GOV 374N Data Visualization Guide

November 12, 2022

1 Quick Introduction

This is a general guide to visualizing your data in such a way as suggests an answer to your research question.

How to use it:

1. Go to the table of contents
2. Find the category that your research project fits into
3. Go to that section and read it (you might also find it helpful to browse the advice on data visualization for other types of projects too)
4. If you need help plotting something Excel, go the final section of the guide and find the tutorial for making that figure in Excel.

If some of the terms in the table of contents are unfamiliar, there's a list of definitions under the table of contents.

2 Longer Introduction

This is a general guide to visualizing your data in such a way as suggests an answer to your research question.

That isn't the only purpose you might have for visualizing data. You might also, for example, want to visually represent your data to develop or establish context around your question. To develop the example more, imagine your project is the one outlined in the good proposal that Prof. Henson had you practice grading. That project was to see if Texas counties that lost polling locations (on account of a recent state law) subsequently had a decrease in voter turnout for the next election compared to other Texas counties. You might, in presenting that question, want to include a graphic that communicates to readers how many counties were affected by the law, how great the effects were, and how varied the effects were. But you'd also want to include a graphic suggestive of the relationship (or the absence of any clear relationship) you found between law-mandated poll closures and voter turnout, and it's *that* sort of graphic that this guide will focus on.

Knowing how to visualize data well is probably more like riding a bike than like doing long division. Learning long division is learning to follow a small number of steps that apply to every pair of numbers out there; learning to ride a bike is a mysterious process. It helps quite a bit to give someone tips, but most of the learning comes down to trying again and again until, somehow, you're no longer falling off the bike. The relevance of this to data visualization is this: This guide contains general advice about does and doesn't work to effectively communicate your findings, but that advice often won't apply perfectly, and almost all of you will be able to find ways to tweak my advice to better suit your particular projects.

To use this guide, go to the table of contents, find the category that your research project fits into, and go to that section (you might also find it helpful to browse the advice on data visualization for other types of projects too). I've tried to include in these categories most of y'all's projects. Not everyone's project, however, fits neatly into one of these, so if you're one of those people, you'll have to think about how the recommendations here do and don't apply to you.

If some of the terms in the table of contents are unfamiliar, I've put a list of definitions under the table of contents.

This guide is not about how to make this or that figure in Excel, but about choosing how to visualize your data. At the end of this guide are links to tutorials that will show you how to use Excel to make all the figures I mention here.

Unless I say otherwise, all the data I use in these examples is made up

3 Table of Contents

3.1 Definitions

Correlational Research Correlational research questions are of the form "Does X increase or decrease as Y increases," or "How are X and Y related," where X and Y are variables you're measuring. Usually, you have a lot of observations in your data (lots of people, counties, school districts, cities, states, or other things). See **Visualizing Correlations (without Time Series Data)** for examples. #### Treatment/Intervention Research This is research about questions of the form "Did Y change after X happened?", where X is some one-time event. In this class, X is usually a law being passed. See **Visualizing Data for Treatment/Intervention Research** for an example. #### Time Series Data If your research question has anything to do with change over time, then you probably have time series data. In fact, you might have time series data even if your question isn't about change over time. You have time series data if 1) your data points come from different periods of time and 2) the difference in time is significant, is part of the data. See **Visualizing Time Series Data** for examples of time series data. #### Continuous and Discrete Variables Continuous variables are variables that take numerical values from a continuous or effectively continous set of numbers. Discrete variables are variables that take numerical values from a limited set of numbers. Another way to think of this is that discrete variables measure the number of something and continuous variables measure the amount of something. Someone's annual income is a continuous variable because it can have any value, up to two decimal points of precision, from 0.00 dollars to infinite. Someone's number of children is a discrete variables, since one can only have whole numbers of children.

In practice, numerical variables can be treated as continuous unless they have a very small number of possible values (say, much less than 100 possible values). For example, Age is a discrete variable when it's measured in whole years, but it can usually be treated as a continuous variable.

For why this distinction matters, see below, section **Categorical or Discrete Variables**. For more on the distinction, see https://en.wikipedia.org/wiki/Continuous_or_discrete_variable. #### Categorical Variable Categorical variables are variables that take non-numerical values. Gender, race, and opinion of the Republican Party (measured on a sclae from "Strong Dislike" all the way to "Strong Like") are all categorical variables. #### Ordinal Variable If a categorical variable's possible values can be put in a sort of order (so, if you could say that one value is "more" or "less", in some sense, than the others), then that categorical variable is ordinal. #### Binary Variable If a categorical variable has only two values (these usually can be represented as yes or no), it is a binary variable. Whether you voted in the 2022 midterms is, for example, a binary variable.

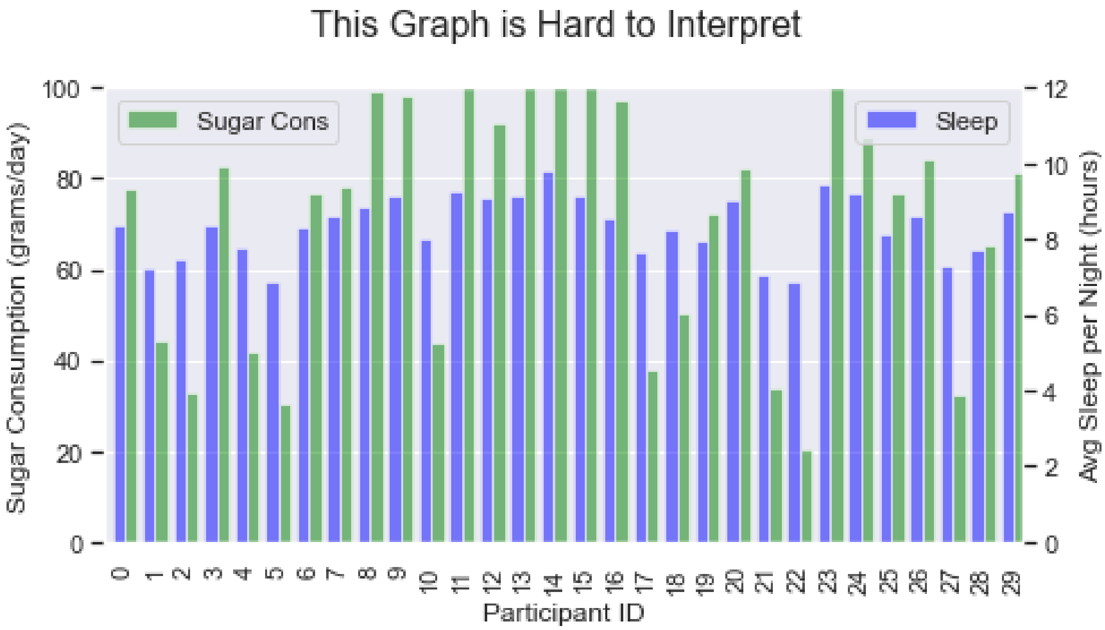
# Correlational Research

## Continuous Variables

In general, if your question is about the correlation between two continous variables, you're going to use a scatter plot to visualize that relationship. The main error you want to avoid here is "dumping" your data into a bar graph or line graph. Let's say you're interested in the relationship between how much sugar someone eats and how many hours they sleep per night on average. Your data would look like this:

```
    Daily Sugar Cons (g)  Avg Daily Sleep (hours)
id
0              77.771841                 8.396169
1              44.170054                 7.242491
2              32.930617                 7.486108
3              82.748577                 8.395889
4              41.646096                 7.786778
```

For each participant in our study (there are 30; I've only displayed the first 5), we have their daily sugar consumption recorded (in grams) and their average daily amount of sleep (in hours). Here's a **very bad** way to visualize our data:

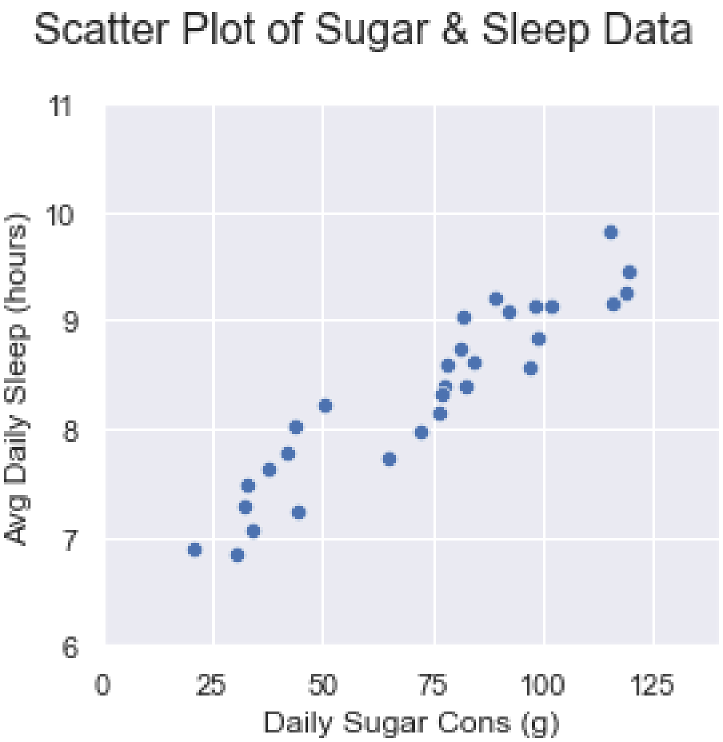


This bar graph tells us almost nothing about the relationship the sleep and sugar consumption might have. Our eyes glaze over in confusion and fear as soon as they're near it. **Don't use bar graphs to show the relationship between two continuous variables**.

In the following sections, I'll give examples of better ways to visualize this sort of data.

### Two Continuous Variables

The research question I used for the bad example above is asking about the relationship between one continuous variable and another, so I'll here give an example of a good visualization of that data:

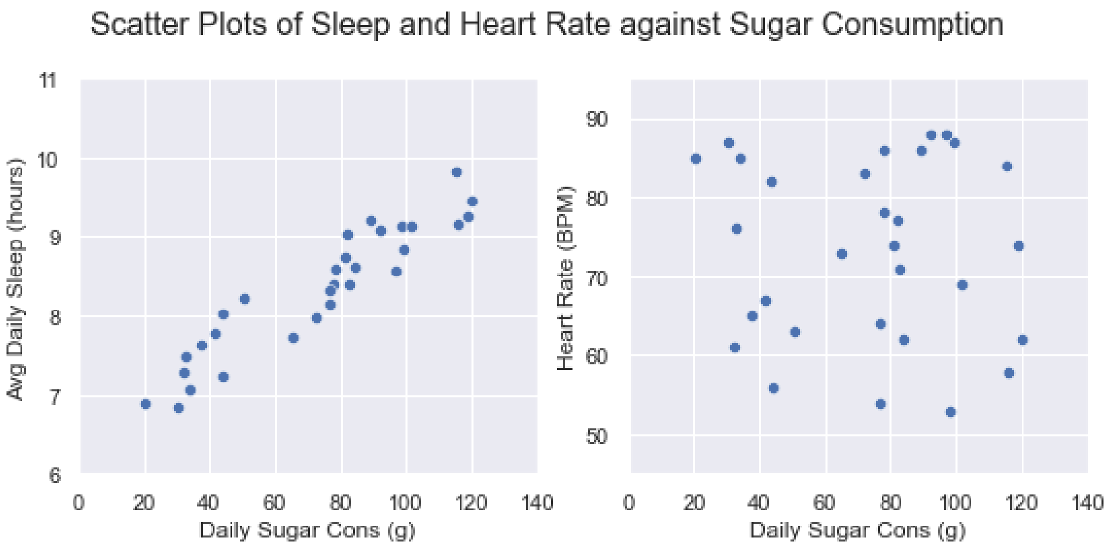


This is a **scatter plot**. This scatter plot is far superior to the bar graph above because it gives a visual clue to the (fictitious, in this case) relationship between sleep and sugar consumption: participants who consume more sugar seem, on average, to sleep more than participants who consume less sugar. In other words, this graphic, unlike the bar graph, shows visually what you're trying to show, the relationship between sugar consumption and sleep.

Note: this fictitious data shows a very strong correlation between sugar consumption and sleep. In fact, sugar consumption "explains" about 70% of the variance in sleep. You often won't find correlations this high between two variables you're interested in, and as a result, your scatterplots won't be as pretty. That doesn't mean there isn't a significant relationship between your variables. The graph is only a visual cue.

### More than Two Continuous Variables

Suppose that instead of just looking at the relationship between sugar consumption and sleep, you're interested in how sugar consumption affects sleep *and* resting heart rate. Again, don't use a bar graph to visualize the relationship between these variables. Instead, use two scatterplots, like this:



The scatter plots again do the trick here: the one on the left (it's the same as the one in the previous section) indicates that sleep and sugar consumption are positively correlated while the one on the rate suggests that heart rate and sugar consumption aren't significantly correlated

Note: again, we wouldn't conclude from the plot alone that heart rate and sugar consumption are uncorrelated. In this case, I've constructed the fake data so that they in fact aren't significantly correlated, but scatter plots can be deceiving.

## Categorical or Discrete Variables

Categorical variables are variables that take non-numerical values.

If the values can be put in a sort of order (so, if you could say that one value is "more" or "less", in some sense, than the others), then the categorical variable is **ordinal**. Some survey questions responses are like this. Suppose respondents are asked "How often do you go to church?", and can answer in the following ways:

1. At least once a day
2. At least several times per week
3. At least once a week
4. At least once a month
5. At least once a year

These answers aren't numerical answers, but they are ordered: If you got to church once a month, you go to church *less* often than someone who goes to church once a week, and *more* often than someone who goes once a year.

Often, the categorical variable's values can't be put in order. Some people call these sorts of variables **nominal** variables, but I don't know if that's commonly used. In any case, variables like this include race (black, white, asian, etc.), political party (Republican or Democrat*), and many others.

A special, because very common, subtype of the nominal variable is the **binary** variable. Binary variables are categorical variables that have only two possible values, and they're common because they represent yes/no questions. They're often used to measure outcomes (Did the patient recover? Did the citizen vote?) and to represent "treatments" (Did the patient receive the experimental medicine? Does the state have the death penalty? Did the county allow early voting?).

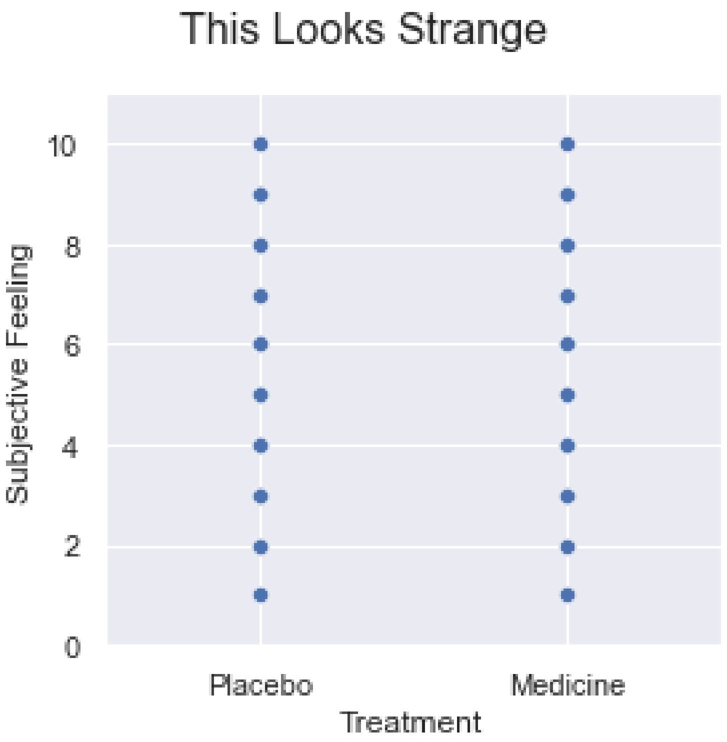
Finally, **discrete** variables are numerical but can only be whole numbers. Any variable that measures things that don't have parts, in some sense or another, is a discrete variable. Hence the number of polling locations in a county is a discrete variable, and the number of votes cast in a county is a discrete variable.

All of these variable types can cause problems in scatterplots. Suppose you're conducting a clinical trial, and your want to see the correlation between taking a medication (instead of a placebo) and how the patient feels (measured on a scale of 1-10). The independent variable, whether the patient took the medication or the placebo, is a **binary** variable. The dependent variable, how the patient feels, is a **discrete variable**. We lost our data on how the patients felt before they took the medicince or the placebo. Here's our data:

```
*If you asked someone how far they lean Republican or Democrat, that would be ordinal, since so
```

```
   Treatment  Subjective Feeling
0          0                   8
1          0                   6
2          0                   3
3          1                   7
4          1                   5
```

There are 60 patients in our study (I've only displayed the first 5). Patients who have a 1 for Treatment received the medication, and patients who have a 0 for treatment received a placebo pill. Let's see how this looks in a scatterplot.



What happened here? Since there are only ten possible values for subjective feeling, and that point on the graph will have a dot on it as long as one person in the treatment or placebo group feels that way, our plot has quickly "saturated." There are in fact more patients who feel 9/10 in the treatment group than in the placebo group, but you can't see that, since there can only be one dot there.

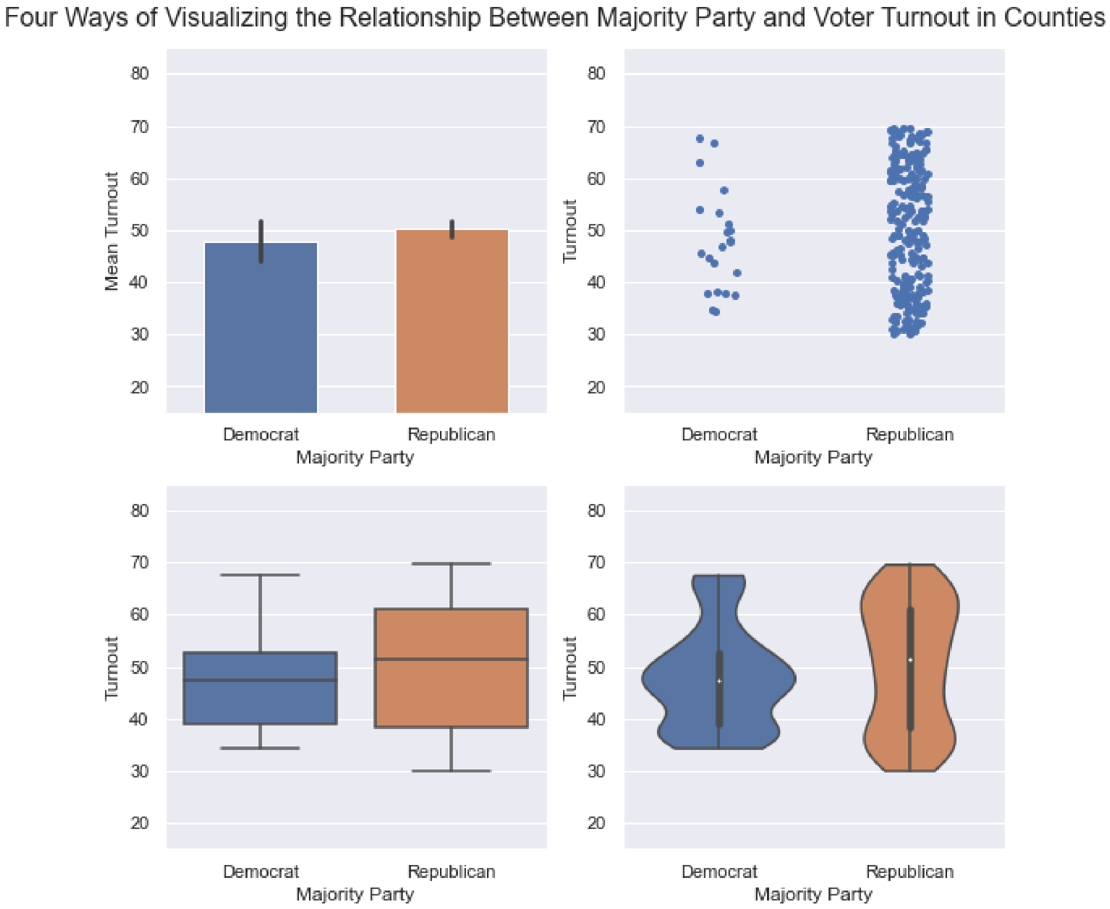
What this means is that if you have a categorical variable or a discrete variable with a small number of possible values, you can't graphically represent your data very effectively with a scatter plot. I'll go through examples of how to solve problems like this in the next three sections. If you want to see how you can better visualize this example, go to the last section in this chapter, **A Categorical or Discrete Variable's and a Categorical or Discrete Variable**.

### A Continuous Variable and a Categorical or Discrete Variable

If you're interested in the relationship between a continuous variable and a categorical or discrete variable, you have a few options. Imagine you have the following data.

```
   Turnout Majority Party
0    47.65       Democrat
1    37.52       Democrat
2    63.09       Democrat
3    51.18       Democrat
4    46.92       Democrat
```

Each observation in our made-up data is a Texas county, so we have 254 data points (I've only shown the first 5). Turnout is the percent voter turnout in the 2020 presidential election, and Majority Party is the party of the presidential candidate that won the most votes in the county. Turnout is a continuous variable, while Majority Party is a categorical variable. We want to see if there's a relationship between Majority Party and Turnout. Here's how we could visualize that relationship:



All four of these graphs tell us something about the difference in Turnout between the two groups of counties.

The **bar plot** on the top left shows the average turnout for the counties in each group. That line on the top of each bar graph is also informative: it's a 95% confidence interval for the means. To learn more about what that is (and isn't), see https://en.wikipedia.org/wiki/Confidence_interval.

The **jittery scatter plot** on the top right is a scatter plot with 'jitter'. By adding 'jitter' to the points, we spread them out a little bit so we can see differences in the frequency of each value. This gives us more information about the spread of the turnout values in the groups of counties. We could enhance it by putting a line where the average is for each group.

The figure on the bottom left is a **box plot**, or a box and whiskers plot. The line in the middle of each box is the median Turnout for the group of counties, and the top and bottom lines of each box are the 75th and 25th percentiles for the group of counties. The advantage of this graphic over the bar plot of the means is that it communicates more information about the distribution of turnout rates for each group of counties. The bar plot only tells us the average turnout for each group, while the box plot gives us a sense of how the counties vary from the average, but it does so in a simpler way than the scatter plot with jitter.

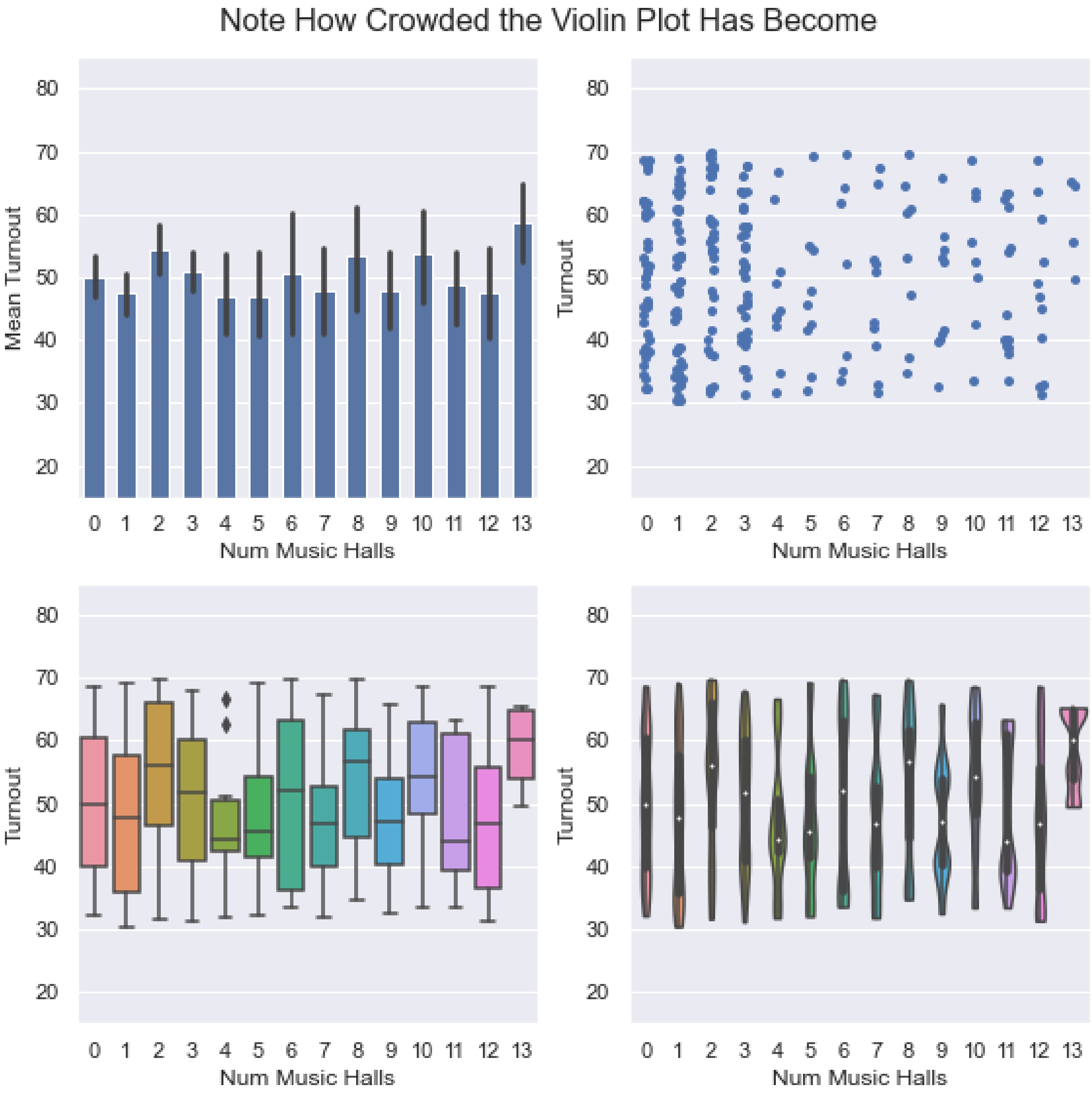
The bottom right graphic is a **violin plot**. That tiny black line in the middle is the same as the box plot: the white dot is the median, the top and bottom of the thick part of the line is the top and bottom of the box, and the top and bottom of the thin part of the line is the top and bottom whisker. But in addition to that, the violin plot gives a smoothed estimate of the distribution for the turnout variable. At wider points in the shape, we observed that value more often, and at narrower points in the shape, we observed that value less often. This is often very useful, since we get even more information about the variable than we get from the box plot, but we still have a *cleaner* presentation than the scatter plot with jitter.

Which of these you choose depends on what you want to show. Do you want your reader to focus on the difference in the average between the two groups? In that case, use the **bar plot**. Do you want to show the reader the median of the two groups as well as some other basic information about the distribution of the measured variable for each group? In that case, use the **box plot**. The **violin plot** shows more information about the distribution of the measured variable in each group than the previous two, but at the cost of visual complexity. Finally, the **jittery scatter plot** shows you exactly what the data is, but at the cost of not showing you any simple statistics about the two groups, like the average or the standard deviation. This can be remedied, however, by adding a few markers on the plot.

A Second Example The violin plot, however, takes up a lot of visual space, and if you have a lot of groups to compare, then it's a bad choice. In a second, I'll show you how each of these graphics look when, instead of having just two groups (Republican and Democrat counties), we have a discrete variable with 14 different values. Here's the data:

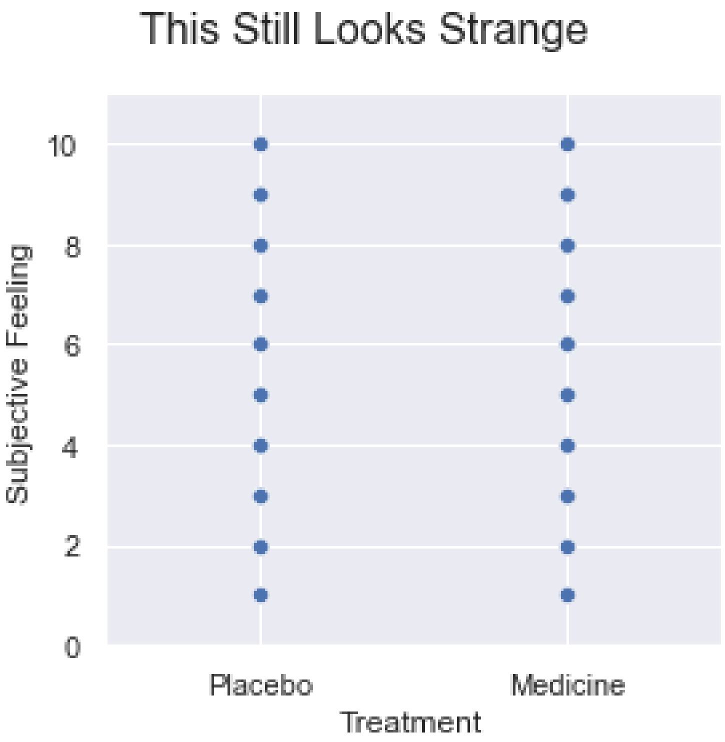
```
   Turnout  Num Music Halls
0    50.97                0
1    63.63                3
2    48.28                1
3    38.04                0
4    32.15                0
```

Each observation in our made-up data is a Texas county, so we have 254 data points (I've only shown the first 5). Turnout is the percent voter turnout in the 2020 election, and "Num Music Halls" is the number of places in the county that host musical performances. We want to find out if there's a relationship between the number of music halls in a county and its voter turnout in the 2020 election. Here's how the graphics above look when used to visualize this data. Notice how the fact that we have 14 different groups (because counties have anywhere from 0 to 13 music halls) affects how these look.

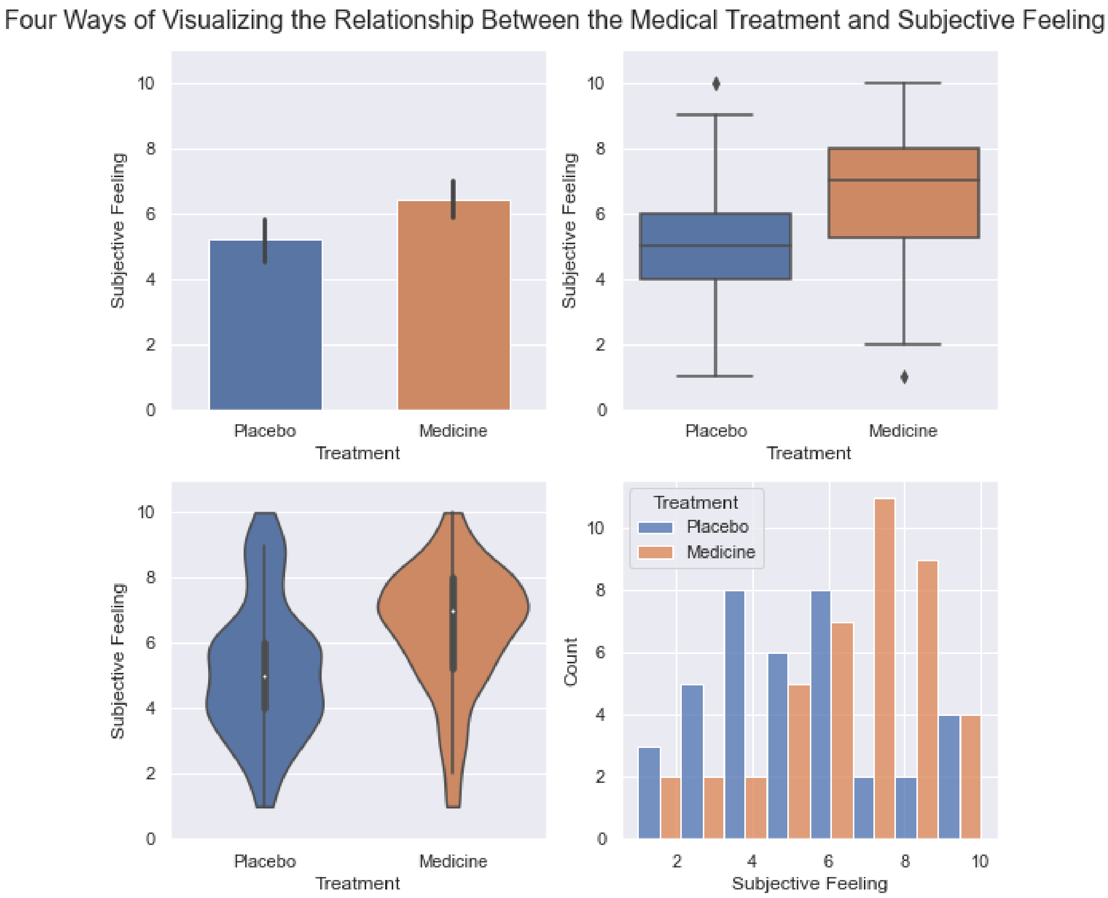


### A Categorical or Discrete Variable and another Categorical or Discrete Variable

I'm picking up on the example given above at the beginning of the **Categorical and Discrete Variables** section. We saw up there that a scatterplot with subjective feeling on the y-axis and treatment (placebo or medication) on the x-axis was useless. I'll reproduce it here:



What can we do about this? You *could* use a bar graph to plot the means of the two groups, but better options are the box plot and the violin plot. I've plotted all three below.



All four of these graphics are better than the scatter plot, in this case, because they indicate that patients who received the treatment are, on average, better off.

The **bar plot** on the top left is showing the average subjective feeling for each group. That line on the top of each bar graph is also informative: it's a 95% confidence interval for the means. To learn more about what that is (and isn't), see https://en.wikipedia.org/wiki/Confidence_interval.

The top right is a **box plot** or a box and whiskers plot. The line in the middle of each box is the median subjective feeling for the group, and the top and bottom lines of each box are the 75th and 25th percentiles for the group. The advantage of this graphic over the bar plot of the means is that it communicates more information about the distribution of subjective feeling reports for each group. The bar plot only tells us the average feeling of each group, while the box plot gives us a sense of how members of each group vary from the average.

The bottom left graphic is the **violin** plot. That tiny black line in the middle is the same as the box plot: the white dot is the median, the top and bottom of the thick part of the line is the top and bottom of the box, and the top and bottom of the thin part of the line is the top and bottom whisker. But in addition to that, the violin plot gives a smoothed estimate of the distribution for the subjective feeling variable. At wider points in the shape, we observed that value more often, and at narrower points in the shape, we observed that value less often. This is often very useful, since we get even more information about the variable, but there's a problem here: subjective feeling is a discrete variable, but the smooth lines of the violin plot suggest that it can take any value between 1 and 10 (like 4.65). Hence the violin plot **isn't** an excellent choice here, as it is in the case where you're measuring a continuous variable.

Finally, the bottom right graphic is a **histogram** of the subjective feeling variable, with colors coding whether the patient received the treatment or the placebo. Like the last two graphics we've discussed, this one gives us information about the distribution of the subjective feeling variable for the two groups. This isn't a very pretty graph, but it gives a lot of low level information about what's going on, so it would be good to pair with the bar graph showing the group averages.

Which of these you choose depends on what you want to show. Do you want your reader to focus on the difference in the average between the two groups? In that case, use the **bar plot**. Do you want to show the reader the median of the two groups as well as some other basic information about the distribution of the measured variable for each group? In that case, use the **box plot**. The **violin plot** and **histogram** show more information about the distribution of the measured variable in each group than the previous two, but at the cost of complexity.

A Second Example Let's look at one more case. Say you want to find out whether people who go to church more often also go to the basketball games more often. Your data for both church attendance and basketball attendance is of the form

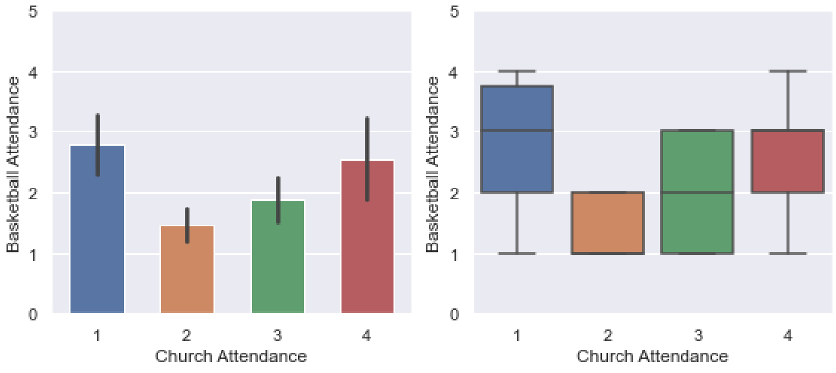
1. I go more than once a week
2. I go at least once a week
3. I go at least once a month
4. I go at least once a year

Hence you have two ordinal variables. Here's the data table:

```
   Church Attendance  Basketball Attendance
0                  3                      3
1                  3                      3
2                  3                      1
3                  1                      3
4                  2                      2
```

You could still use a bar graph for this, but the caveat would be that the averages represented for each group in your bar graph would be a little weird since your ordinal variables aren't numbers.

Two Ways of Visualizing the Relationship Between Church Attendance and Basketball Game Attendance



Pardon my laziness. The axes should have words instead of numbers (‘once a week’ instead of 2, for example).

Both of these work to communicate the relationship here, namely that those who attend church every day don’t have time to go to basketball games very much, but that those who attend church less than once a day attend more basketball games if the attend church more often.

Note, however, that the box plot is less useful if your y-axis variable is categorical or discrete with a small number of values. When that’s the case, there are fewer values the lines in the box plot can take, and sometimes you’ll find that the boxes for two x-values are the same, even though the distribution of y-values is importantly different. In this case, I would go with the *bar plot* if I wanted a simple visualization of the result.

# Treatment/Intervention Research

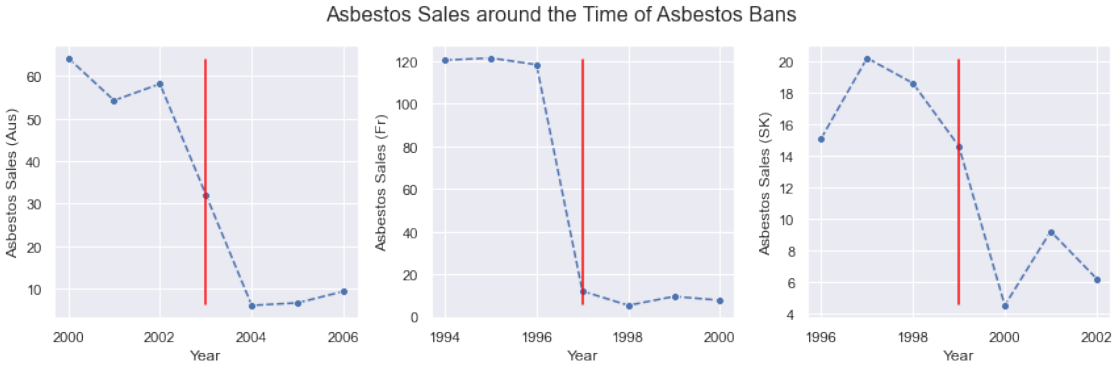
A few of you have research questions of the form “Did Y change after X happened?”, where X is some one-time event, usually a law being passed.

If your question is like this, you probably have time-series data, and the times your data covers are before and after the event you’re interested in happened. I’m treating you separately from the time-series data folks, however, because you’ll probably end up having at least one graphic that looks like a time-series graphic, and at least one graphic that looks like a correlational graphic. Let’s walk through an example.

Say your research question is “How much did asbestos use decrease as a result of asbestos ban laws?”. You decide to look at asbestos sales data in three countires that have banned asbestos, South Korea, France, and Australia. Since the countries banned asbestos at different times (1999, 1997, 2003, respectively, although this is oversimplifying matters), you’ll look at asbestos sales data for three different periods in the countries. In the case of South Korea, you’ll look at sales in 1996-2002, in the case of France, you’ll look at 1994-2000, and in the case of Australia, you’ll look at 2000-2006. Here’s the made up data.

```
   Year  Asbestos Sales (Aus)
0  2000                  64.3
1  2001                  54.3
2  2002                  58.2
3  2003                  32.1
4  2004                   6.0
5  2005                   6.6
6  2006                   9.3
```

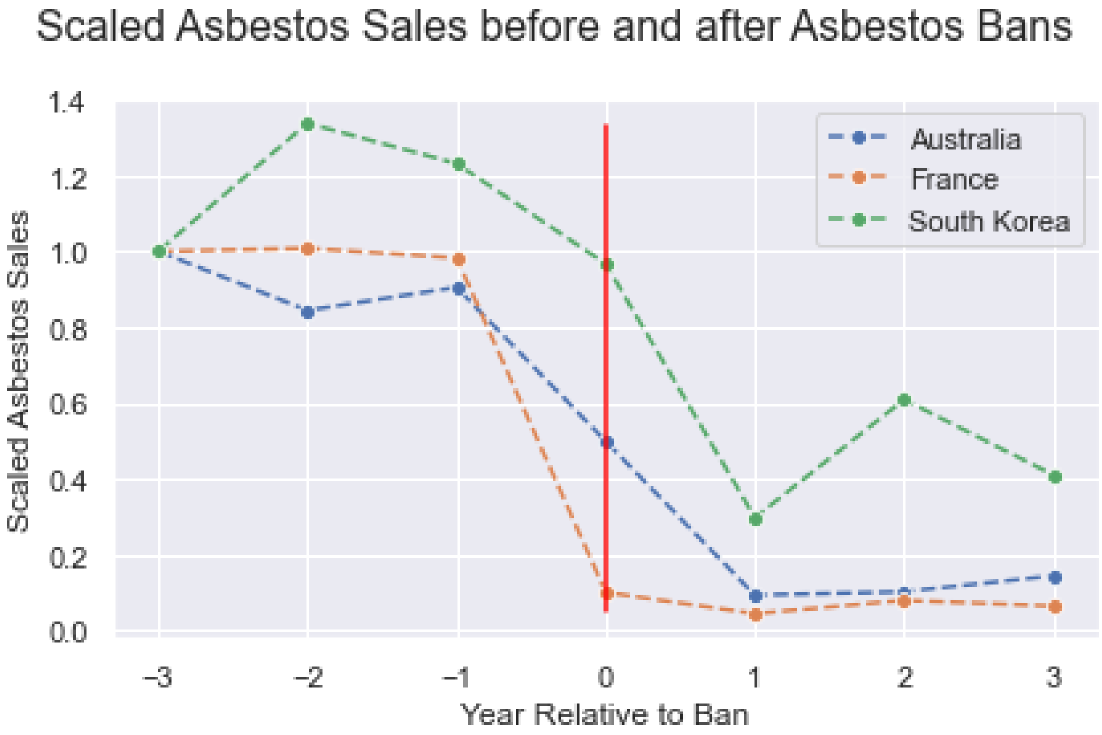
I’ve only shown the Australia data, but the France and South Korea data would be of the same sort. One type of graphic you’ll want to include is one that just shows the trend over time of asbestos sales for each country. There are two main ways to do this. One would be to make a graphic for each country. That would look like this:



Pardon my laziness. These graphs should have units for the y-axes, and they should have legends indicating that the red line is placed on the year the asbestos ban was passed into law in each country.

You'd include graphics like these if you want to give your readers a starting intuition for how your variable of interest (asbestos sales) changed around the time of the ban in each country.

Suppose, however, that you had data from 8 countries. It's probably a waste of space to include 8 separate graphics of this sort, so what you could do instead is plot them all on one graph. See below, for what this would look like in this case (I'm sticking with just 3 countries, because making up new data is a pain, but the graph below should get the point across).



Note that to make this work I had to do two things. First, I scaled the asbestos sales values for each country by dividing the values by the value from the first year in the dataset. That's why each country's asbestos sales starts at 1.0. Second, I rescaled the year values so that the year of the ban was 0.

Also note that if we had 20 countries, we would need to have 20 lines, which would look very messy. If you're in that situation, you might choose to plot a couple of the countries at random, or you might plot the average for each year across all the countries. Use your judgment.

The **second sort of graphic you'll want to include** is one that shows broadly the difference before and after the asbestos ban took place in each country. To my eyes, the most attractive way to do this is to take the average of each countries before-ban sales and post-ban sales, and then to plot these two groups against a binary treatment variables. This might sound complicated. All I mean is that you'll turn your data into something like this:

```
Country   |   Pre-Ban Scaled Sales Avg  |   Post-Ban Scaled Sales Avg

Australia |                .  .  .      |                        .  .  .

France    |                .  .  .      |                        .  .  .

S Korea   |                .  .  .      |                        .  .  .
```

You'd of course replace the '. . .' in each case with the relative values. Note that the averages should be of the scaled values. If you don't scale the values, you could end up with a very wonky looking graph, because France used a lot more asbestos than South Korea before the ban (according to my fictitious data).

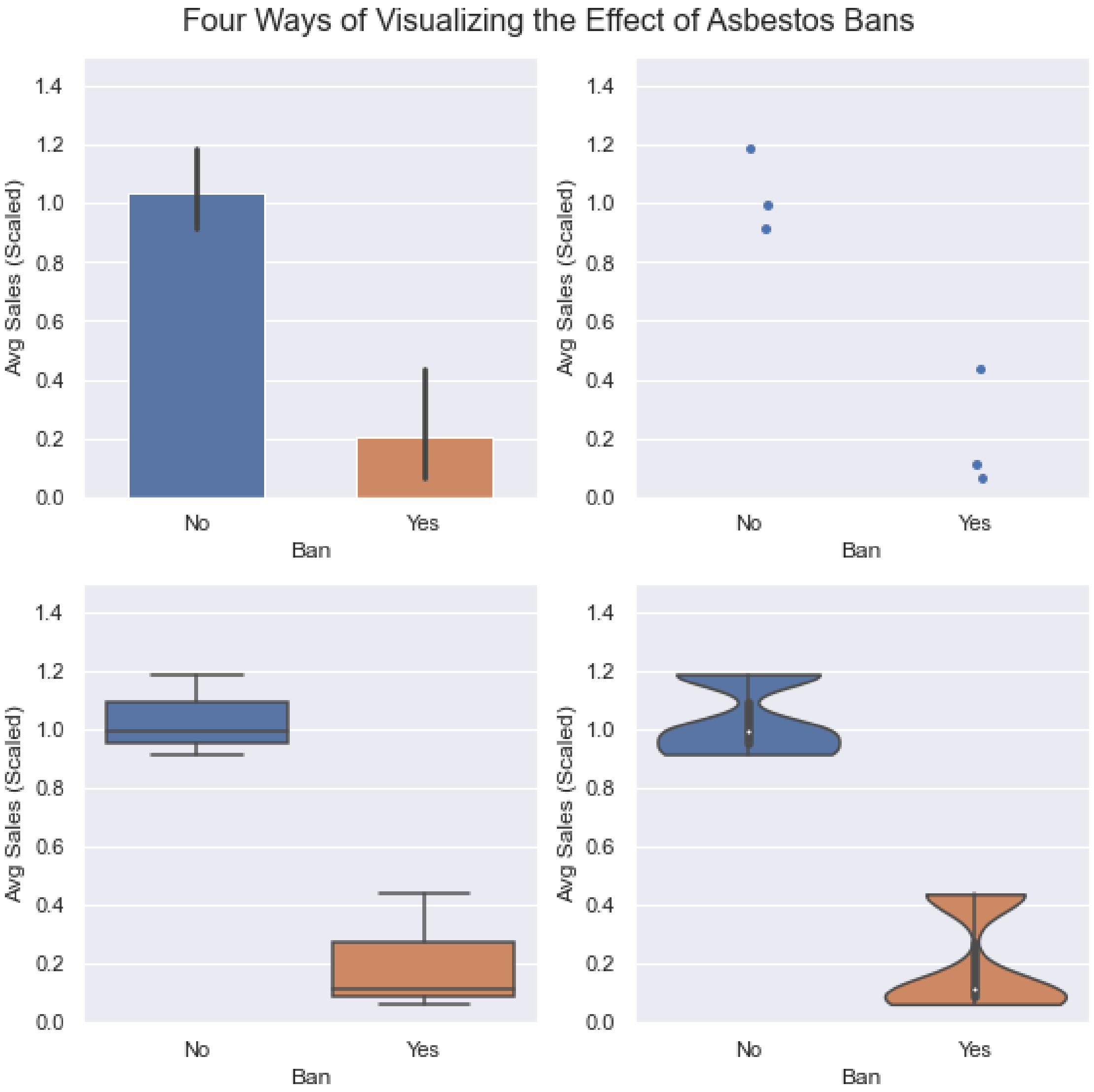
You'll then visualize this data in the same way that you'd visualize the relationship between a binary variable and a continuous variables. For more on this, see the section **A Continuous Variable and a Categorical or Discrete Variable** above.

I'll show what this looks like here in the case of our asbestos data. Here's our new data table:

```
  Country  Ban  Avg Sales (Scaled)
0     Aus   No            0.996952
1      Fr   No            0.916537
2      SK   No            1.189845
3     Aus  Yes            0.063175
4      Fr  Yes            0.113530
5      SK  Yes            0.439294
```

Note that I've formatted the new data table differently than I described it above. This is to make it clearer that we have a new binary variable that records whether the avg sales value is from before or after the asbestos ban became law.

Here's how you could visualize the effect of the bans overall:



For information about what these graphics are called and what their pros and cons are, see **A Continuous Variable and a Categorical or Discrete Variable**.

In this example, where we have only 3 data points for each group (pre-Ban and post-Ban), I think the bar plot and scatter plot are the superior visualizations. The scatter plot gives us a direct look at the data, and the bar plot tells us the averages of the pre-ban and post-ban groups, which is the most important single statistic for comparing two groups. Because there is so little data, the box plot and violin plot are somewhat misleading. We don't have enough data to estimate the 25th and 75th percentile markers for each group, and we certainly don't have enough data to estimate what the distribution of avg sales is for each group, but the box plot and the violin plot present those respective estimates as if there weren't a problem. If, however, we had data for 30 countries, then the box plot and violin plot would be very good choices. In general, you shouldn't use the box or violin plots if you have a very small number of data points (< ~10 each) for your pre- and post-intervention groups.

# Time Series Data

If your research question has anything to do with change over time, then you probably have time series data. In fact, you might have time series data even if your question isn't about change over time. You have time series data if 1) your data points come from different periods of time and 2) the difference in time is significant, is *part of the data*.

Examples of time series data include voter turnout in a county over time, prison population over time, and water consumption over time. Any data that you would describe as something "over time" is time series data.

Time series data is a little bit tricky because for each object or person we're looking at (e.g. citizens of Texas, Texas counties, or corporations), we have one or more variables recorded at two or more points in time. This means that visualizations of time series data can get messy if we aren't careful about what we plot and how we plot it.

Before getting into visualization recommendations, I want to introduce the vocabulary I'll be using to describe time series data. Consider these two example data sets:

Example 1: You have data that includes avg income in each state in the US, recorded yearly from 1970 to 2022.

Example 2: You have data that includes number of early voting locations the voter turnout in Travis county for each presidential election from 2004 to 2020.

In **example 1**, you have one variable, measured for each of the 50 states, over 52 years. In **example 2**, you have two variables, measured in only one county, over 5 years (since prez elections are once every 4 years). The unit of analysis in **example 1** is states, while the unit of analysis in **example 2** is counties. But while there are 50 instances of your unit of analysis in **example 1**, there is only 1 instance of your unit of analyis in **example 2**, namely Travis county. I'm going to call the number of instances of your unit analysis the *number of objects* described in your data. For each of the objects in your data (whether you have one or several), you have records of the value of each variable in your data at each time step. Thus, when I talk about *variables*, I don't mean *objects*. Your time series data could have one object and multiple variables just as well as it could have multiple objects but only one variables. Just to be sure, I'll give one more example of time series data. This one includes multiple objects and multiple variables.

Example 3: You have data on voter turnout and partisanhip for Black Texans, White Texans, Latino Texans, and Asian Texans for each election year from 2000 to 2016.

Each race/ethnic group is an *object* in your data. Your data thus has 4 objects. For each of those objects, at each point in time, you have the values of the two *variables*, partisanship and voter turnout.

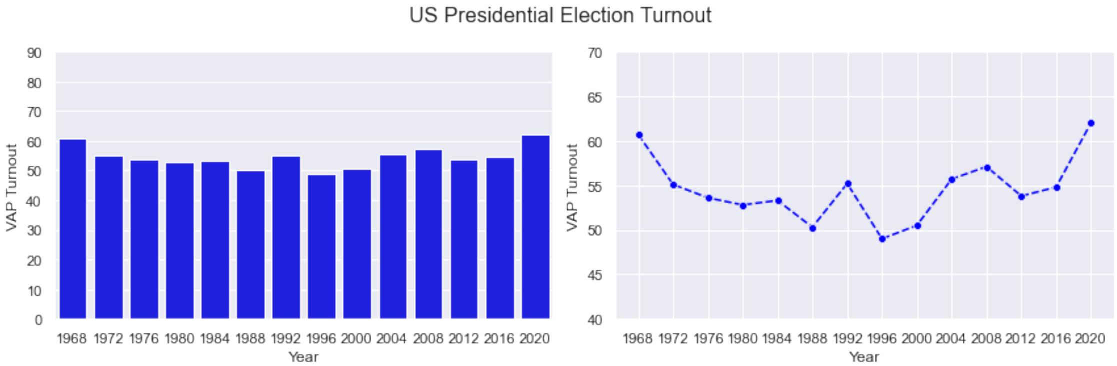
The next two sections will give recommendations for visualizing time series data that only has **one object**. The final section will give advice on how to visualize time series data with **multiple objects**. Statistical inference about time series data is generally trickier than inference about other types of data, so I'm going to limit my discussion here to visualizations that are useful simply for displaying your data neatly.

## Plotting One Variable

Plotting one variable with one object is pretty straightforward. Consider the data below.

```
   Year  VAP Turnout  Sugar Consumption
0  1968         60.7                325
1  1972         55.1                333
2  1976         53.6                328
3  1980         52.8                330
4  1984         53.3                340
```

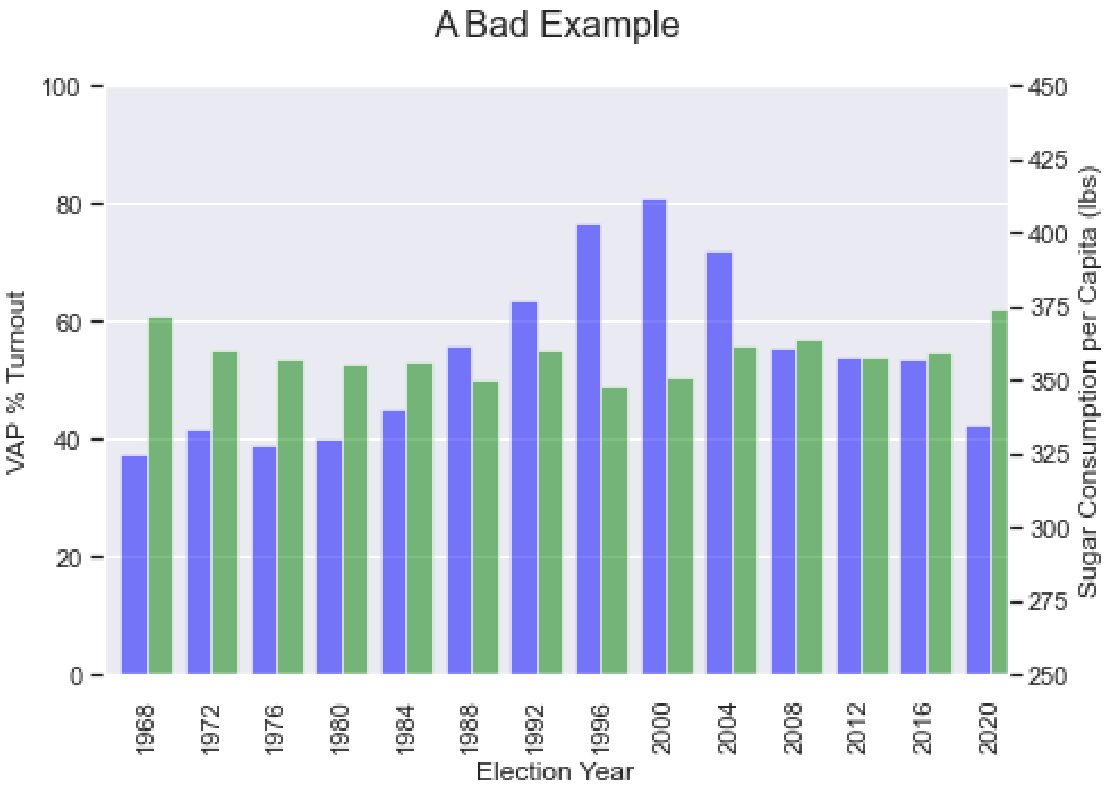
For each US presidential election year from 1968 to 2020, we have voter turnout as a percent of the US voting age population and the average sugar consumption (in calories per day) for US citizens. I've displayed the first 5 data entries. Suppose we want to plot just the turnout figures. Either of the following works:



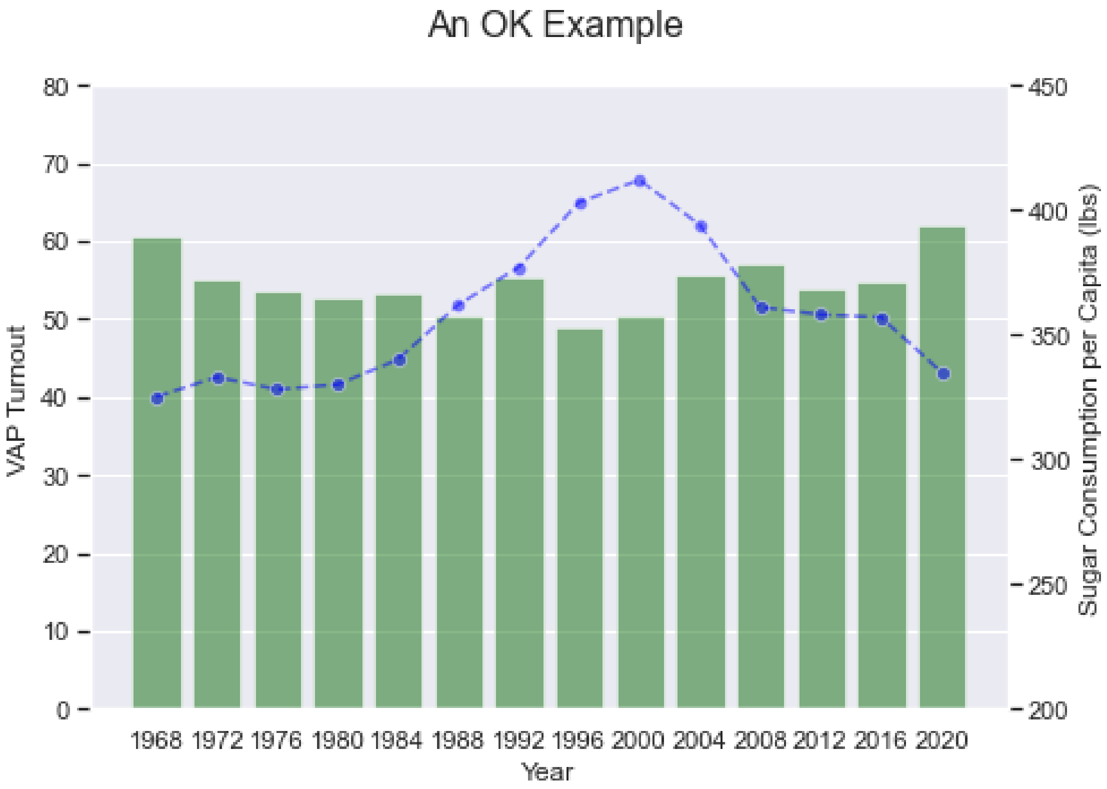
Note the difference in scales for the y-axes of the figures. Choosing a scale is up to you. Note that, when comparing the figures, the scale on the line graph gives a clearer view of the variation in voter turnout but can also, for that very reason, give the impression that voter turnout in presidential elections has varied over the years more than it in fact has.

## Plotting Multiple Variables

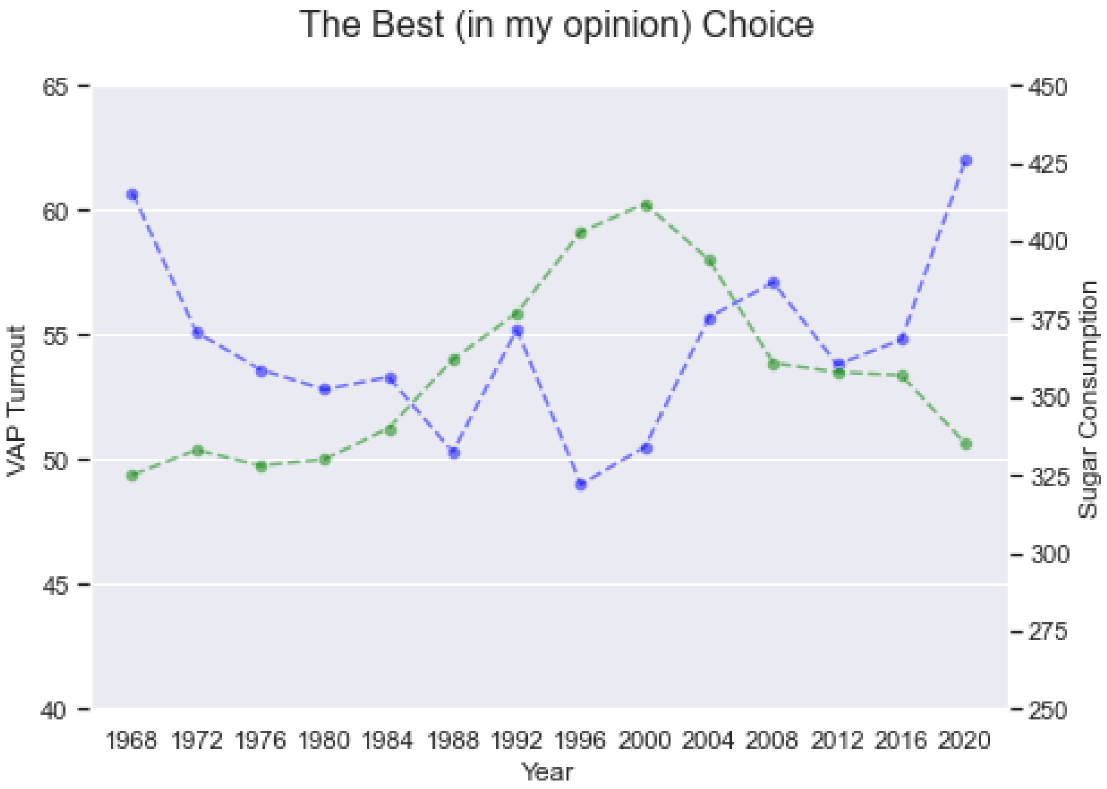
Now let's plot the sugar consumption variable in the same graph with the voter turnout variables. We don't want to do this with a bar graph. See what this looks like below.



This isn't awful, but it's a bit confusing. The following two visualizations are usually better.



This (the bar + line graph above) is better. It's easier for the eye to follow each variable separate over the years. These sorts of plots can be somewhat misleading, however, since they plot in different ways (as a line and as a bar) variables that are of the same type. I think the next visualization is best of these three.



This (the double line graph above) is what I recommend if you're plotting two or more variables from time series data. Both variables here are represented as lines, so one isn't led into thinking that they're serving fundamentally different purposes in the graph, and the eye has no trouble following the lines separately and looking at their relationship.

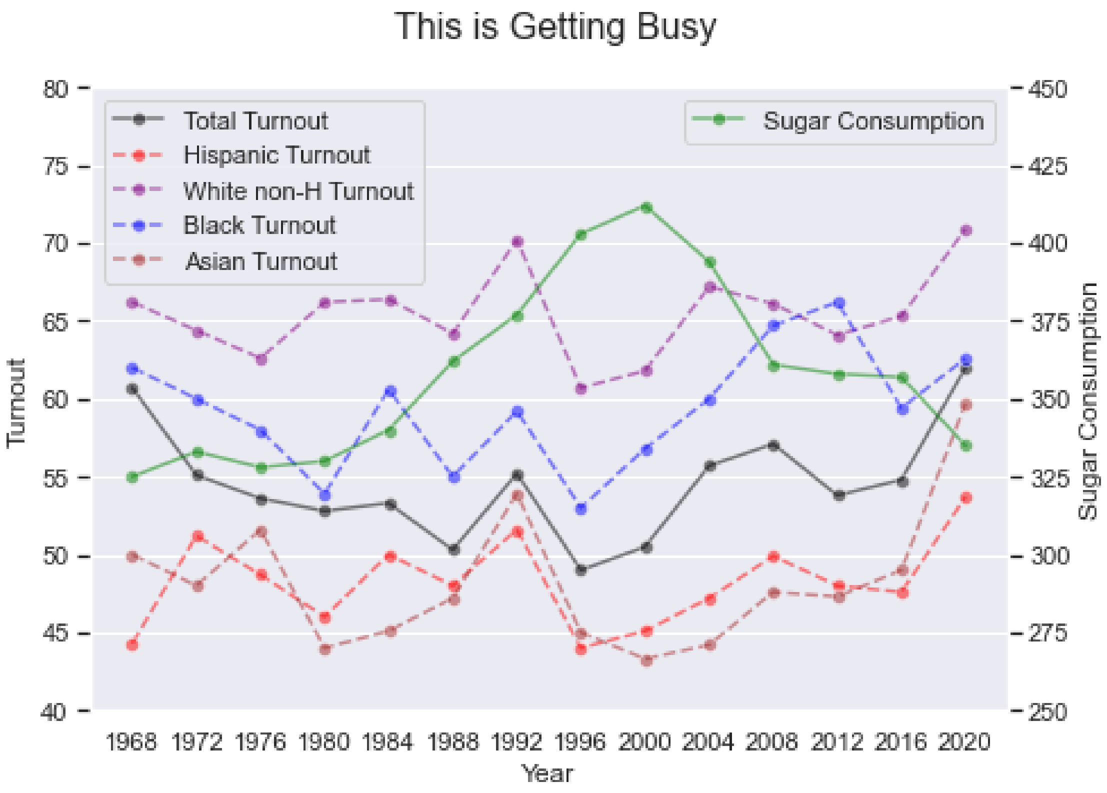
Big Note of Caution: When you plot two or more variables that use different scales (in this case, Turnout has one scale, on the left, and sugar consumption has another, on the right), your choice of scale for each variable has a large effect on the relationship the eye picks up between the plotted values. I don't think this is as dangerous as some make it out to be, but be aware of it, and please don't use absurd scales to paint a misleading picture of things. To illustrate what I'm talking about: consider that if I had made the sugar consumption scale go from 0 to 600, the sugar consumption line would move quite a bit less in the plot, and it would seem to have a different relationship with the voting variable (keep in mind that *any* relationship you might see here is almost certainly spurious–I haven't encountered any theories of higher sugar consumption decreasing voting, voting decreasing sugar consumption, or something else affecting them both in a predictable way).

## Plotting One or More Variables with Multiple Objects (e.g., multiple counties or demographic groups)

Visualizing time series data becomes more difficult if you have multiple objects. Suppose you had the same data as above (American voter turnout and average American sugar consumption for each US presidential election year), but instead of just voter turnout, you had voter turnout for Black, White, Latino, and Asian Americans. How should you plot that? Let's see what happens if we naively plot that data with line graphs, like in the previous example (see the most recent graphic above). Here's our fictitious data:

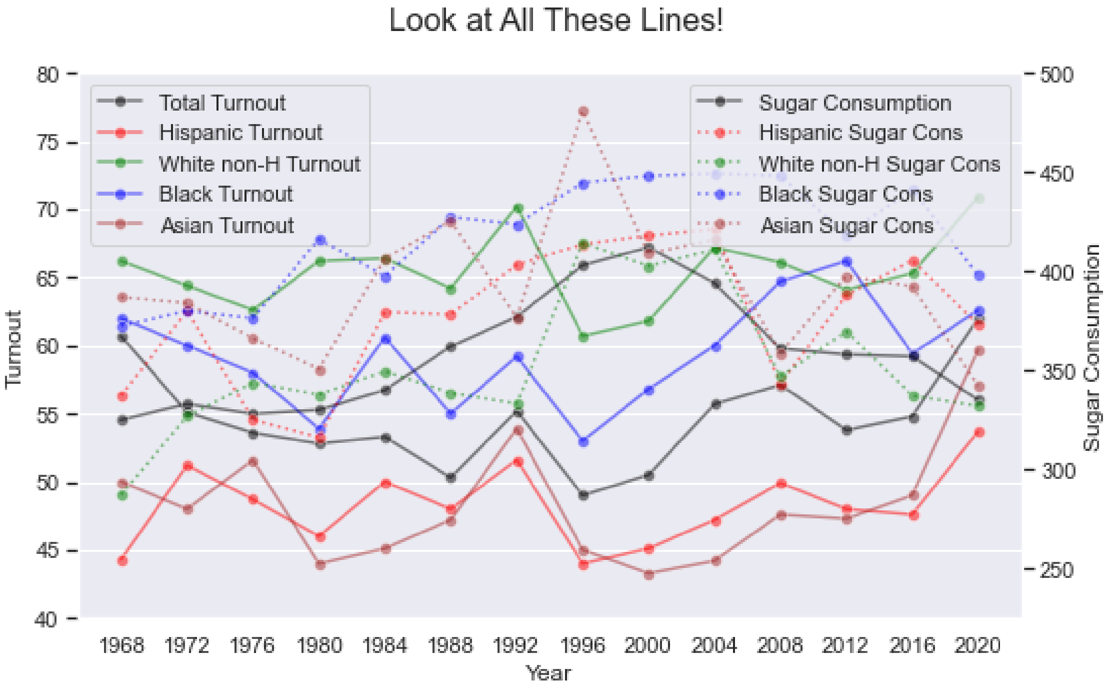
```
   Year  VAP Turnout  Sugar Consumption  White H Turnout  White non-H Turnout  \
0  1968         60.7                325             44.3                 66.2
1  1972         55.1                333             51.2                 64.4
2  1976         53.6                328             48.7                 62.6
3  1980         52.8                330             46.0                 66.2
4  1984         53.3                340             50.0                 66.4

   Black Turnout  Asian Turnout
0           62.0           50.0
1           60.0           48.0
2           58.0           51.6
3           53.9           44.0
4           60.6           45.1
```

And here's what it looks like if we plot it the same way we plotted the simple voter turnout and sugar consumption data:



Even 6 lines has made our graph a little difficult to view. To be sure, we can get useful information from it. For example, after our eyes adjust, we get the impression that non-hispanic white Americans tend to vote the most, followed by Black Americans, with Asian and Hispanic Americans generally voting the least.

Imagine how disorienting this graph would be, however, if we were measuring sugar consumption for each demographic group: we'd have 4 objects with 2 variables each, which would mean 8 lines on our graph, and if we continued to plot total turnout and sugar consumption, we'd have 10 lines! Let's see what that looks like.



At this point, getting information from the graph is difficult without studying it. Imagine how disorienting it would be if we had three variables for each group. What should we do in these sorts of cases?

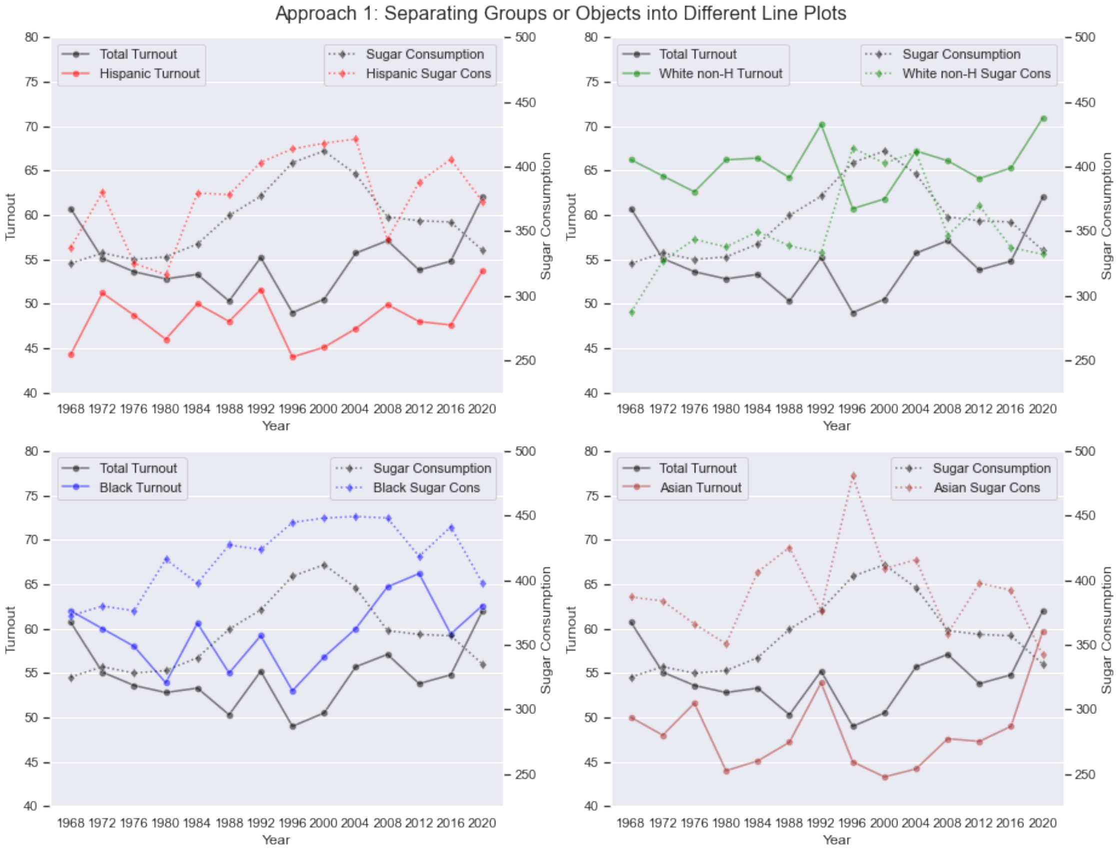
In general, how you should plot high-dimensional data depends on how big the different dimensions are (that is, how many points in time your data includes, how many objects or groups your data is tracking, and how many variables you are tracking for each object or group) and what exactly you want to focus your reader on. The line graph approach above fails when we have both a large number of snapshots in time (in this case we have 14) and a medium or large number of variable or objects (in this case, we're plotting 6 lines). On top of this, the more your lines overlap, the more difficult your graph will be to visually comprehend.

The first thing you should consider is whether you need to graph all of your data in the first place. Could you make the point you want to make to your reader by plotting fewer of the years, or by plotting fewer of the demographic groups? If you can, then do that, and protect your readers from eye strain.

Along with this, consider what it is you want your reader to focus on. Do you want your reader to focus on how the groups have been changing over time relative to each other? If so, then you probably should stick with a line graph of the form given above, and do your best to reduce the complexity of the graph. If, however, you want your reader to focus more on how each group is changing over time than on how they are changing relative to each other, then you could try one of these two approaches:

1. Plot separate line graphs for each group (including all of the variables measured for each group)
2. Plot bar graphs by group (This is only viable if you have a small number of time steps in your data)

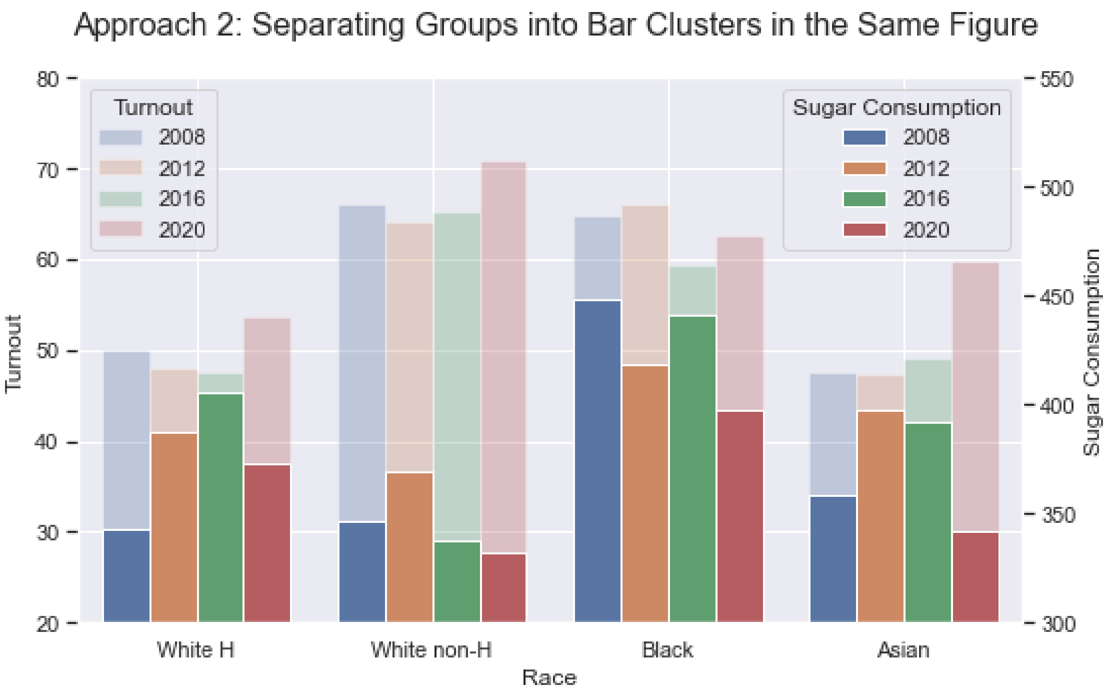
Here's what approach 1 looks like for this data.



Approach 1: Separating Groups or Objects into Different Line Plots

You sacrifice comparison between groups by taking approach 1, but each individual graph is cleaner. If you didn't care about intergroup comparison at all, you could even remove the Total Turnout and Total Sugar Consumption lines.

Here's approach 2.



Note that to make approach 2 work, I only plotted four of the election years (2008, 2012, 2016, and 2020). Approach 2 won't work if you have many more than 3 times steps because the groups of bars will get very dense, and your graph will become difficult to read. It also won't work if you're measuring more than 2 variables per group.

Generally, approach 2 is worth considering if you have a small number of time steps. If you compare approach 2 to approach 1, you'll see that they're not so different: approach 2 squeezes all the figures in approach 1 into one figure and replaces the lines with bars. While you sacrifice the detail and timescale of approach 1, you have the neatness of only needing one graph.

# How to Make These Plots in Excel

Line Plot: https://support.microsoft.com/en-us/topic/present-your-data-in-a-scatter-chart-or-a-line-chart-4570a80f-599a-4d6b-a155-104a9018b86e#OfficeVersion=Windows

Scatter Plot: https://support.microsoft.com/en-us/topic/present-your-data-in-a-scatter-chart-or-a-line-chart-4570a80f-599a-4d6b-a155-104a9018b86e#OfficeVersion=Windows

Box Plot: https://support.microsoft.com/en-us/office/create-a-box-and-whisker-chart-62f4219f-db4b-4754-aca8-4743f6190f0d#OfficeVersion=Windows

Histogram: https://support.microsoft.com/en-us/office/create-a-histogram-85680173-064b-4024-b39d-80f17ff2f4e8#OfficeVersion=Windows

Bar Plot: https://support.microsoft.com/en-us/office/present-your-data-in-a-column-chart-d89050ba-e6b6-47de-b090-e9ab353c4c00

Violin Plot (this requires a 3rd party add-on): https://help.xlstat.com/6377-violin-plot-excel#:~:text=What%20is%20a%20violin%20plot,in%20R%20(Wickham%20H)