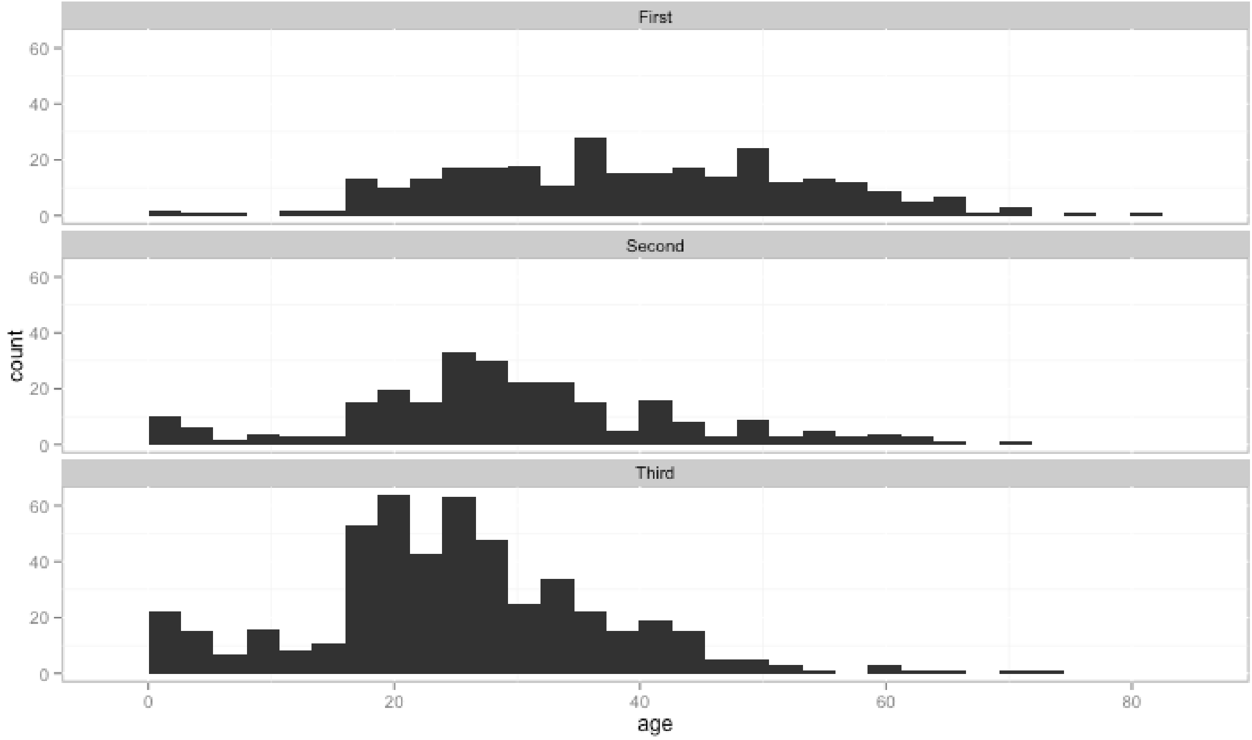
Distributions

September 18, 2013

Jason Bryer (jason@bryer.org)
epsy530.bryer.org

Comparing Histograms

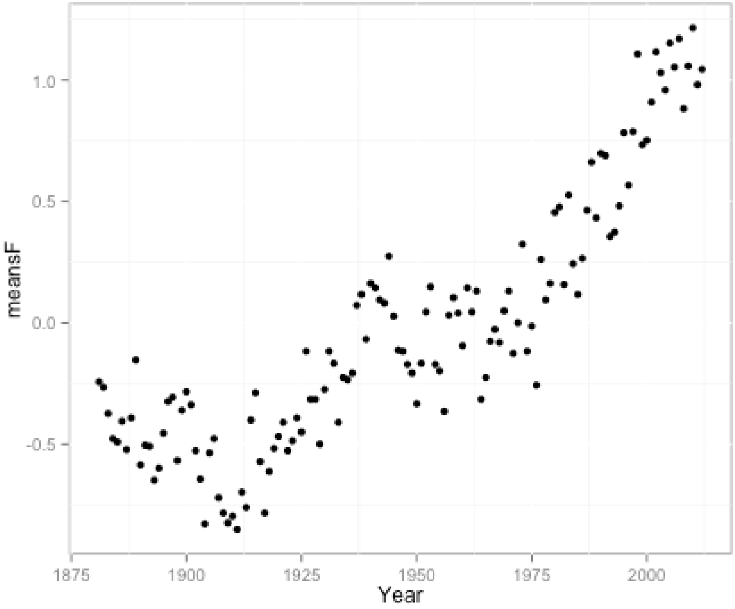
```
ggplot(titanic, aes(x=age)) + geom_histogram() + facet_wrap(~ pclass, ncol=1)
```



Timeplots

- Timeplots display every data value on a timeline.
- Great for spotting trends

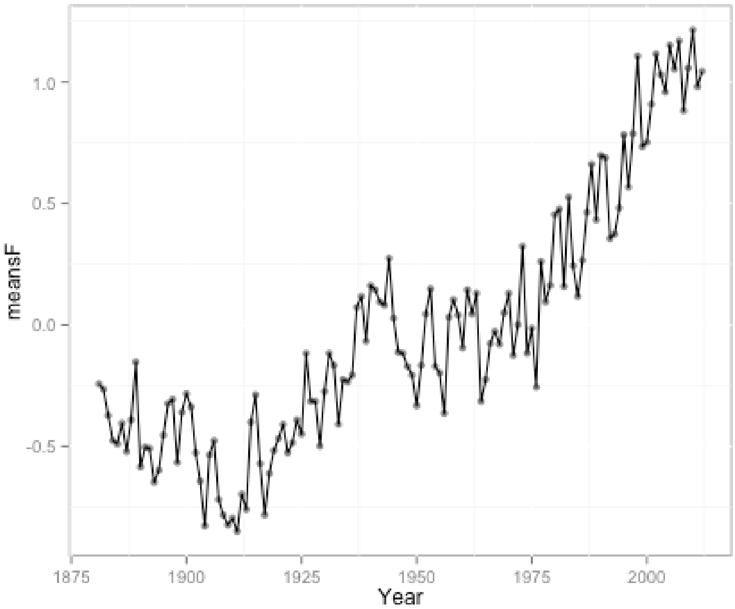
```
ggplot(temp, aes(x=Year, y=meansF)) +
    geom_point()
```



Connecting the Dots

- Connecting the dots of a timeplot can sometimes better illustrate the trends.
- This example has so many dots that this graph is busy and not that illustrative.
- Connecting the dots is better for either fewer data values or data with less variation.

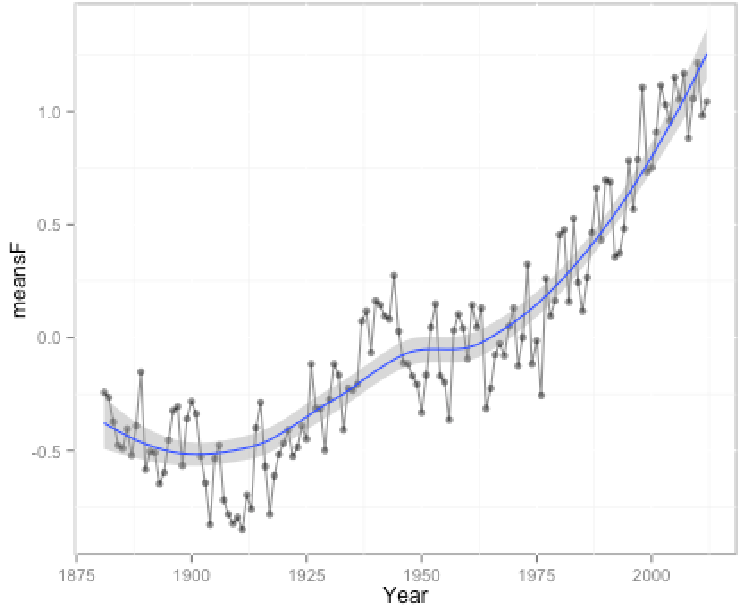
```
ggplot(temp, aes(x=Year, y=meansF)) +
    geom_point(alpha=.5) +
    geom_line()
```



Smoothing the Data

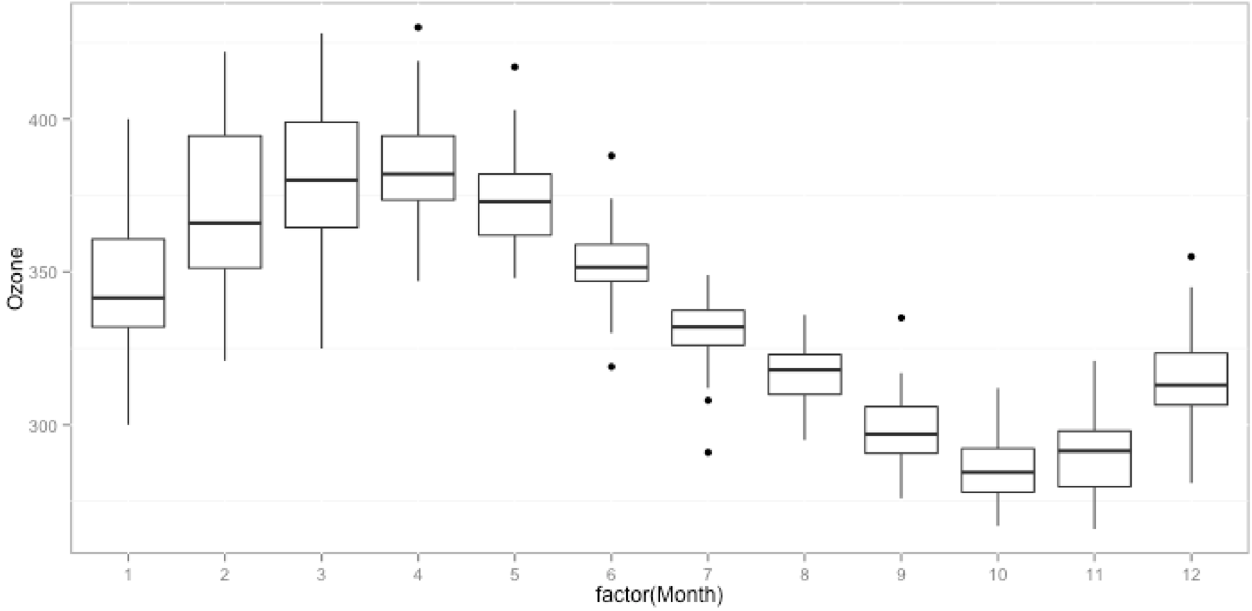
- Drawing a curve of typical values in the neighborhood can sometimes tell the story better.
- There are many ways of doing this and a computer can be used to create this curve.
- The curve, called the lowess (or loess) curve, helps the eye follow the main trend and spot the outliers.

```
ggplot(temp, aes(x=Year, y=meansF)) +
    geom_point(alpha=.5) +
    geom_line(alpha=.5) +
    geom_smooth()
```



Boxplots

```
ggplot(ozone, aes(x = factor(Month), y = Ozone)) + geom_boxplot()
```



Outliers

How to Approach Outliers

- Check to see if there may have been an error in the data collection or data input.
  - If the reported heights of students includes a student that is 170 inches tall (14 feet), maybe that student was measured in centimeters.
- Check to see if there was an extraordinary outcome.
  - The median number of daily customers at the Punxsutawney, PA, gift store may be 42 with an IQR of 12, but on February 2, there were 831 customers.

Common Errors Causing an Outlier

- Transposing the digits
- A respondent not understanding the survey question
- Misreading results
- Confusion about units
- Cheating

However, Outliers Can be the Most Interesting Data Values

- Income Data: The CEO
- Student Height: The basketball team's center
- Snowfall: The great blizzard of '98
- Exam Score: The curve breaker
- Milk Purchased: Octomom!

Always comment on the outliers.

Transforming Data

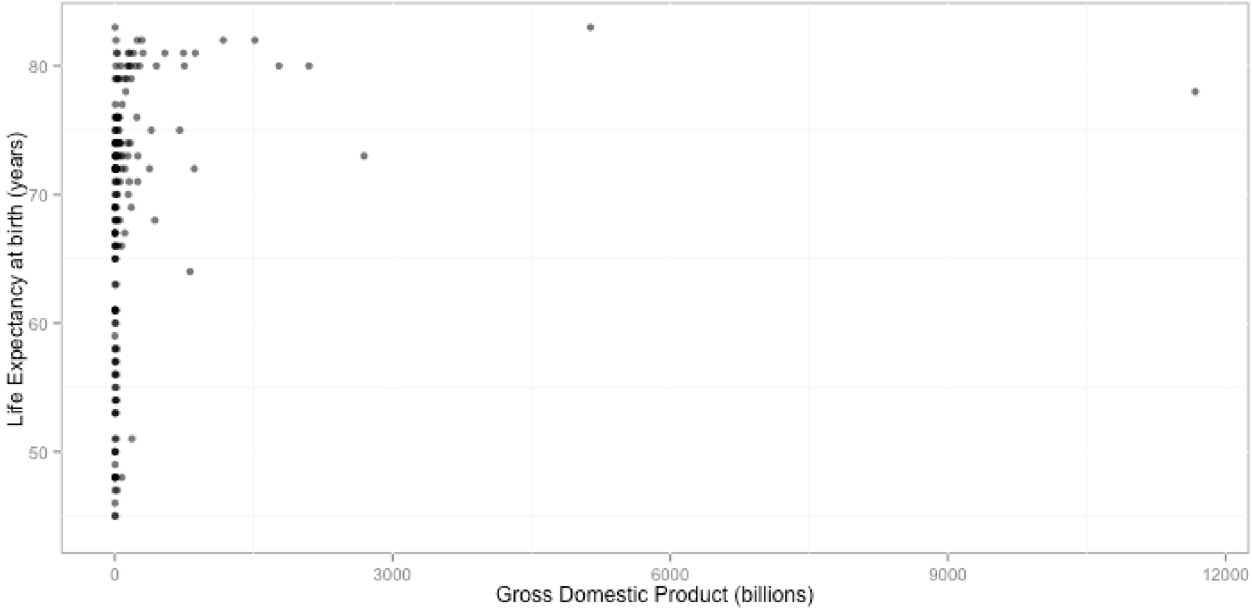
- When data is skewed it becomes difficult to interpret measures of center and spread.
- Transforming data is an approach to make skewed data more symmetric.

Common Transformations

- Skewed Right: Use log, ln, or $\frac{1}{x}$
- Skewed Left: Use x^2

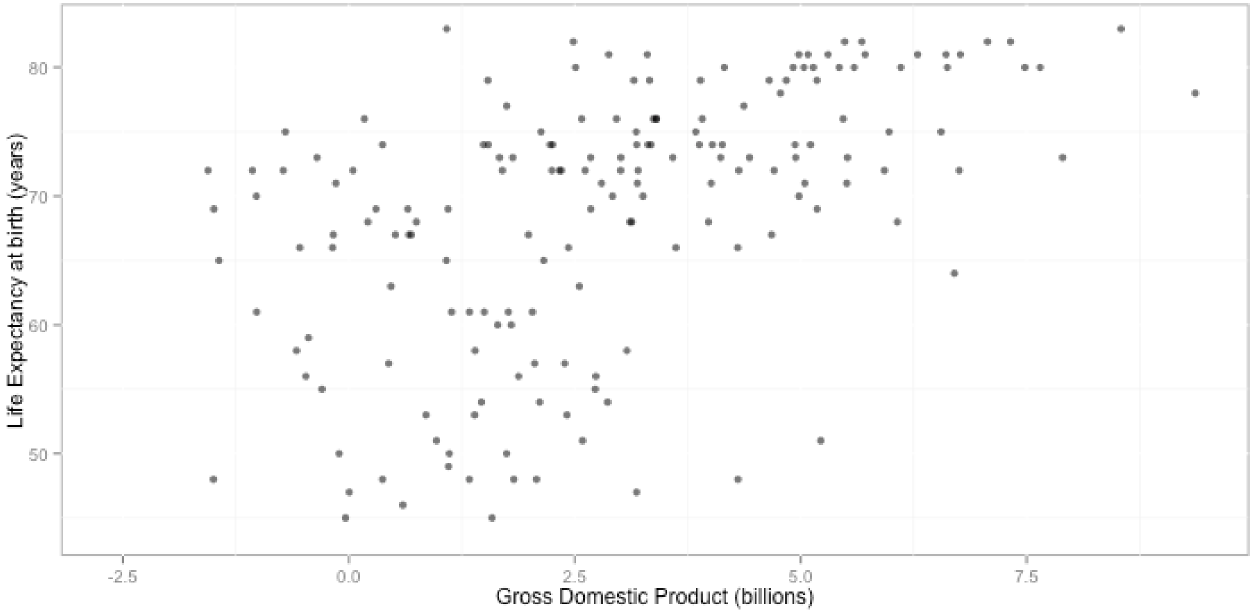
Example: World GDP and Life Expectancy

```
ggplot(worldData3, aes(x=GDP, y=Life.Expectancy)) +
    geom_point(stat='identity', alpha=.6) +
    xlab('Gross Domestic Product (billions)') +
    ylab('Life Expectancy at birth (years)')
```



Example: (log of) World GDP and Life Expectancy

```
ggplot(worldData3, aes(x=log(GDP), y=Life.Expectancy)) +
    geom_point(stat='identity', alpha=.6) +
    xlab('Gross Domestic Product (billions)') +
    ylab('Life Expectancy at birth (years)')
```



On Comparing Distributions

Choose the right tool.

- Use histograms to compare two or three groups.
- Use boxplots to compare many groups.

Treat outliers with attention and care.

- Local or global, especially in a time series
- Investigate if the outliers are errors or remarkable.
- Use a timeplot to track trends over time.

Re-express or transform data for better understanding.

- Can transform skewed distributions to symmetric ones
- Can help to compare spreads of different groups