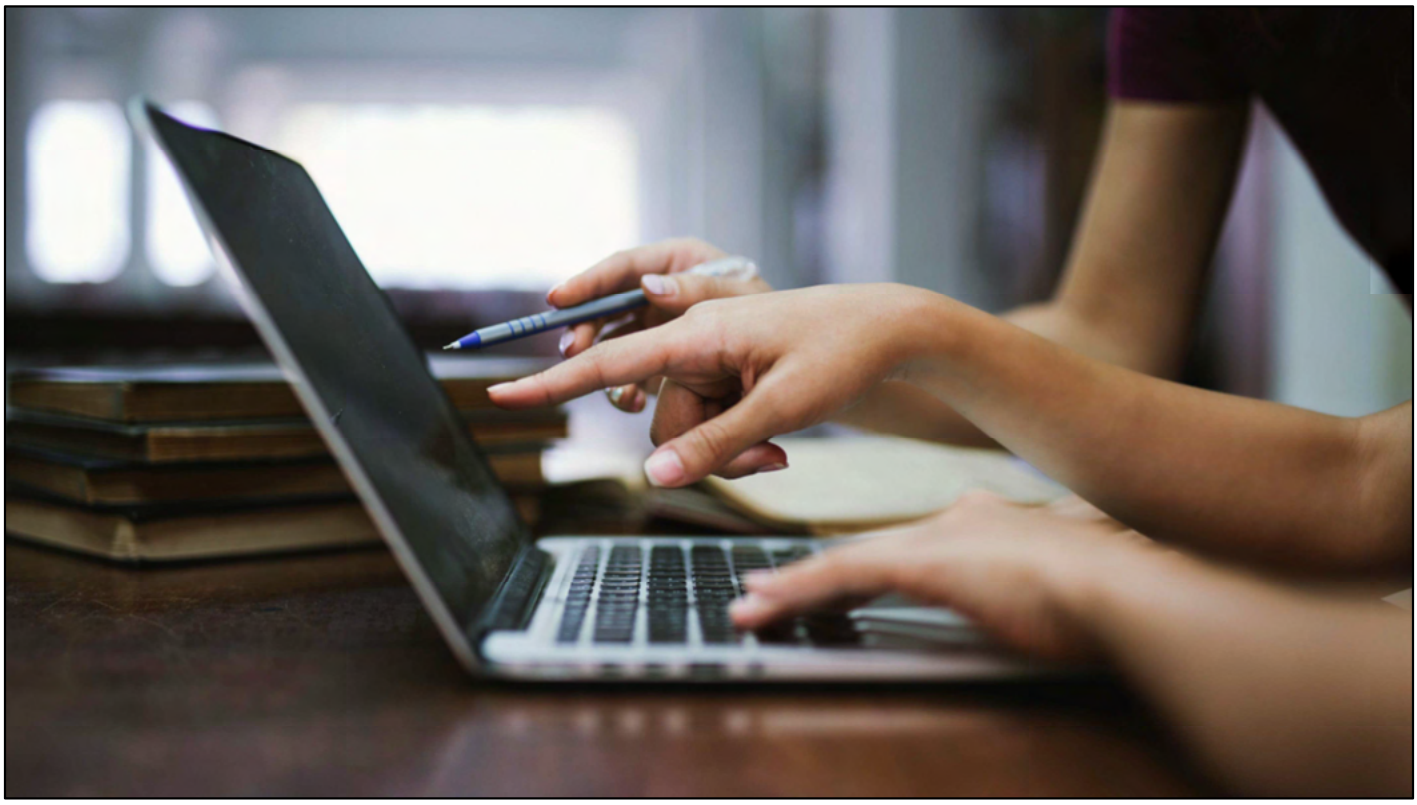
Building a SAS Data Viz Toolkit

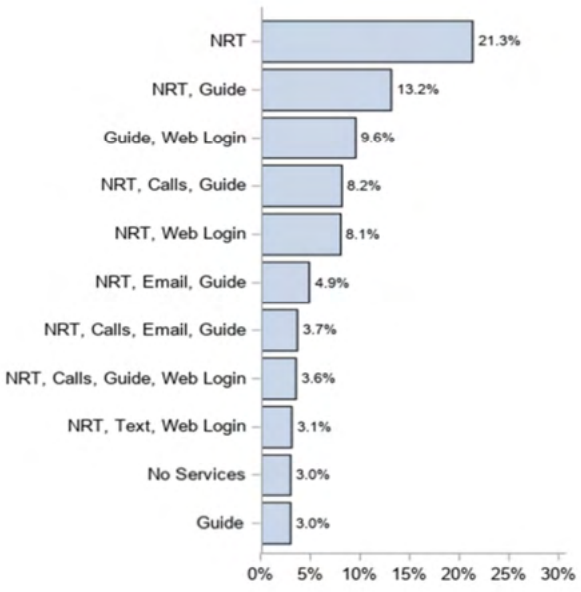
Sara Richter, MS
MWSUG Annual Conference 2024
#BB050



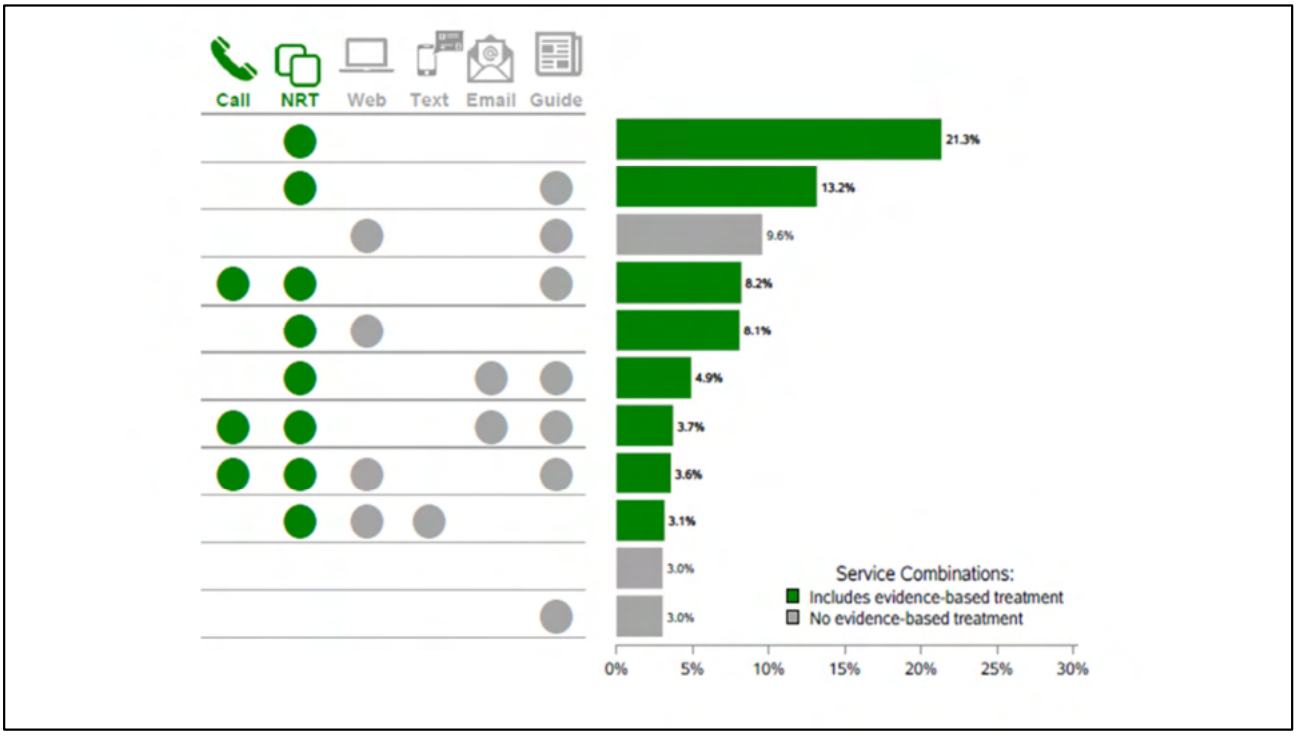
Data visualizations are really important when conveying analysis results, especially if you have a broad audience. Personally, I work with a variety of clients, some very data-savvy and some not. Even the ones that are often forward my results to colleagues, so my results and visualizations really need to stand alone.

Understanding the purpose of the visualization and the audience can help tailor the image. Applying some simple best practices can take the visualization to the next level.

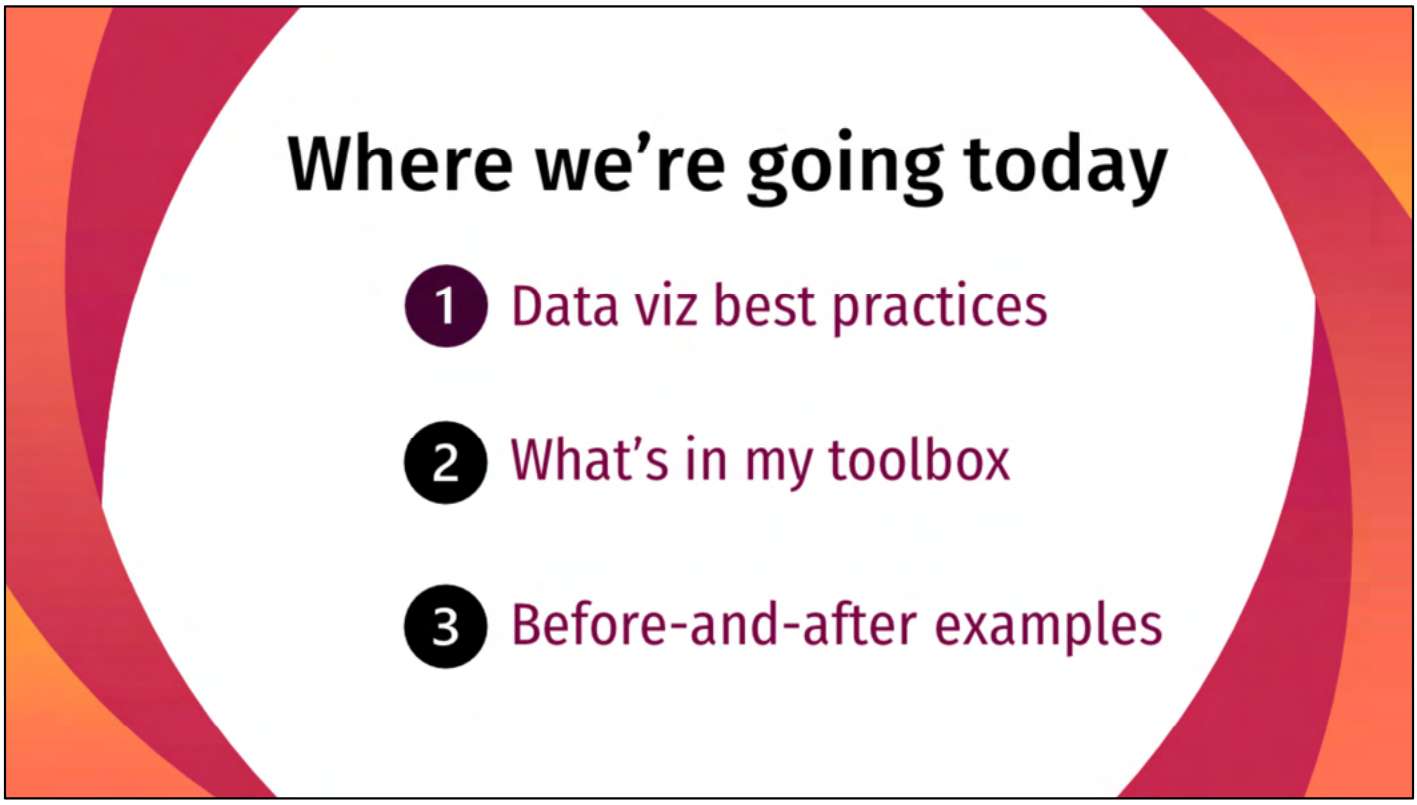
SAS Viya and Visual Analytics are great visualization tools, though not universally available. SAS/GRAPH can provide a great alternative, though it does take some extra tools in your visualization toolkit.



I'm going to admit I was not always an advocate of graphing in SAS. In fact, when I first started learning SAS version 9.1, I didn't graph in it. I'd export data and to graph it in another program else because the visualizations in SAS were difficult and never quite turned out right. Eventually, I needed to start creating reports from SAS, which included graphs. Thankfully, SAS's graphing capabilities grew by leaps and bounds, letting me turn this basic bar chart. Into this…

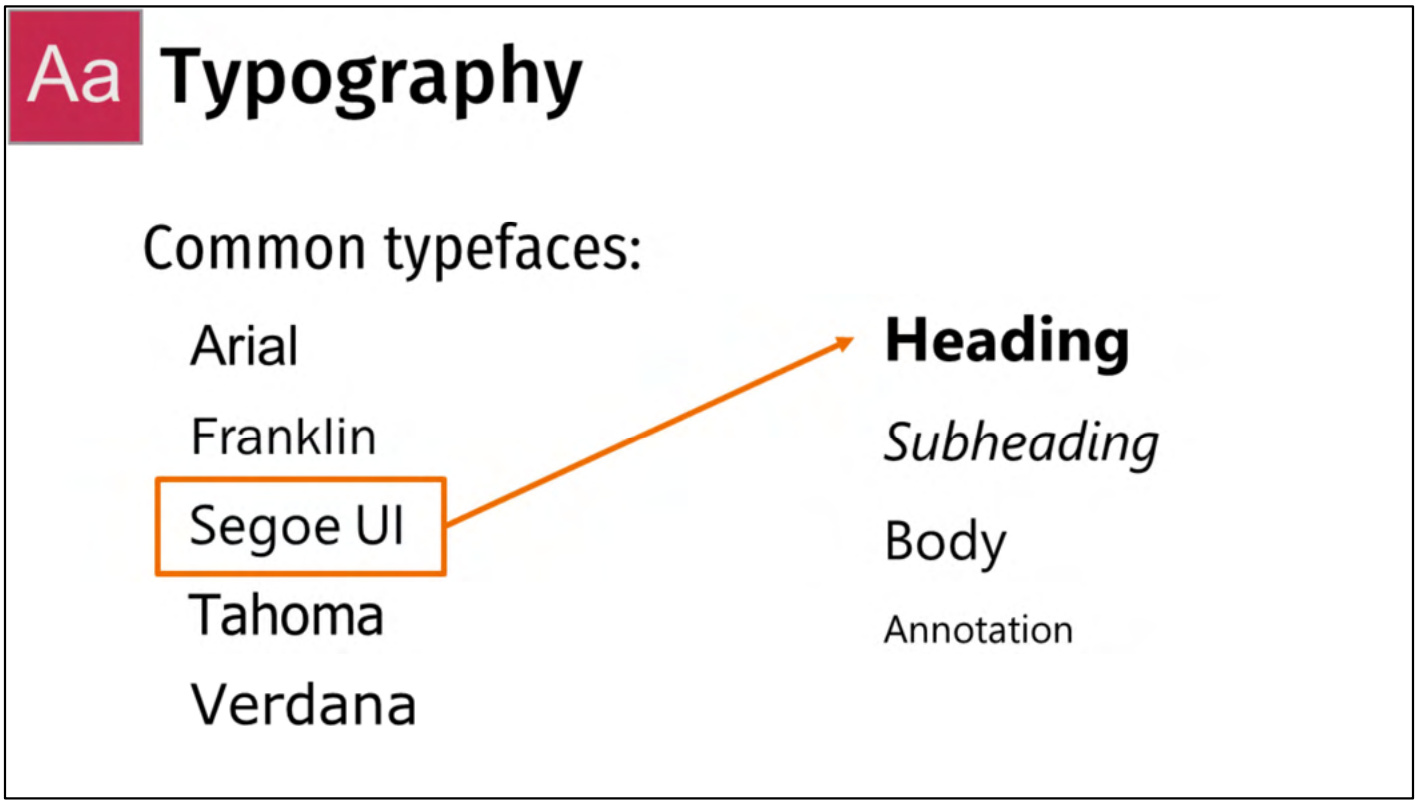


more visually engaging, easier-to-read version of the same chart. And best of all, I was able to do it in Base SAS. By the end of my presentation today, you'll learn how I put this visualization together.


Where we're going today
1 Data viz best practices
2 What's in my toolbox
3 Before-and-after examples

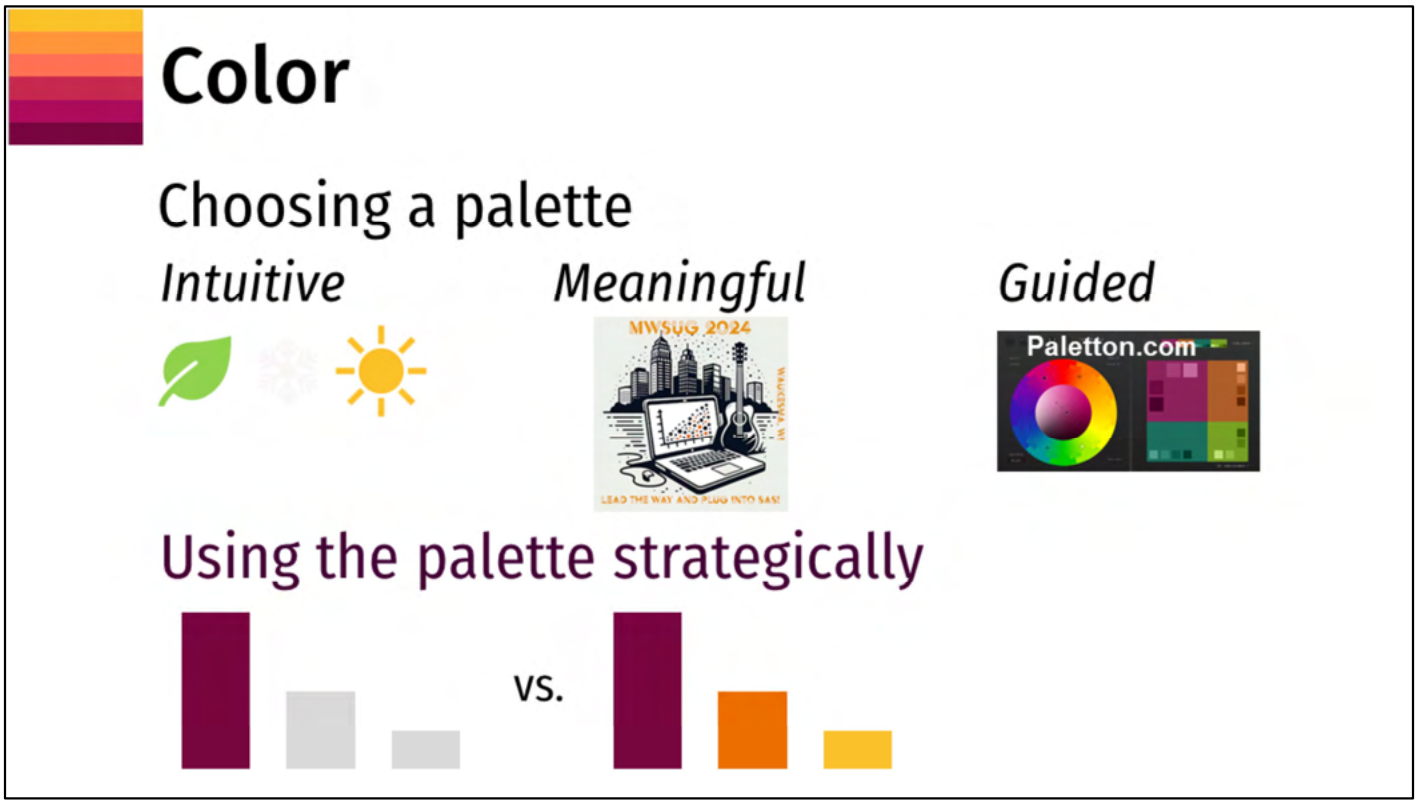
This presentation is meant to be a sampling or an overview of the tools I use for data viz. Sure, there will be some code, but that's not focus here. That said, if you're interested in more of the code, talk to me after and I'll be happy to discuss.

Let’s get started!



First – typography. Typefaces matter. Different fonts can portray the same idea a little differently depending on the weight of the text, the direction, the embellishments, the spacing of the letters – it can all add to or detract from the impression you're trying to make. Above all, the text needs to be easily readable. In today's world, that means it needs to be fairly universal – a font that reads well across devices (PCs, Macs, tablets, phones, etc.), softwares (Microsoft products, SAS, LaTeX, R, etc.), software versions (e.g. Microsoft Word 2010 and 2016), and that reads well online, if applicable. If it's not universal, the formatting will be messed up when the client opens your document.

Once you have picked a font – use it to your advantage. Modify the size, weight, and styling to help convey your messages.



There is lots of guidance related to choosing a color and I'm sure you've heard most of it – don't pick colors that are too similar, consider colors work for all types of color impairments, with enough contrast to print well in black and white. That still leaves a lot of options – so here are some ideas for ideas for choosing a color palette:

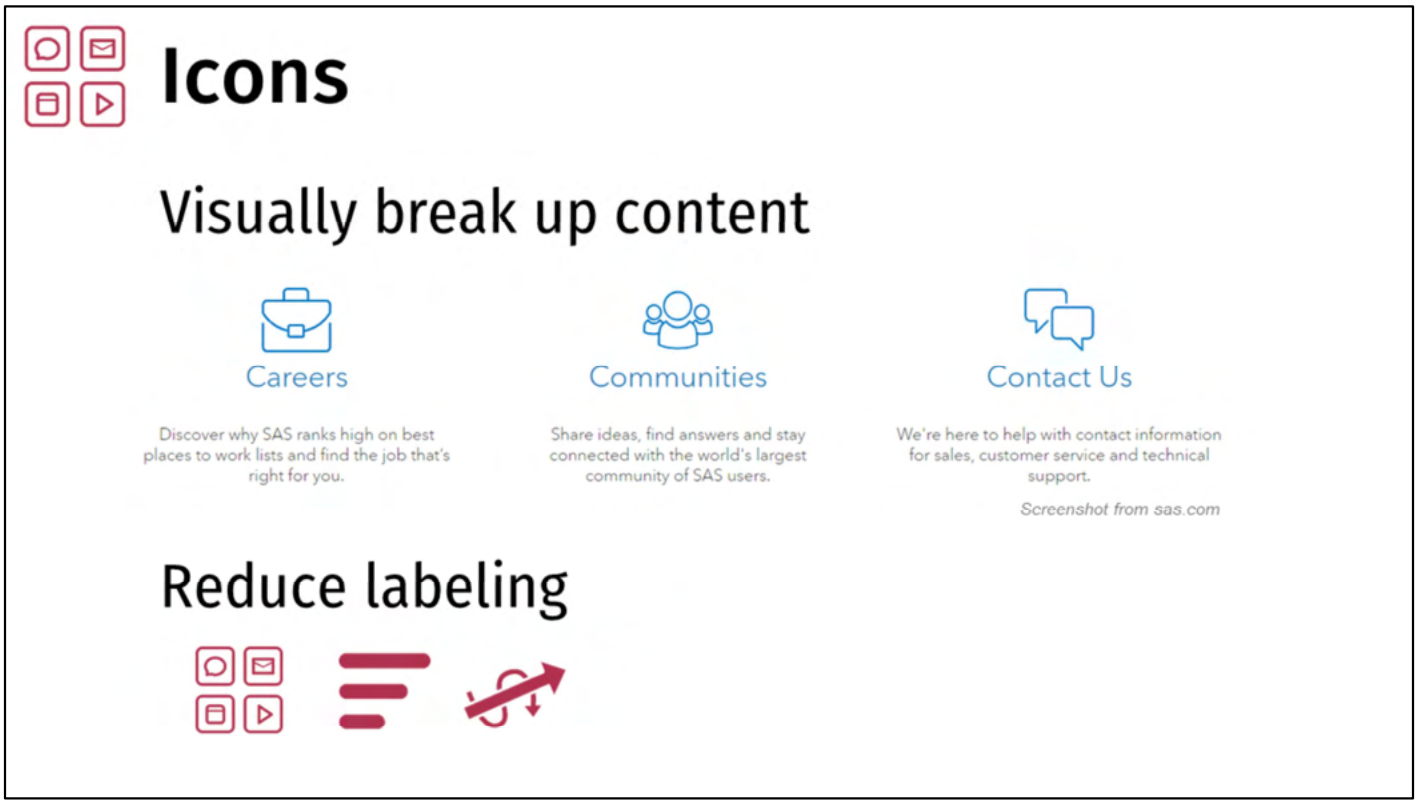
1) Use the intuitive color. Talking about new growth? Consider a spring green. Sunny outlook? Marigold. Democrats vs. Republicans – blue vs. red. If there are already colors associated with the topic you are visualizing, use those colors. Using other colors will only confuse the reader.

2) There's not always an intuitive option for the subject matter. Another option is to use a color scheme that's meaningful to the client – use colors from their logo, their branding, or their project colors. If there is a salient image or photo for the organization or project, you could use colors from that image. You could also consider using colors from your own logo or branding, as I've done with this presentation.

3) When all else fails, use an online tool to help you pick. There are lots of options. One of the first ones that comes up on Google and is pretty easy to use is paletton.com. You can pick 3+ contrasting or congruent colors. It even gives you examples using the colors. You can find a color scheme or two that you like and save them for future use. Or, if that's not for you, there are sites that provide color

palettes. I have a couple of those listed in my Resources slide.

Once you've picked your colors, you need to use them consistently and strategically. Use the colors consistently – if you assign a color to a certain subgroup, only use it for that subgroup. If there are not consistent groupings, you can use color to delineate different topics/sections of your report. Don't use the same color to represent multiple subgroups or a subgroup and a topic. Using the colors strategically - no matter how you assign colors, knowing when NOT to use color is just as important as when to use it. Color can help tell your data story – use it to highlight a trend or to call attention to a particular result, as in the bar graph on the left and then use grey to de-emphasize the data that does not tell you something/is not of interest. If you use color all over, as in the graph on the right, your eye doesn't really know where to go.



Next up: icons – if I'm conflicted about any of these suggestions, it's icons. When they're done well they can visually break up content, making it easier to digest, and easier to have as a reference point. Icons can also be used to reduce labeling – you can associate an icon with a particular topic or theme and then use the icon to denote that theme throughout the report.

However, used too often or not in the right way, they can be very distracting and take away from what you're trying to say. Icons really need to be obvious – if you have to stretch to find a relevant icon, then you probably don't want to use one.

Order

Ascending/
Descending

Alphabetical

Sequential

Client importance

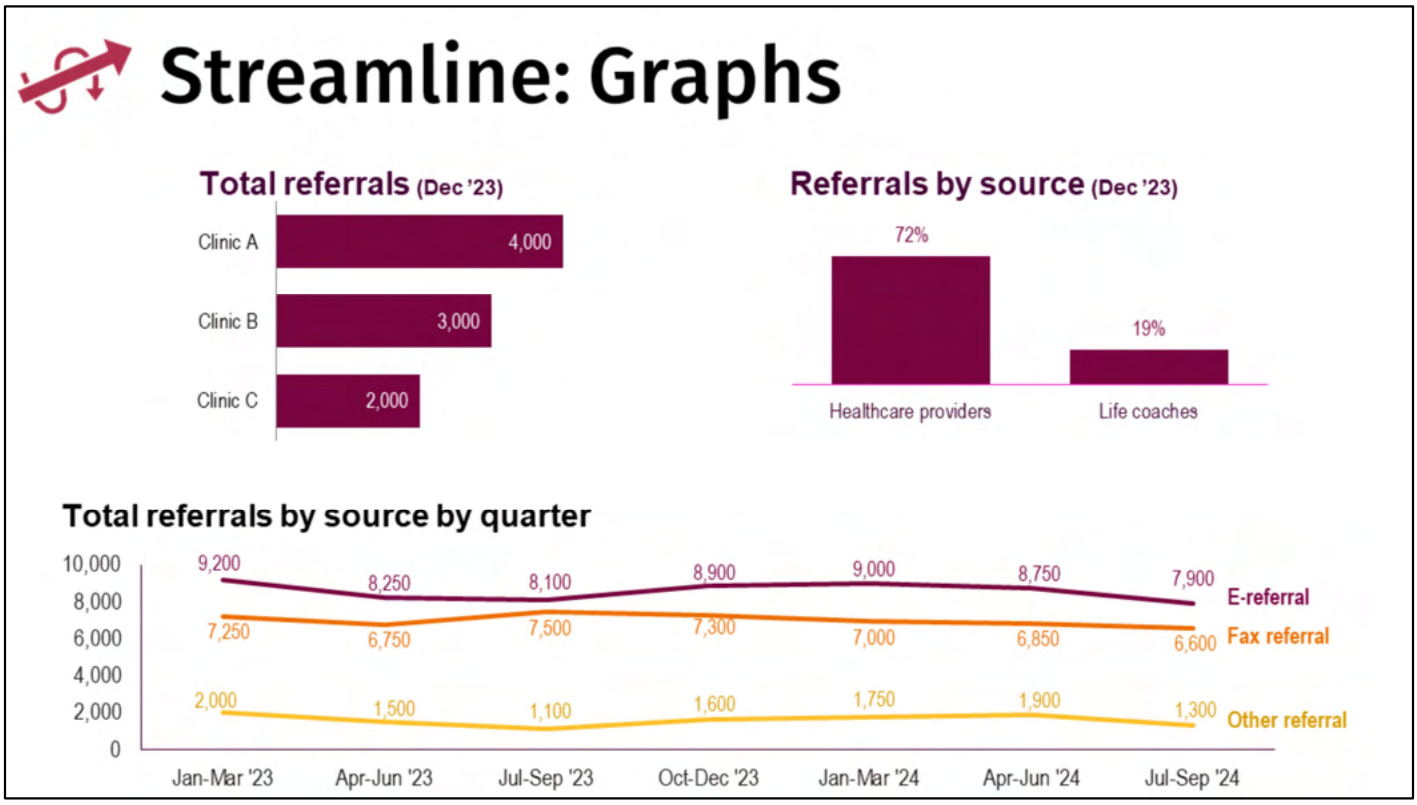
Order is important and something we don't think enough about – we should choose the order we want our results in every time. It's such an easy thing to do to help guide our readers/clients through a document, table, and graph. There are lots of options for how you order and depending on the situation, one may work better than another.

Streamline: Tables

	Oct-Dec '17		Jan-Mar '18		Apr-Jun '18		Jul-Sep '18		Oct-Dec '18	
eReferrals	*Site A*	*Site B*	*Site A*	*Site B*	*Site A*	*Site B*	*Site A*	*Site B*	*Site A*	*Site B*
Healthcare providers	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Life coaches	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Other	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Total eReferrals	**xxx**	**xxx**	**xxx**	**xxx**	**xxx**	**xxx**	**xxx**	**xxx**	**xxx**	**xxx**
Fax referrals										
Healthcare providers	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Life coaches	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Total fax referrals	**xxx**	**xxx**	**xxx**	**xxx**	**xxx**	**xxx**	**xxx**	**xxx**	**xxx**	**xxx**
Total referrals	**xxx**	**xxx**	**xxx**	**xxx**	**xxx**	**xxx**	**xxx**	**xxx**	**xxx**	**xxx**

Last, streamlining – less is more! Edward Tufte said "Above all else show the data." He defines a data-ink ratio – a ratio of the amount of ink used to portray data : the amount of ink used for non-data (borders, shading, etc). This data-ink ratio should be really high.

One of the first distinctions I'll make is knowing when to use a table vs. chart. Use a table if you need more precision, you have lots of values, the values have different units, or you're showing a detail & summary view. You want to remove as many of the borders as you can and only shade what you have to. Right-align numeric fields and remove unnecessary decimal points. One good way I've found increase the data-ink ratio is to remove all borders and shading (start with a "bare" table) and add back in just enough elements to make it readable – that way the elements in the table are purposeful.



For graphs, again you want to remove any unnecessary ink – the graph's frame, extra gridlines, sometimes data labels, etc. For bar graphs, pay attention to bar width vs. spacing. Try to put the data label inside the end of the bar, though sometimes it's not possible, as in the graph on the right.

For a line graph, try to directly label the lines. It looks cleaner and then your eye doesn't have to travel to a legend and back to the data. It's all together. Remove unnecessary tick marks and axis marks. Be thoughtful about now many values you need on the y-axis and whether or not you need labels on the graph. In this case, for this client, we chose to keep them.

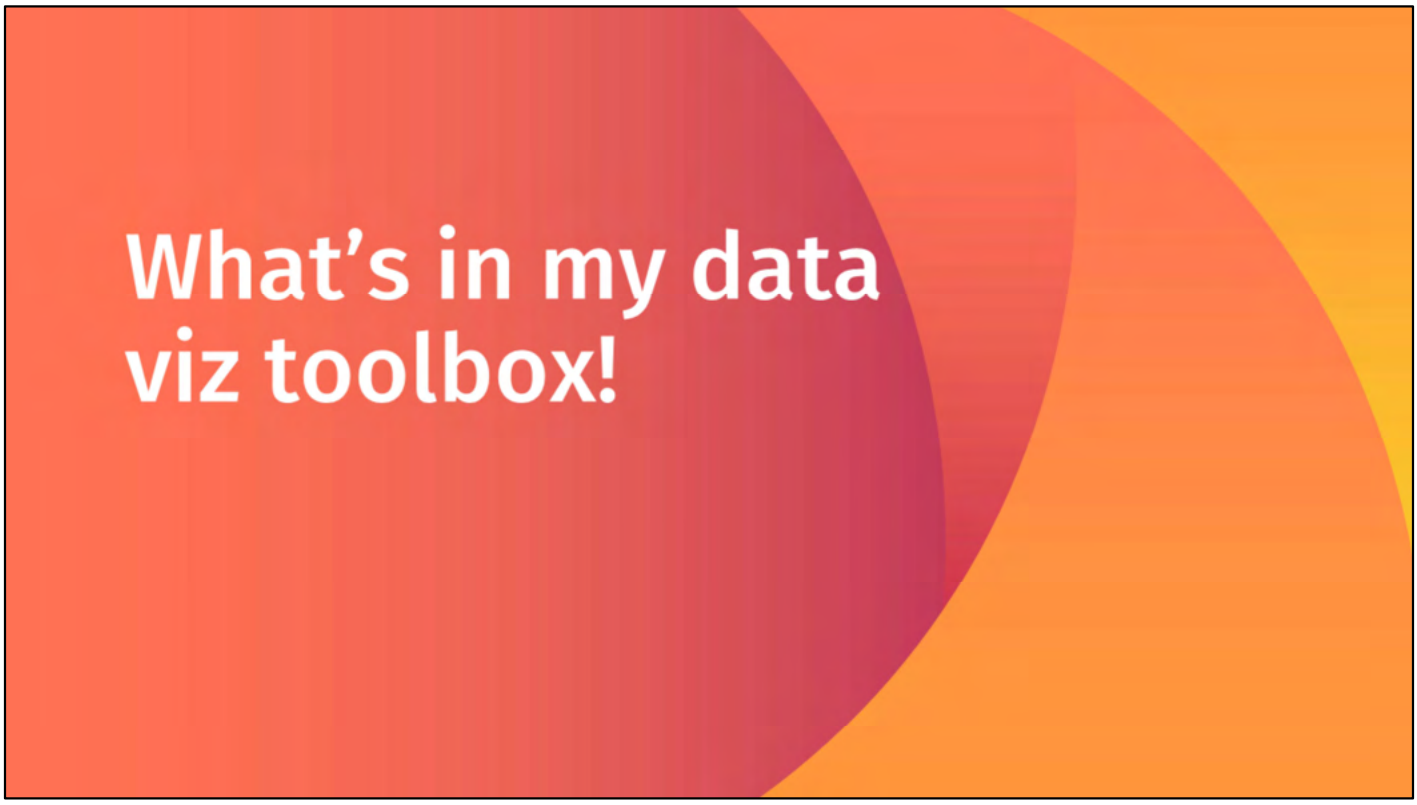
I think it's important to balance time invested with outcome/need. Before I make a report/results document I'm trying to gather:
What is the purpose of this dashboard / report / thing? Awareness / understanding / monitoring / Is it for monitoring purposes (needs to be more at-a-glance) vs. in-depth report / other reporting with detailed tables.

What type of content? Static or dynamic? Dynamic graphs or text or both? How often do they need it?

Who is this report for? Drives level of detail needed, how much text/explanation is incorporated.

- What will they do with it? Maybe they just want charts to monitor. Maybe they share with their leadership. Maybe they need to copy/paste/manipulate the data into another format.

These questions will help me understand if SAS is the right tool and how much effort to put into it.



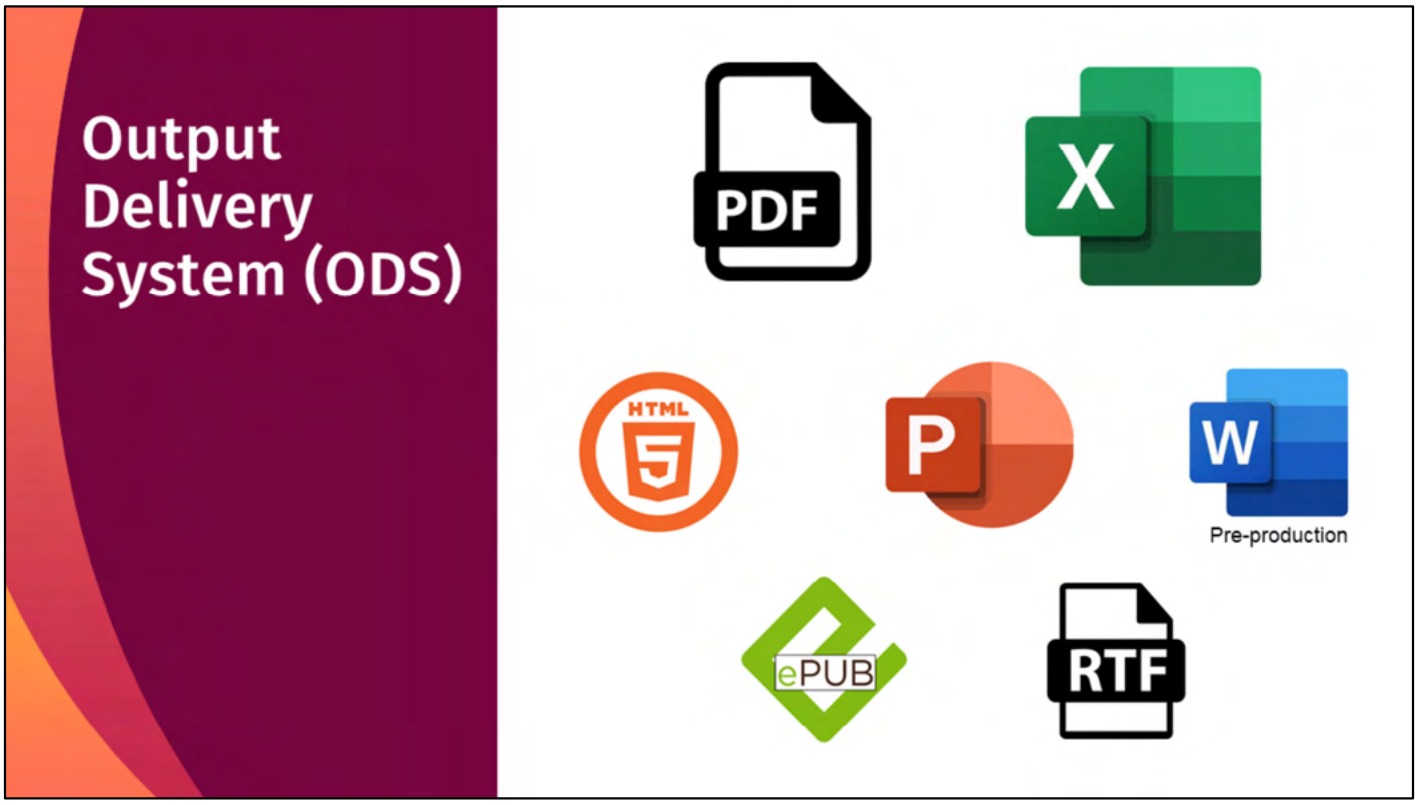
So how do I implement these strategies in SAS? Let's look through my toolbox.

Using the output delivery system, especially the PDF and Excel destinations…

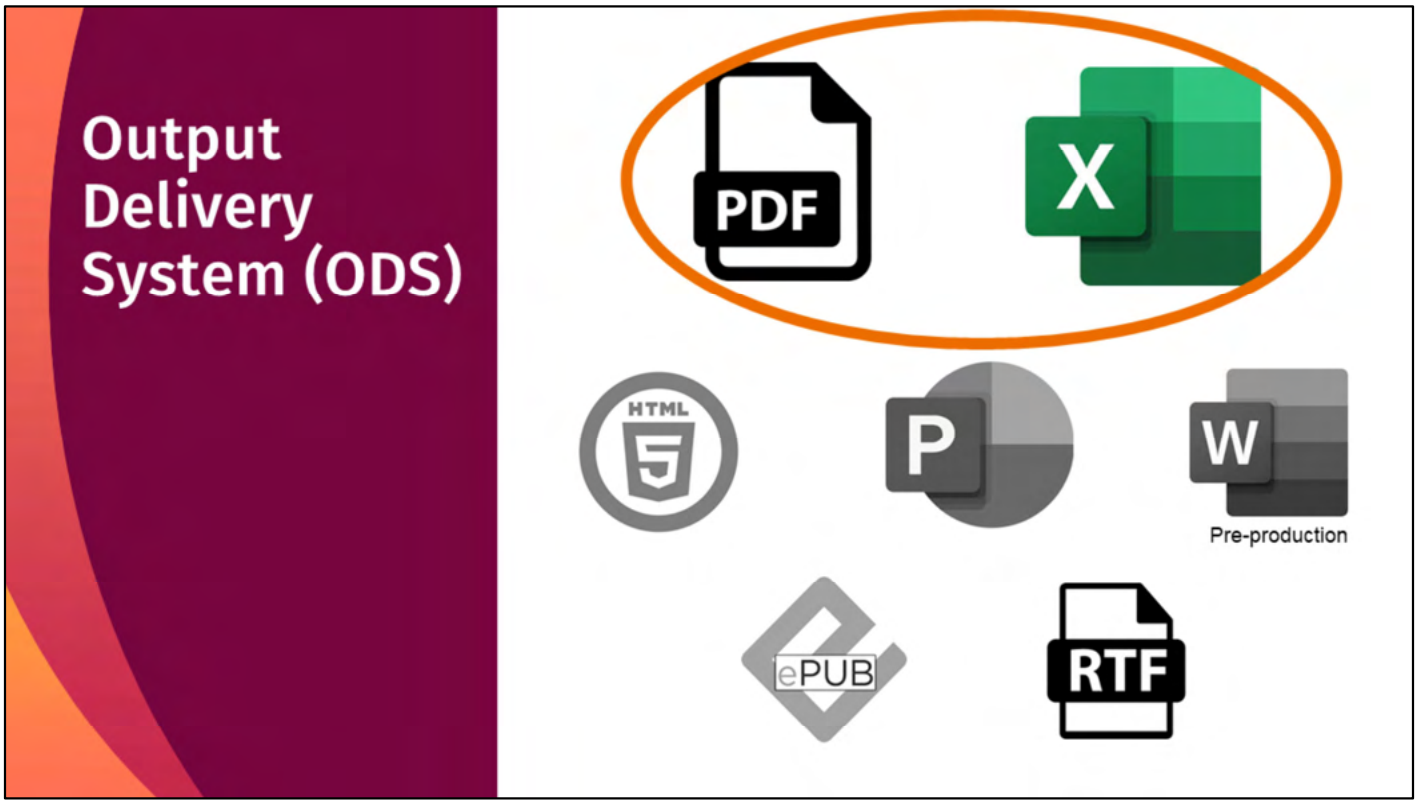
Using a few key procedures – for me, those are the REPORT procedure and the statistical graphing procedures…. And even some graph template language

And then to put the finishing touches on, utilizing user-defined formats, annotations, the ODS escape character and macro variables. There are obviously a lot more tools in SAS, but these are the ones in my toolkit.

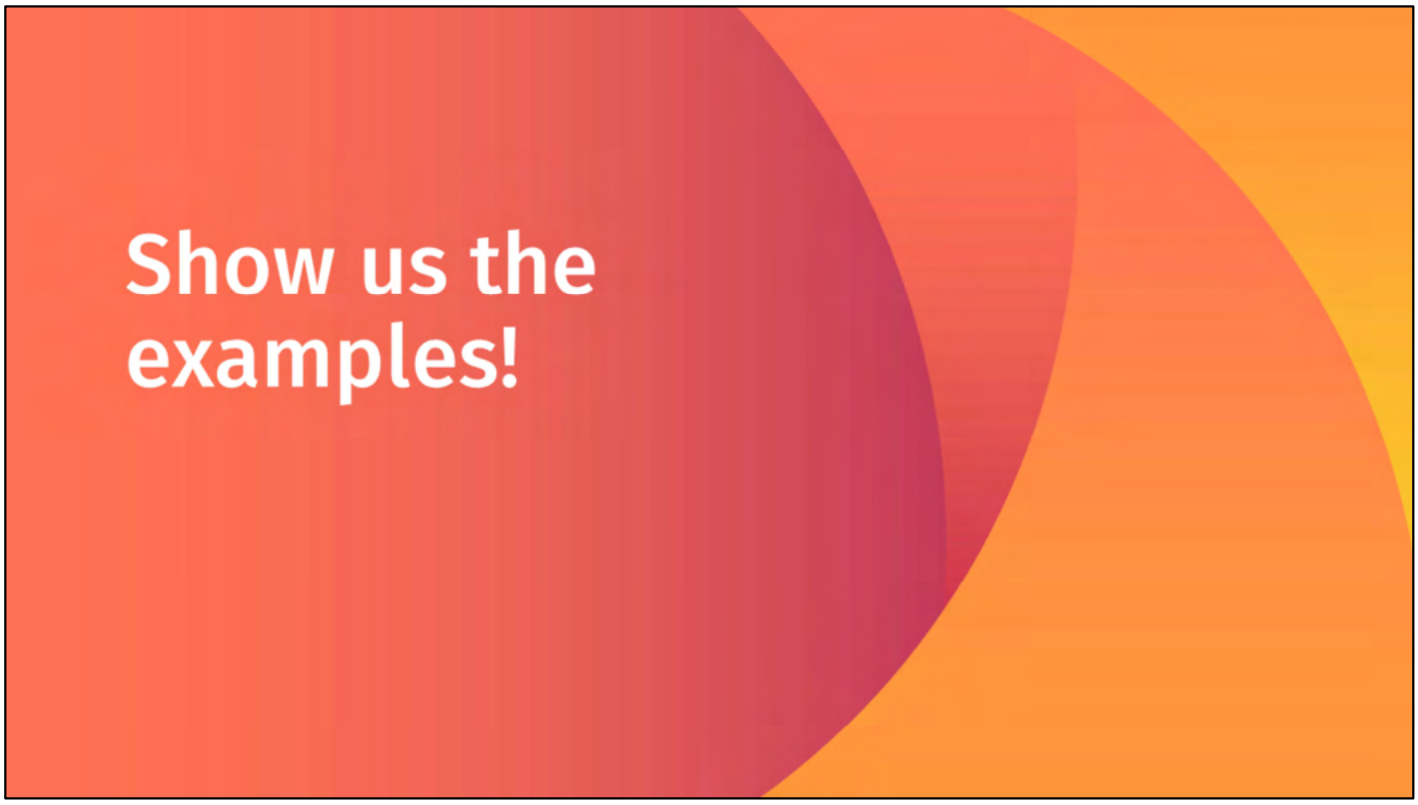
The tool I most often pull out is user-defined formats! For everything from PROC FREQ to graphing to automated reports, most everything gets a user-defined format. In fact, I have a separate talk on that.



One thing I tend to think about is the Output Delivery System – where is SAS sending the results?



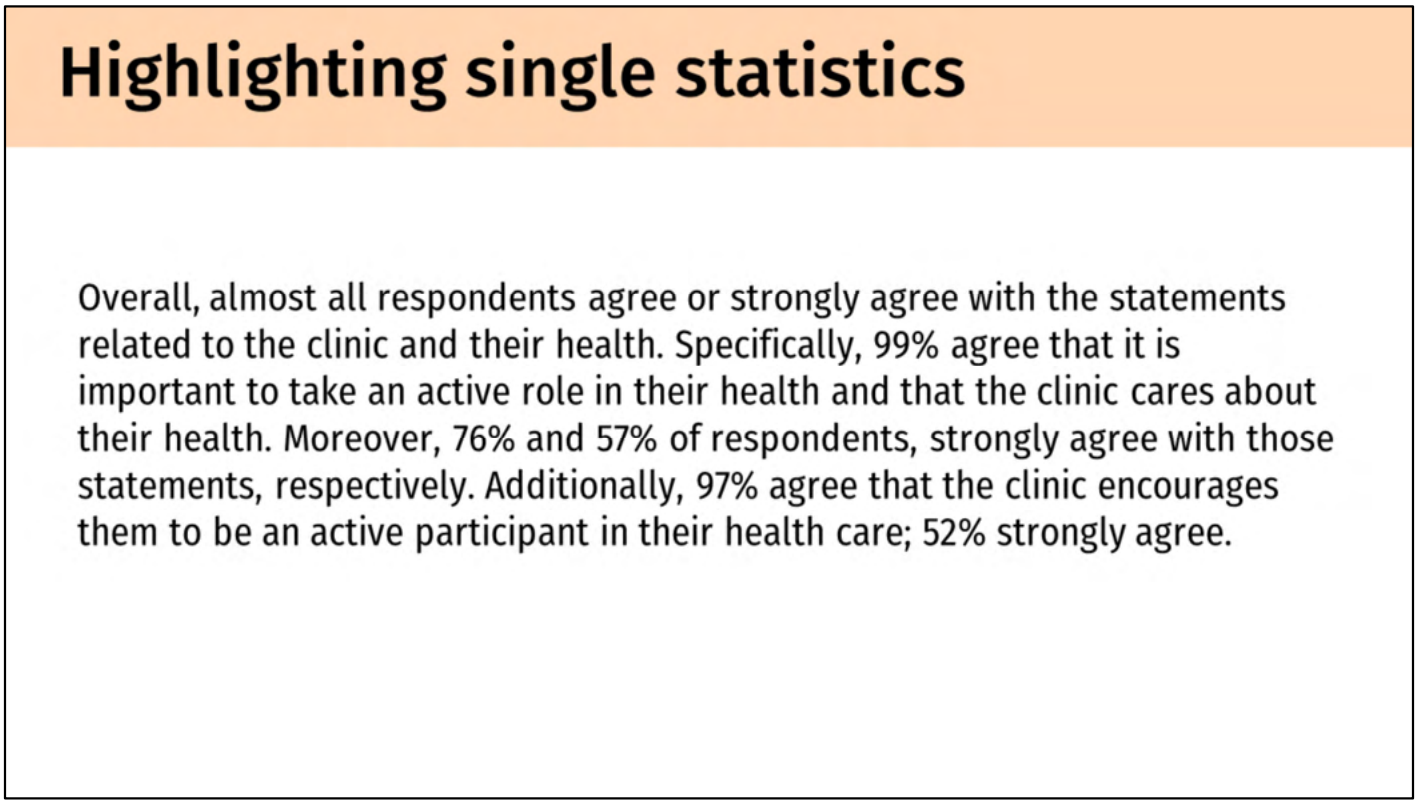
I tend to use Excel for 80% of my work and PDF for 20%.


Show us the
examples!

Health Club Survey

	Strongly Agree	Agree	Disagree	Strongly Disagree
1. My clinic cares about my health.	☐4	☐3	☐2	☐1
2. My clinic encourages me to be an active participant in my health care.	☐4	☐3	☐2	☐1
3. It is important for me to take an active role in my health and well-being.	☐4	☐3	☐2	☐1

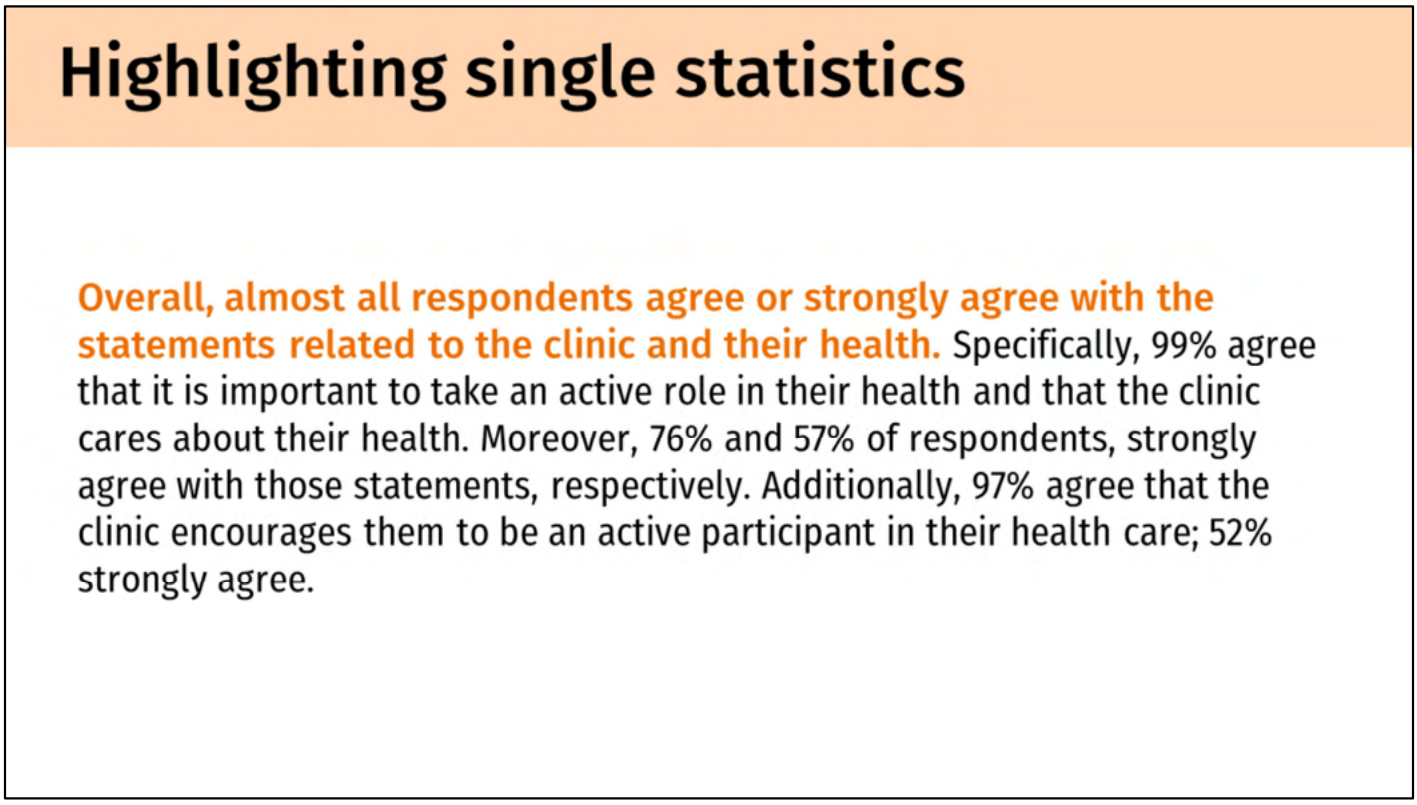
We'll start with the health club survey. A local health care clinic started running a health club within their clinic. It had typical gym facilities as well as wellness classes (i.e. cooking classes, stress reduction classes, etc). They wanted to assess perceptions of clinic and health, awareness of the health club, and barriers to using the health club. They designed this survey and administered it to patients . Once the data were collected, they came to me to analyze their data. This is a screenshot of part of their survey.



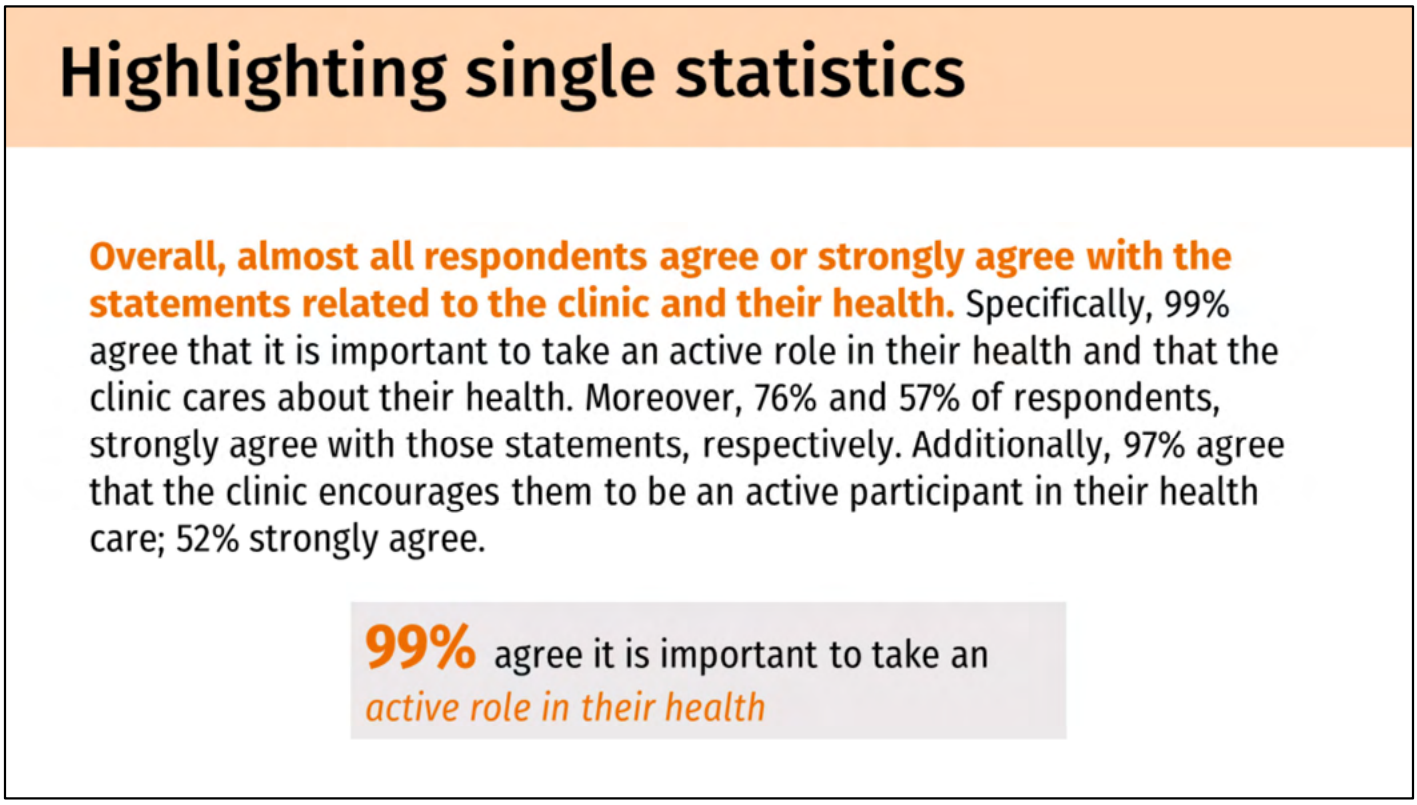
After running the data analysis, one of my first instincts when putting together a report for the client was to write out my findings. I want to be explicit about the methods I've used, what I've found, and explain the nuances of the data. Now, it's important to know that this block of text is dense and hard to read in the back of the room. The point I want to make is that it's a big block of text – nothing stands out. It doesn't really matter what it says. It's boring. And, our brains don't pay attention to boring things. It describes the percent of respondents that support or strongly support the three questions I just showed you.

So, my starting point looked something like this: {remove shaded box}

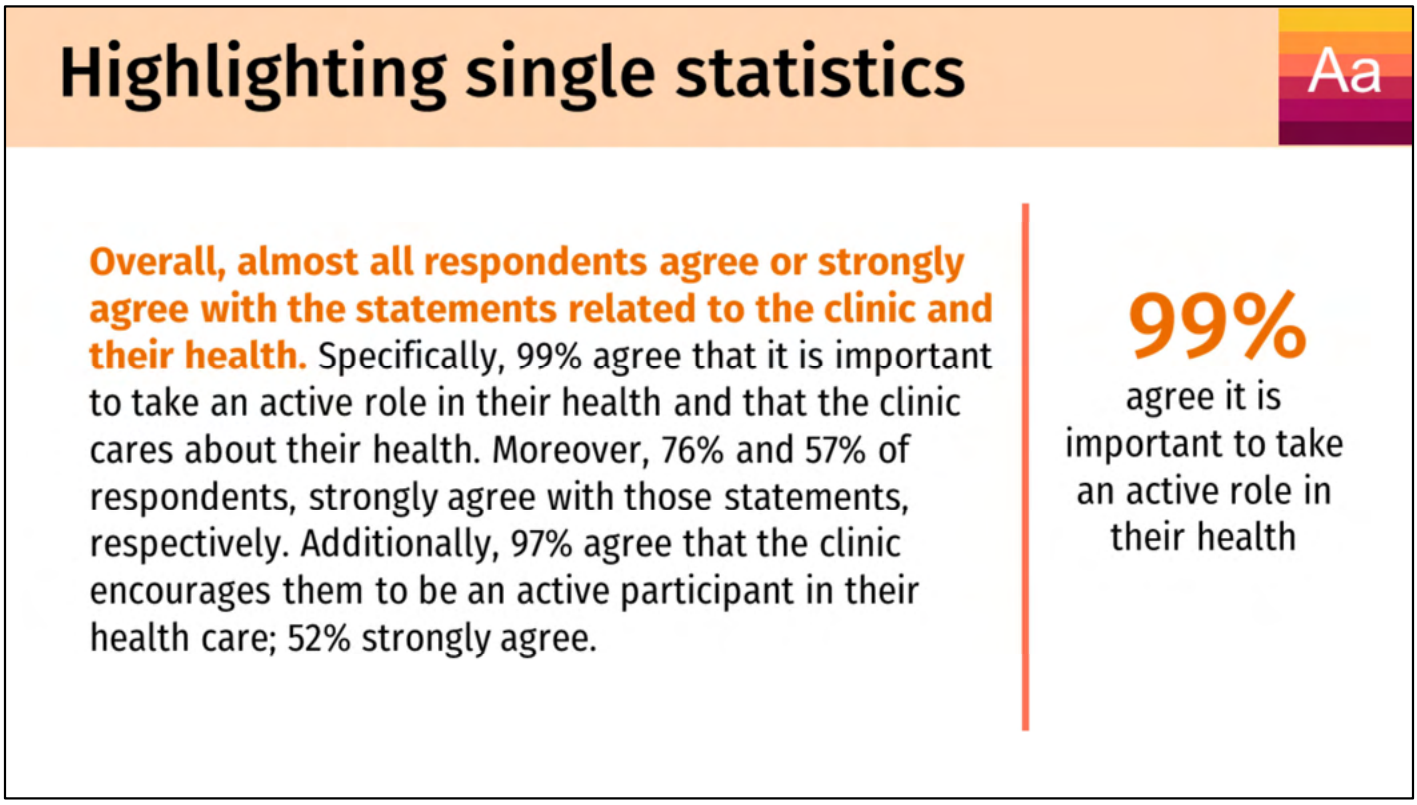
That is boring for anyone. My clients in this case are clinic administrators and health club staff. A big block of text is not the best way to go. As we move through this example, pay attention to what pieces draw your eye.



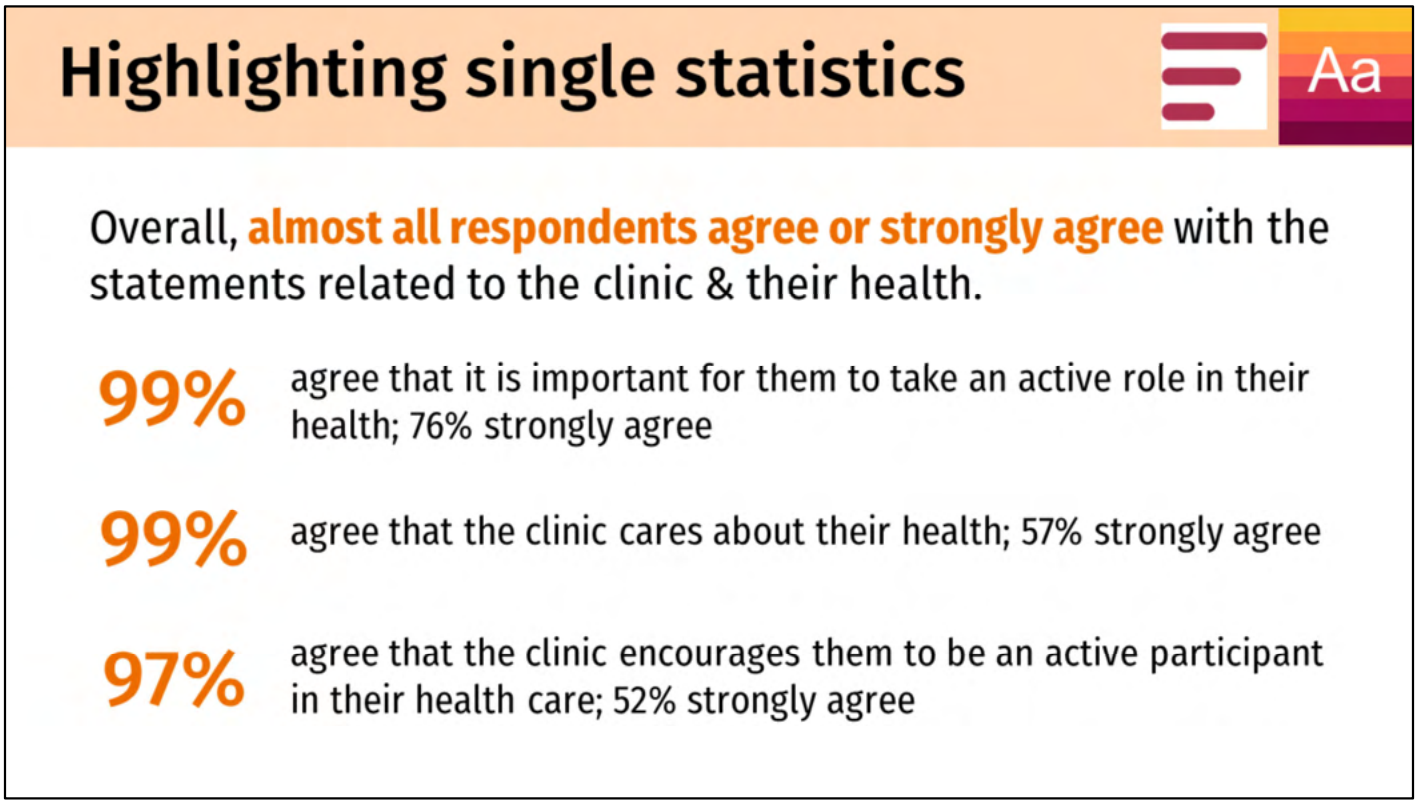
One easy way to help guide the reader is to use font color and styling to call attention to the summary sentence of the paragraph or to important phrases within the paragraph. Then, the client can skim the results and have a reference point if they need to go back to find something. While it's not ideal, at least I'm guiding their skimming as opposed to the reader finding their own nuggets while skimming. It's better, though there are still many findings within the paragraph.



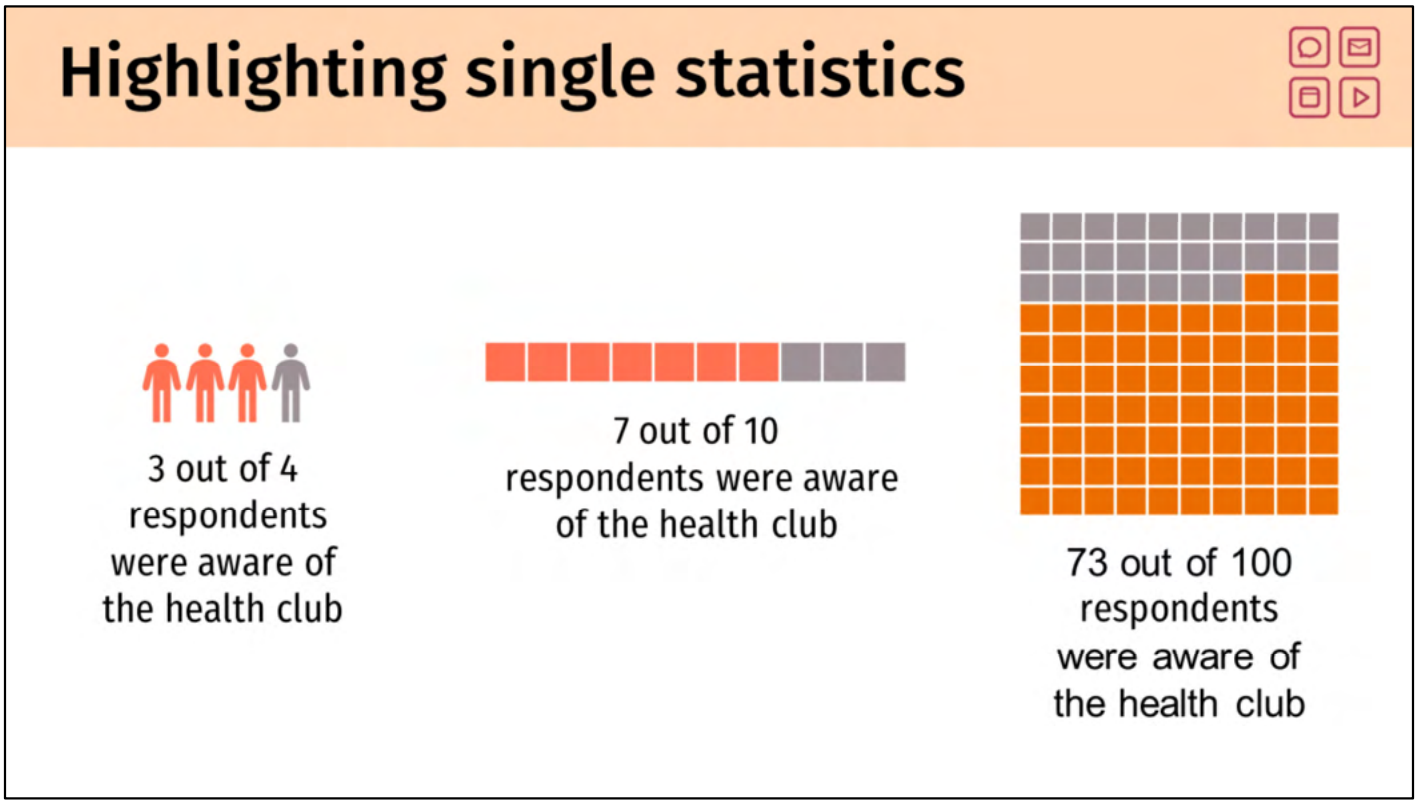
Another easy way to call attention to a single statistic is to use a call-out box. Pull out an important finding, a surprising finding, a finding that supports the hypothesis. Call-out boxes visually break up the page and can be a good way to highlight a single statistic.



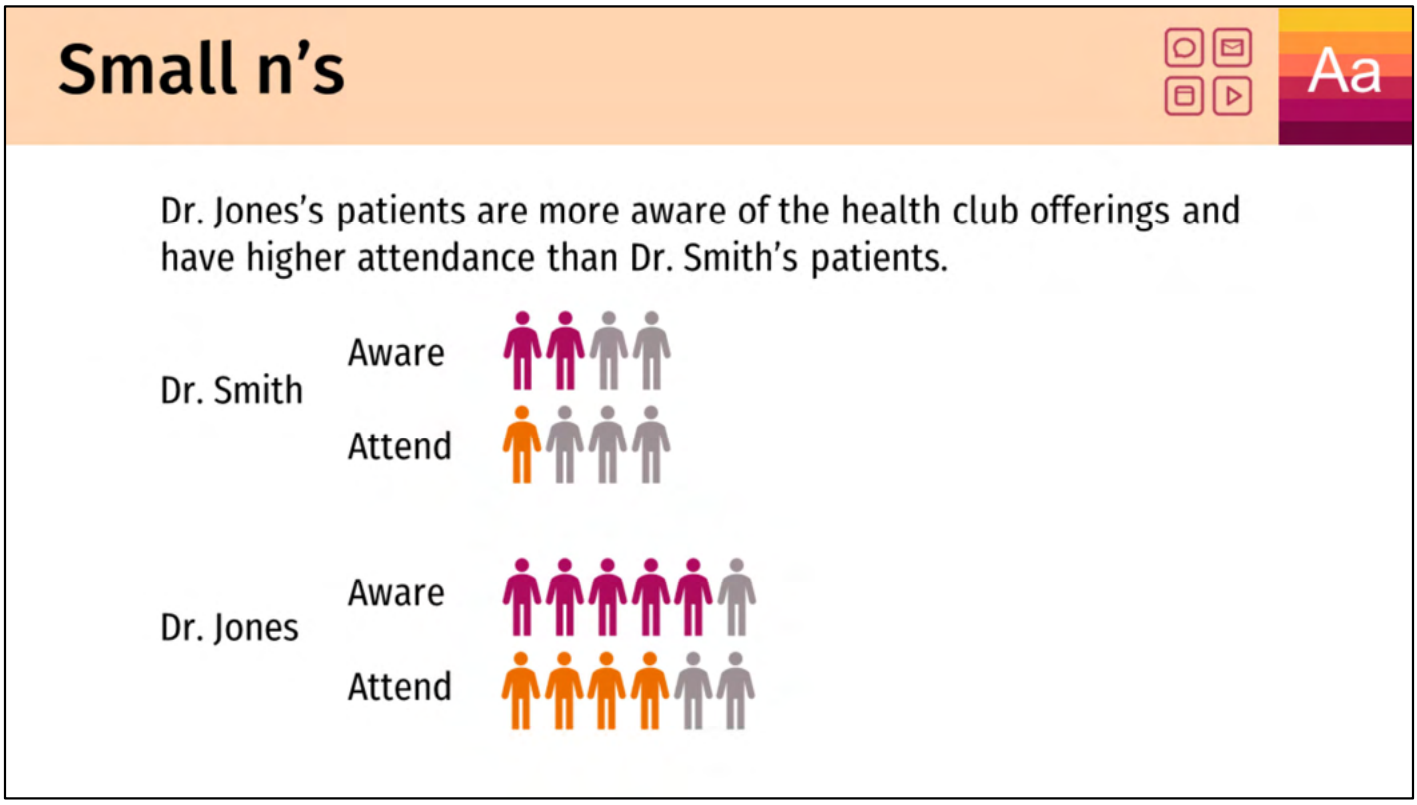
Another way to highlight a single statistic is to offset it – put the statistic itself in bigger, bolder font and then have the text block. It's another version of option instead of a call-out box. Depending on the layout, the statistic, etc. you may prefer one over the other. This was getting better for my client, but I still felt like it was too text heavy.



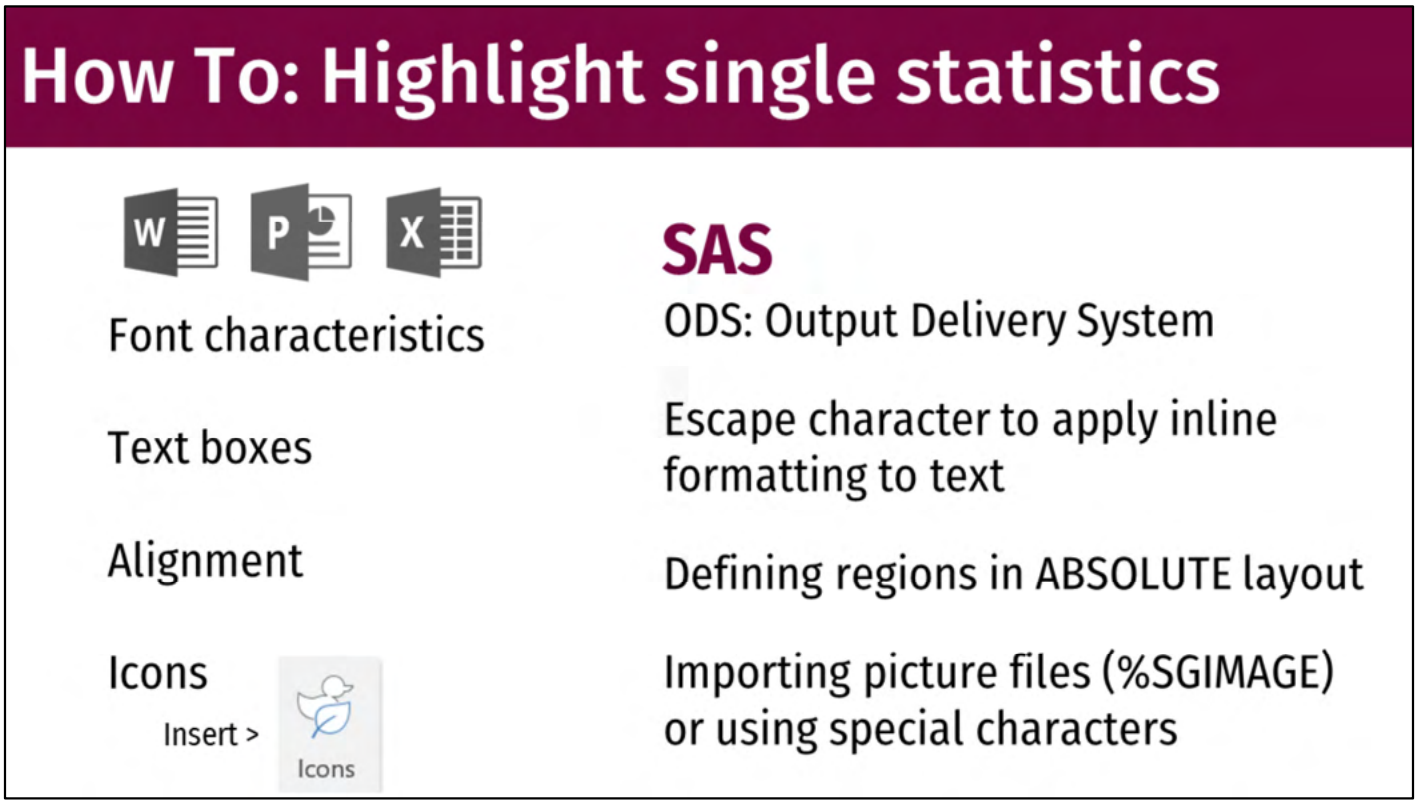
When I thought about how to make it more readable, I considered what types of information people tend to read – bulleted lists. So I took the idea from the last slide and applied it to the three statistics, making almost a bulleted list of the results, ordered by the percent that agree. Ultimately, this is what I used for the client. They thought it was great! In fact, I've used this a few times and it's always gone over well!



If you'd like to add more of a visual element, consider highlighting single statistics are with unit charts or waffle charts. These are a little controversial as far as data viz efficacy. You either need to count the units or read the text, so it doesn't save time or provide enhanced meaning. However, they are very popular in journalism and infographics. Personally, I don't mind them and clients seem to like them. They convey an overall sense of how many and provide an easy alternative to the call-out box or bulleted lists. However, you should not have half units, so you need to decide how precise to be – if the true percent is 73 – is 75% close enough? Is 70% close enough? Do you need the reader to know that it's exactly 73%?



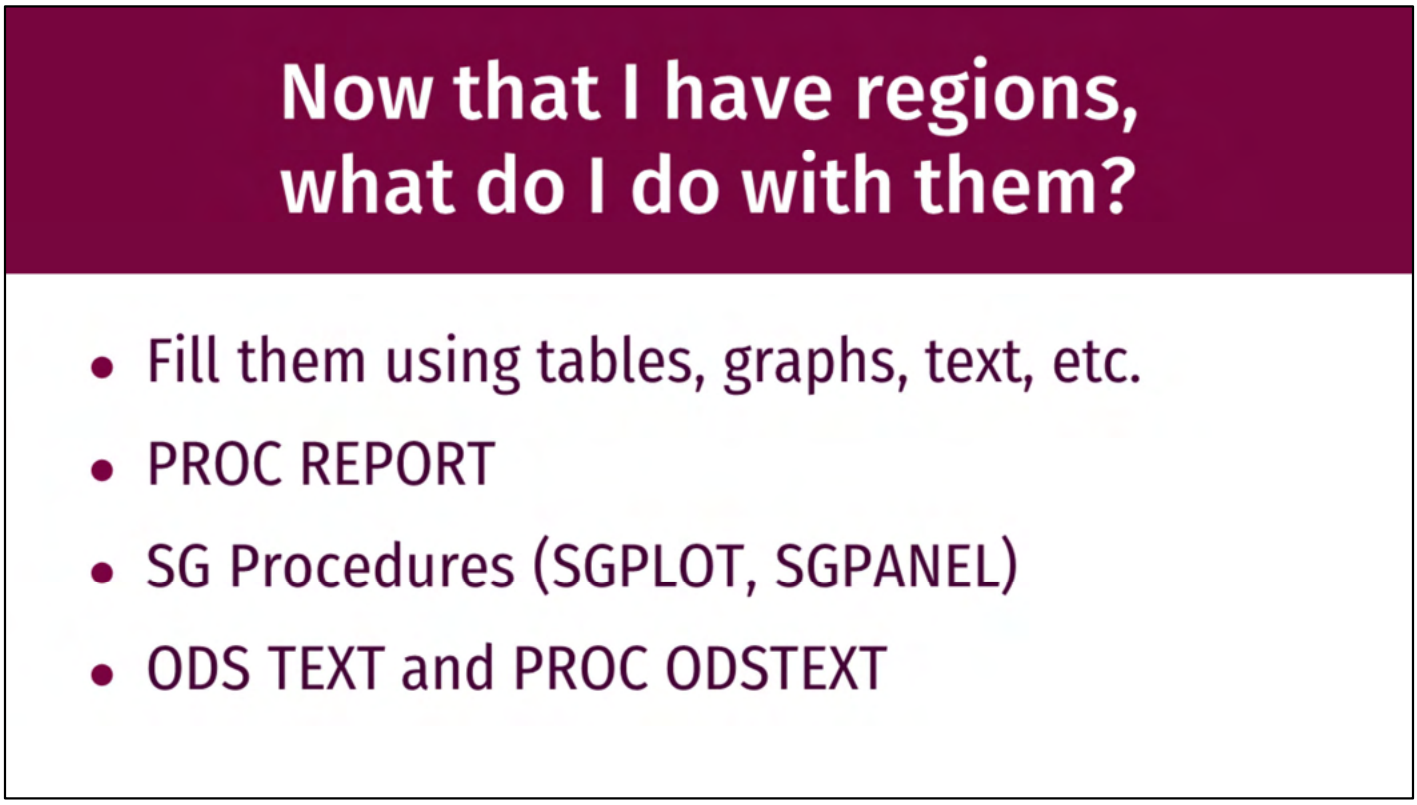
One example where I really do like these unit charts is for emphasizing results with small n's. This again depends on your audience – I'm wary because I've had investigators grab onto percentages from small n's and run with them, forgetting that it's from a really small sample of their patients, so it might not be representative. (Generally I wouldn't provide percentages with n's this low – but I think the example suits the purpose.) I've found that having the little icons there helps keep the underlying count in perspective. Again, it has worked for me, might not work for you.



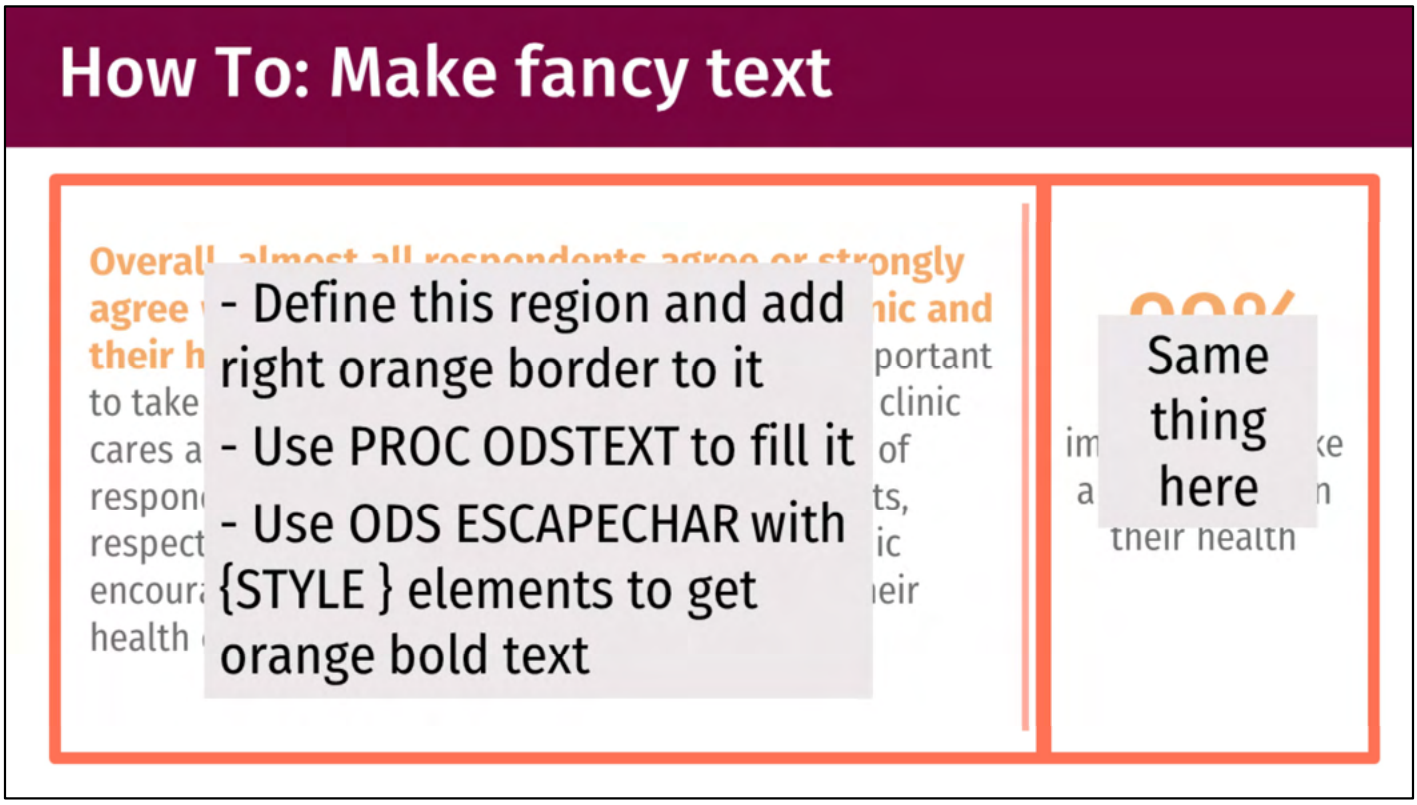
Let's pause – what have we really done so far to highlight single statistics? We've used some of the most basic functions of word processing software like font characteristics, text boxes, alignment, and inserting icons. That's easy!

While it takes a little more work in SAS, you can get the same results using the Output Delivery System (ODS), inline formatting, and importing pictures. If you haven't used the ODS functionality of SAS, I highly recommend looking into it! In fact, once you start using it, you'll have a hard time going back!

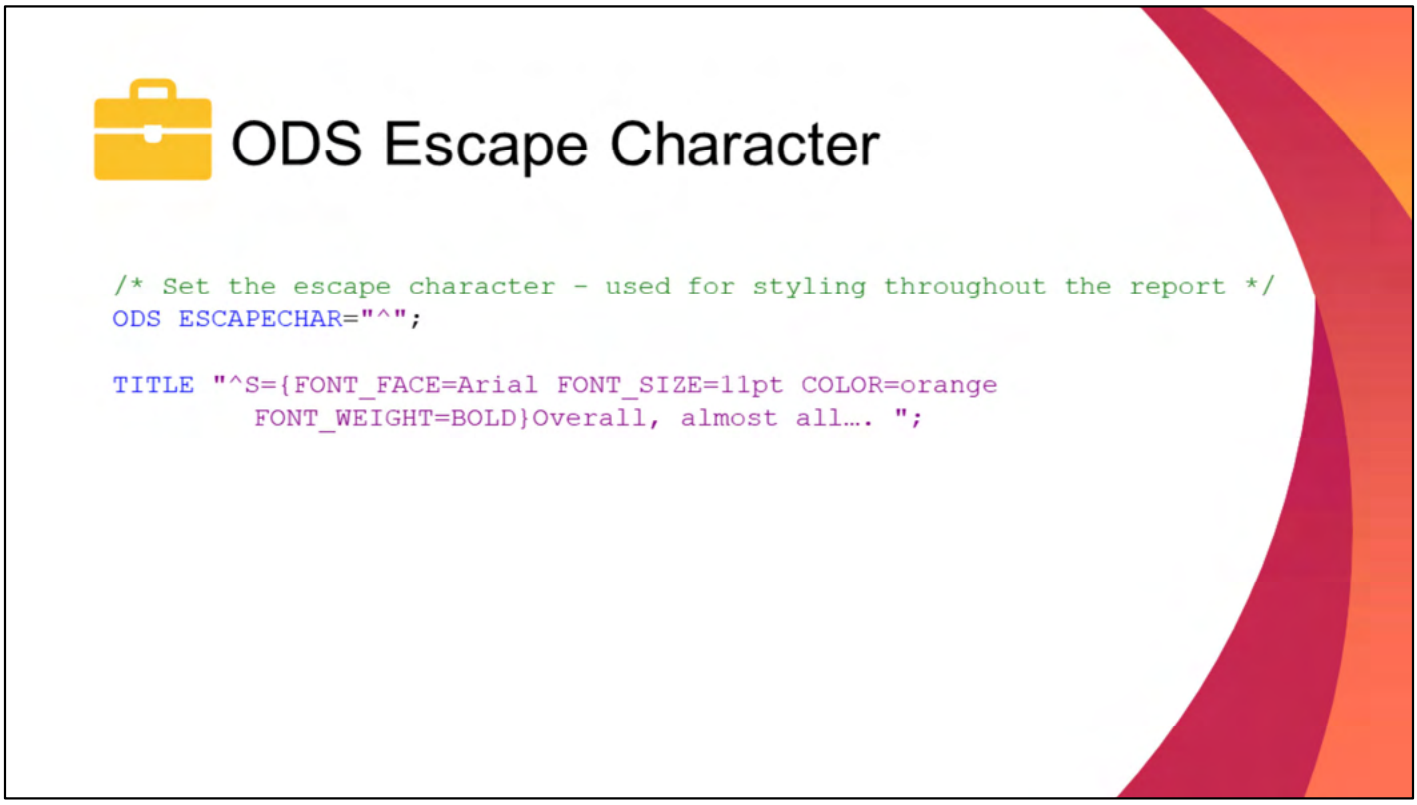
Deeper dive into the output delivery system: Generally, if I'm creating complex visualizations, I'll output to a PDF using the absolute LAYOUT. This layout allows me to create regions anywhere I wish on the page, of any size. For instance, I could create region in the top left to hold an itemized list, a region in the top right to hold a chart, a big region in the middle, and two other little regions on the bottom. I can even add littler regions over the top of big regions to get the effect I want. Think of it as layering post-it notes. Whatever you put on top will cover up whatever is underneath.



Once your region is established, you fill it with output from a procedure. That could include text, tables, graphs, or a combination of these items. The procedures listed here are my go-tos.



To make alterations to text/typography, you'll need to use the output delivery system. In this example, I might do it by defining a region on the left, making sure that the region has an orange right-hand border, then filling it with text using PROC ODSTEXT. I'll use the escape character to get the orange bold. Then define a region and do the same thing on the right.



I've mentioned the escape character a couple times – let's take a quick look. We use ODS ESCAPECHAR to set a special character. When SAS then sees this character it says, "Aha! I've been waiting for this character and I know you want me to do something special now." In this example, we're defining the caret symbol as the special character. It's my go-to, though you can define whatever you want. Another good option is the tilda (~). You just want a character that's not used otherwise. For instance, you could define 'a' as the escape character, but then every time SAS runs into the letter 'a' it's going to be looking for special instructions.

How To: Icon charts

3 out of 4 respondents were aware of the health club

One approach:

- PROC SGPLOT to make a blank graph (or white lines)
- %SGIMAGE to insert an orange or grey icon or PROC GSLIDE
- %SGTEXT to insert text

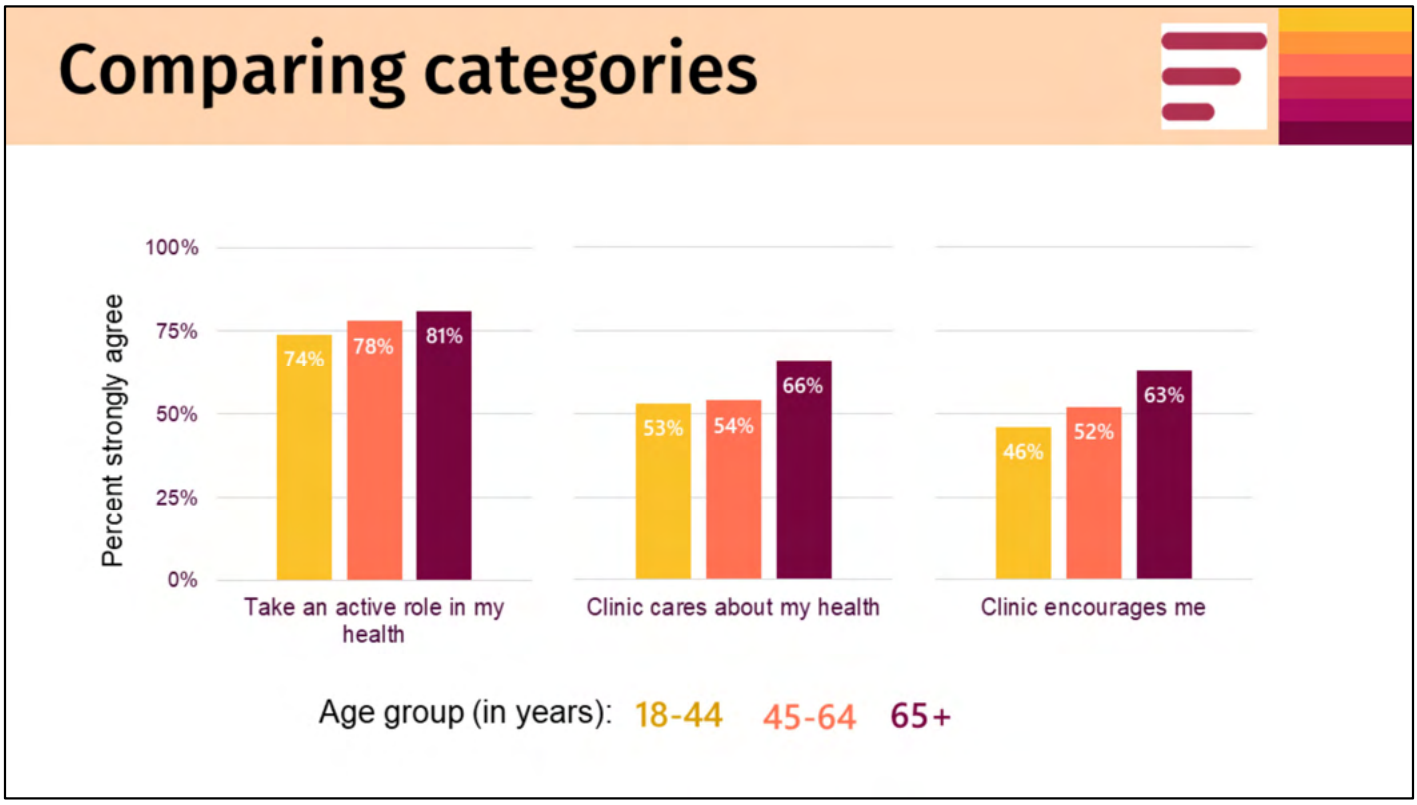
One other thing I like to do is create empty graphs and then use annotations to create the graphic I want.

One approach to adding icons is to make a blank graph – use dummy data if you need to and make a graph with no data to plot (or have the data outside the display range of the graph). Turn all the graph elements (axes, etc.) off or turn them white (if printing to a PDF). The use either PROC GSLIDE or %SGIMAGE to insert either orange or grey icons. Use a %SGTEXT to insert the text.

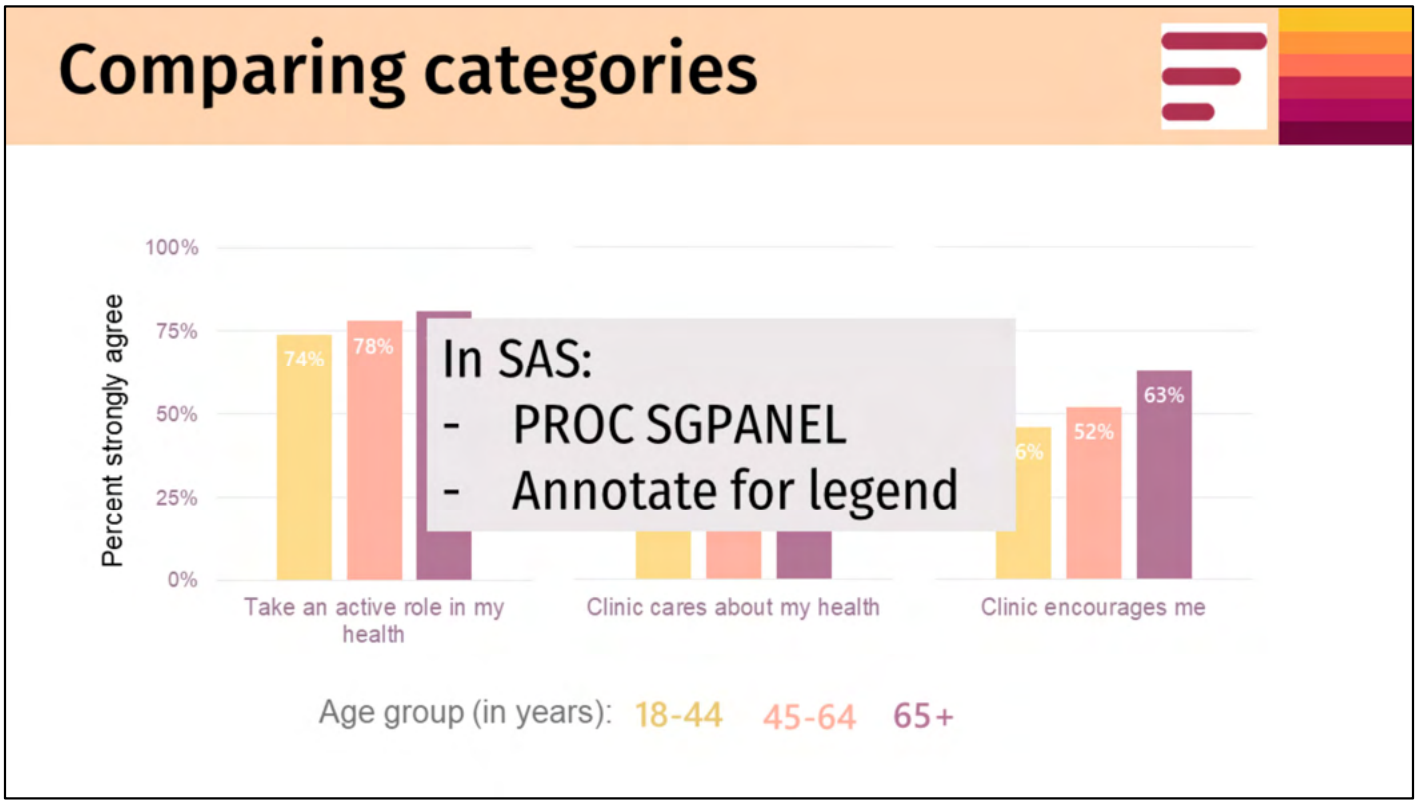
Comparing categories

Age category	Percent that strongly agree		
	Take active role in my health	Clinic cares about my health	Clinic encourages me
18-44 yrs	74%	53%	46%
45-64 yrs	78%	54%	52%
65+ yrs	81%	66%	63%

Going back to those three questions – the client wanted to know if there were differences in the strongly agree responses by age group. Looking at this table, we can see the older age group has higher agreement than low age groups, but you have to work a little bit to find it. Again, there's no obvious place for your eye to go.



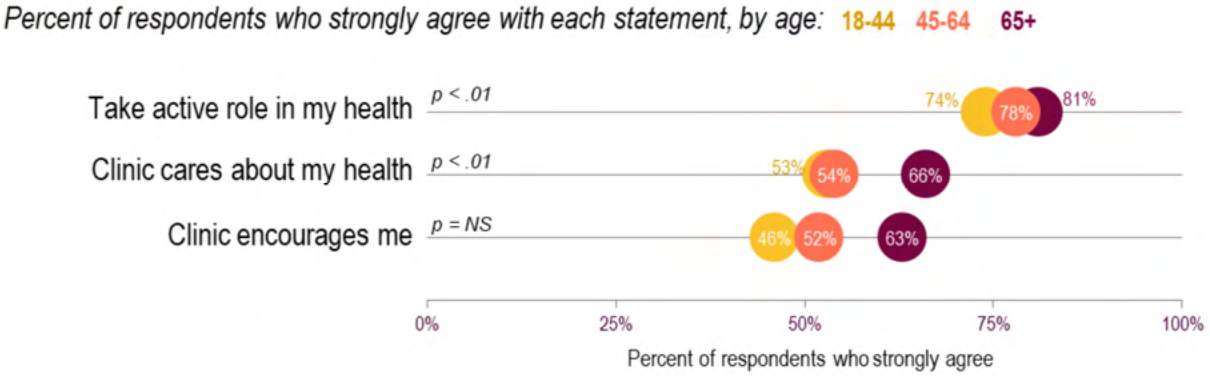
To more easily visualize the trends, we might use a small multiples bar chart. One chart for each question with colored bars for the age groups. Notice that we've removed the y-axis from the second and third charts because it's redundant (the graphs are all on the same scale) and having the axis would add extra clutter or non-data ink.



One way to do this in SAS is to use the SGPANEL procedure with VBAR options and annotations for the legend.

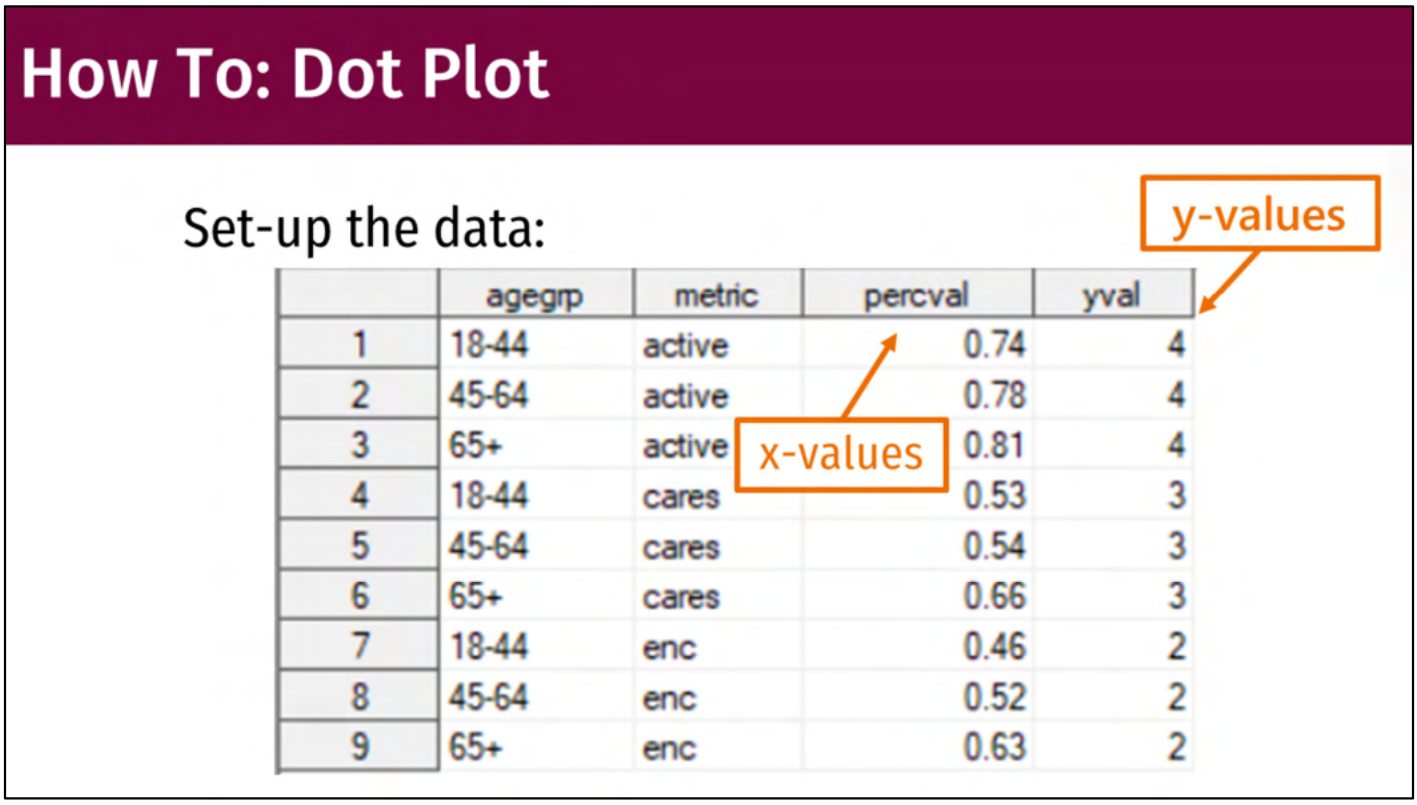
Comparing categories

Older respondents (**aged 65+ years**) more strongly agreed to with the clinic statements than those **18-44 years** or those **45-64 years**.



Another option to display this data is the dot plot. How many people have seen a dot plot? Typically, we see them with one value per line. In this case, 3 questions each with 3 age groups, I would need 9 lines, each with 1 dot. To save space and make the comparisons easier, I put all 3 age groups on the same line, identified by color. The difference in the older group comes out, the higher agreement with taking an active role in my health.

A note: This plot really works well for highlighting differences. Too many similar results and your dots all pile up, as the dots in the top line are starting to do. If all of the results are similar, consider using another type of chart (bar chart). Also, you can't use this graph effectively for more than 3 groups. More than 3 and there's too many dots to keep track of. It works really well for 2 groups; for instance, you can also use this to show demographic differences between your cohort and a reference group; or differences between responders and non-responders.



	agegrp	metric	percval	yval
1	18-44	active	0.74	4
2	45-64	active	0.78	4
3	65+	active	0.81	4
4	18-44	cares	0.53	3
5	45-64	cares	0.54	3
6	65+	cares	0.66	3
7	18-44	enc	0.46	2
8	45-64	enc	0.52	2
9	65+	enc	0.63	2

We'll do a quick run-down of how to create this dot plot in SAS:

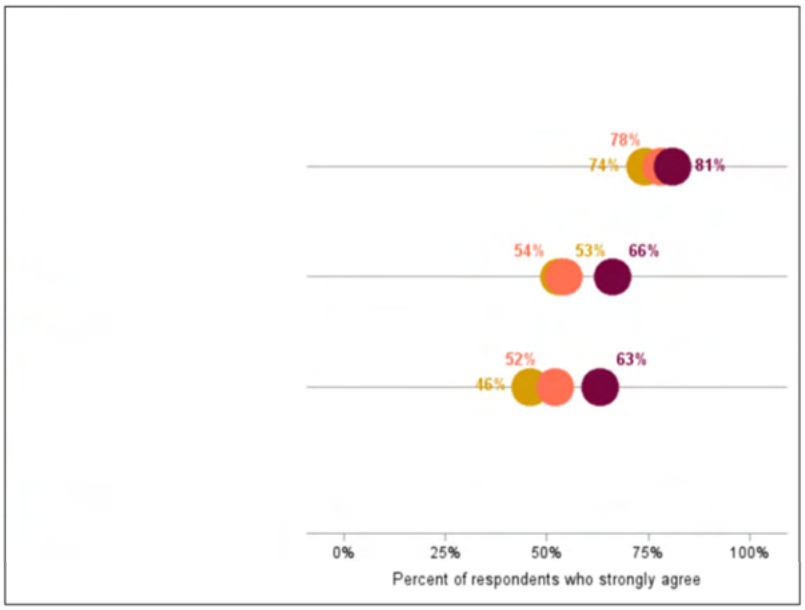
1) Set-up the data so that age group, the questions, and the values are in their own column in long format. The values will be the x-values. Then add dummy Y column – this is going to be used to force the spacing between the lines.

SAS code to make dataset:

```
DATA temp;
 INPUT agegrp $ metric $ percval yval;
 DATALINES;
 18-44    active     0.74       4
 45-64    active     0.78       4
 65+      active     0.81       4
 18-44    cares      0.53       3
 45-64    cares      0.54       3
 65+      cares      0.66       3
 18-44    enc        0.46       2
 45-64    enc        0.52       2
 65+      enc        0.63       2
 ;
RUN;
```

How To: Dot Plot

PROC SGPLOT
- Added padding on left for labels
- Defined colors
- SCATTER statement
- Increased marker size
- Added data labels
- Inserted REFLINEs
- Modified x and y axis



Then:
- Use PROC SGPLOT with SCATTER statement – GROUP by agegrp
- Define colors using STYLEATTRS statement
- Edit the marker size
- Add REFLINEs at y=2, y=3, and y=4
- Add data labels
- And format your axes, suppressing the y-axis

SAS code to generate the bones of the plot:

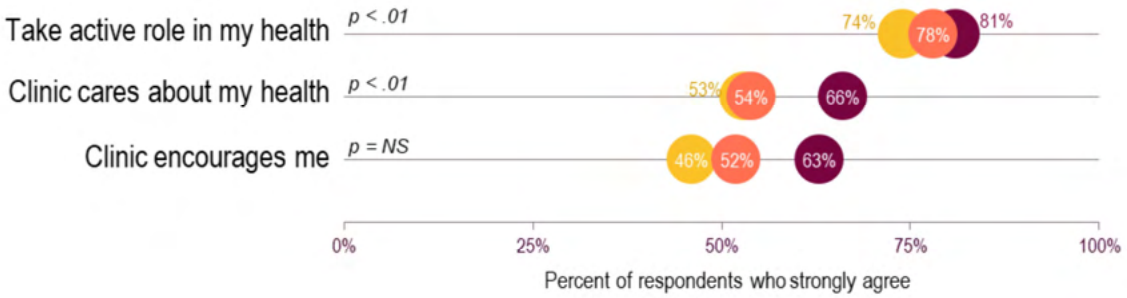
```
PROC SGPLOT DATA=temp NOBORDER NOWALL NOAUTOLEGEND
PAD=(LEFT=50pct);
          * Set the group colors;
          STYLEATTRS DATACONTRASTCOLORS=(CXD79D05 CXFF7154
CX77063F);
          * Add reference lines;
          REFLINE 2 3 4 / AXIS=Y LINEATTRS=(COLOR=CXa6a6a6);
          * Create initial plot;
          SCATTER X=percval Y=yval / GROUP=agegrp
MARKERATTRS=(SIZE=30 SYMBOL=CircleFilled)
                    DATALABEL=percval
DATALABELATTRS=(FAMILY="Arial" SIZE=10 WEIGHT=Bold);
          * Format values to XX%;
```

```
        FORMAT percval PERCENT7.0;
        * Set X- and Y-axis ranges and clean up labels,
including suppressing the Y-axis;
        XAXIS VALUES=(0 to 1 BY 0.25) LABEL="Percent of
respondents who strongly agree";
        YAXIS VALUES=(1 to 5 BY 1) DISPLAY=NONE;
RUN;
```

How To: Dot Plot

Older respondents (**aged 65+ years**) more strongly agreed to with the clinic statements than those **18-44 years** or those **45-64 years**.

Percent of respondents who strongly agree with each statement, by age: 18-44 45-64 65+



To finish, you need to add the text with annotations and adjust the dimensions of the graph.

When you want/need to keep the table....

	A	B	C	D	E	F	G	H
43	**Detail Report - Missing Data by Subgroup**							
44			CY20					
45	**Variable**	**Subgroup**	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**
46		Group 1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
47		Group 2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
48	Variable 1	Group 3	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
49		Group 1	4.0%	4.2%	5.1%	0.7%	3.9%	1.6%
50		Group 2	0.1%	0.5%	0.0%	0.0%	0.3%	0.0%
51	Variable 2	Group 3	0.0%	0.0%	3.8%	11.8%	0.0%	6.7%
52		Group 1	4.0%	4.9%	5.1%	0.7%	3.9%	0.8%
53		Group 2	0.1%	0.5%	0.2%	0.0%	0.3%	0.0%
54	Variable 3	Group 3	0.0%	0.0%	3.8%	11.8%	0.0%	6.7%
55		Group 1	13.7%	11.1%	12.3%	14.7%	11.7%	12.8%
56		Group 2	6.1%	10.0%	5.1%	6.7%	6.8%	6.5%
57	Variable 4	Group 3	6.1%	6.3%	7.7%	11.8%	0.0%	6.7%

Let's get back to examples. There are times when you want or need to keep data tables. If the table is dense, try to call out values that are meaningful or values that are above/below a threshold. In this case, we're monitoring monthly data dumps and the amount of missing data. Variables with >5% missing get light orange, variables with >10% get a darker orange. This way, it's a little clearer where the missing data issues are and if they're expected or not. For instance, variable 4 always has a lot of missing data, but variables 2 and 3 do not, so missing there is more worrisome.

```
PROC REPORT DATA = MissSummaryv3 … SPANROWS
    …
    STYLE(header)=[BACKGROUND=&Primary1. FOREGROUND=white];

    COLUMN Variable Subgroup &ColumnText;
    DEFINE Variable / ORDER ORDER=internal "Variable";
    DEFINE Subgroup / ORDER ORDER=internal "Subgroup";
    DEFINE &YrMo1 /DISPLAY "&MoLabel1" RIGHT FORMAT=PERCENT8.1
        STYLE={BACKGROUND=Color.} STYLE(HEADER)=[BACKGROUND=&Primary1.];
    DEFINE &YrMo2 /DISPLAY "&MoLabel2" RIGHT FORMAT=PERCENT8.1
        STYLE={BACKGROUND=Color.} STYLE(HEADER)=[BACKGROUND=&Primary1.];
    …
RUN;
```

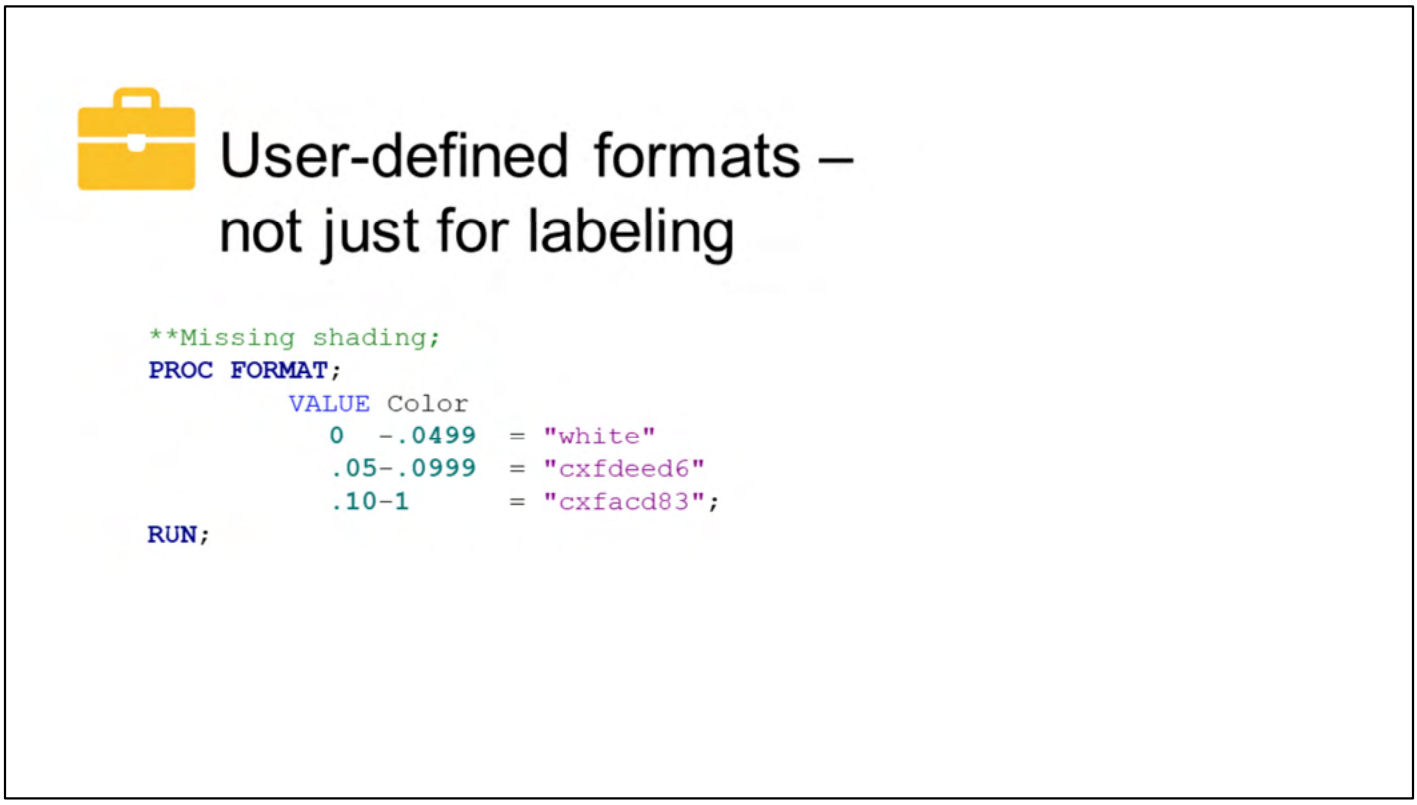
Let's take a quick stroll through the code to see how this was accomplished. This table was generated using PROC REPORT – it's one of my favorite table-producing procedures. In fact, it's mu second favorite procedure. (The first is PROC FREQ!) The SPANROWS option was used to merge the rows in the 'Variable' column. (Terrible name to use in an example, I know.)

```
PROC REPORT DATA = MissSummaryv3 … SPANROWS
    …
    STYLE(header)=[BACKGROUND=&Primary1. FOREGROUND=white];

    COLUMN Variable Subgroup &ColumnText;
    DEFINE Variable / ORDER ORDER=internal "Variable";
    DEFINE Subgroup / ORDER ORDER=internal "Subgroup";
    DEFINE &YrMo1 /DISPLAY "&MoLabel1" RIGHT FORMAT=PERCENT8.1
        STYLE={BACKGROUND=Color.} STYLE(HEADER)=[BACKGROUND=&Primary1.];
    DEFINE &YrMo2 /DISPLAY "&MoLabel2" RIGHT FORMAT=PERCENT8.1
        STYLE={BACKGROUND=Color.} STYLE(HEADER)=[BACKGROUND=&Primary1.];
    …
RUN;
```

Then we add styling to the header rows – specifically, filling the background of the cells with navy blue and using the FOREGROUND option to add white text. Note that the navy blue hex code is stored in the macro variable “&Primary1.” I like to use macro variables to hold color codes in case my client changes their mind on the color palette – then I only have to change the color code one place and re-run the code instead of changing it a dozen places and re-running the code.

```
PROC REPORT DATA = MissSummaryv3 … SPANROWS
    …
    STYLE(header)=[BACKGROUND=&Primary1. FOREGROUND=white];

    COLUMN Variable Subgroup &ColumnText;
    DEFINE Variable / ORDER ORDER=internal "Variable";
    DEFINE Subgroup / ORDER ORDER=internal "Subgroup";
    DEFINE &YrMo1 /DISPLAY "&MoLabel1" RIGHT FORMAT=PERCENT8.1
        STYLE={BACKGROUND=Color.} STYLE(HEADER)=[BACKGROUND=&Primary1.];
    DEFINE &YrMo2 /DISPLAY "&MoLabel2" RIGHT FORMAT=PERCENT8.1
        STYLE={BACKGROUND=Color.} STYLE(HEADER)=[BACKGROUND=&Primary1.];
    …
RUN;
```

Then to add the dynamic highlighting for the missing data, I add another style element. This time, it's added to the column within the DEFINE statement and it's done using a user-defined format called 'Color.' (Again, terrible name…. I know, I know…) This is one of my favorite uses of user-defined formats! Let's take a closer look.



Here is how we define the format – if the missing value is less than 5, then we set the format to white (with quotes because that's how the style needs it formatted), 5-10% is light yellow, and 10 or higher is darker yellow.

```
PROC REPORT DATA = MissSummaryv3 … SPANROWS
    …
    STYLE(header)=[BACKGROUND=&Primary1. FOREGROUND=white];

    COLUMN Variable Subgroup &ColumnText;
    DEFINE Variable / ORDER ORDER=internal "Variable";
    DEFINE Subgroup / ORDER ORDER=internal "Subgroup";
    DEFINE &YrMo1 /DISPLAY "&MoLabel1" RIGHT FORMAT=PERCENT8.1
        STYLE={BACKGROUND=Color.} STYLE(HEADER)=[BACKGROUND=&Primary1.];
    DEFINE &YrMo2 /DISPLAY "&MoLabel2" RIGHT FORMAT=PERCENT8.1
        STYLE={BACKGROUND=Color.} STYLE(HEADER)=[BACKGROUND=&Primary1.];
    …
RUN;
```

You can see that this color format is applied to the monthly columns of data in the DEFINE statements. How easy was that?! One user-defined format and we've got dynamic highlighting in our table. So easy!

	A	B	C	D	E	F	G	H
43	**Detail Report - Missing Data by Subgroup**							
44			CY20					
45	**Variable**	**Subgroup**	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**
46		Group 1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
47		Group 2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
48	Variable 1	Group 3	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
49		Group 1	4.0%	4.2%	5.1%	0.7%	3.9%	1.6%
50		Group 2	0.1%	0.5%	0.0%	0.0%	0.3%	0.0%
51	Variable 2	Group 3	0.0%	0.0%	3.8%	11.8%	0.0%	6.7%
52		Group 1	4.0%	4.9%	5.1%	0.7%	3.9%	0.8%
53		Group 2	0.1%	0.5%	0.2%	0.0%	0.3%	0.0%
54	Variable 3	Group 3	0.0%	0.0%	3.8%	11.8%	0.0%	6.7%
55		Group 1	13.7%	11.1%	12.3%	14.7%	11.7%	12.8%
56		Group 2	6.1%	10.0%	5.1%	6.7%	6.8%	6.5%
57	Variable 4	Group 3	6.1%	6.3%	7.7%	11.8%	0.0%	6.7%

Pretty cool, huh?

Combining tables and graphs

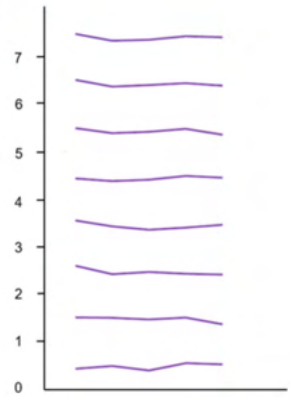
DMA	Program	FY17 Q4	FY18 Q1	FY18 Q2	FY18 Q3	FY18 Q4
Western area	AAAA	1,167	817	868	1,050	998
	BBBB	1,262	896	983	1,084	950
	CCCC	603	467	505	585	425
	DDDD	1,845	1,597	1,728	2,097	1,925
Eastern area	AAAA	157	119	97	111	129
	BBBB	237	158	179	162	154
	CCCC	91	89	82	90	61
	DDDD	135	157	121	180	170

Another thing I like to do with data visualizations is combine data tables and charts. Here we're going to embed spark lines in a table. The context of the data is not important – we're focusing on the data viz. The end goal is to have a bunch of small line charts within cells of a table. To create this, I envision 2 elements – the table element and the graph element. The elements overlap, so I'm thinking about sending them to PDF destination with ABSOLUTE layout.

Extend PROC REPORT table with an empty column

DMA	Program	FY17 Q4	FY18 Q1	FY18 Q2	FY18 Q3	FY18 Q4
Western area	AAAA	1,167	817	868	1,050	998
	BBBB	1,262	896	983	1,084	950
	CCCC	603	467	505	585	425
	DDDD	1,845	1,597	1,728	2,097	1,925
Eastern area	AAAA	157	119	97	111	129
	BBBB	237	158	179	162	154
	CCCC	91	89	82	90	61
	DDDD	135	157	121	180	170

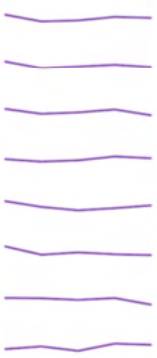
Build "stacked" line chart



We start by creating a table using PROC REPORT. We use some data viz best practices like right-aligning all of the numbers and printing only the borders that we need to. We can extend the PROC REPORT table to include an empty column that will hold the mini line charts. Then, we create a stacked line chart using the same dimensions as the table.

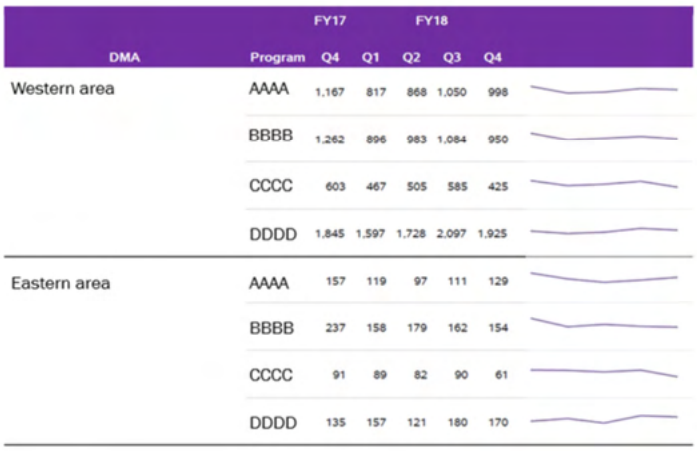
Overlay REGIONs within ODS PDF destination

		FY17		FY18		
DMA	Program	Q4	Q1	Q2	Q3	Q4
Western area	AAAA	1,167	817	868	1,050	998
	BBBB	1,262	896	983	1,084	950
	CCCC	603	467	505	585	425
	DDDD	1,845	1,597	1,728	2,097	1,925
Eastern area	AAAA	157	119	97	111	129
	BBBB	237	158	179	162	154
	CCCC	91	89	82	90	61
	DDDD	135	157	121	180	170



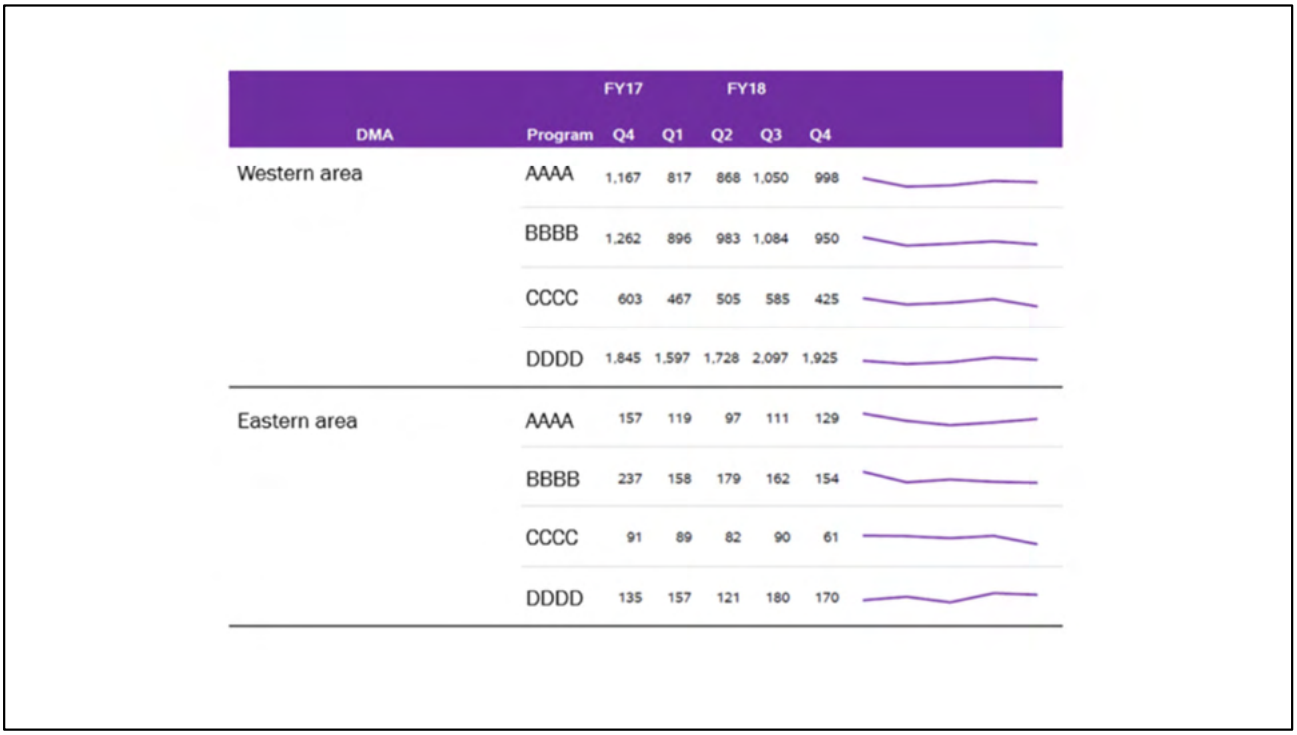
To get the line charts into the table, we are going to send the output to PDF, using ODS PDF with ABSOLUTE layout.

Overlay REGIONs within ODS PDF destination



		FY17		FY18		
DMA	Program	Q4	Q1	Q2	Q3	Q4
Western area	AAAA	1,167	817	868	1,050	998
	BBBB	1,262	896	983	1,084	950
	CCCC	603	467	505	585	425
	DDDD	1,845	1,597	1,728	2,097	1,925
Eastern area	AAAA	157	119	97	111	129
	BBBB	237	158	179	162	154
	CCCC	91	89	82	90	61
	DDDD	135	157	121	180	170

In this case, it lets us create overlapping REGIONS, so we can put our line chart over the top of the table.

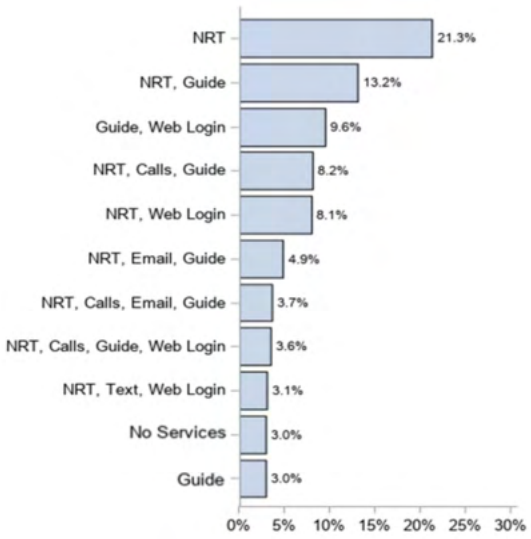


		FY17		FY18		
DMA	Program	Q4	Q1	Q2	Q3	Q4
Western area	AAAA	1,167	817	868	1,050	998
	BBBB	1,262	896	983	1,084	950
	CCCC	603	467	505	585	425
	DDDD	1,845	1,597	1,728	2,097	1,925
Eastern area	AAAA	157	119	97	111	129
	BBBB	237	158	179	162	154
	CCCC	91	89	82	90	61
	DDDD	135	157	121	180	170

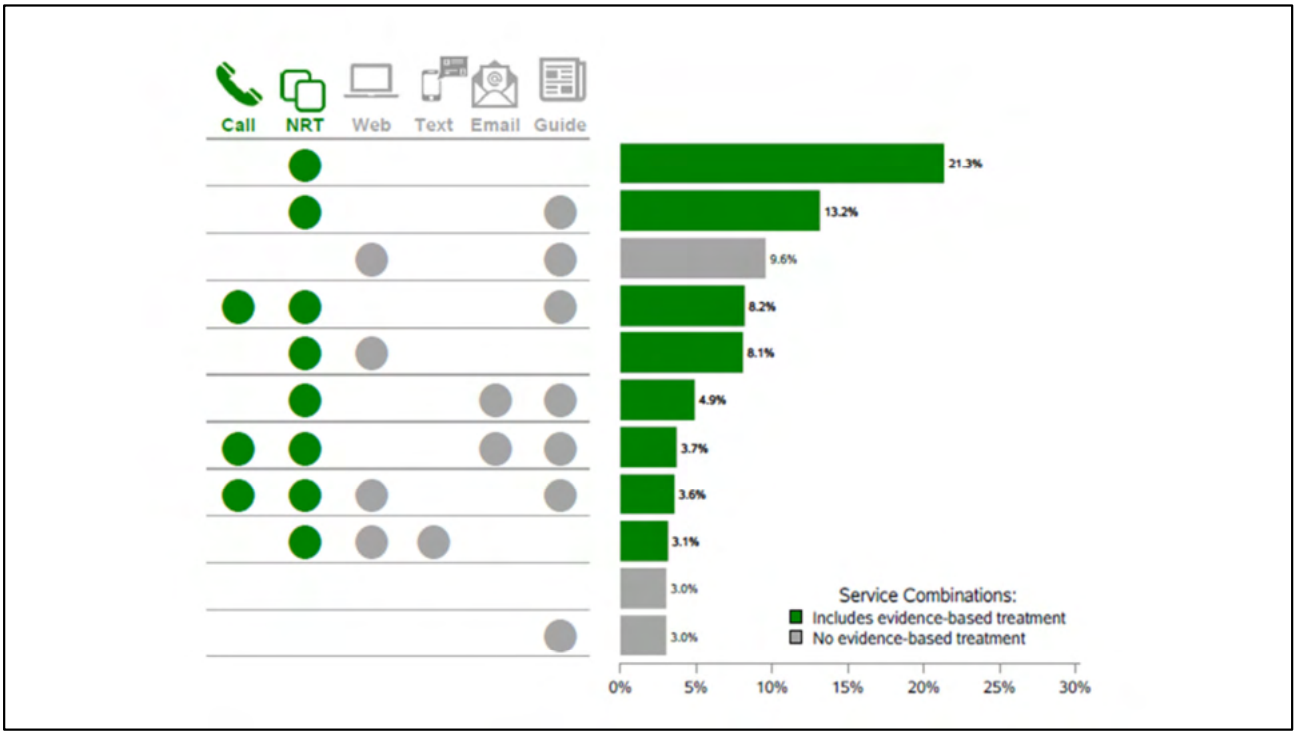
Yielding our finished product.

Combining tables and graphs #2

Which service combinations are most popular?



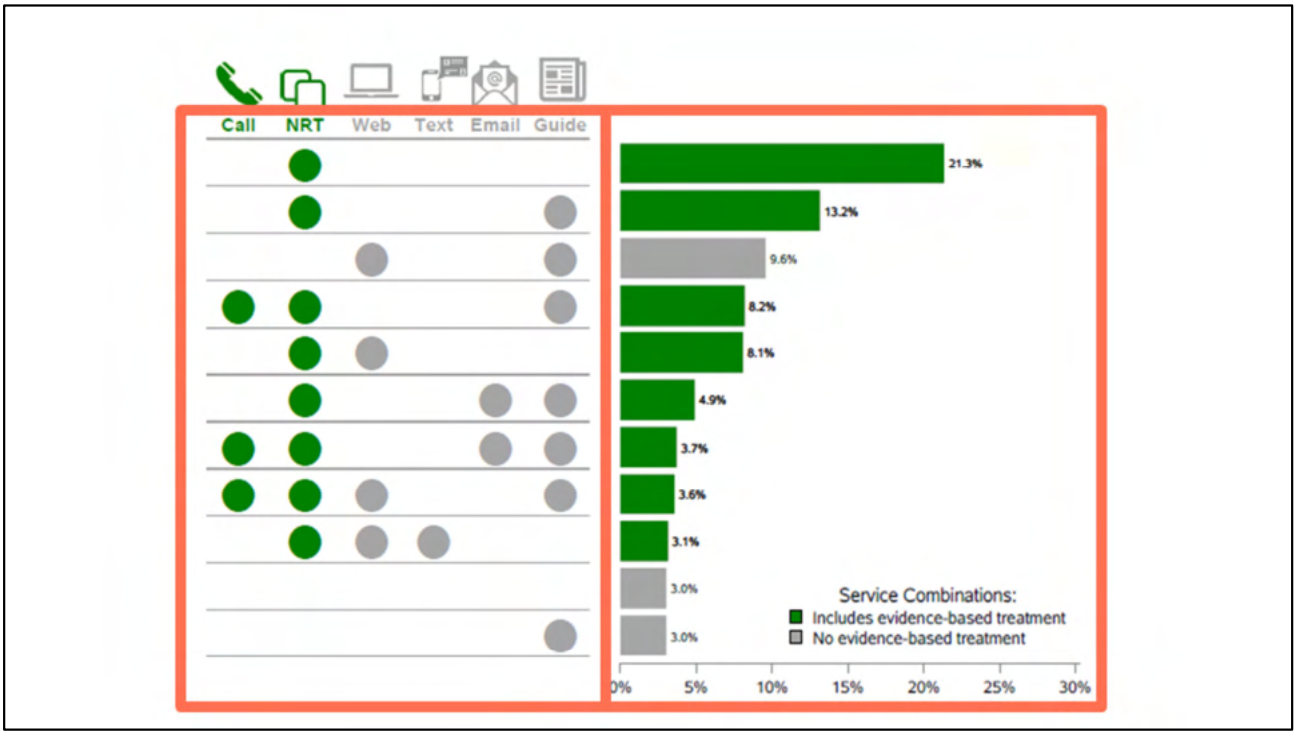
In this next example, I'm going to show how the graph from my introduction was created. This graph was created as part of a report to track combinations of services received by people enrolling in a tobacco cessation program. Even though the bar chart is ordered in a descending manner, it's difficult to tell which services are being used. I asked one of my colleagues, a data viz wizard, if she could help.



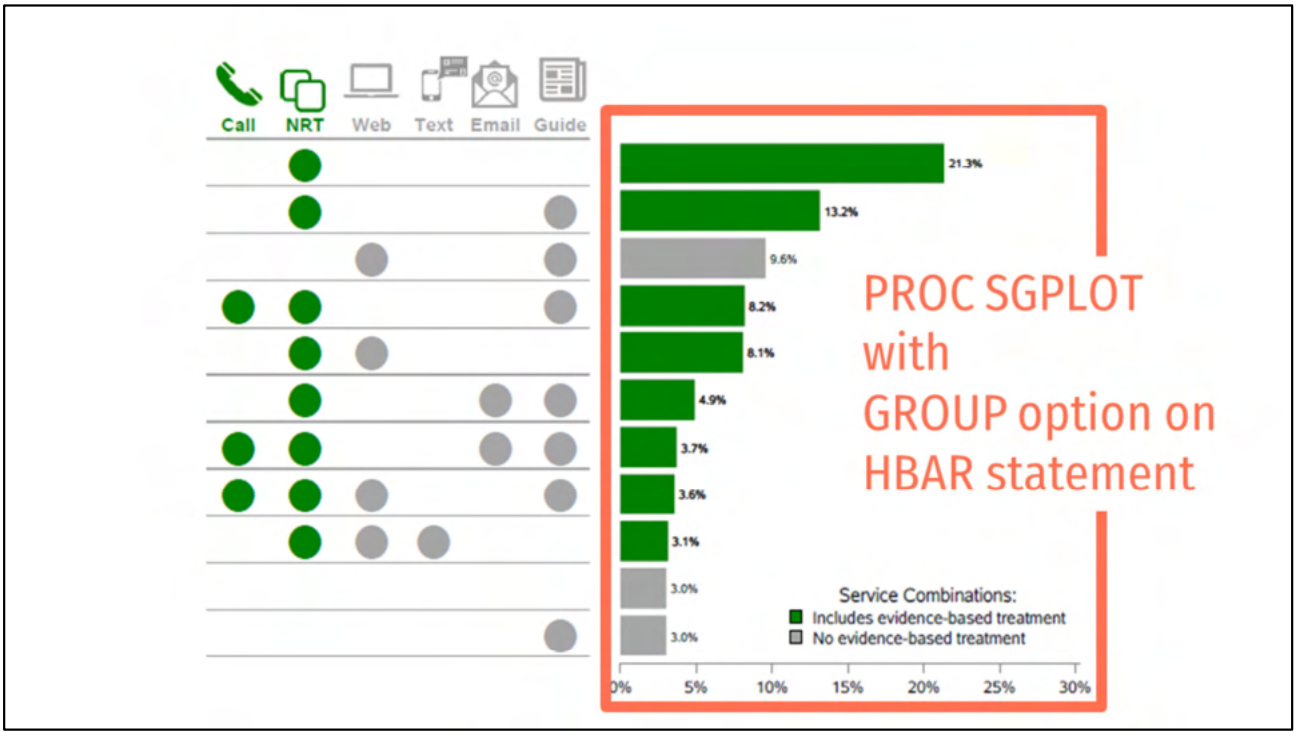
She came back with this mock-up. She
- kept the ordering, and
- turned the text labels into a table with 1 column for each service,
- added icons to help quickly identify the service,
- And colored the bars based on whether or not it included an evidence-based treatment (calls and NRT)

It's a lot easier to see which services are most popular and it's more visually engaging.

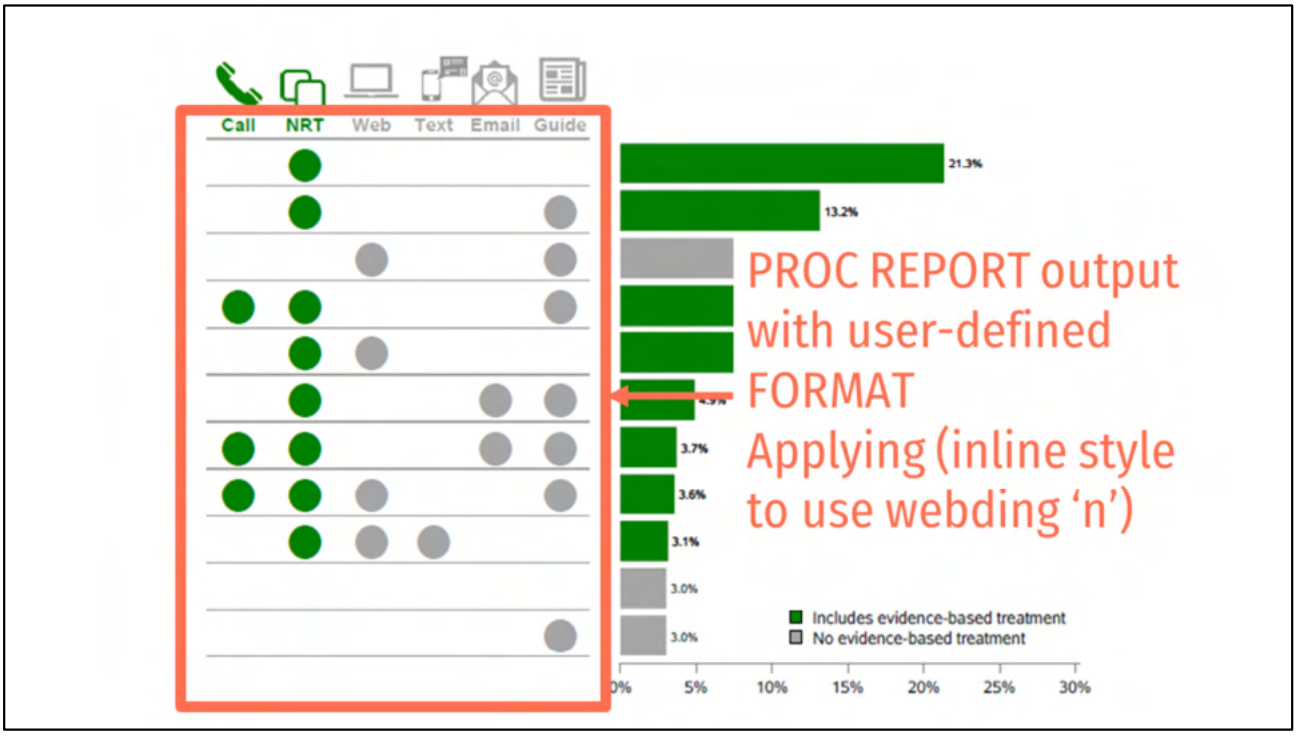
Then she asked if I could do this in SAS. Of course! If you had to do it, where would you start? What components do you see? How would you add the icons? We've talked about all of these today.



First – I see 2 main components. The graph component and the table component.



There are 2 main components – a chart and a table. For the chart, we're going to use the SGPLOT procedure with HBAR statement. We'll create 2 groups – 1 for each color.



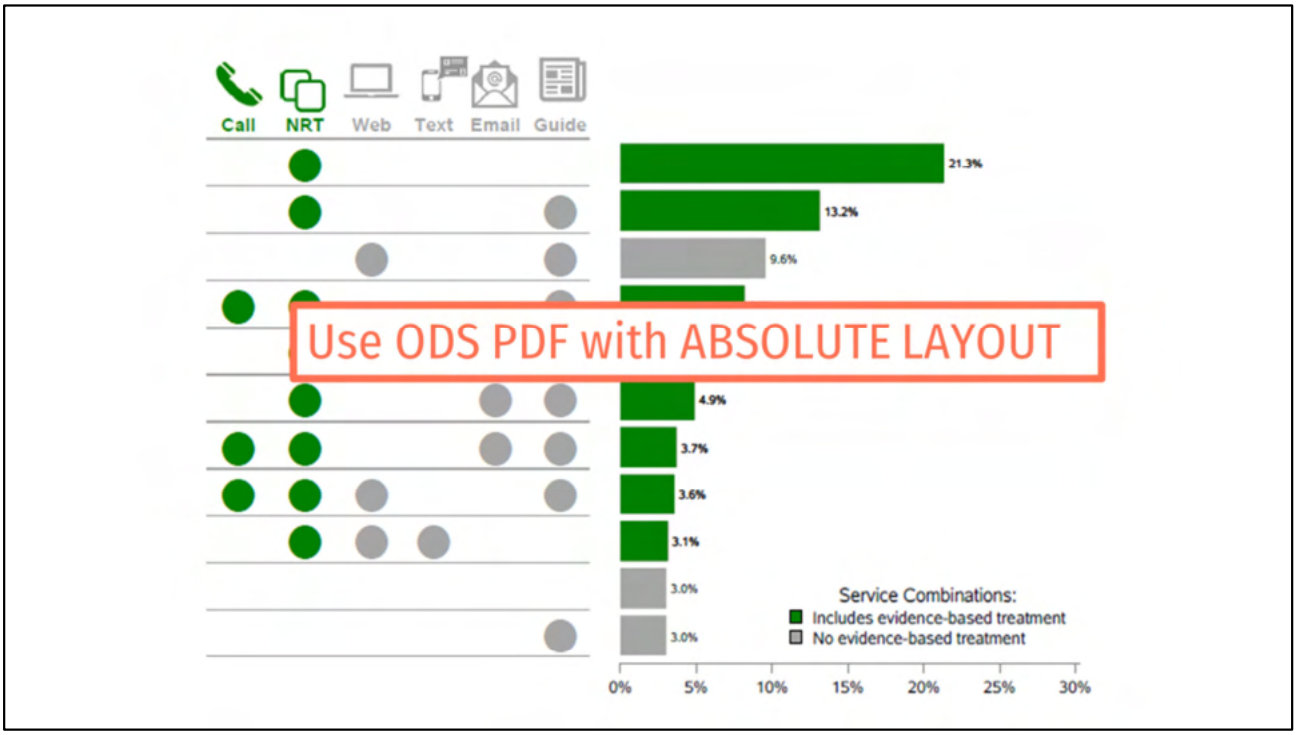
Then, to create the table we'll use a PROC REPORT to make the table. In this case, the underlying dataset is a table with this layout that has a 1 if the service was included, a 0 if it was not. Then, we create a custom FORMAT applying an inline style to turn the 1's into dots, and the 0's into blanks. PROC REPORT is being used as a fancy PROC PRINT.

```
PROC FORMAT;
        VALUE Evfmt 0=" " 1="^S={COLOR=green FONT_FACE='Webdings'}n";
        VALUE NoEvfmt 0=" " 1="^S={COLOR=cxa6a6a6 FONT_FACE='Webdings'}n";
RUN;
ODS REGION WIDTH=3.6in HEIGHT=7.05in x=0in y=2.9in;
PROC REPORT DATA = ServCombo_v2 NOWINDOWS NOCENTER SPLIT='*' CONTENTS=""
        STYLE(REPORT)=[RULES=none FRAME=void CELLSPACING=0 CELLPADDING=4.0pt]
        STYLE(COLUMN)=[BORDERBOTTOMCOLOR=cxa6a6a6 BORDERBOTTOMWIDTH=0.5pt]
        STYLE(HEADER)=[BORDERBOTTOMCOLOR=cxa6a6a6 BORDERBOTTOMWIDTH=0.5pt];
        COLUMN CallFlag NRTFlag WebFlag TextFlag EmailFlag QGFlag;

        DEFINE CallFlag /DISPLAY "Call" CENTER FORMAT=EvFmt.
                STYLE(COLUMN)=[FONTSIZE=&Fntsz CELLWIDTH=0.55in]
                STYLE(HEADER)=[COLOR=green FONTSIZE=12pt];
        DEFINE NRTFlag /ANALYSIS "NRT" CENTER FORMAT=EvFmt.
                STYLE(COLUMN)=[FONTSIZE=&Fntsz CELLWIDTH=0.55in]
                STYLE(HEADER)=[COLOR=green FONTSIZE=12pt];
        DEFINE WebFlag /ANALYSIS "Web" CENTER FORMAT=NoEvFmt.
                STYLE(COLUMN)=[FONTSIZE=&Fntsz CELLWIDTH=0.55in]
                STYLE(HEADER)=[COLOR=cxa6a6a6 FONTSIZE=12pt];
RUN;
```

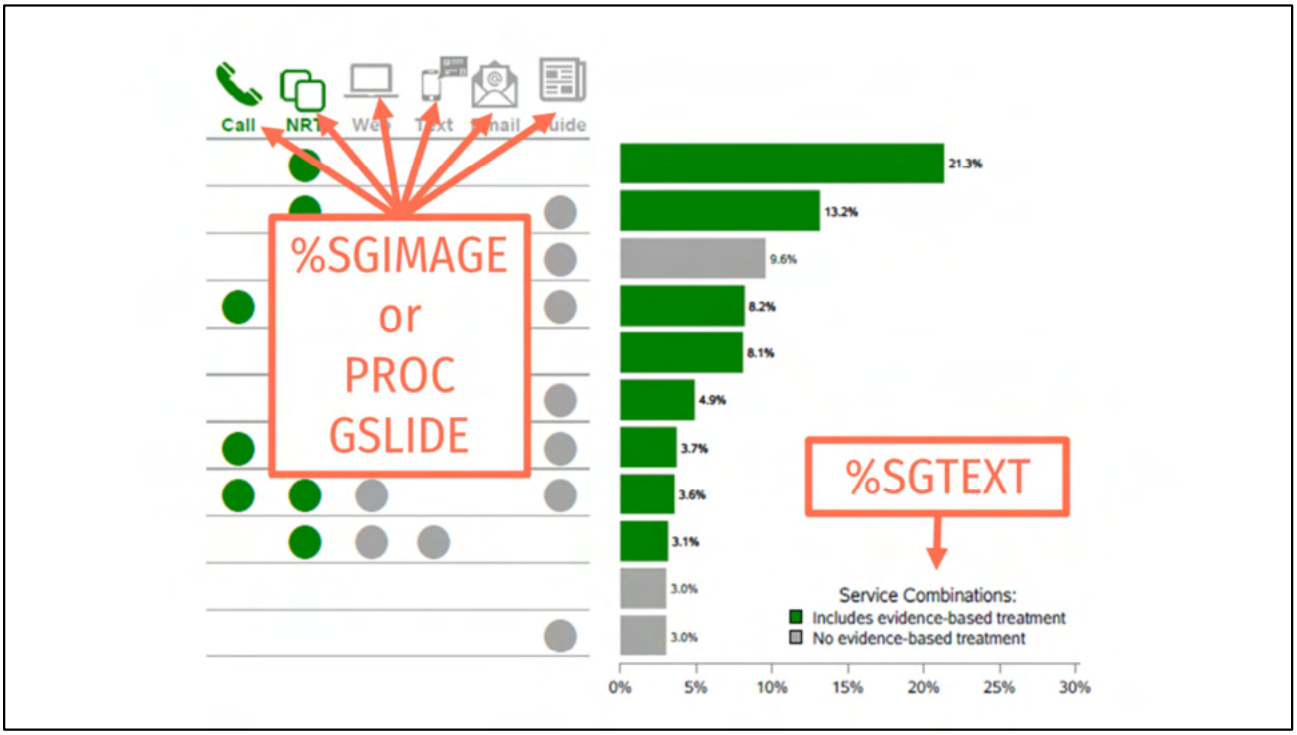
Here's a quick look at the user-defined formats and how they were applied to the PROC REPORT. The format was applied to each column in the PROC REPORT. The only difference between the two formats is color – one is green, one is grey. The webding is just a circle.

```
ODS REGION WIDTH=4.6in HEIGHT=7.4in x=3.4in y=2.93in;
ODS GRAPHICS / RESET=ALL BORDER=OFF HEIGHT=7.4in WIDTH=4.7in;
PROC SGPLOT DATA = ServCombo_v2 NOBORDER NOWALL;
        FORMAT Perc Perc2 PERCENT7.1;
        FORMAT TextString $35.;
        HBAR TextString /RESPONSE=Perc CATEGORYORDER=RESPDESC NOOUTLINE
                DATALABEL=Perc FILLATTRS=(COLOR=cxa6a6a6)
                LEGENDLABEL="No evidence-based treatment" NAME="B"
                BASELINEATTRS=(THICKNESS=0);
        HBAR TextString /RESPONSE=Perc2 NOOUTLINE DATALABEL=Perc2
                FILLATTRS=(COLOR=green)
                LEGENDLABEL="Includes evidence-based treatment" NAME="A"
                BASELINEATTRS=(THICKNESS=0);
        YAXIS DISPLAY=NONE;
        XAXIS DISPLAY=(NOLABEL) VALUES=(0 to 0.3 BY 0.05);
        KEYLEGEND "A" "B"/TITLE="Service Combinations:" NOBORDER ACROSS=1
                POSITION=BOTTOMRIGHT LOCATION=INSIDE;
RUN;
```

Since we're looking at code anyway, here's the PROC SGPLOT code. Notice how I've lined up the regions with the code in the slide before. Lining up these regions takes a lot of trial and error!



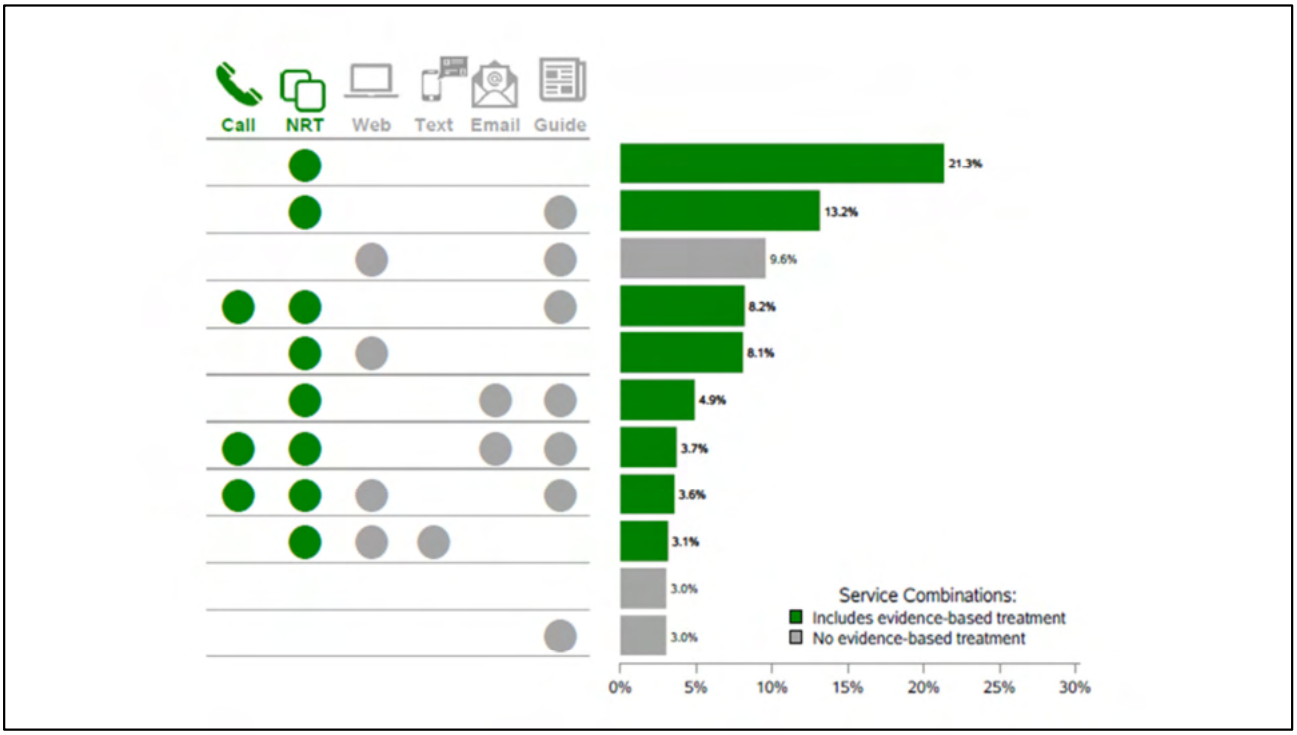
To get these elements lined up, we'll use the ODS PDF destination with ABSOLUTE LAYOUT



To add the icons, we'll add IMAGE annotations and add a custom label using a TEXT annotation.

```
ODS REGION WIDTH=0.45in HEIGHT=0.45in x=0.06in y=2.45in;
GOPTIONS DEVICE=SASPRTC NOBORDER TRANSPARENCY;
FILENAME icon1 "&imagepath\PhoneIconv2.jpg";
PROC GSLIDE IFRAME=icon1 CFRAME=white WFRAME=0 IMAGESTYLE=fit;
RUN;
QUIT;

.
.
.

ODS REGION WIDTH=0.35in HEIGHT=0.35in x=0.67in y=2.5in;
GOPTIONS DEVICE=SASPRTC NOBORDER TRANSPARENCY;
FILENAME icon2 "&imagepath\NRTIconv2.jpg";
PROC GSLIDE IFRAME=icon2 CFRAME=white WFRAME=1 IMAGESTYLE=fit;
RUN;
QUIT;
```

Here's a quick look at the PROC GSLIDE code that I used to insert the icons. I've done it by defining a region for each icon, then inserting either a green or grey icon.



That's it. Easy peasy.

Regression

Which characteristics were associated with awareness?

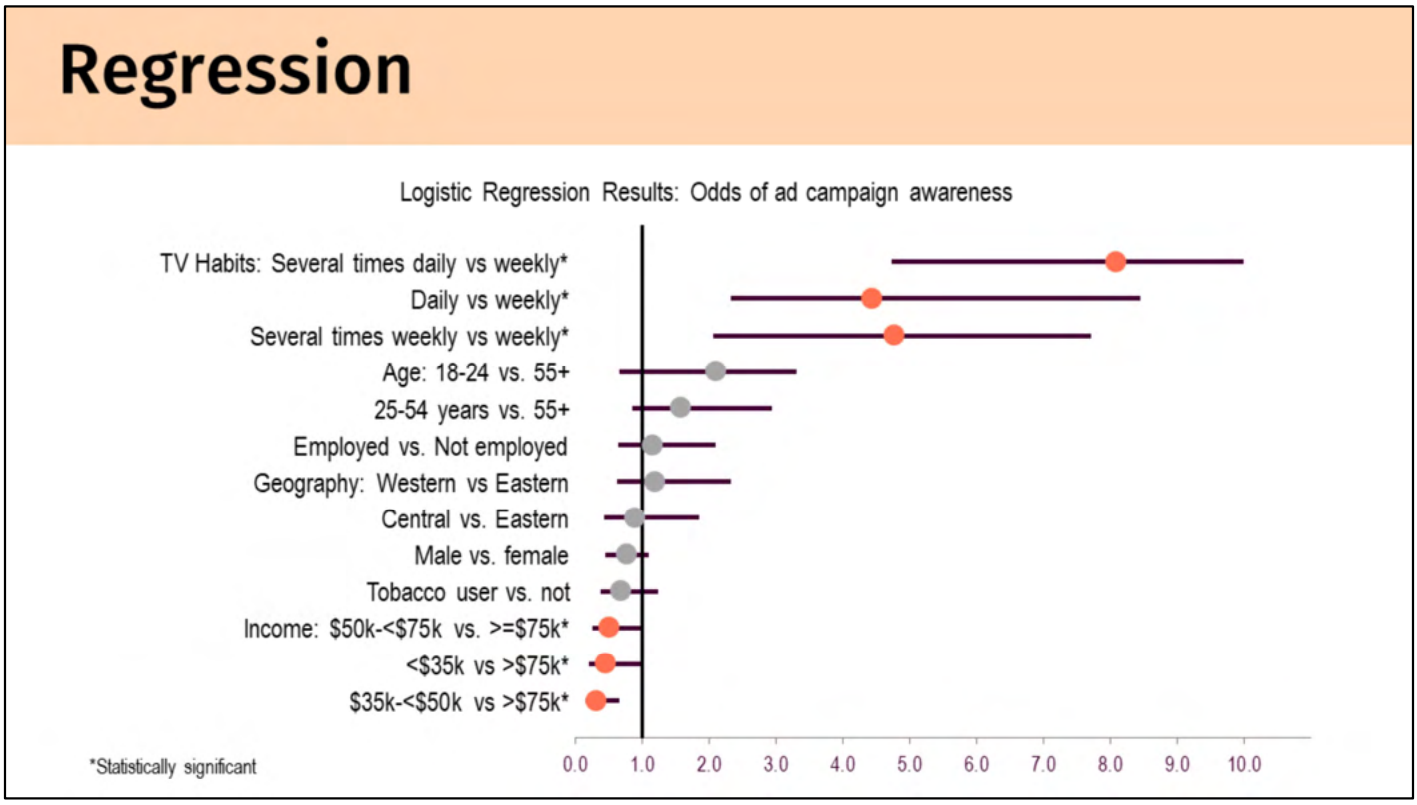
Logistic regression model
Outcome = aware of ad campaign

Variable		B	S.E.	Wald	df	P-value	Odds Ratio
Gender	Male (vs. female)	-0.263	0.277	0.903	1	0.342	0.769
Age group	Reference = 60+ years			2.705	2	0.259	
	18 - 24 years	0.743	0.592	1.573	1	0.210	2.102
	25 - 59 years	0.461	0.315	2.135	1	0.144	1.586
Employment Status	Employed (vs not employed)	0.144	0.304	0.224	1	0.636	1.155
Income	Reference = $75,000+			10.211	3	0.017	
	< $35,000	-0.802	0.393	4.169	1	0.041	0.449
	$35,000 - < $50,000	-1.168	0.387	9.103	1	0.003	0.311
	$50,000 - < $75,000	-0.691	0.340	4.123	1	0.042	0.501
TV Habits	Reference = < 1x /week			37.881	3	0.000	
	Several times/day	2.389	0.425	31.529	1	0.000	10.904
	Daily	1.489	0.329	20.494	1	0.000	4.432
	Several times/week	1.561	0.428	13.327	1	0.000	4.766
Geographic Area	Reference = Eastern DMA			0.494	2	0.781	
	Oil corridor	-0.120	0.373	0.103	1	0.748	0.887
	Central corridor	0.179	0.341	0.275	1	0.600	1.196
Tobacco Status	Tobacco user (vs no)	-0.388	0.308	1.586	1	0.208	0.678
	Constant	0.787	0.472	2.782	1	0.095	2.198

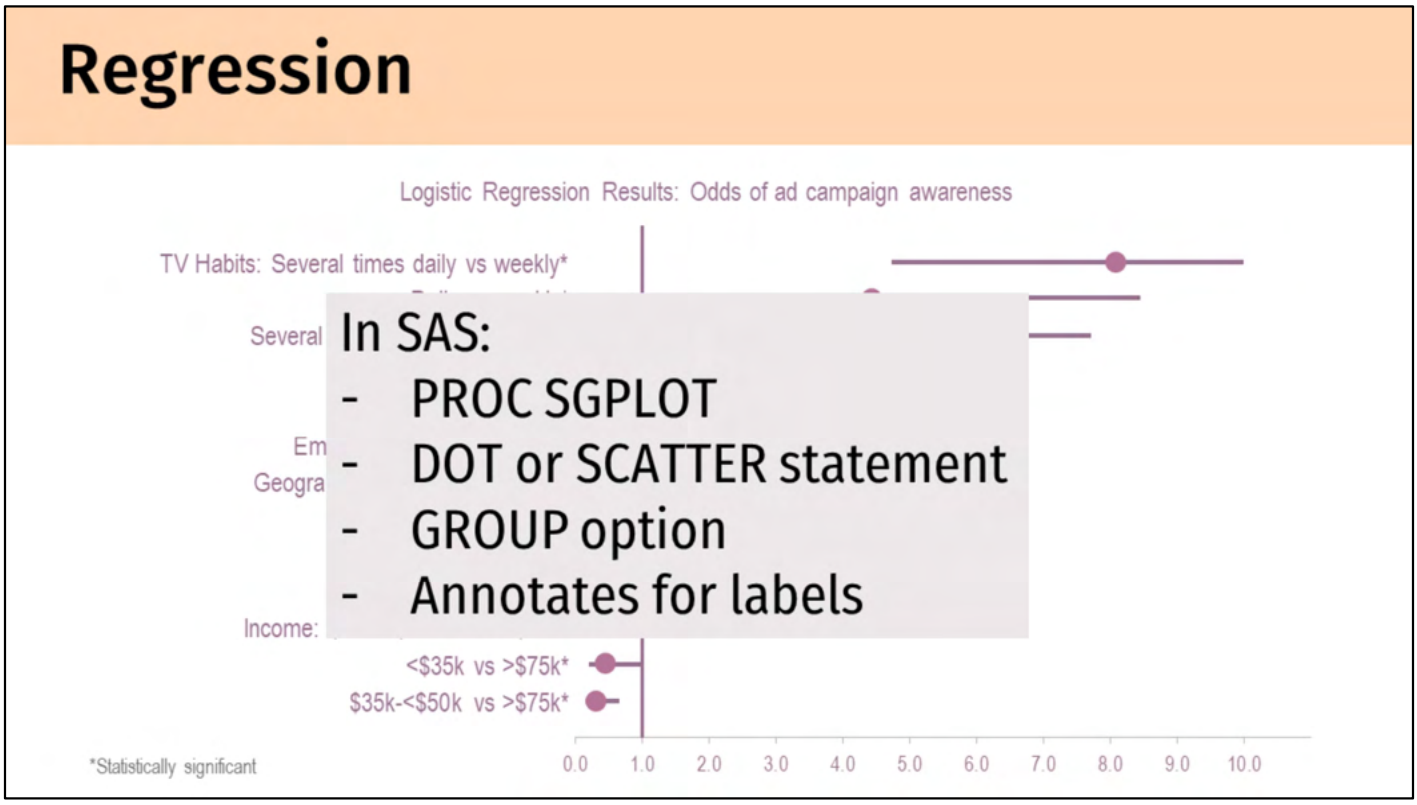
The more complicated the method, the more ways you can display data. I'm not going to say a lot about regression results, other than we can use the same principles we've been talking about to clean up plots related to more sophisticated analyses.

This is a logistic regression output table from SPSS. Pretty raw – pretty dense. We're highlighted significant findings, but this is still hard to read and overwhelming if the reader does not have a solid grasp of regression output. Does the client still need this level of detail? Absolutely! Does the client need this table in the body of the report? Probably not! Put this table in an appendix and use graphs and descriptions in the body of the report.

Regression

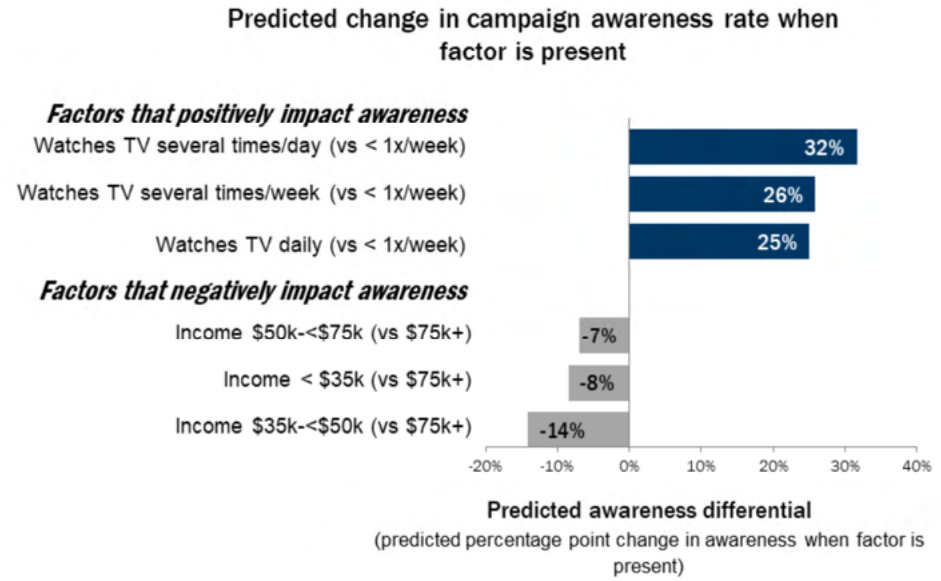


Here's a forest plot – just odds ratios with their 95% CIs. I ordered it by odds ratio value, took out the extra grid lines, colored the significant values/made the non-significant ones grey, removed the end caps from the 95% CI lines.

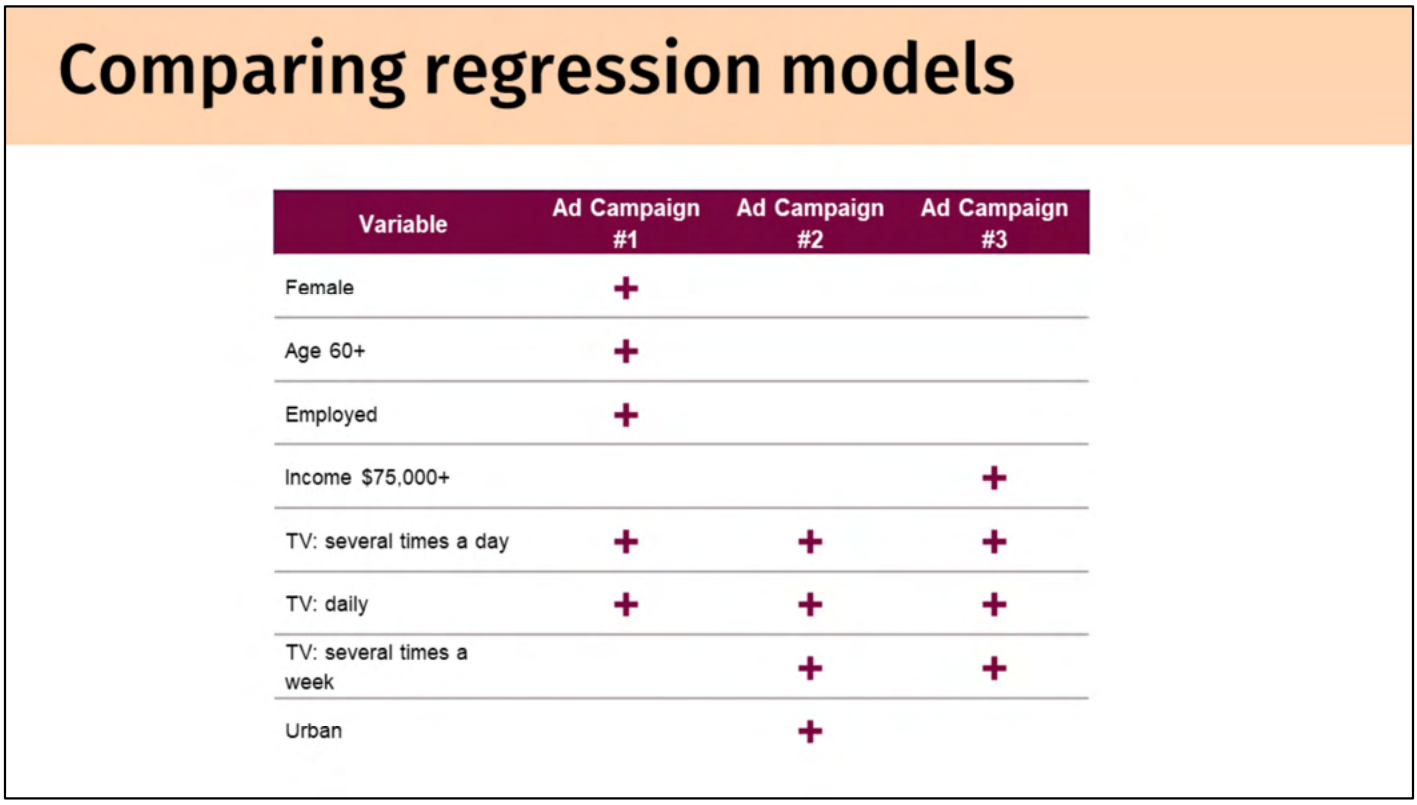


One way to approach this in SAS is to use PROC SGPLOT with either a DOT or SCATTER statement and a GROUP option. The labels can be added with annotate statements.

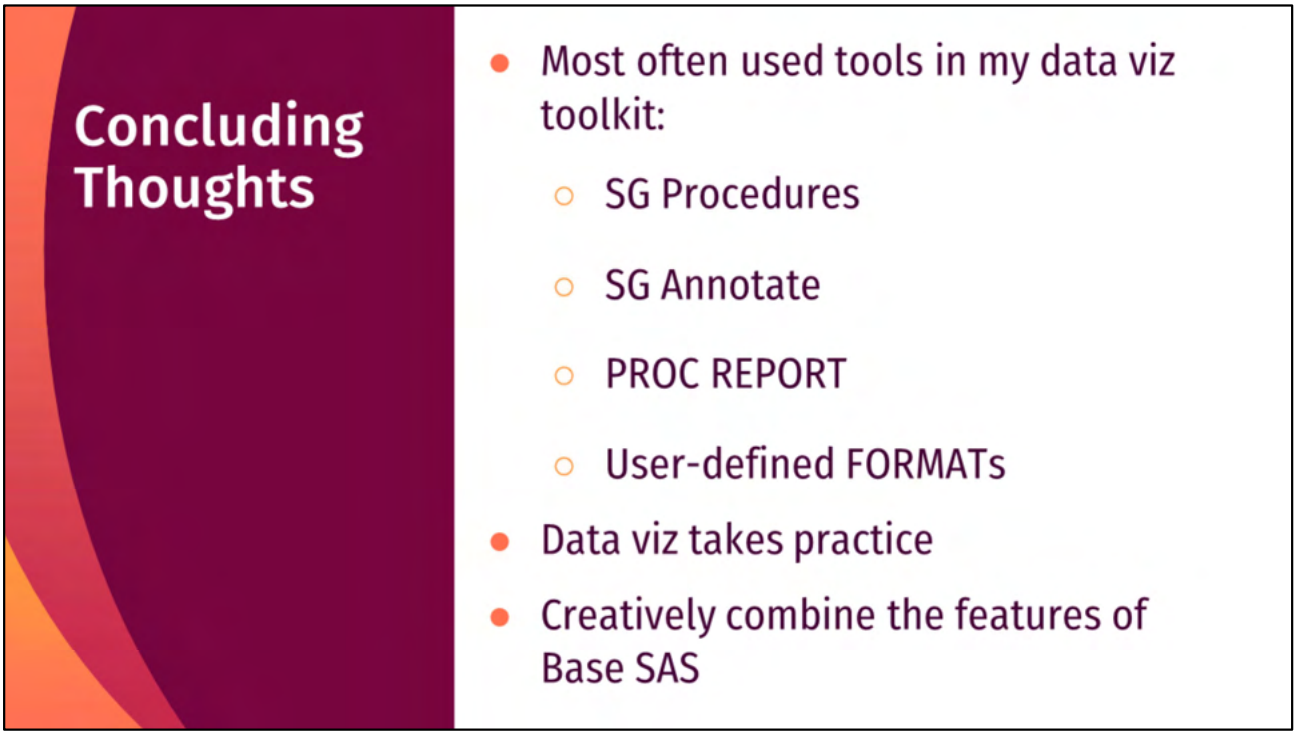
Regression



Another way to present these results is in the risk differential. For variables that are significant in the model, calculate the predicted percent point change in awareness when the factor is present. Then order your results and use split bar graphs. Again, we've cleaned up the results by removing grid lines and coloring bars. SAS can calculate these values for you if you run the model with PROC GENMOD then use PROC SGPLOT with HBAR statement to graph



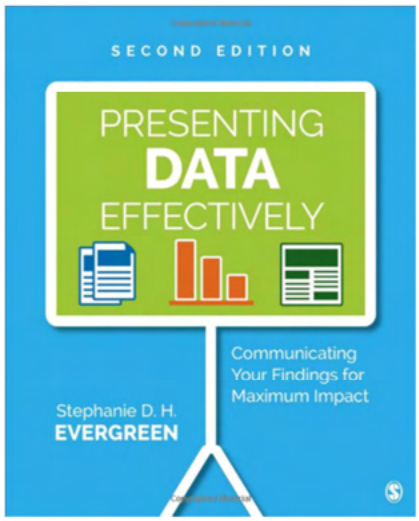
The last thing I'll mention is if you need to compare regression models that used the same set of predictors, consider using an at-a-glance table like this one where the significant relationships are called out by symbols instead of comparing regression coefficients or odds ratios. Here we used the same set of predictors for the 3 ad campaigns and we see that the more TV you watch, the more likely you are to have seen all 3 ad campaigns, as you'd expect.

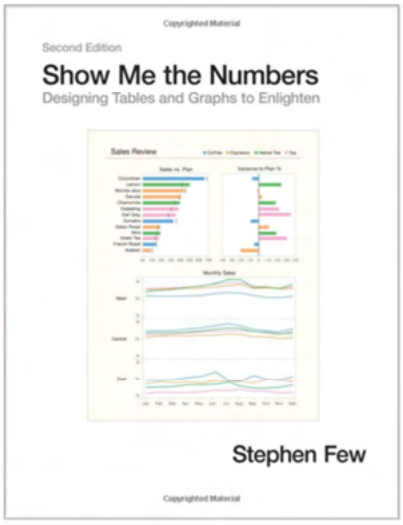


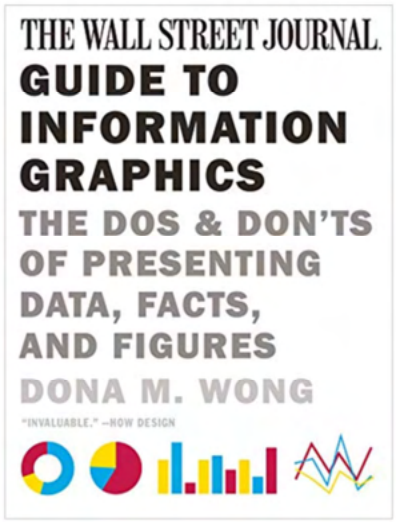
In the last 45 minutes, you've learned about some of my most-often used tools in my data viz toolkit. Building these skill sets don't happen overnight, it does take practice, but it's possible and really satisfying when you figure it out. The parting thought that I hope you take away today is that: With a lot of creativity, and not so much work, we can create really awesome visualizations in Base SAS.


Thanks!
Any questions?
You can find me at:
SRichter@RichterStats.com

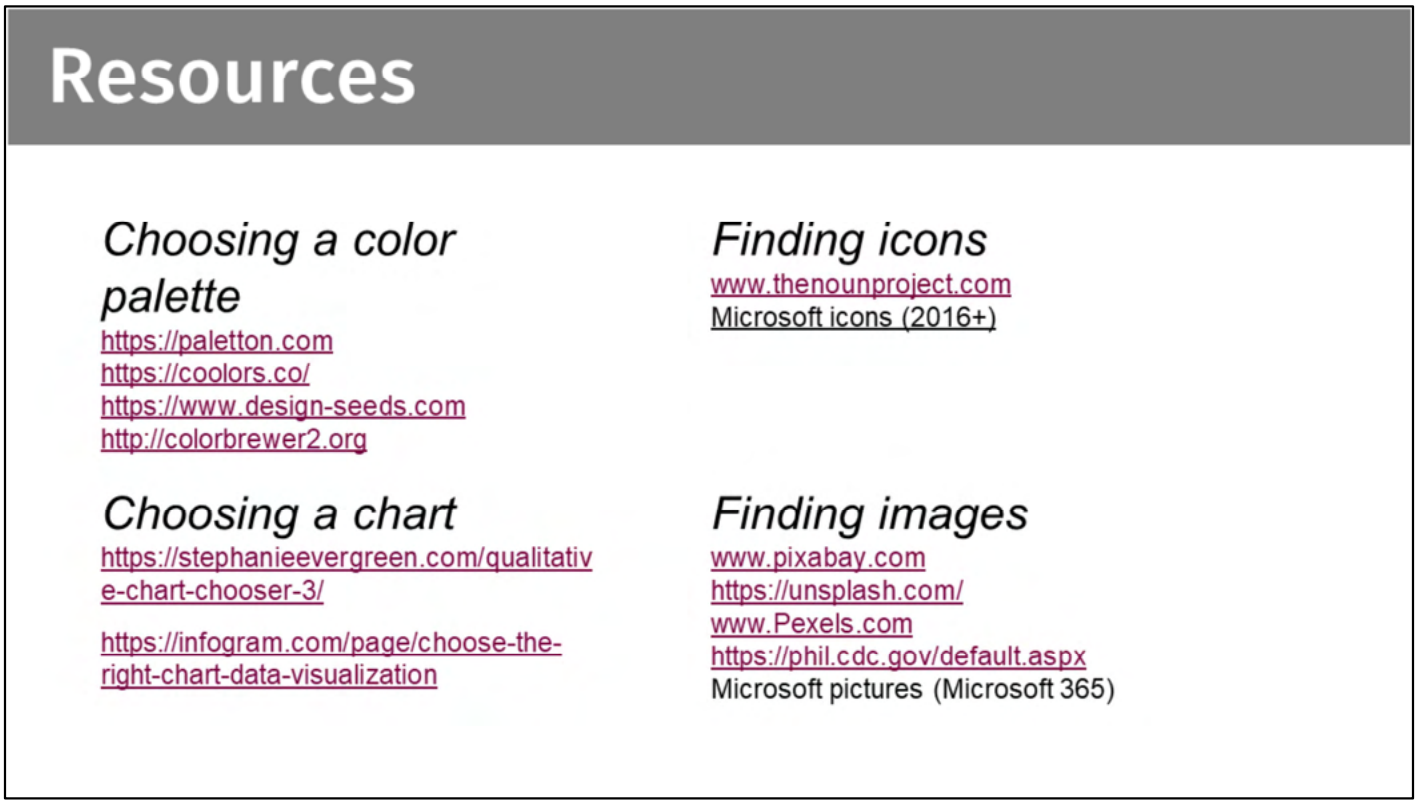
References







If you're looking for more resources, here are the main references for my talk. I recommend Stephanie Evergreen's book Presenting Data Effectively. She also has a blog with a lot of good information.



Here are resources for choosing color and finding icons and images.

Resources

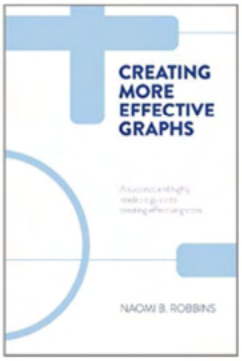
Blogs

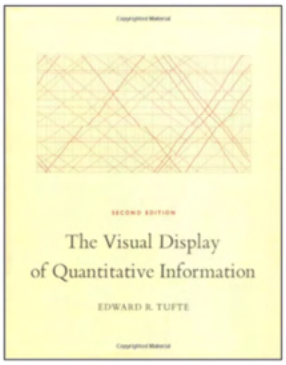
Stephanie Evergreen
https://stephanieevergreen.com/blog/

Ann Emery
https://depictdatastudio.com/blog/

Nancy Duarte
https://www.duarte.com/

Books





Even more resources!