Access



40/41

Summary

- We can compare numeric variables across groups by making multiple histograms or boxplots for each group
- You should never remove outliers without good reason
- Data recorded over time is plotted on a Timeplot
- Patterns over time can be described using a trace
- If the distribution is skewed, or variances are extremely different across groups, we can correct these problems by re-expressing or transforming the numeric variable