# USING SAS® GTL TO VISUALIZE YOUR DATA WHEN THERE IS TOO MUCH OF IT TO VISUALIZE

Perry Watts, Stakana Analytics Elkins Park, PA

Nate Derby, Stakana Analytics, Seattle, WA

# The Challenge

## *An Effective Graph*

Is one that reveals "**patterns**, **differences** and **uncertainty**" in the underlying data.

## *But*

What if your data map to **crowded displays** with overlapping points, lines, or other obstructions that interfere with pattern detection?

## *Our Examples are Challenging*

| | | |
|---|---|---|
| **Framingham Heart Study** | Overlapping points | (n=5,209) |
| **Airlines Data** | Many overlapping lines | (n=6,100) |
| **Barley Data** | Unreadable response axis | (n=120) |
| **Stock Data** | Untraceable interleaving lines | (n=699) |

# Our Approach

***Incremental***

Go from preliminary graphs that are less than optimal

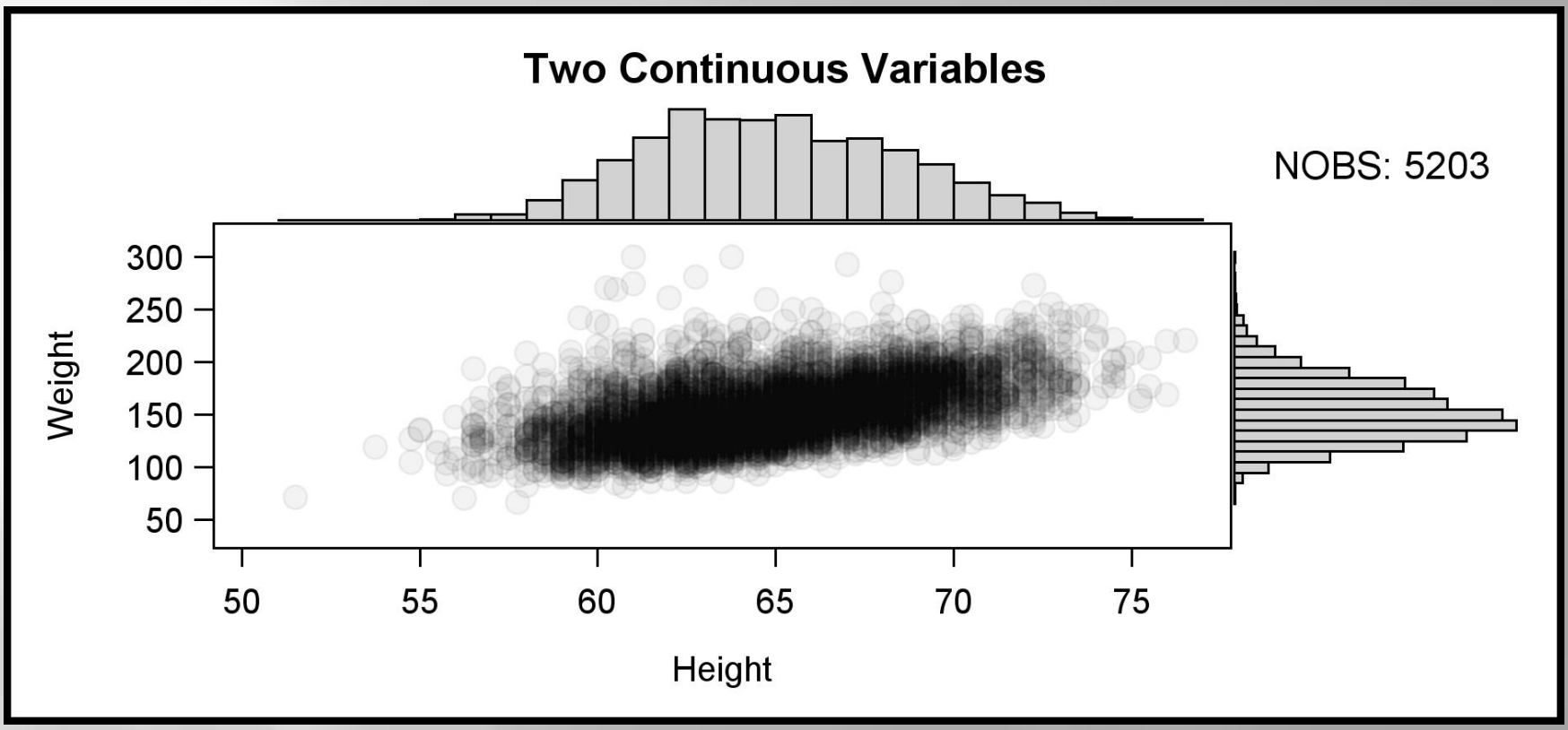
***To***

Output that conveys its message more effectively

***Along the Way:***

- Point out problems | issues.
- Solutions offered take advantage of new features in ODS statistical graphics and the insights of William S. Cleveland.
- Show why GTL must be used instead of a more convenient SG PROC to produce the graph you are looking at.
- We don't spend a lot of time on SAS code, however. Our goal is to define graphics problems and show how to solve them.

# Framingham Heart Study (sashelp.heart)




- SAS Sample #35172 deals with dense data by using 95% transparency in the scatter plot, stretching the graph out, and including marginal histograms.

# Framingham Heart Study (sashelp.heart)

## *Code Outline for SAS Sample #35172*

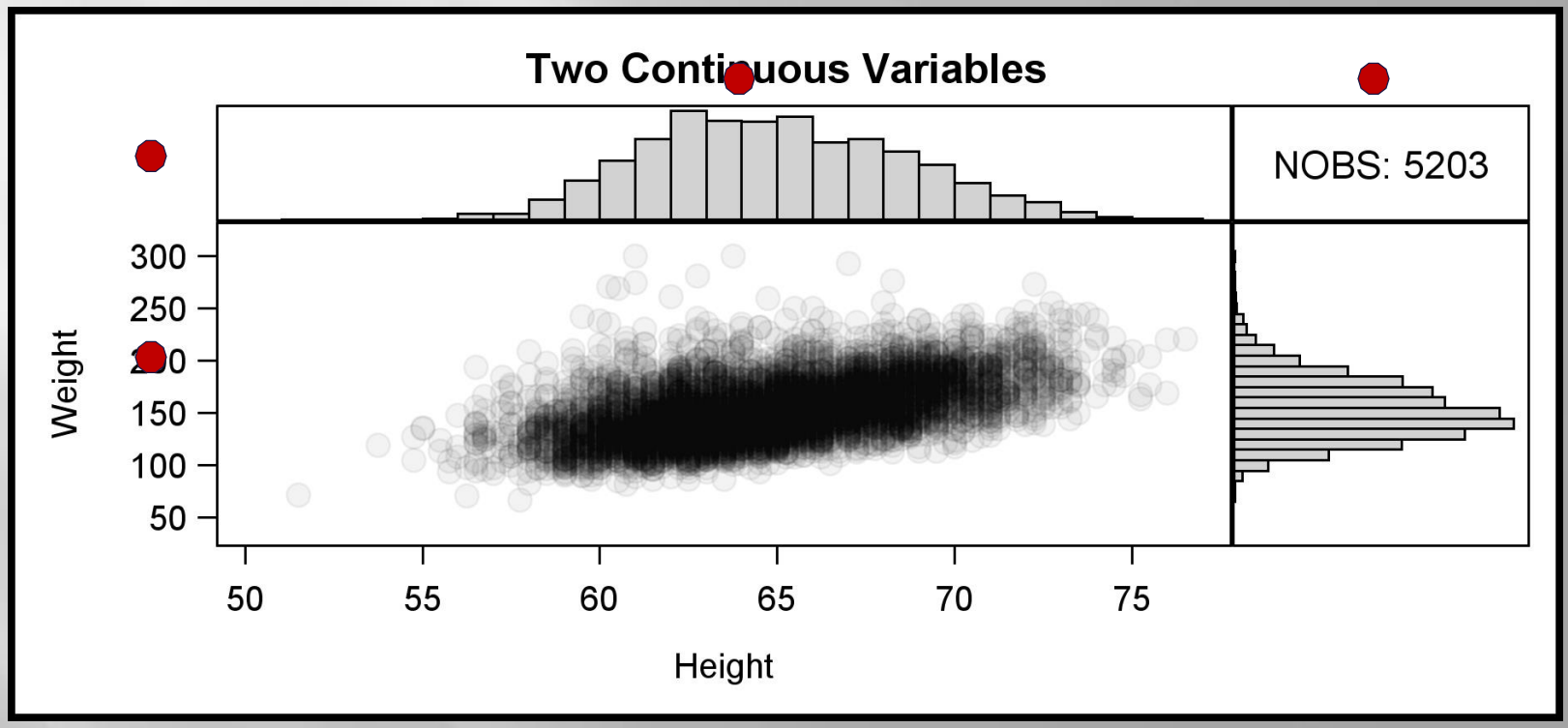
```
PROC TEMPLATE;
❶DEFINE STATGRAPH scatterhist;
   ❷BEGINGRAPH / DESIGNWIDTH=600px DESIGNHEIGHT=400px;
      ENTRYTITLE "Two Continuous Variables";
     ❸LAYOUT LATTICE / ROWS=2 COLUMNS=2;
       ❹LAYOUT OVERLAY; HISTOGRAM  Xvar;  ENDLAYOUT;
        LAYOUT OVERLAY; ENTRY 'NOBS: ' ...; ENDLAYOUT;
        LAYOUT OVERLAY; SCATTERPLOT  Y=Yvar X=Xvar; ENDLAYOUT;
        LAYOUT OVERLAY; HISTOGRAM  Yvar; ENDLAYOUT;
      ENDLAYOUT; /*LATTICE*/
    ENDGRAPH; /*END GRAPH BLOCK*/
  END; /*END DEFINE BLOCK*/
RUN;

PROC SGRENDER DATA=sashelp.heart TEMPLATE=scatterhist;
RUN;
```

# Framingham Heart Study (sashelp.heart)

## *Why PROC SGPANEL Doesn't Work*

```
❸LAYOUT LATTICE / ROWS=2 COLUMNS=2
                  ROWWEIGHTS=(.2 .8) COLUMNWEIGHTS=(.8 .2);
```

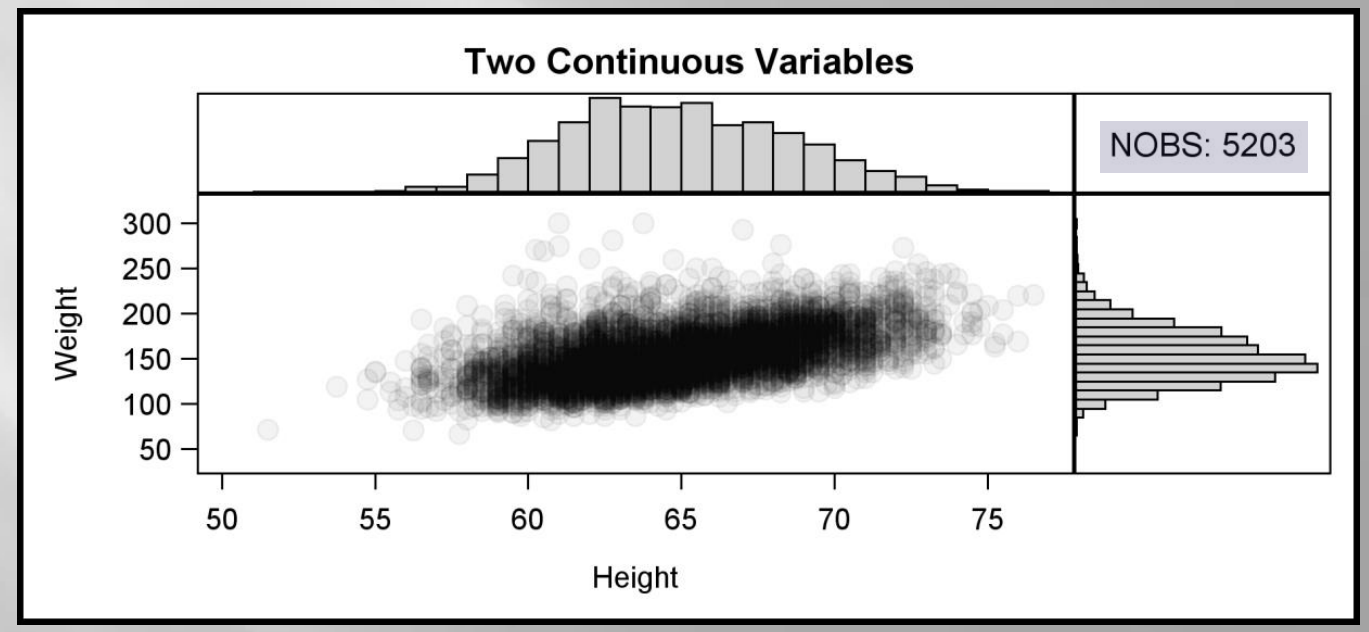



Panels must have **equal dimensions** in PROC SGPANEL

# Framingham Heart Study (sashelp.heart)

## *What's missing from the definition for NOBS?*

```
❹LAYOUT OVERLAY / BORDER=true;
   ENTRY 'NOBS: ' EVAL(N(xvar)) / ...;
 ENDLAYOUT;
```

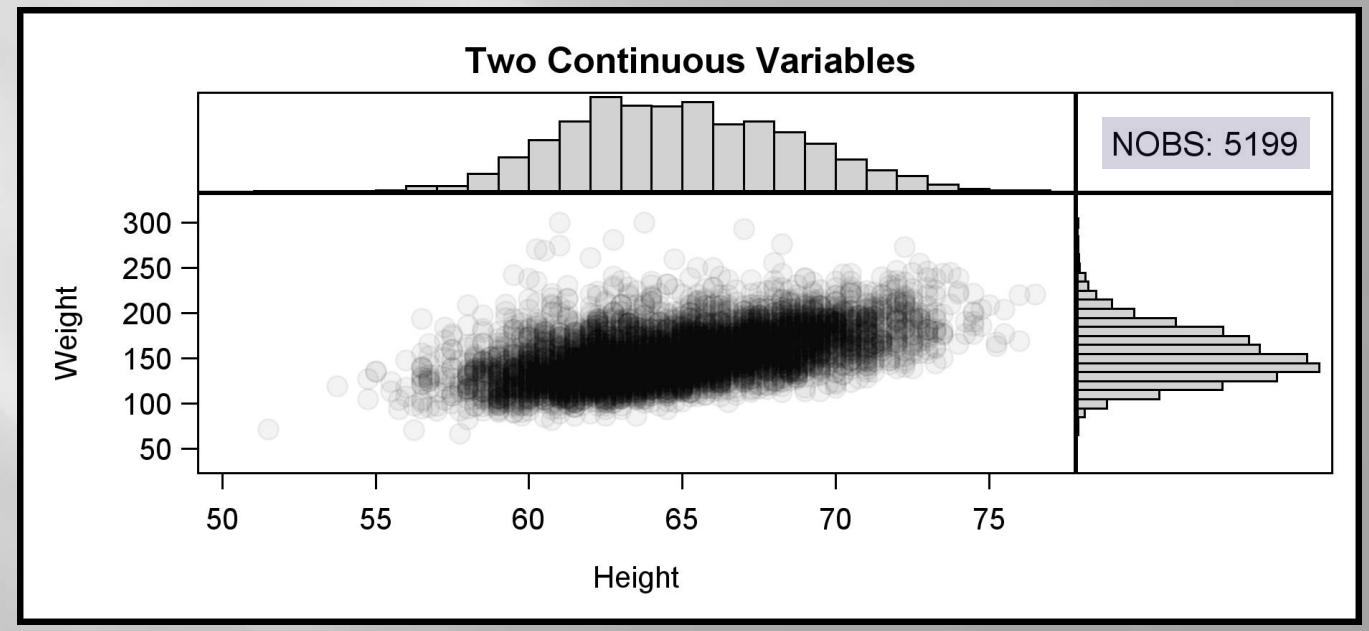



In a **scatter plot** each point references an X **and** a Y coordinate. (Neither can be missing).

# Framingham Heart Study (sashelp.heart)

## *Changing the code gives the right answer*

```
❹LAYOUT OVERLAY / BORDER=true;
   ENTRY 'NOBS: ' EVAL(N(xvar + yvar)) / ...;
 ENDLAYOUT;
```

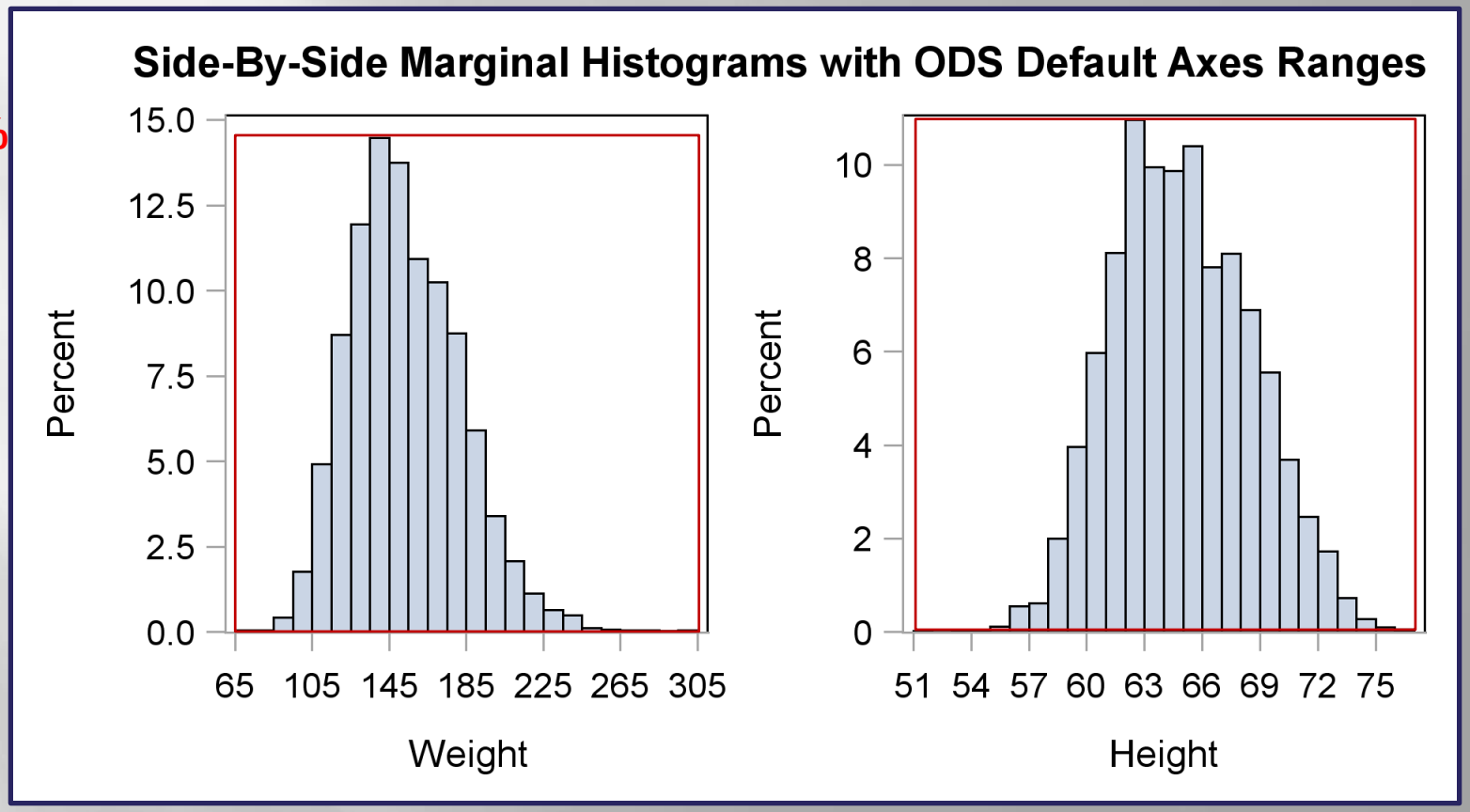



The '+' operator works, because a missing value is returned when **at least** XVAR **or** YVAR is missing. (**SUM** won't work).

# Framingham Heart Study (sashelp.heart)

## *ODS Statistical Graphics Axis Format*

14.5%




11%

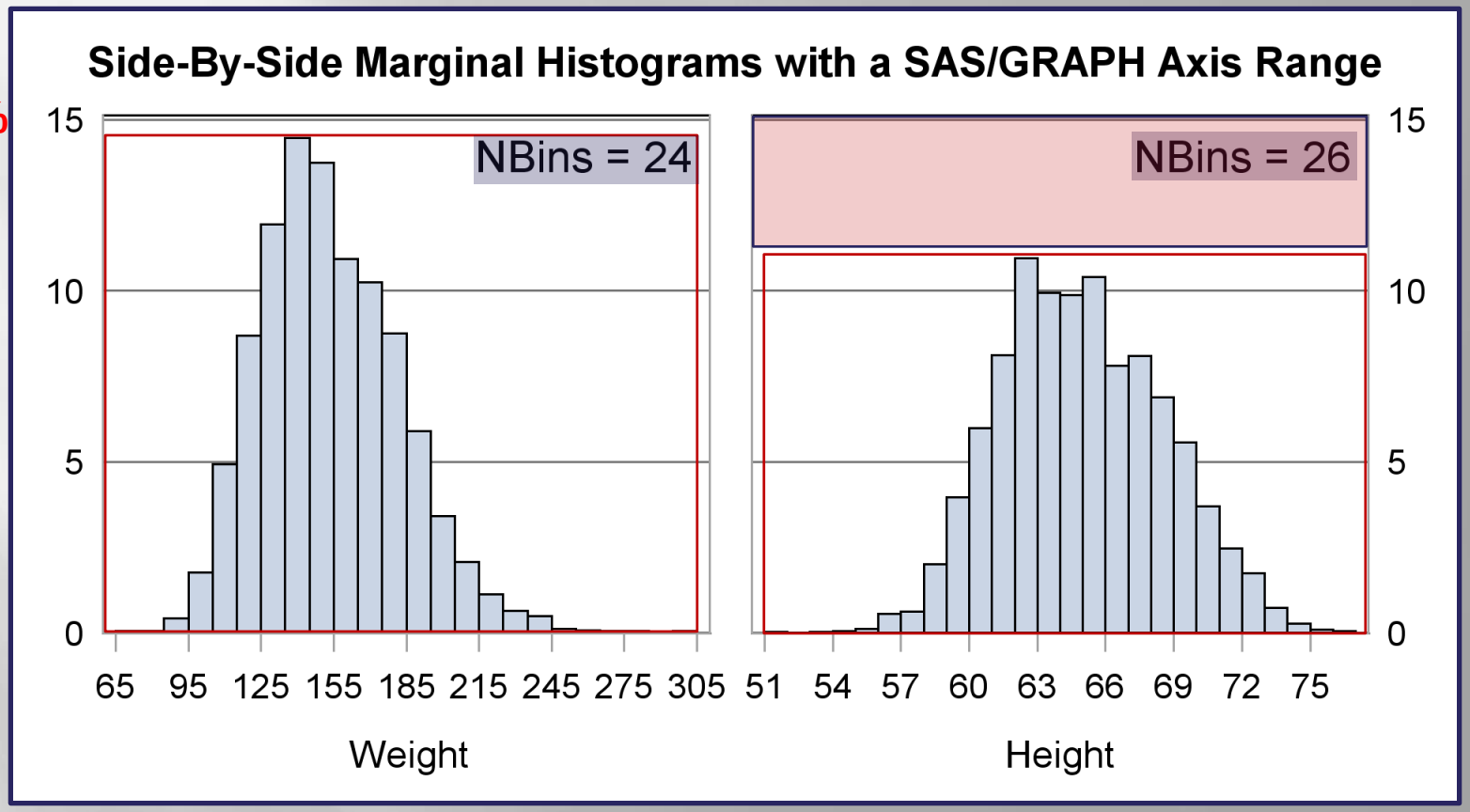
From William S. Cleveland :

"make the ***data rectangle*** slightly smaller than the ***scale-line*** rectangle".

# Framingham Heart Study (sashelp.heart)

## *Conventional SAS/GRAPH Axis Format*

14.5%



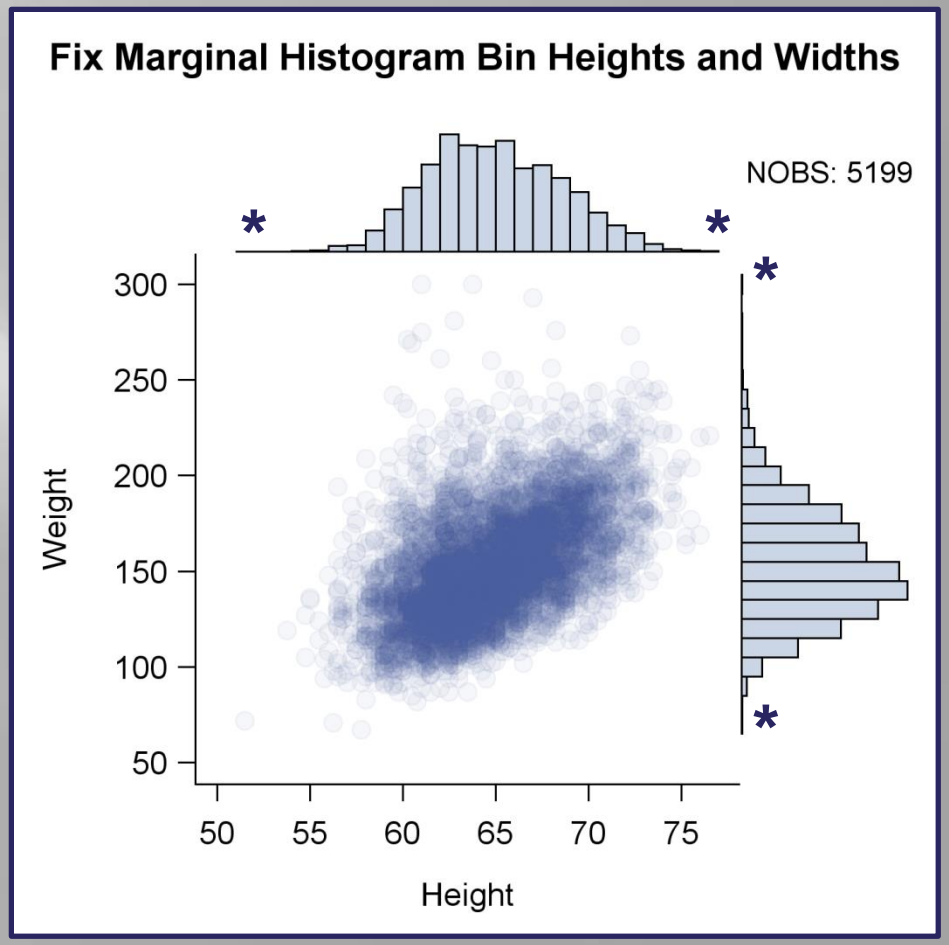

11%

Data points can't appear above the axis maximum tick value.

# Framingham Heart Study (sashelp.heart)

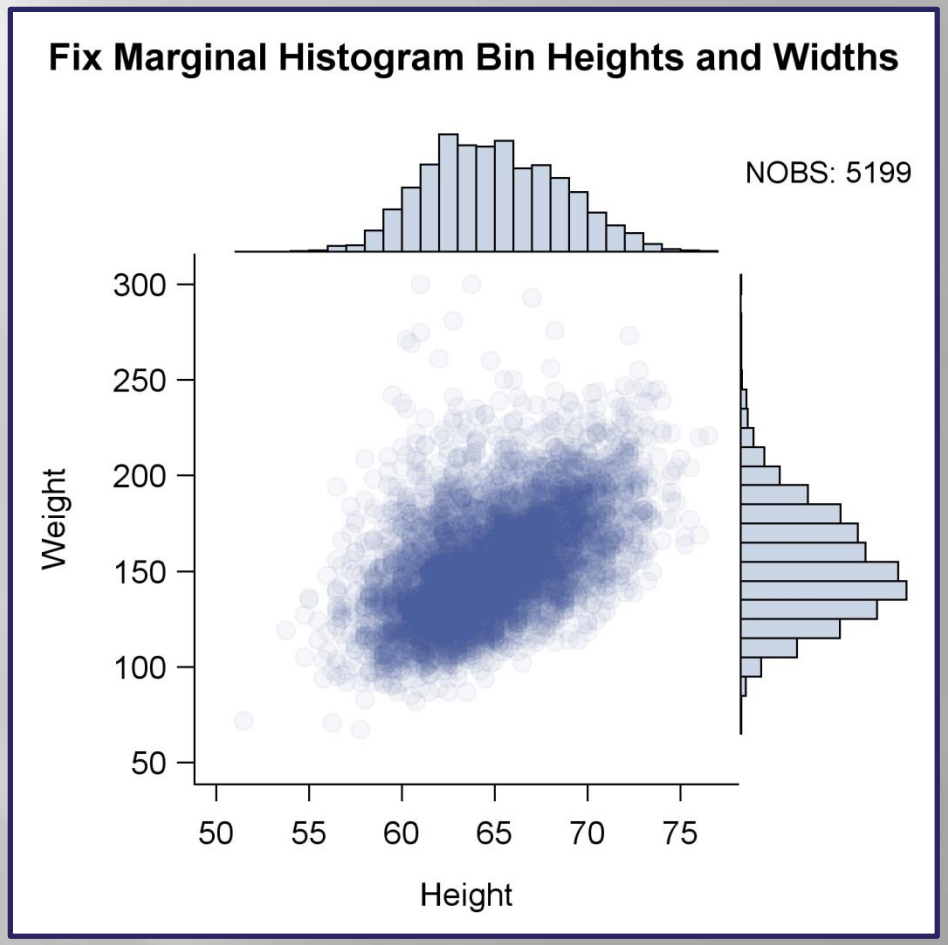
## *The Revised Graph: Histogram Fixes*

- The graph is squared off to eliminate bin-width distortion due to stretching.
- Marginal histogram bin heights are now comparable, because VIEWMAX is set to 15%.
- Borders are removed to make marginal histogram bin ranges more visible.



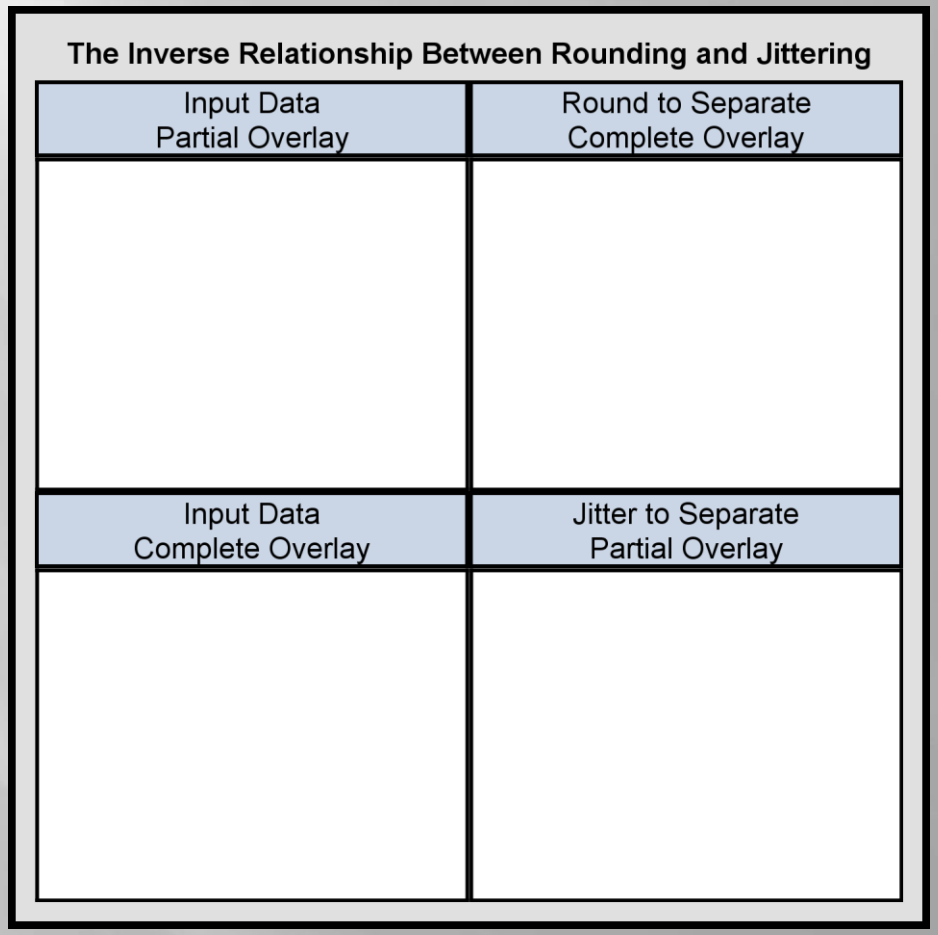
# Framingham Heart Study (sashelp.heart)

## *We Still Have a Problem with the Scatter Plot*



# Framingham Heart Study (sashelp.heart)

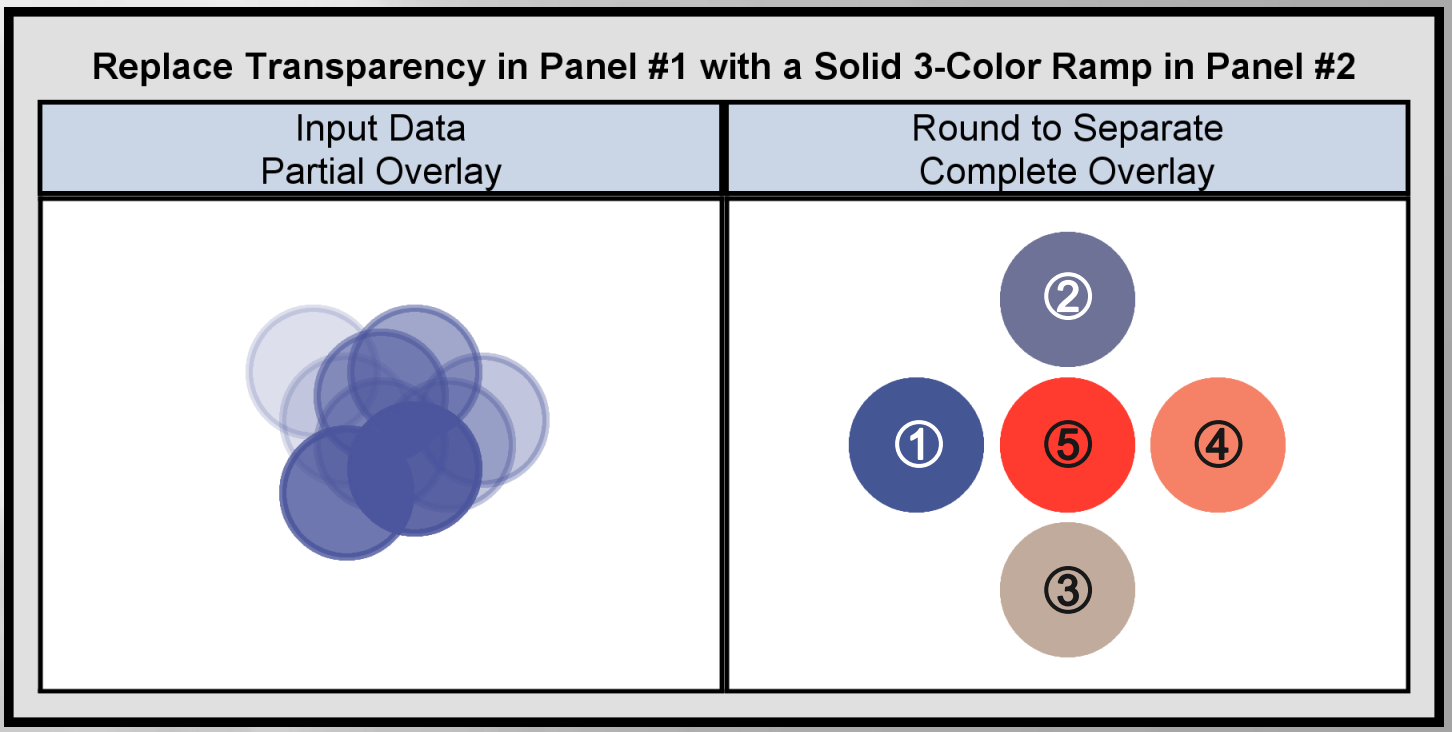
## *Try Rounding related to Cleveland's Jittering*




Jittering adds "random noise" to each point for a slight separation.

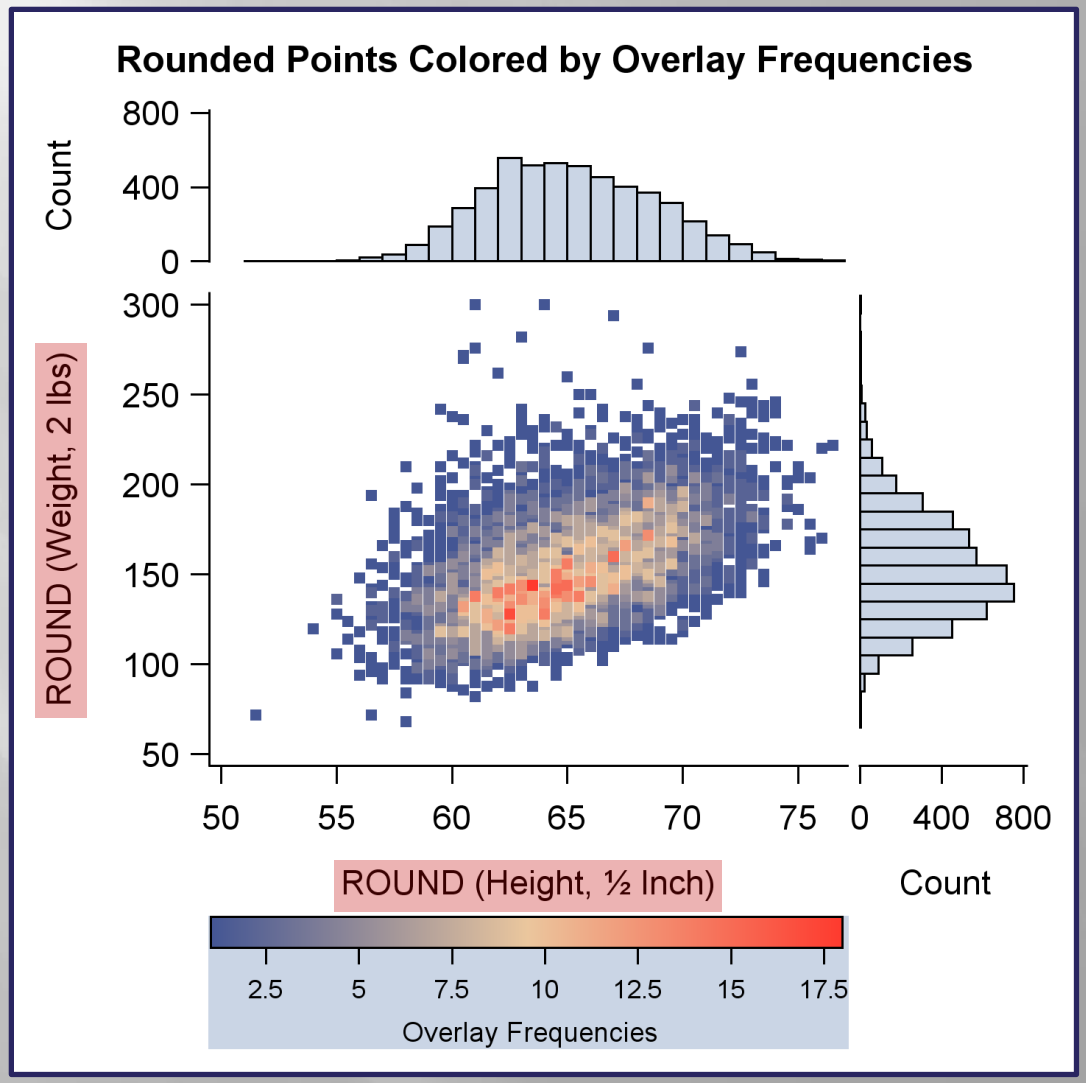
# Framingham Heart Study (sashelp.heart)

## *Try Rounding related to Cleveland's Jittering*



# Framingham Heart Study (sashelp.heart)

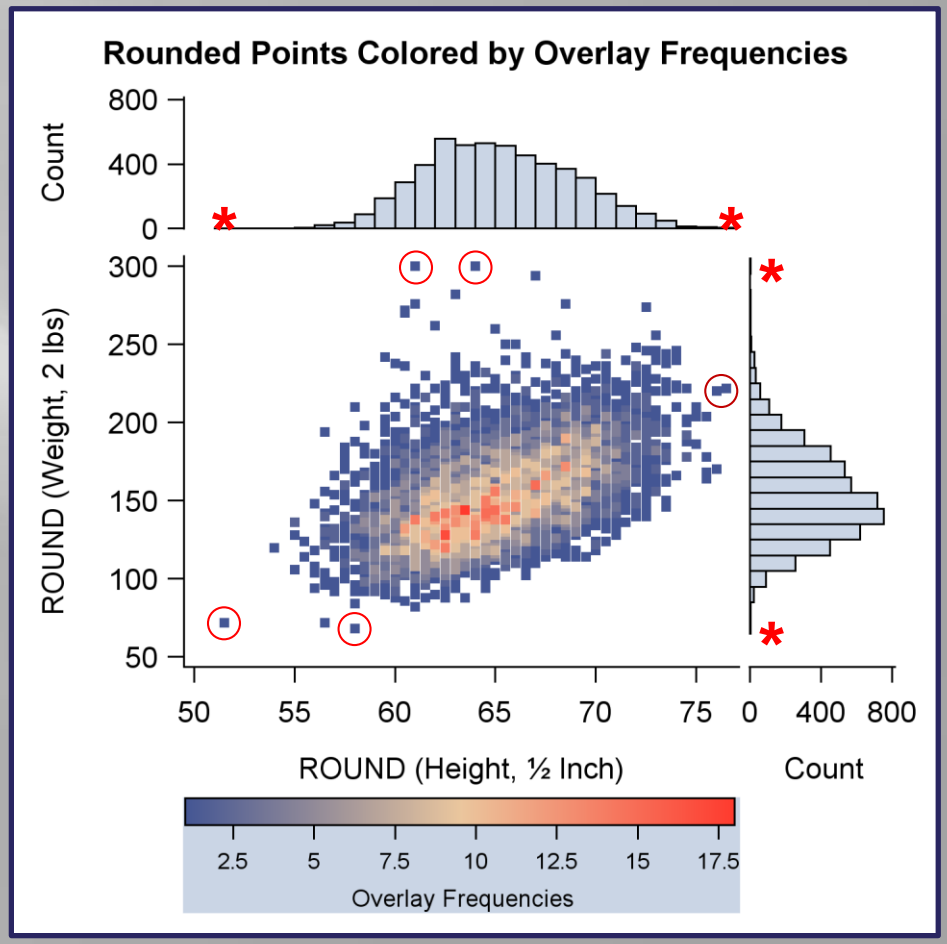
## *Rounding for a 3rd Dimension based on Frequency*




# Framingham Heart Study (sashelp.heart)

## *Rounding for a 3rd Dimension based on Frequency*

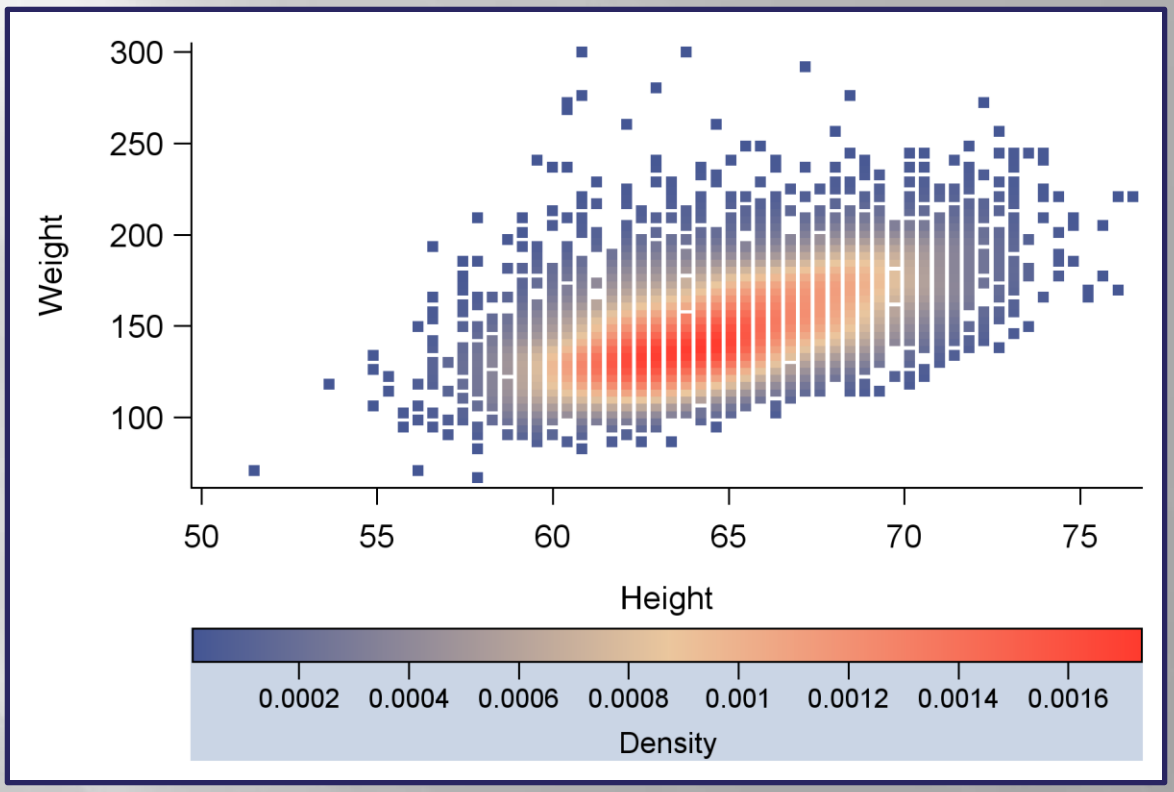
- SQUAREFILLED markers in the scatter plot line up better with histogram bins.
- The legend makes the graph less square. Compensate by labeling histogram axes tick marks.
- With solid color plotting symbols, it is easier to line up histogram end bins with the blue data outliers.




Continuous legends are only available in **GTL**

# Framingham Heart Study (sashelp.heart)

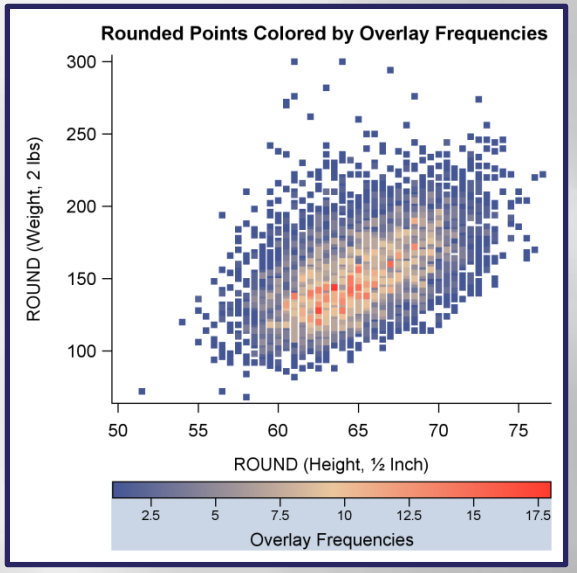
## *Create a Digitized Contour Plot with PROC KDE*

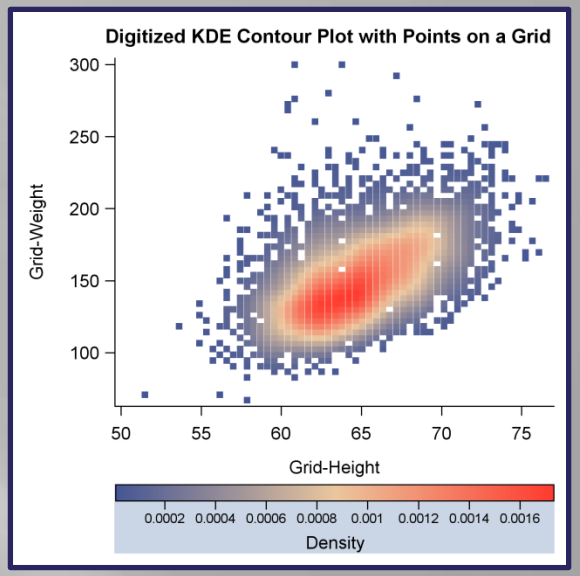



Switch from raw data manipulation ("rounding") to statistical estimation where cell color is based on probability.

# Framingham Heart Study (sashelp.heart)

## *A Rounded vs. Digitized KDE Contour Plot*



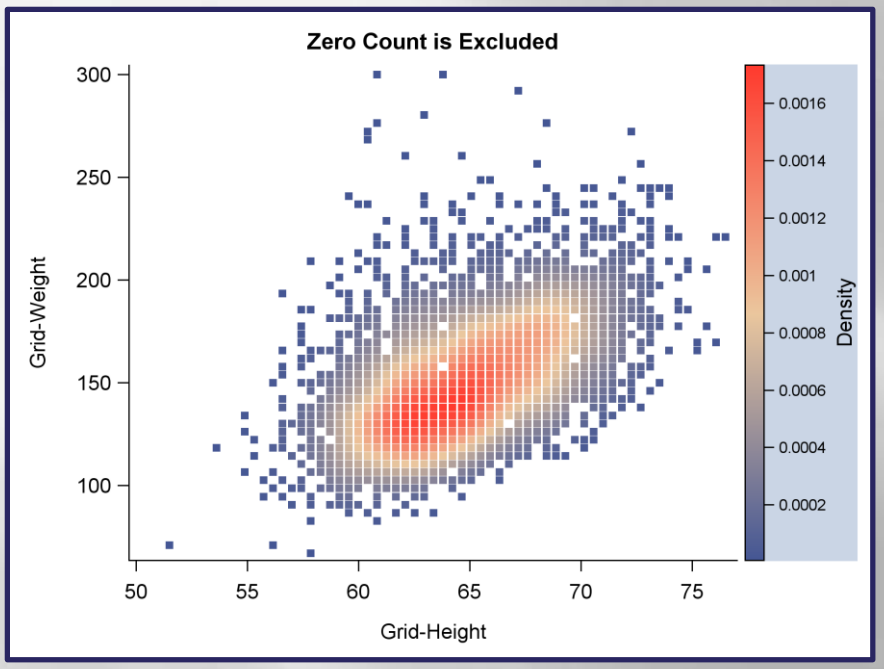

- An adjusted raw data set is plotted.
- X and Y data values are "rounded".
- Z, rendered by color, is the count of tied observations at a given (rounded) point.

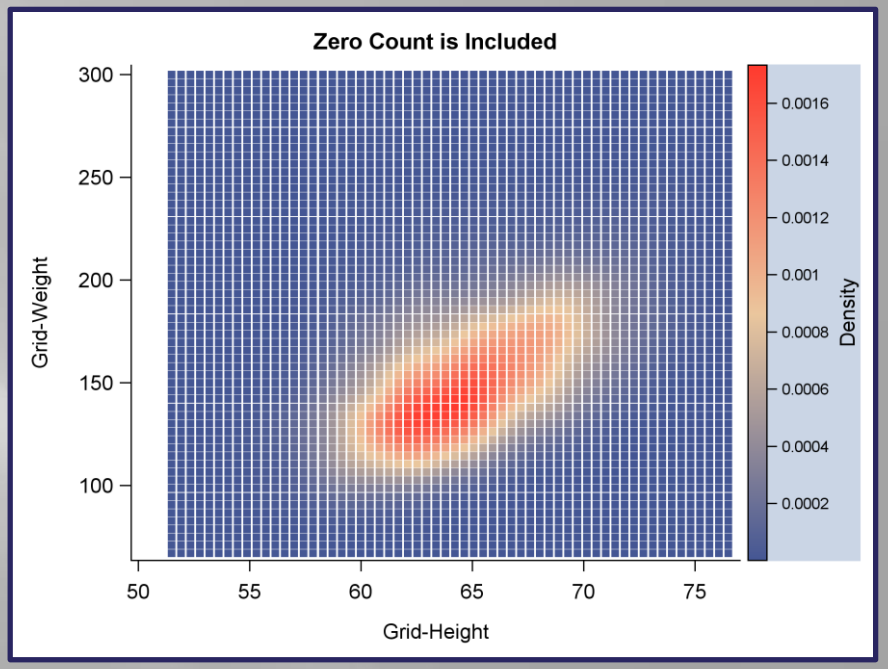
- Output from PROC KDE is plotted.
- The plotting region is divided into a 60X60 grid of cells in X and Y variable units (3,600 obs).
- Z equals DENSITY not Frequency.
- COUNT, another variable, sums to 5,199.

STAKANA

# Framingham Heart Study (sashelp.heart)

## *Generating the Digitized Plot from PROC KDE*

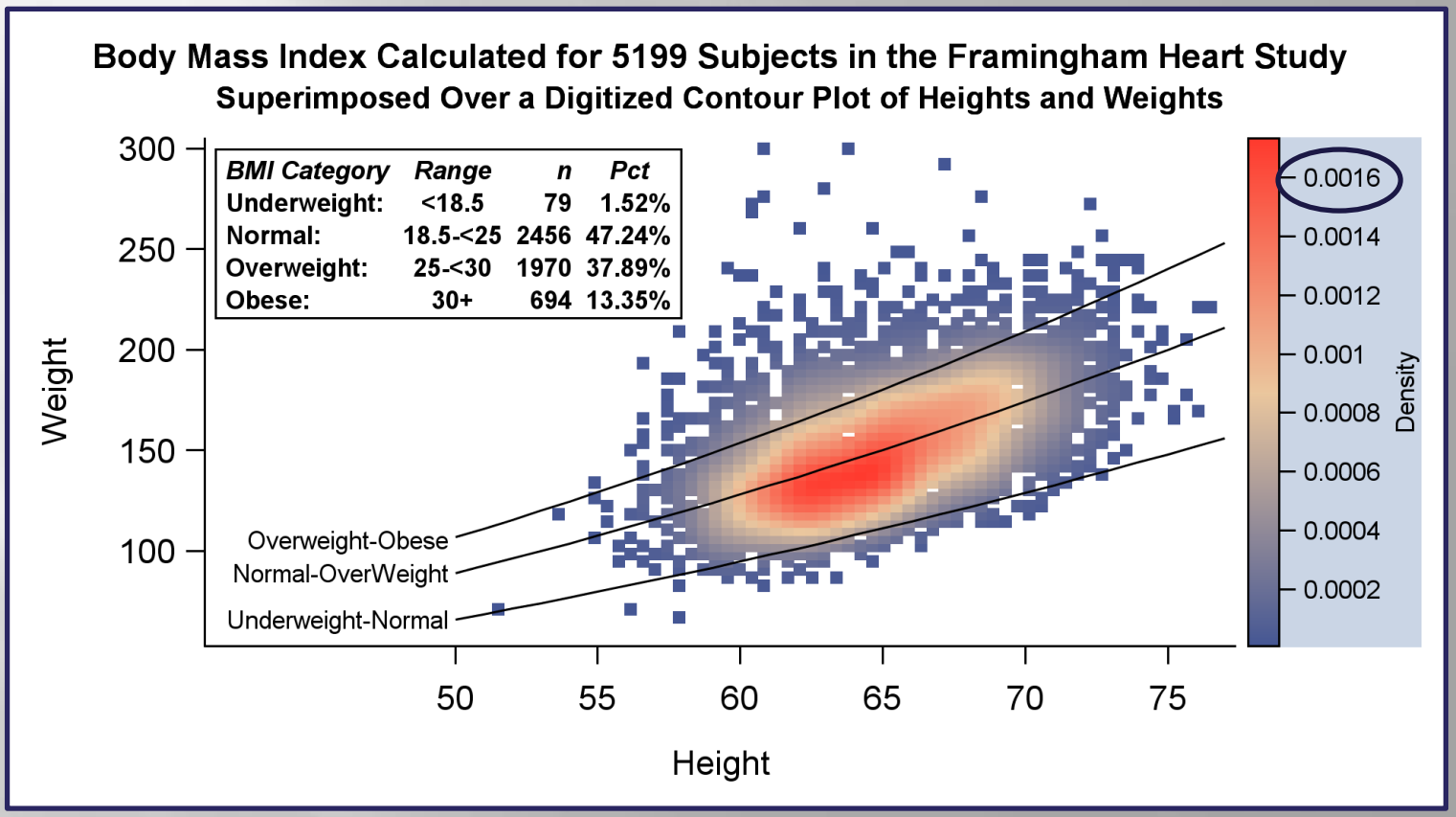



```
proc kde data=sashelp.heart;
  Bivar Height Weight / PLOTS=NONE out=KDEGridded;
run;
proc sgrender template=xTmp data=KDEGridded;
run;
```

# Framingham Heart Study (sashelp.heart)

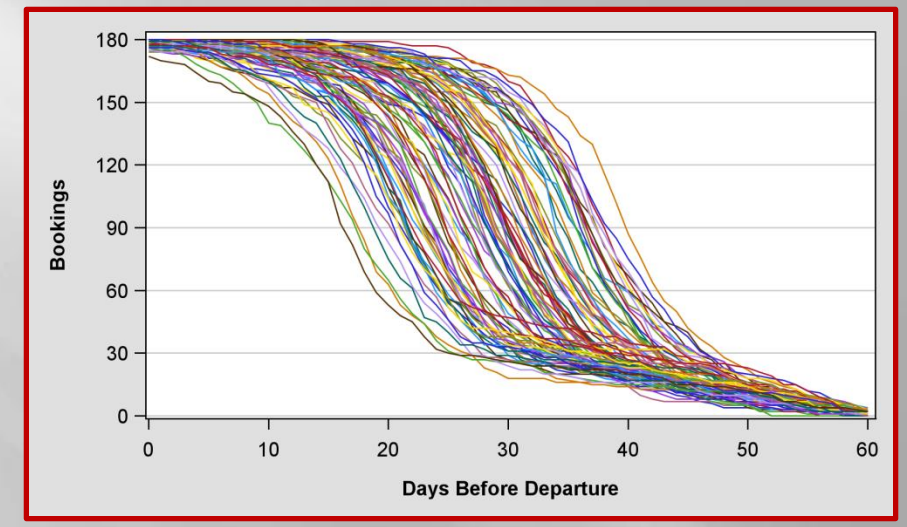
## *Add the BMI to the Digitized Contour Plot*




Complete source code can be found in the ZIP file referenced in the Paper

# Airlines Data

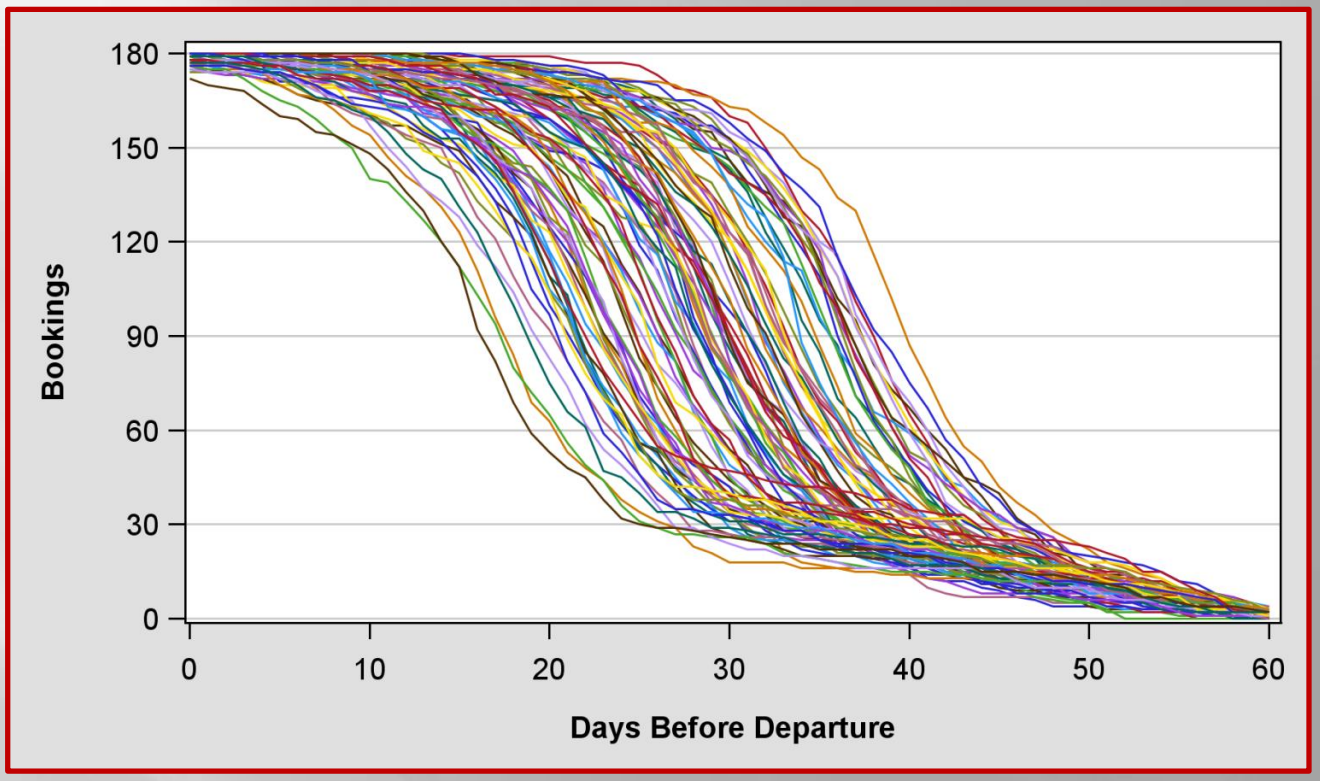
## *A Progression of Time Series Plots*




- This is a ***progression*** of 100 series plots of flights where each flight has a unique departure date.
- The X axis = the number of days before departure a flight is booked.
- The Y axis = the cumulative number of bookings. Each flight accommodates 180 passengers.

# Airlines Data

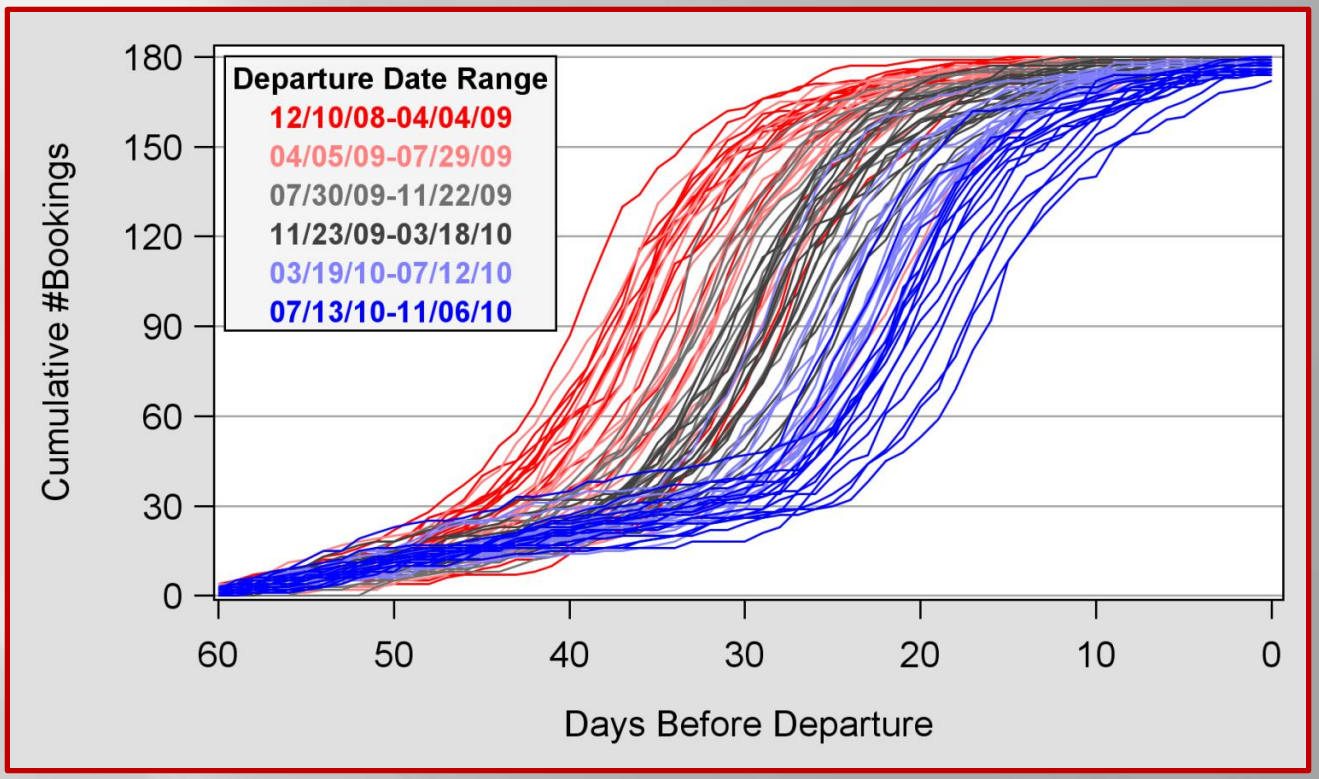
## *A Progression of Time Series Plots*




**Is there a Relationship between *Days Before Departure* and *Departure Date*?**

# Airlines Data

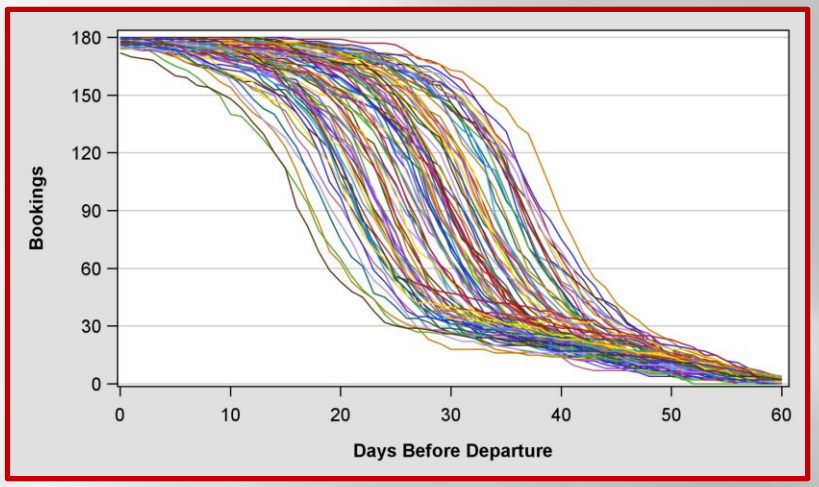
## *A Progression of Time Series Plots*

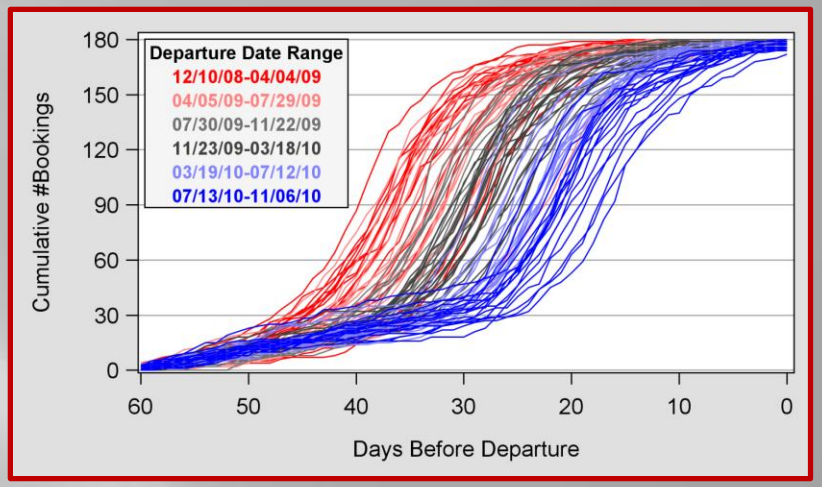



**Add a Color Dimension to see the Connection between Days Before Departure and Departure Dates**

# Airlines Data

## *A Progression of Time Series Plots*




## What's Different?

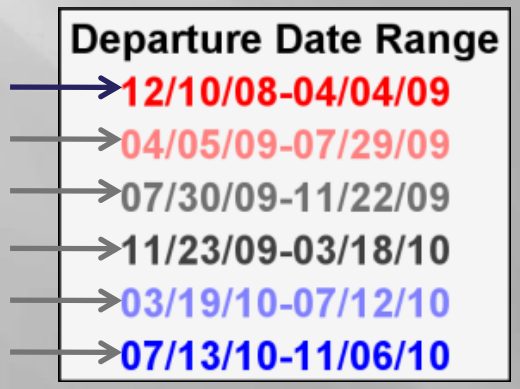
- Time Series plots should cumulate **left to right**. That means the X-axis needs to be **reversed**.
- An **inset replaces** the **legend**, because the legend points to group variable, ***Departure Date*** (100), not ***Date Range*** (6).
- Inset text maps **colors** to plot lines. No legend line-to-line mapping is needed.

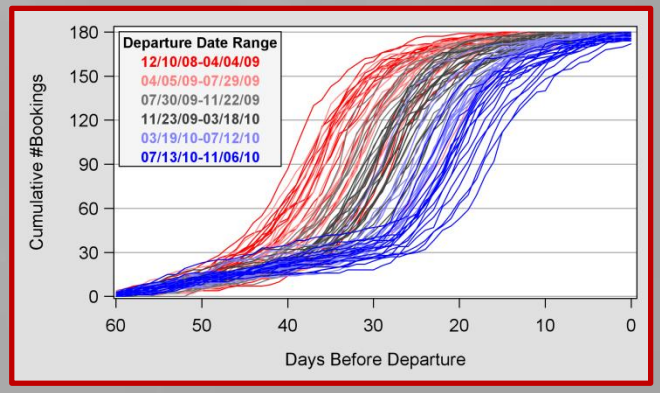
STAKANA

# Airlines Data

## *A Progression of Time Series Plots*

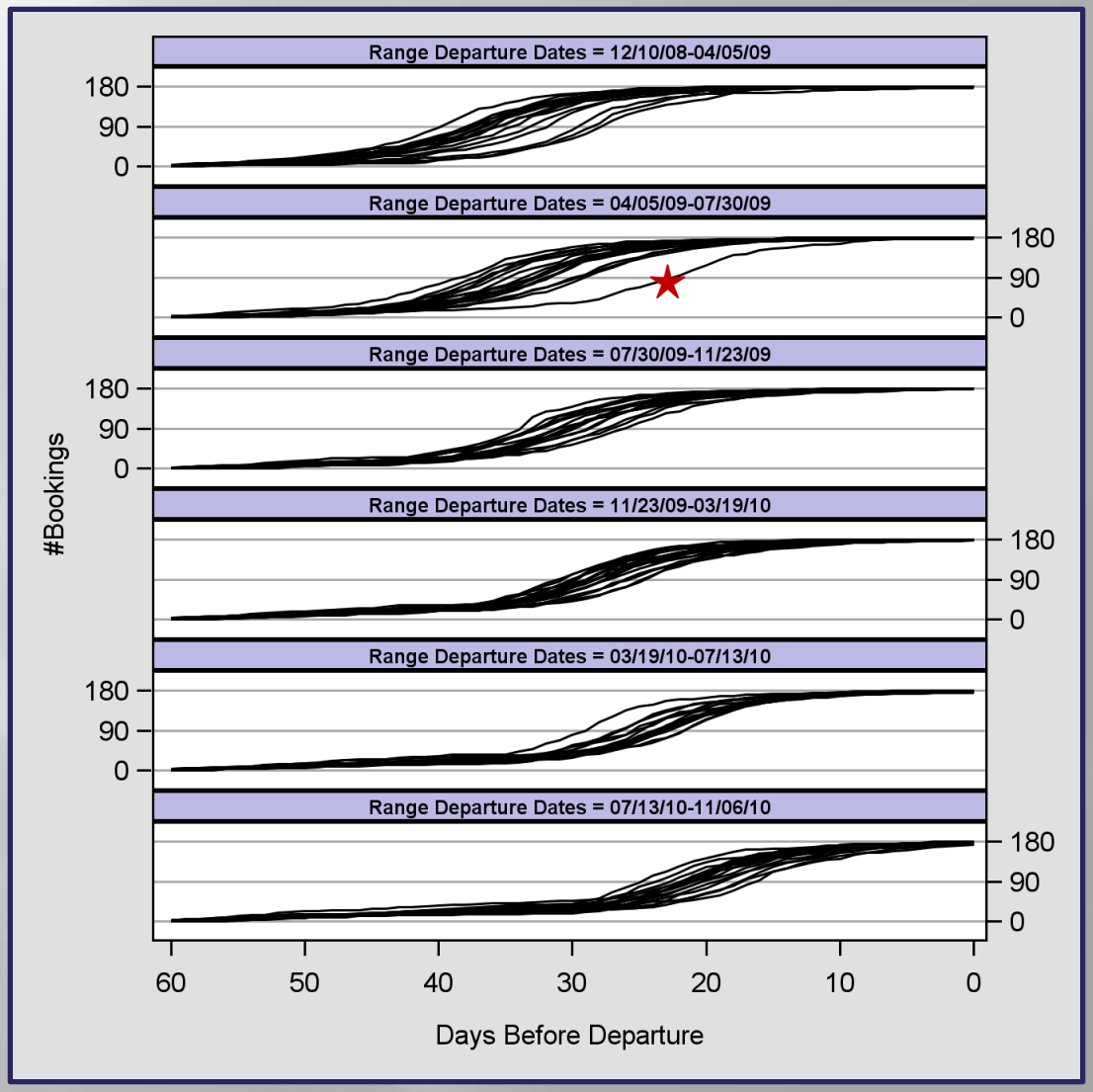
```
❹LAYOUT OVERLAY / ... Xaxisopts=(... reverse=true);
 %do i= 1 %to 6;
   SERIESPLOT X=x&i Y=y&i / GROUP=ddate
                            LINEATTRS=(COLOR=&&color&i);
 %end;
  ❺LAYOUT GRIDDED / COLUMNS=1 ...; ...;
  %do j = 1 %to 6;
    ENTRY TEXTATTRs=(WEIGHT=bold COLOR=&&color&j)"&&Range&j";
  %end;
  ENDLAYOUT; /*gridded*/
 ENDLAYOUT; /*overlay*/
```



# Airlines Data

## *Using LAYOUT DATAPANEL*

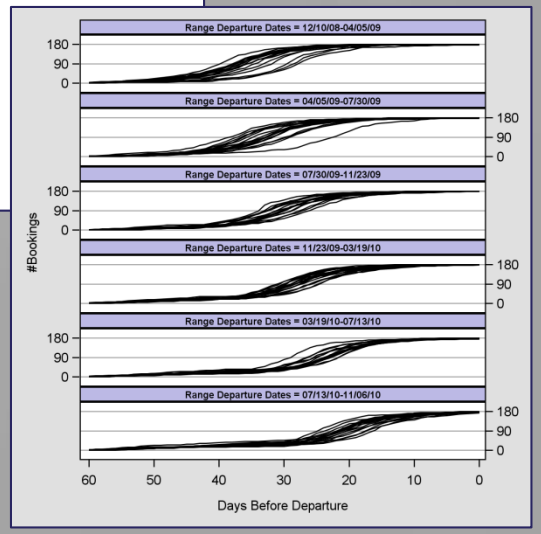


# Airlines Data

## *Using LAYOUT DATAPANEL*

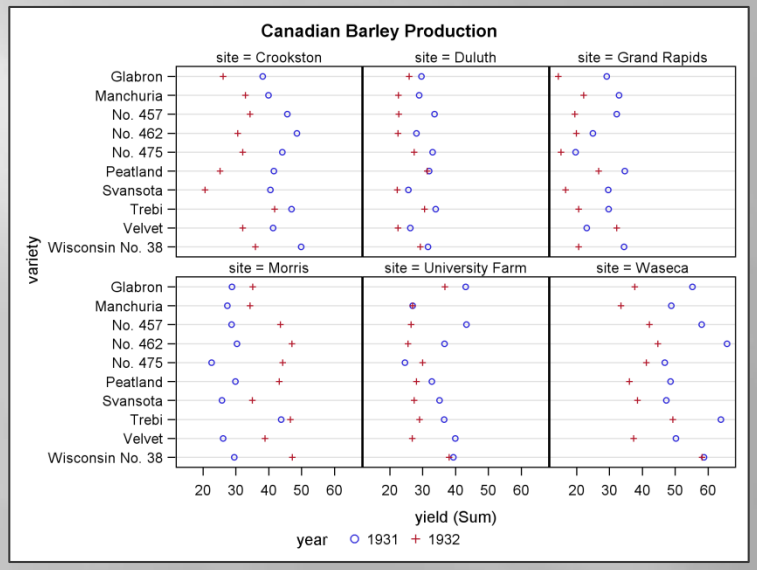
```
❸LAYOUT DATAPANEL classvars=(ByDdateLbl)/
  headerlabelattrs=(weight=bold ...)
  headerbackgroundcolor=CXBCB9E5
  columndatarange=union
  columnaxisopts=(... REVERSE=TRUE)
  rowaxisopts=( ... );

    ❹layout prototype /...;
      seriesplot x=DaysLeft y=bookings/
                 group=ddate ...;
     endlayout; /*prototype*/
endlayout; /*dataPanel */
```



# Barley Data

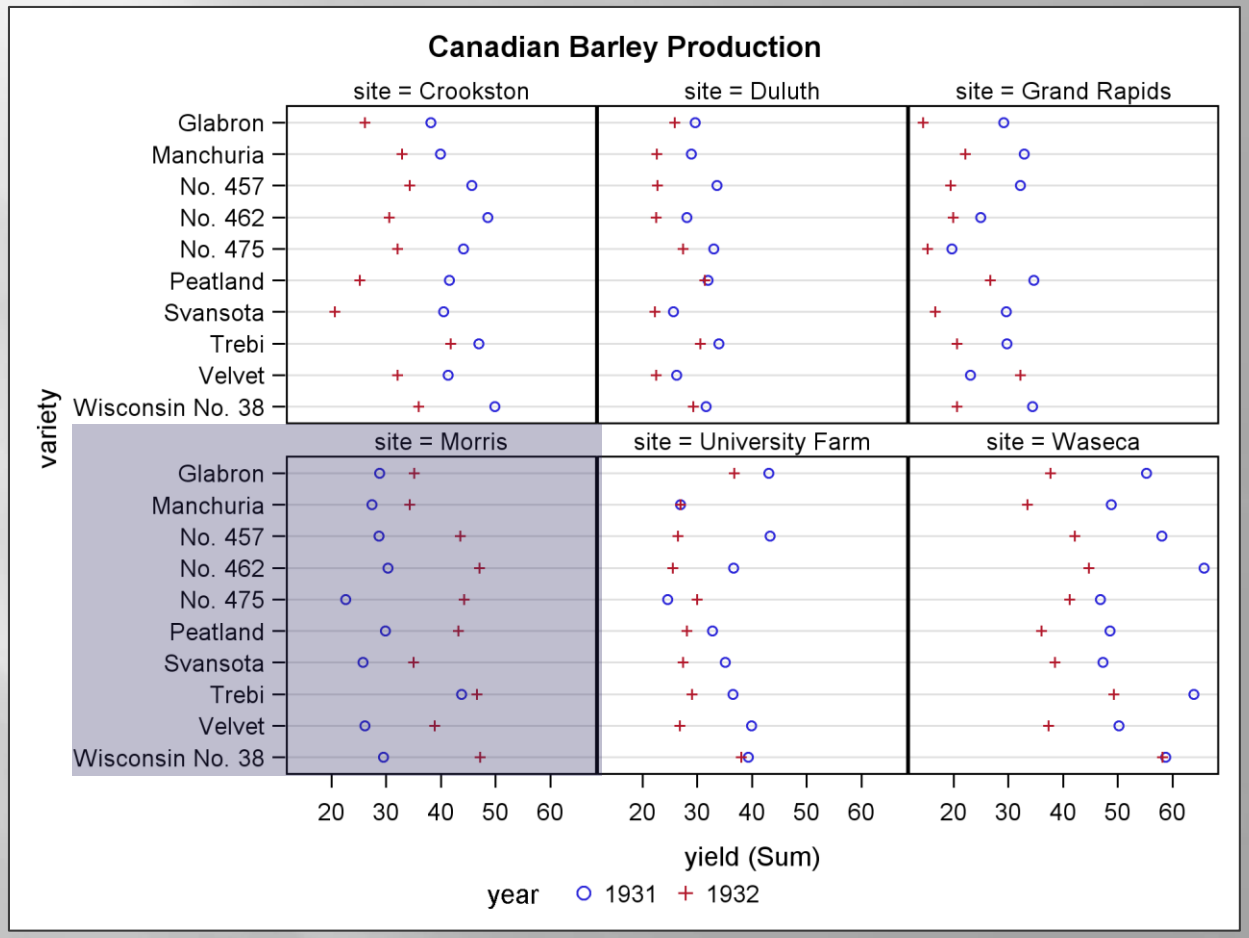
## *Working with "Multi-Way" Dot Plots*




- The named inventor, William S. Cleveland, recommends his dot plot as a replacement for the horizontal bar chart.
- The barley data *"multi-way"* dot plot is famous. R.A. Fisher used the data to illustrate his ANOVA method of experimental design. Years later, Cleveland discovers the data error that ANOVA missed.

# The Barley Data *"Multi-Way"* Dot Plot

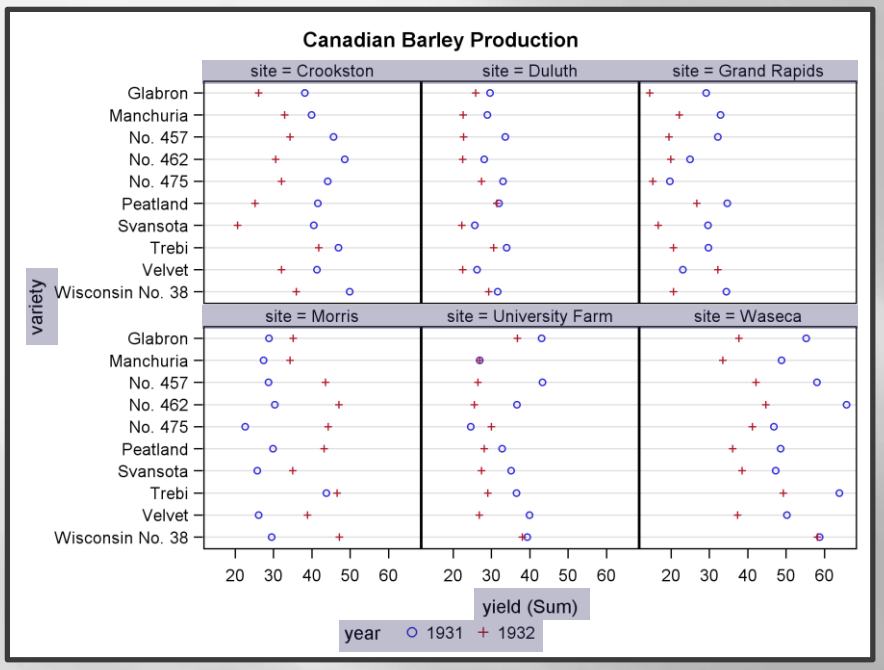
## *The Data Error*

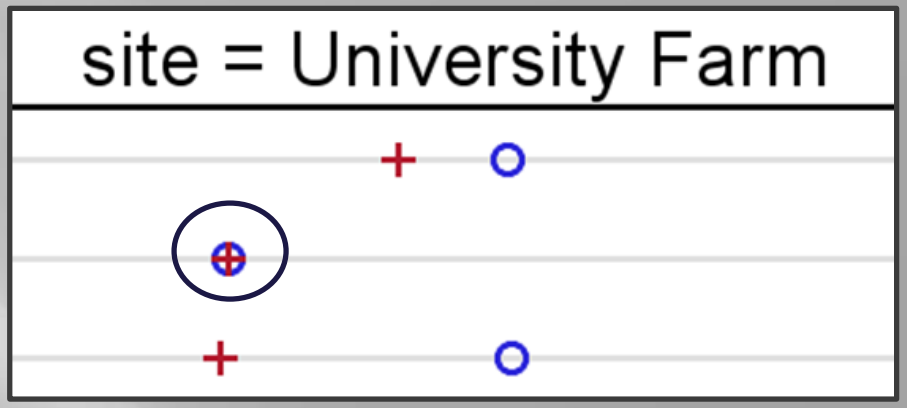



1931 and 1932 YIELDS are reversed at the MORRIS site

# The Barley Data *"Multi-Way"* Dot Plot

## *From the DOT Statement in PROC SGPANEL*



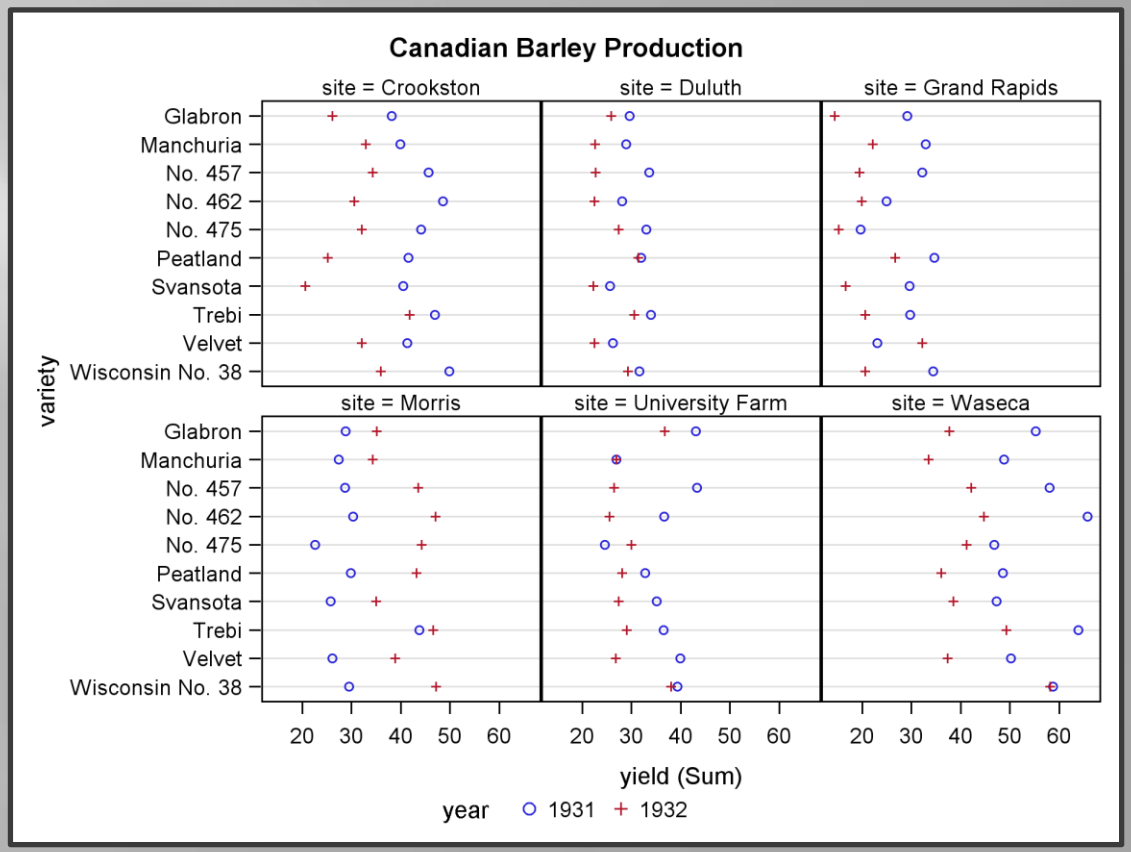

Cleveland supplied the data on STATLIB

```
proc sgpanel data=barley;
  title1 "Canadian Barley Production";
  panelby site;
  dot variety / response=yield group=year;
run;
```

# The Barley Data *"Multi-Way"* Dot Plot

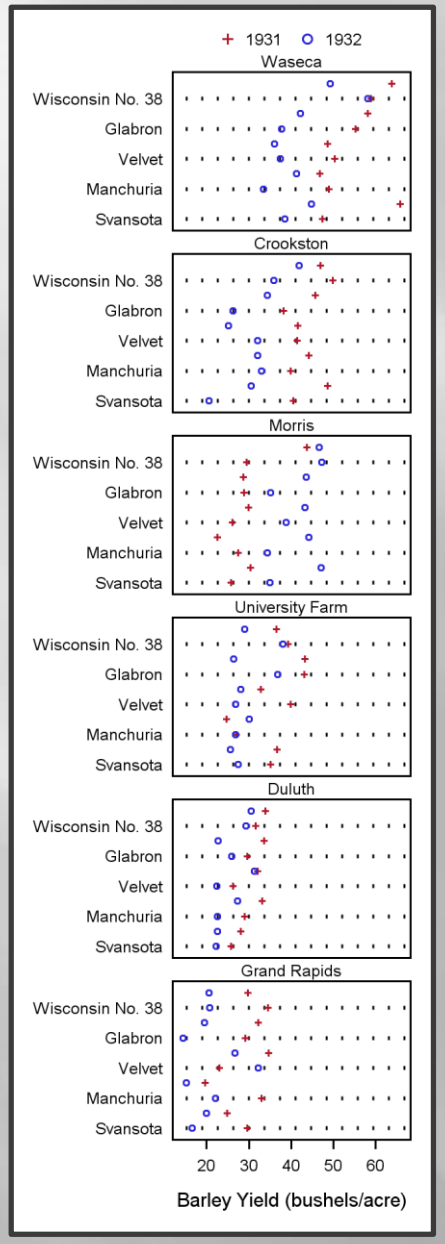
## *From the DOT Statement in PROC SGPANEL*

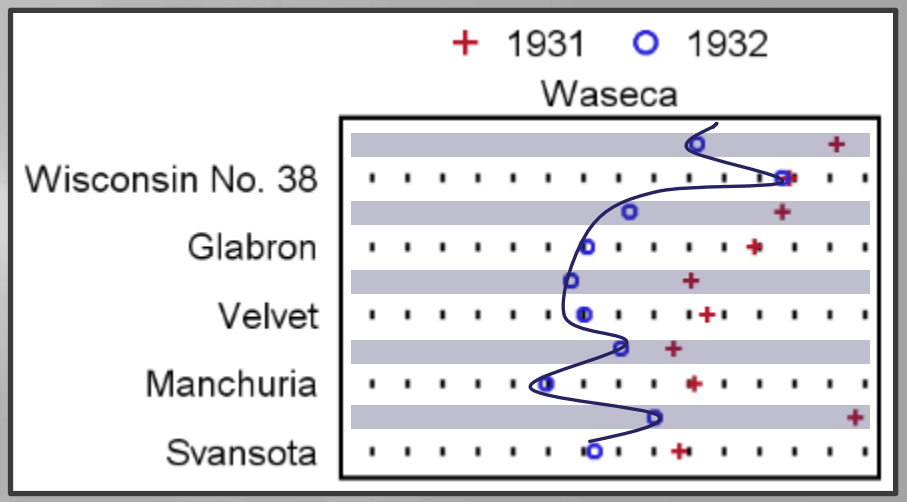
- Sites are in Minnesota.
- Can't see patterns. SITE and VARIETY are ordered alphabetically, not by median.
- To re-order, switch from DOT in SGPANEL to SCATTERPLOT in GTL.




- Plot a 6X1 paneled graph to see the connection between SITE and YIELD.

# The Barley Data *"Multi-Way"* Dot Plot

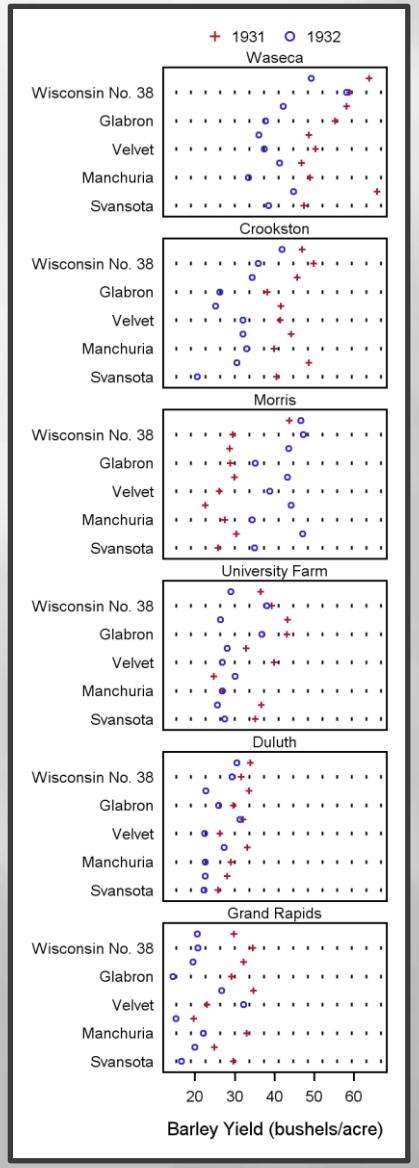


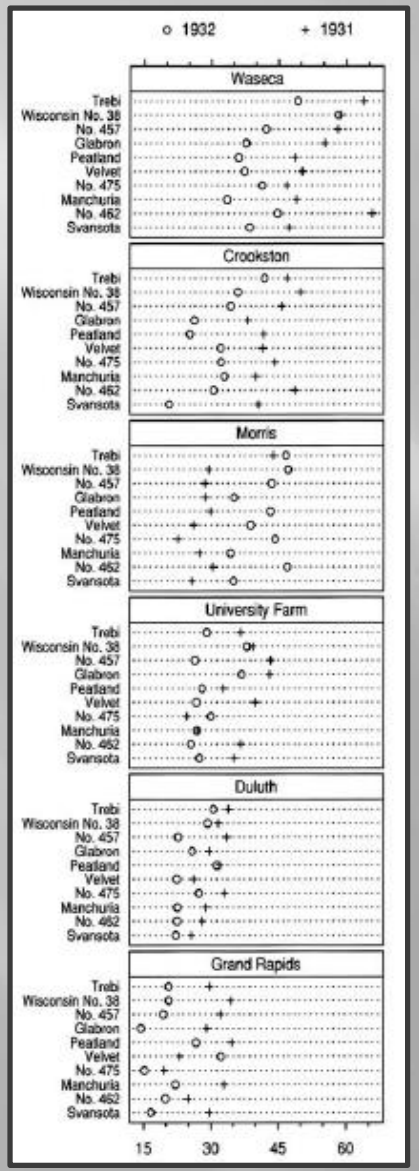

***From the SCATTERPLOT Statement in GTL***

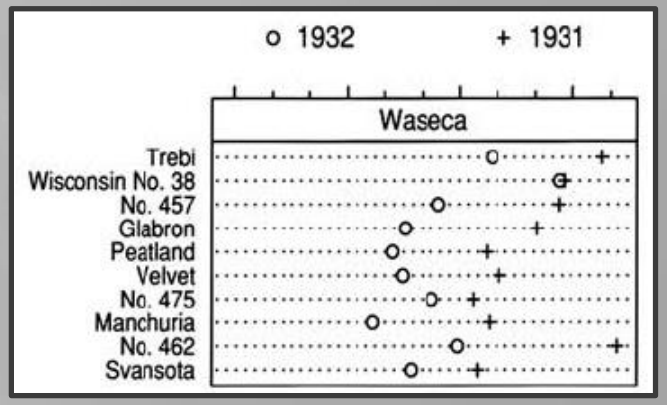
# The Barley Data *"Multi-Way"* Dot Plot

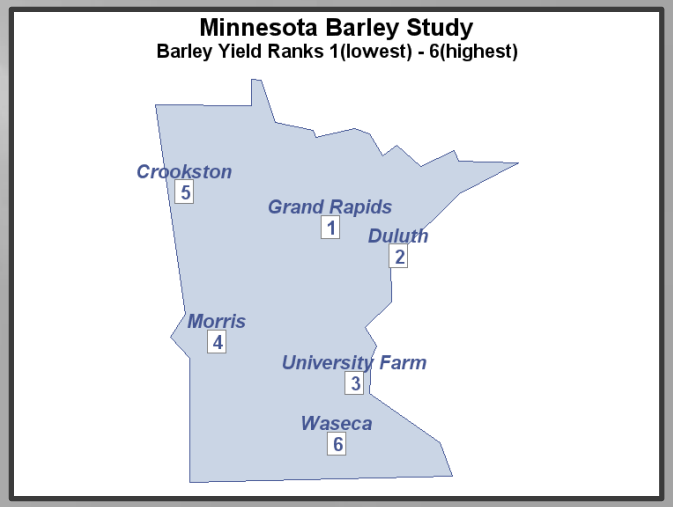
## *GTL*




## *CLEVELAND'S Graph*



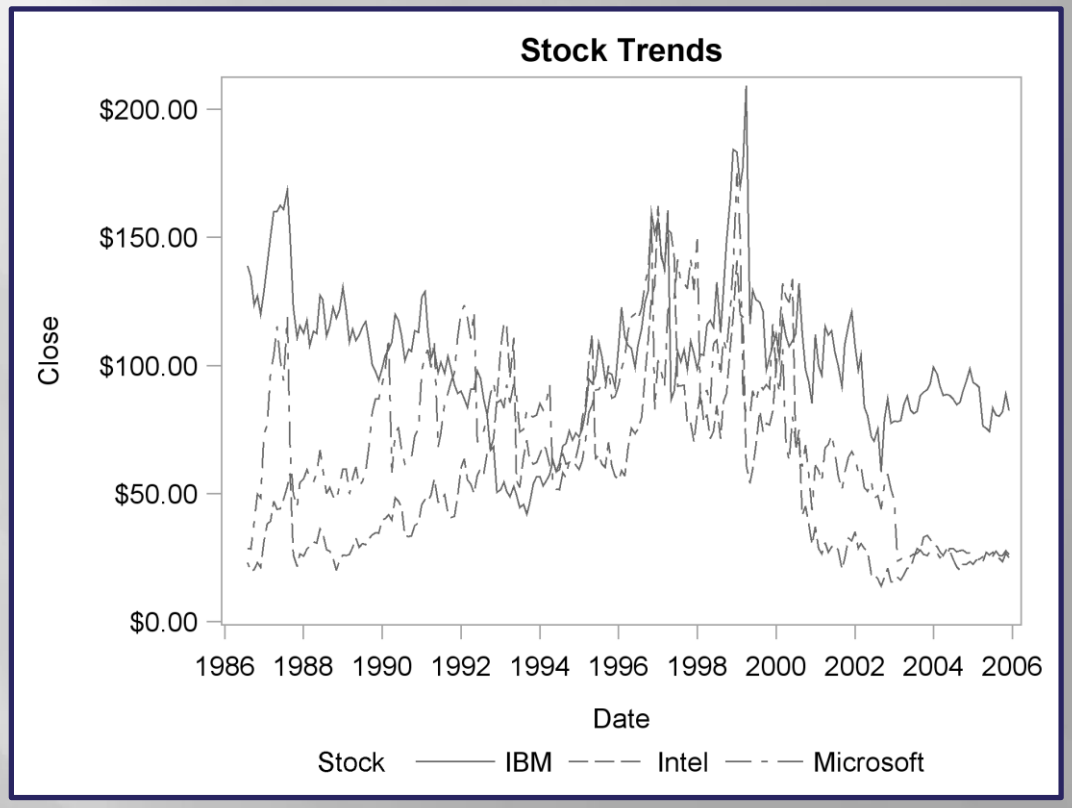




*The Elements of Graphing Data*

# Stock Data (sashelp.stocks)

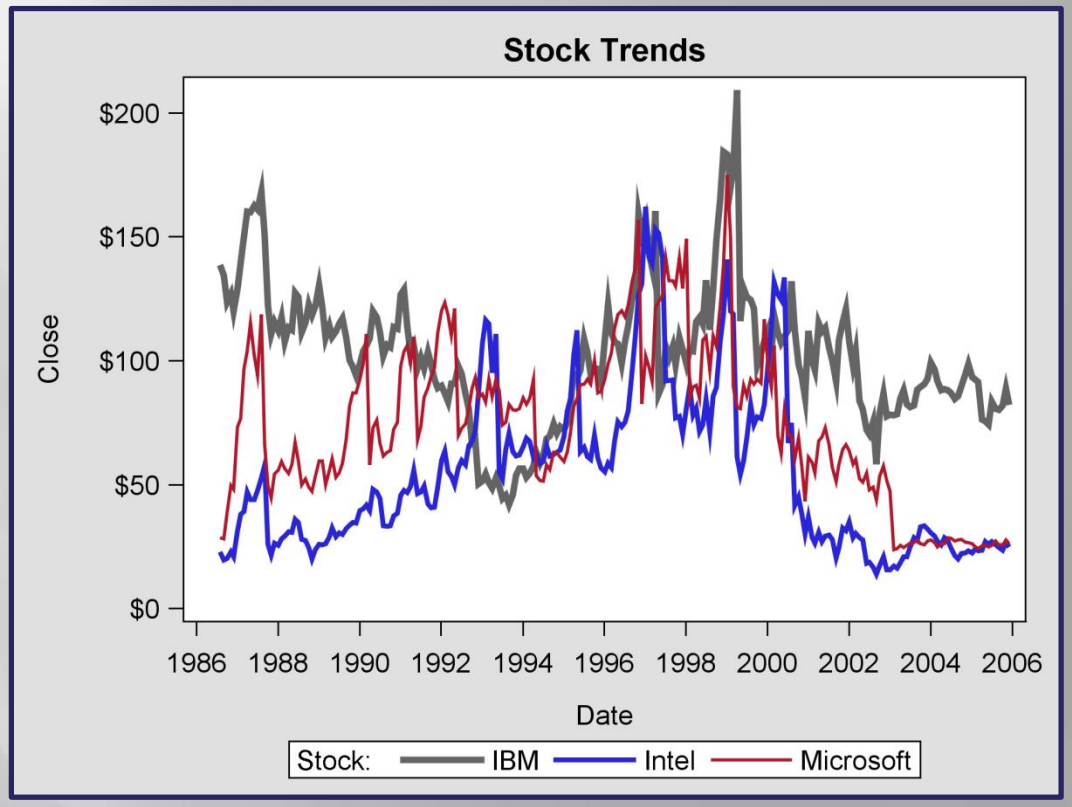
## *Working with Interleaving Time Series Plots*




Stock trends are difficult to see in this graph.
Overlaid dashed lines are difficult to track.

# Stock Data (sashelp.stocks)

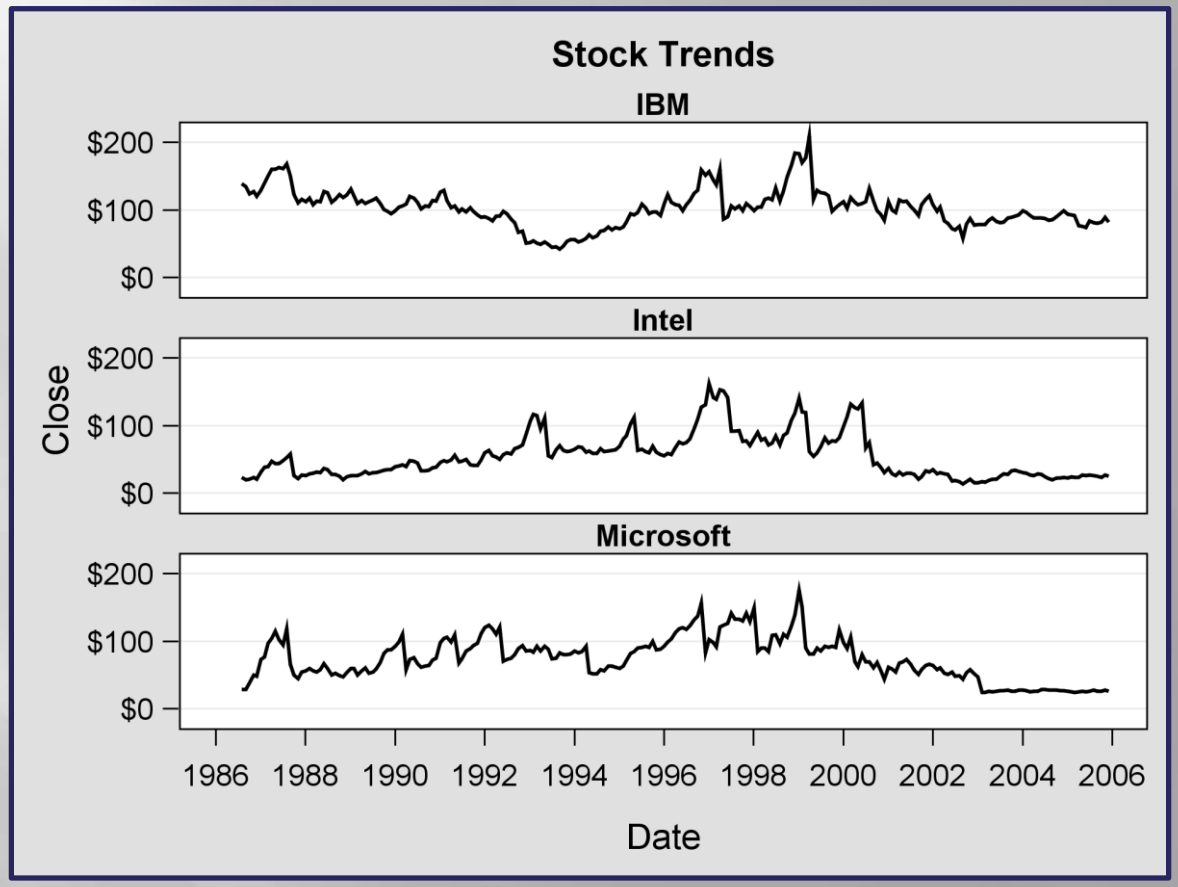
## *Working with Interleaving Time Series Plots*




- Use the Default Style.
- Replace dashed lines with solid ones.
- Use different line widths.
- Use anti-aliasing to improve resolution.

# Stock Data (sashelp.stocks)

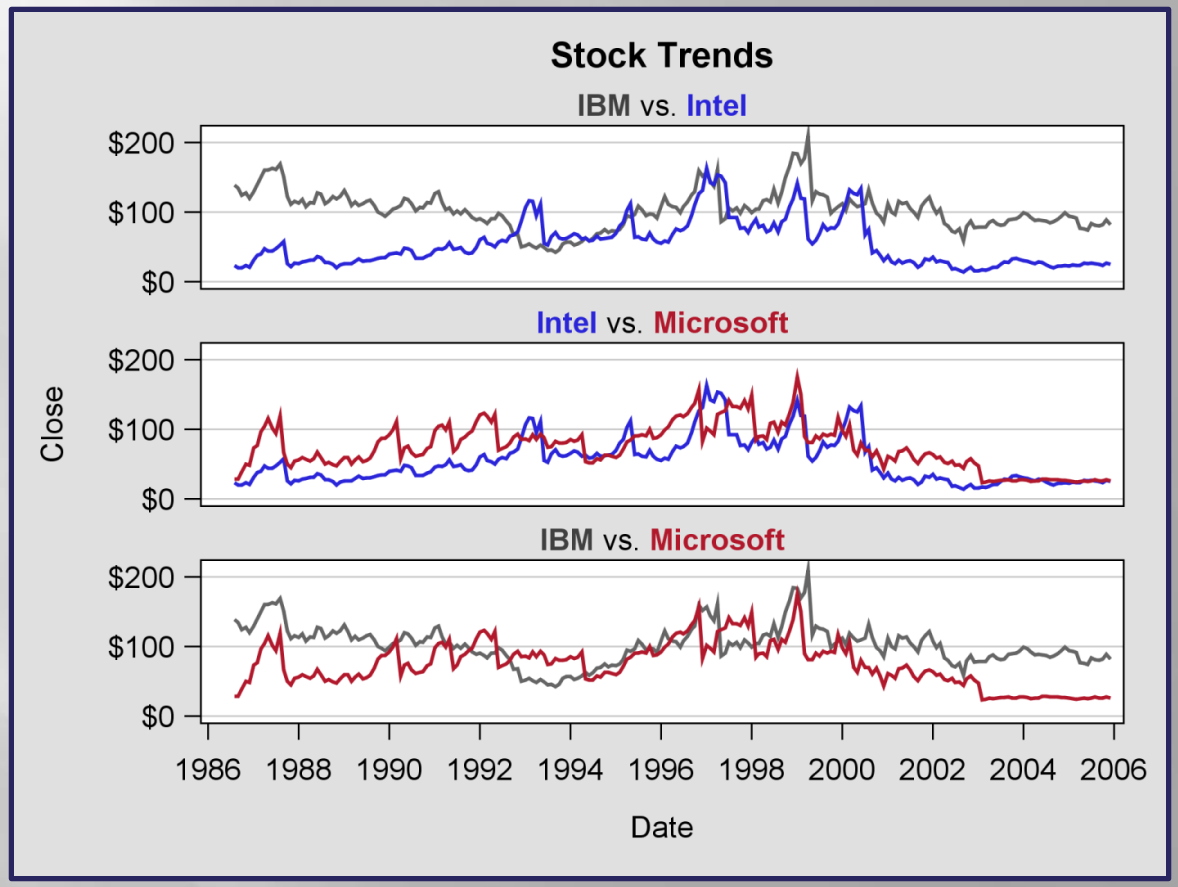
## *Working with Interleaving Time Series Plots*




Naomi Robbins says to place plot lines into separate panels to increase visibility. However, lines are then harder to compare.

# Stock Data (sashelp.stocks)

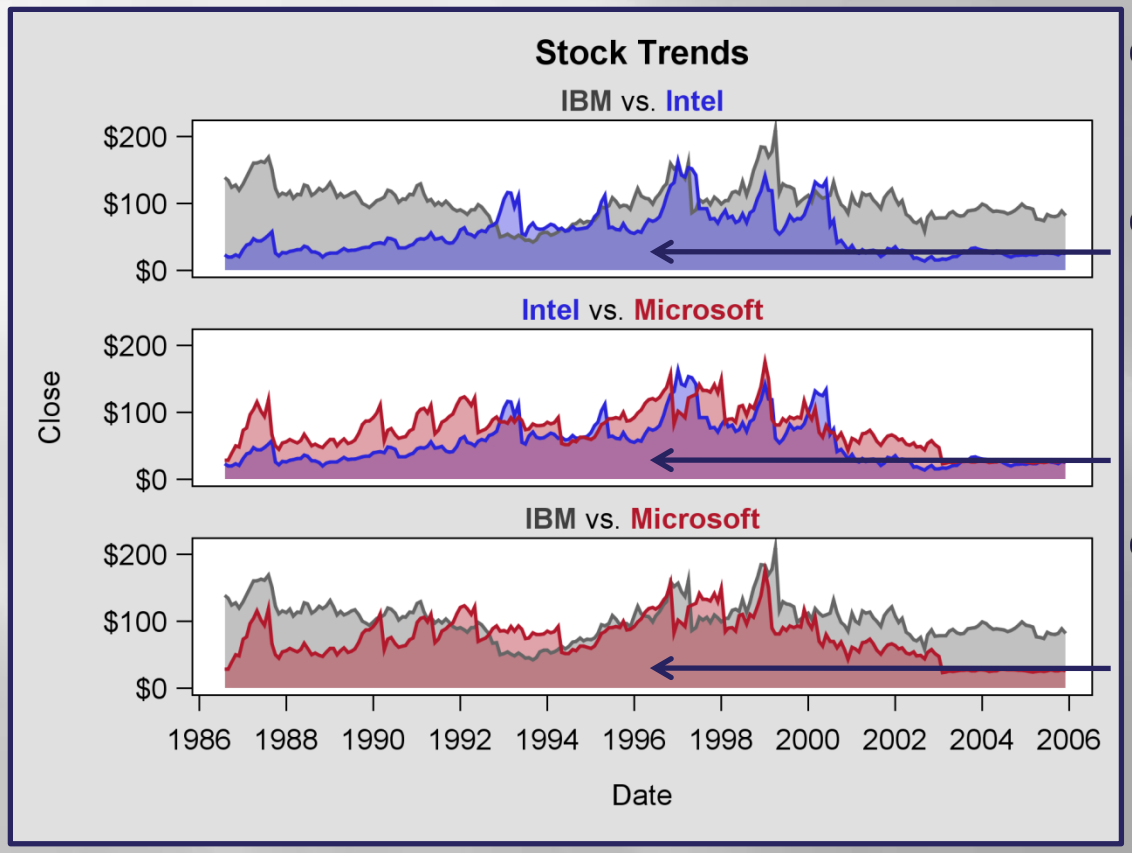
## *Working with Interleaving Time Series Plots*




Display stocks two at a time to increase comparability.

# Stock Data (sashelp.stocks)

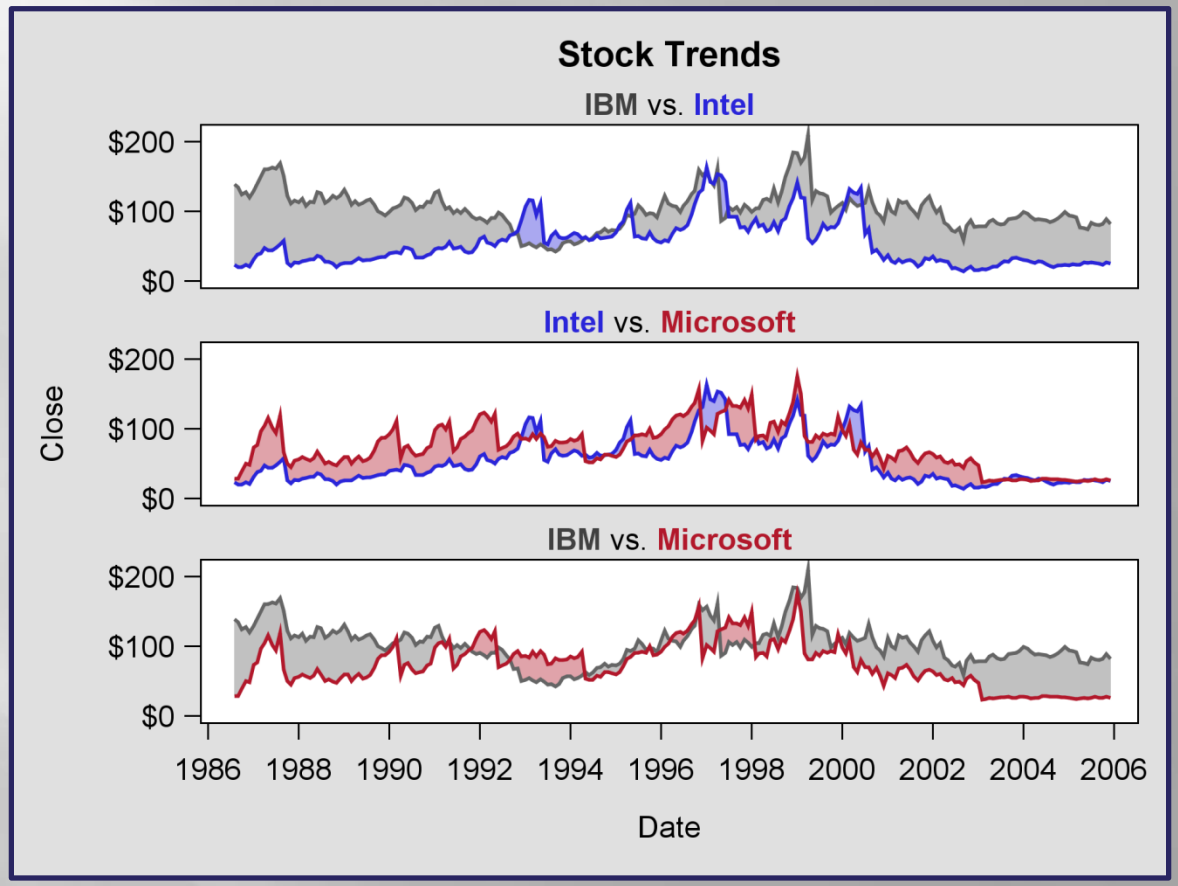
## *Working with Interleaving Time Series Plots*




- Add Band Plots for emphasis.
- Schwartz says the bands represent the Area Under the Curve (**AUC**).
- LIMITLOWER for both curves is set to $0, but this creates unwanted overlay (see arrows).

# Stock Data (sashelp.stocks)

## *Working with Interleaving Time Series Plots*

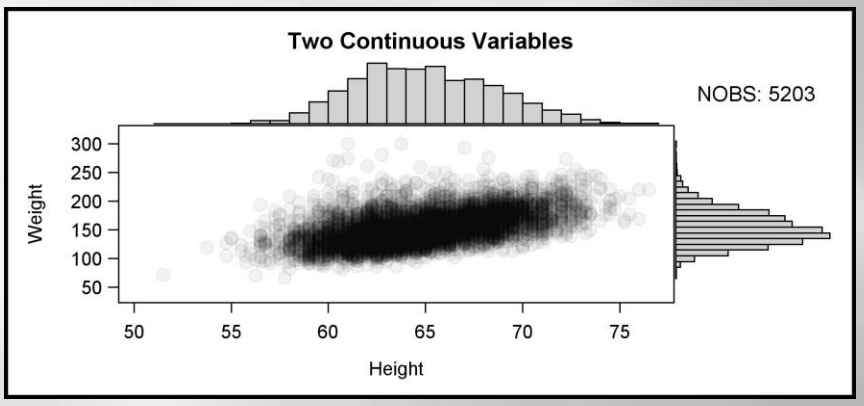



With "interleaved" band plots, the area ***between*** the curves (**ABC**) is emphasized.

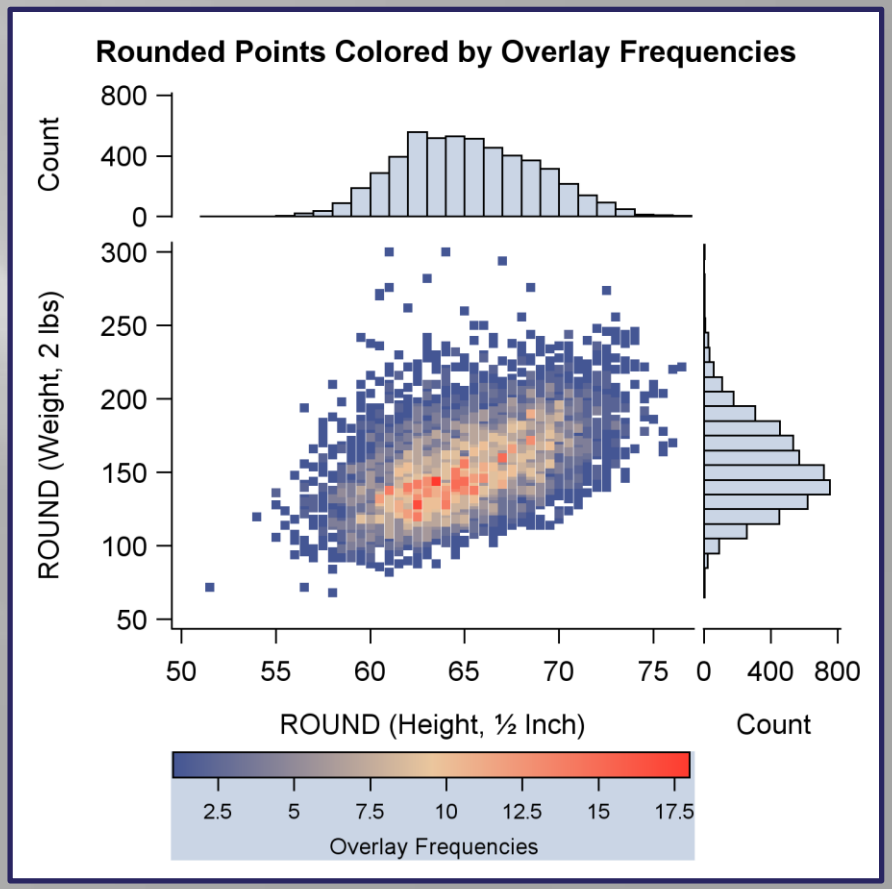
# Summary

## *Heart Data: Overlapping Points (n=5,209)*

**Before**




**After (1)**

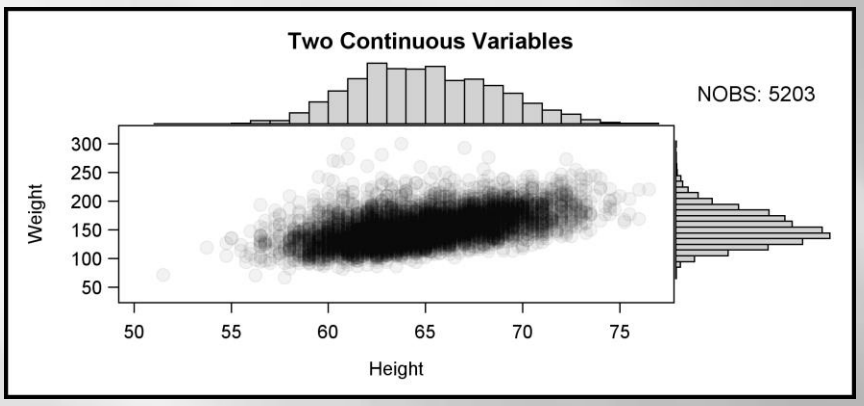



**Color** provides a 3rd Dimension for ***Frequency***

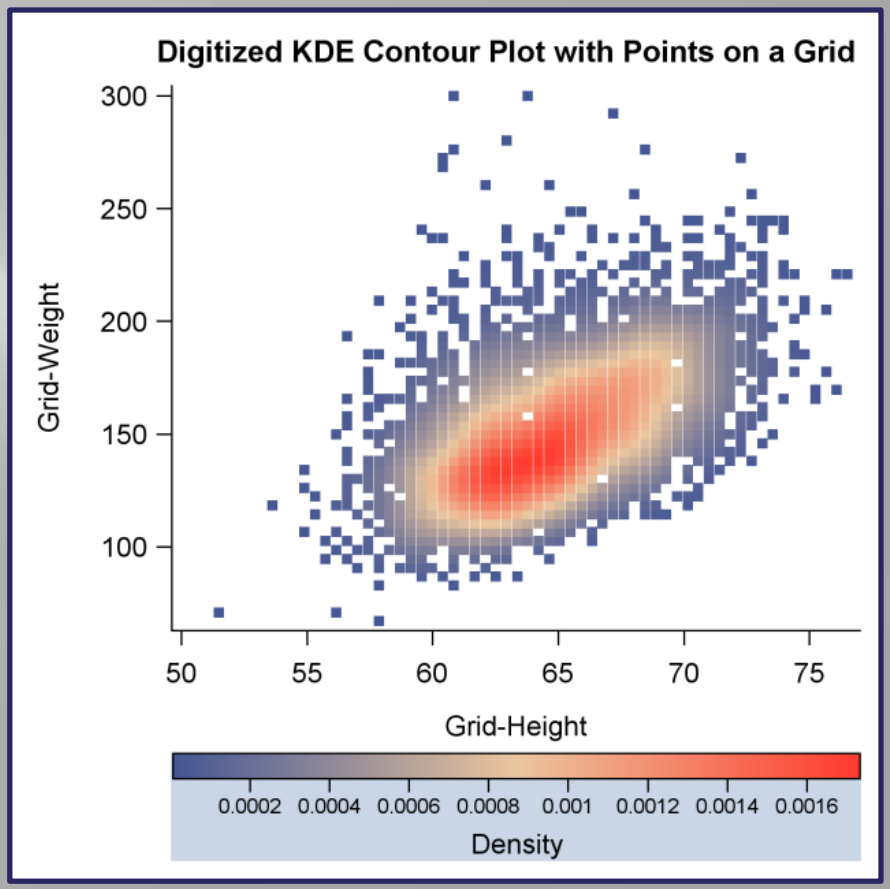
# Summary

## *Heart Data: Overlapping Points (n=5,209)*

**Before**




**After (2)**



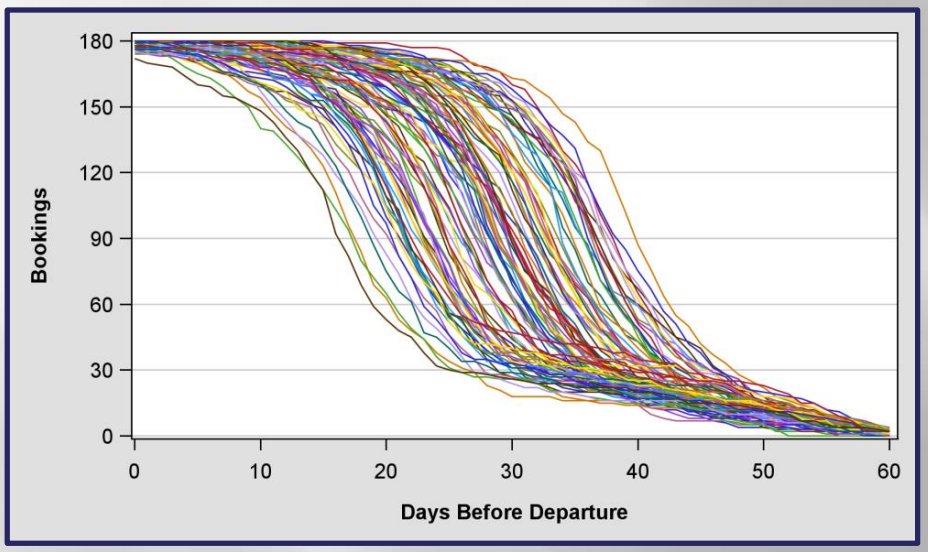

**Color** provides a 3rd Dimension for ***Density***

STAKANA

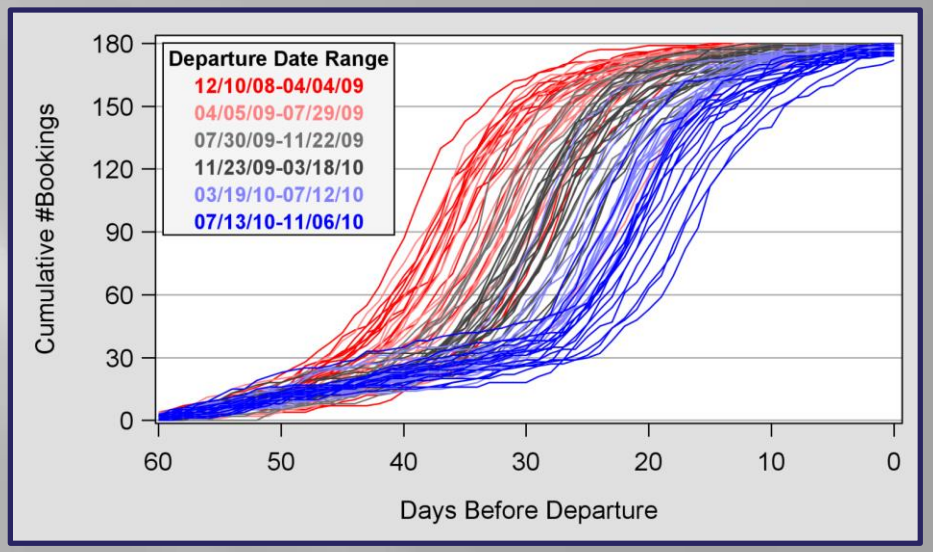
# Summary

## *Airlines Data: Overlapping Lines (n=6,100)*

**Before**




**After (1)**

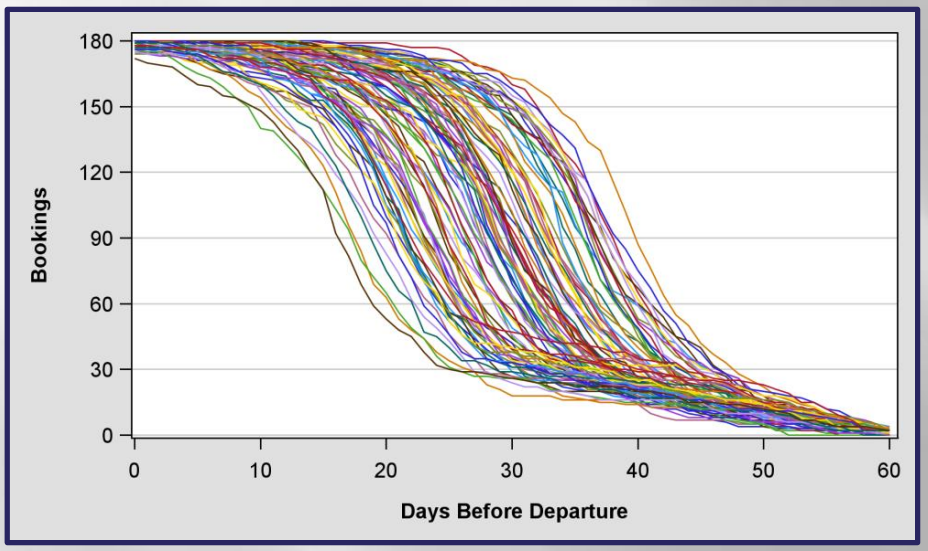



**Color** provides a 4th Dimension for the Departure Date Range

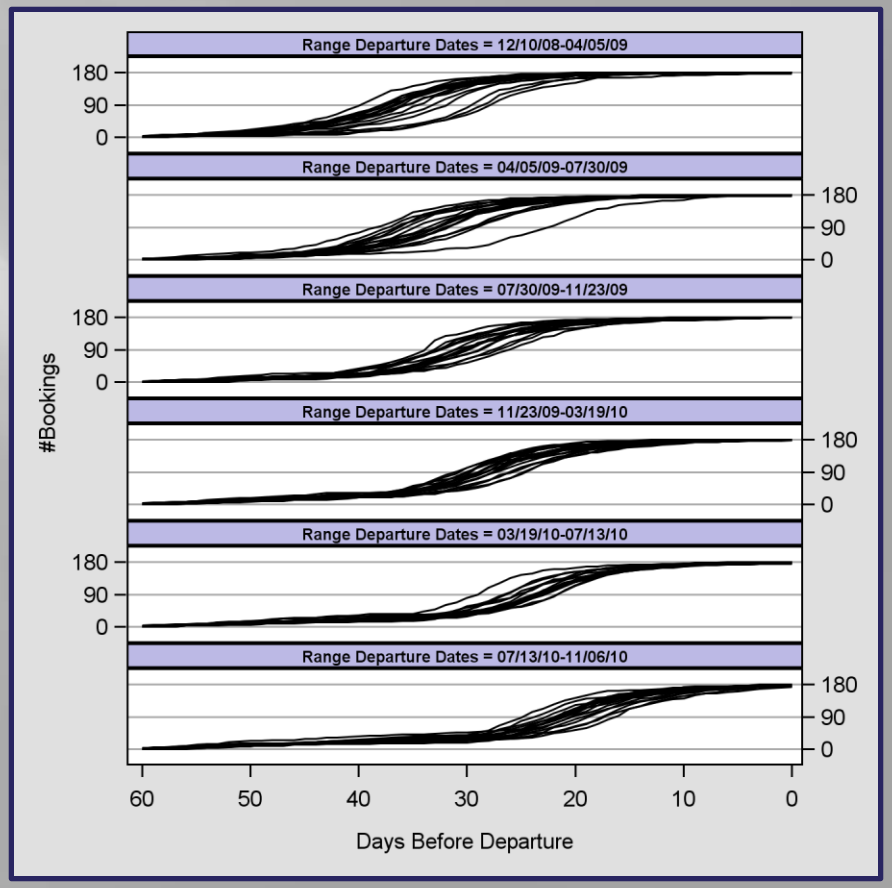
# Summary

## *Airlines Data: Overlapping Lines (n=6,100)*

**Before**




**After (2)**



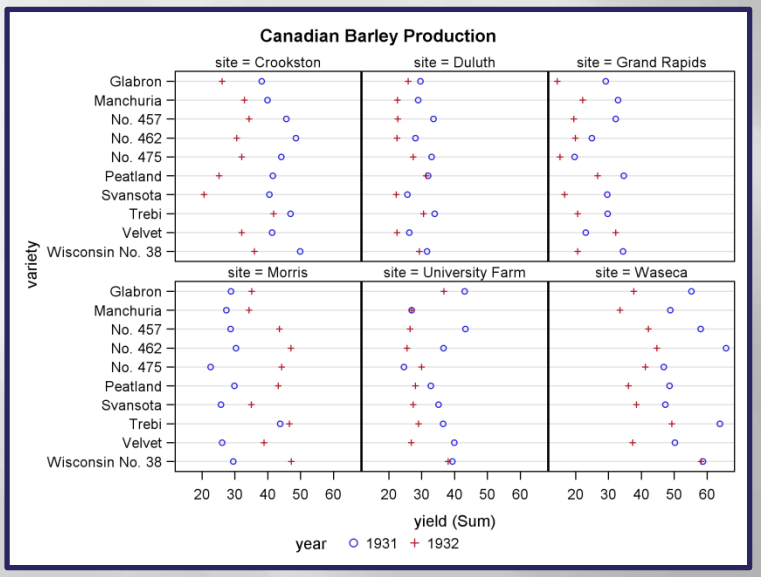

**DATAPANEL** now provides the 4th Dimension

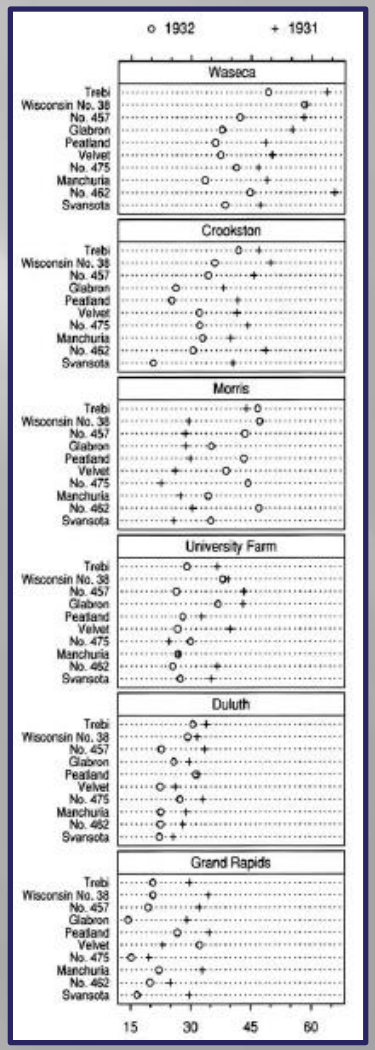
STAKANA

# Summary

## *Barley Data: Overlapping tick labels (n=120)*

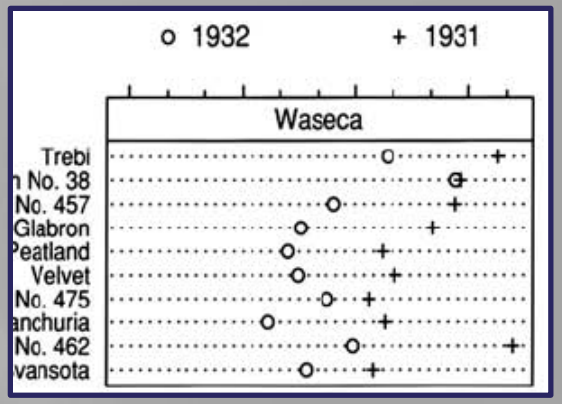
**Before**




**After**

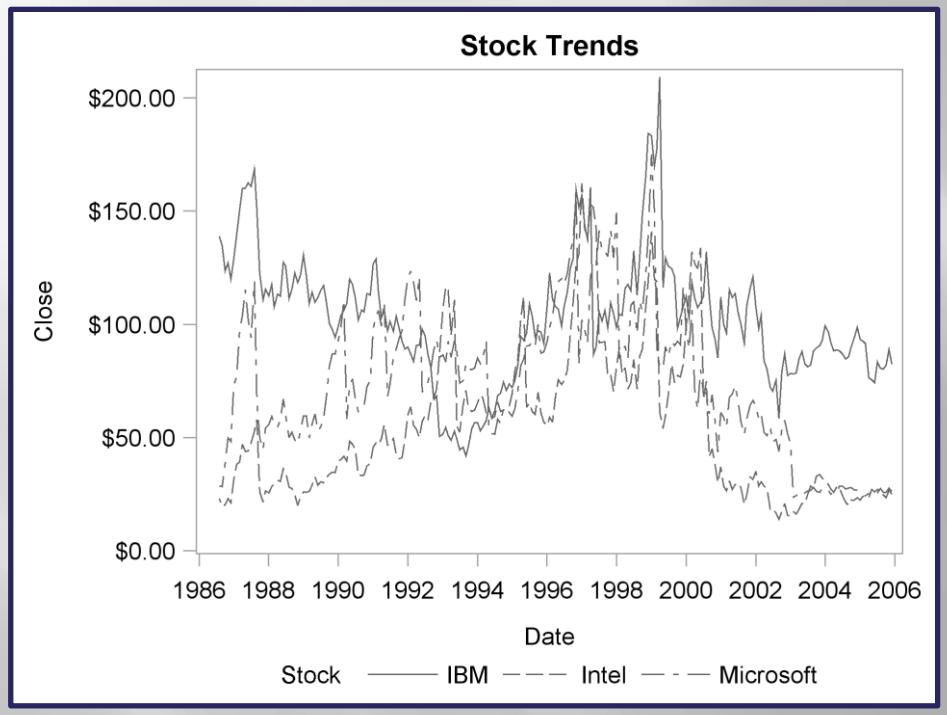
**Did not work in 9.2 SAS**



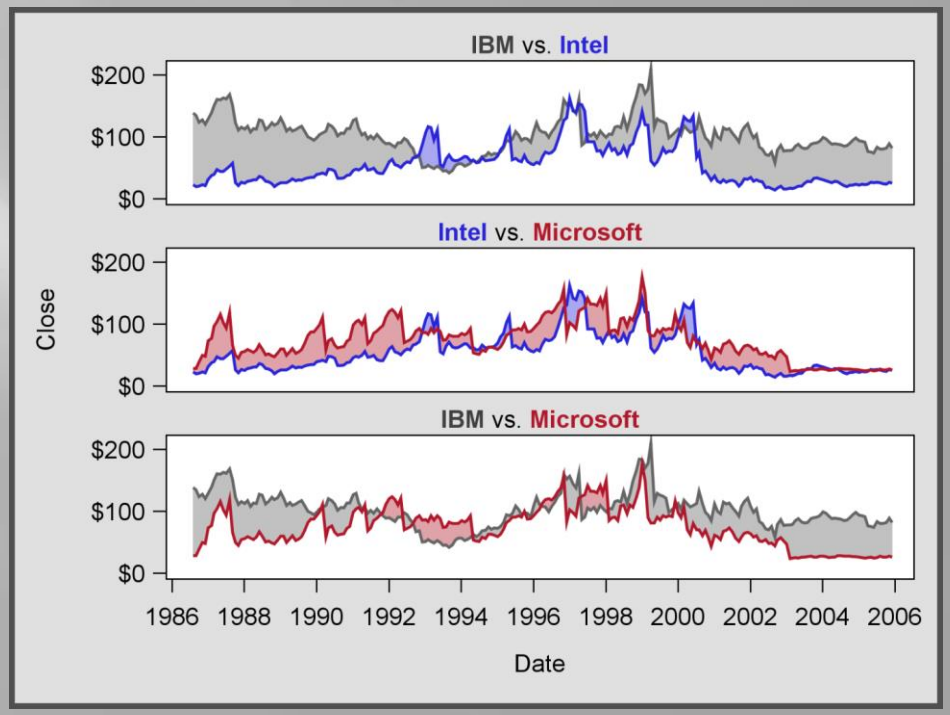
# Summary

## *Stock Data: Interleaving lines (n=699)*

**Before**




**After**



# Contact Information

**Perry Watts**
**Stakana Analytics**
**pwatts@stakana.com**
**www.PerryWatts.org**

**Nate Derby**
**Stakana Analytics**
**nderby@stakana.com**
**www.NDerby.org**