

# Exploratory Vistas

## Ways to Become Acquainted with a Data Set for the First Time

Stephen Few, Perceptual Edge  
*Visual Business Intelligence Newsletter*  
July/August/September 2011

When you lay eyes on an unfamiliar territory for the first time, it always makes sense to get an overview before venturing into the thick of it. An unfamiliar data set is like an unknown land. Unless you have unlimited time to wander (who does?) and don't mind getting lost, it helps to study a map before starting out. If no map exists, then you should head for high ground to get the lay of the land.



Analytical journeys are quite different from vacationing in Italy or France. Leisure travel involves a series of destinations with the hope of enjoying ourselves along the way. We want to add sweet memories to our lives of the great meals, beautiful sites, and interesting people we meet along the way. When we embark on an analytical journey, however, we don't pre-plan all our destinations and we're not just collecting memories. The goal of the analytical journey is a thorough understanding of the territory, often to solve specific problems. We approach the journey as if we're preparing to become tour guides, able to explain each site in a way that ranges from its history to predictions about its future. We must learn to navigate like a native.

I speak often of *Shneiderman's Mantra* for visual information seeking: "Overview first, zoom and filter, then details-on-demand" (Ben Shneiderman, "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations," *Proceedings of the 1996 IEEE Symposium on Visual Languages*). Although this brief expression does not describe all aspects of data sensemaking, it proposes a means of analytical navigation that is always useful. In this article, I'll focus on a particular aspect of data sensemaking that is extremely important: the start of the journey. It is always important to begin with an overview of unexplored territory ("overview first") before diving into the details. I'll propose, describe, and illustrate six useful ways to get that overview, including five sets of data visualizations and one technique for data interaction, which together provide an effective introduction to unfamiliar data. These will help you examine the following fundamental relationships that exist among and within a data set's primary variables, both quantitative (a.k.a. measures) and categorical (a.k.a. dimensions):

- How quantitative values are distributed across their full range (e.g., donations by size)
- How categorical items rank and contribute to the whole when associated with quantitative values (e.g., cities by population)

- How multiple variables (quantitative and categorical) intersect one another (e.g., what combination of region, product, customer type, and month produces the highest profit)
- How categorical items, when associated with quantitative values, contribute to the whole (e.g., total sales broken down by product)
- How quantitative values change through time (e.g., new incidences of a particular disease by day)
- If and how quantitative variables correlate to one another (e.g., whether increases in marketing expenditures correspond to an increase in sales revenues, a decrease in sales revenues, or there is no apparent relationship)
- Where things are located (e.g., employees by zip code)

## Preparation

Before slinging a pair of binoculars over your shoulder and hiking up to the overlook, some preparation is in order. In the realm of physical exploration, it helps to review a list of the sites and to learn a bit about the region's history before beginning the journey. In the realm of data exploration, good preparation involves learning about the domain—not only the data elements that inhabit it but also the processes that produce them. For instance, if you're preparing to explore sales information, you had better understand the sales process, including its goals, historical decisions and events that have influenced it, and the meanings of sales terms that correspond to the data (revenue, bookings, billings, and so on). This context will help you become quickly oriented once you hike up to the overlook and begin to look around.

In this article, we'll use a simple product sales data to illustrate effective ways to get an overview. It consists of the following data elements:

### Categorical Variables (Dimensions)

- Period (four quarters in the year 2008)
- Channel (the channel through which purchases were made)
- Product Line
- Customer ID

### Quantitative Variables (Measures)

- Revenue
- Cost of Sales
- Gross Profit (Revenue – Cost of Sales)
- Gross Profit % (Gross Profit / Revenue)

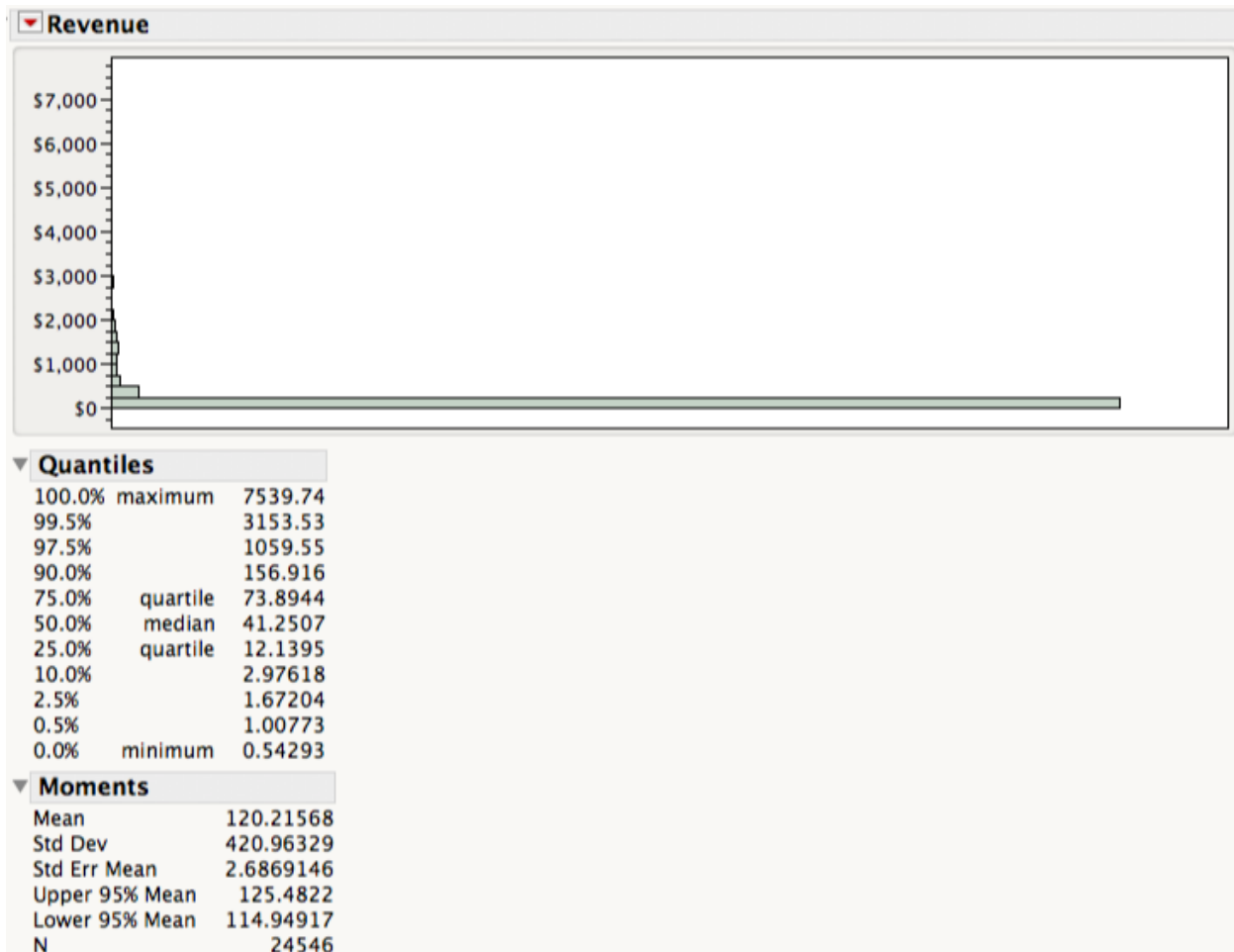
## Visualizations and Interactions for Building an Overview

The five sets of visualizations and one interaction technique consist of the following:

1. Histograms for viewing distributions
2. Brushing and linking for viewing interconnections among variables
3. Treemaps and bar graphs for viewing parts of the whole
4. Table lens displays for viewing rankings and correlations
5. Line graphs for viewing change through time
6. Geographical maps for viewing locations

## Histograms for Viewing Distributions

Perhaps the best place to start is to get a sense of the spread and shape (i.e., the distribution) of the values that make up each of the data set's primary quantitative variables. The graph that is most frequently used for seeing how a quantitative variable's values are distributed is the *histogram*. A histogram is a bar graph that's specifically designed to display a distribution by partitioning the full range of values (i.e., the spread) into intervals of equal size along one axis, and then counting the number of values in each of those intervals. Each bar represents the count of values that fall within an interval. In the figure below, we're looking at revenues represented in two ways: a histogram above and a list of statistics that describe various aspects of the distribution.

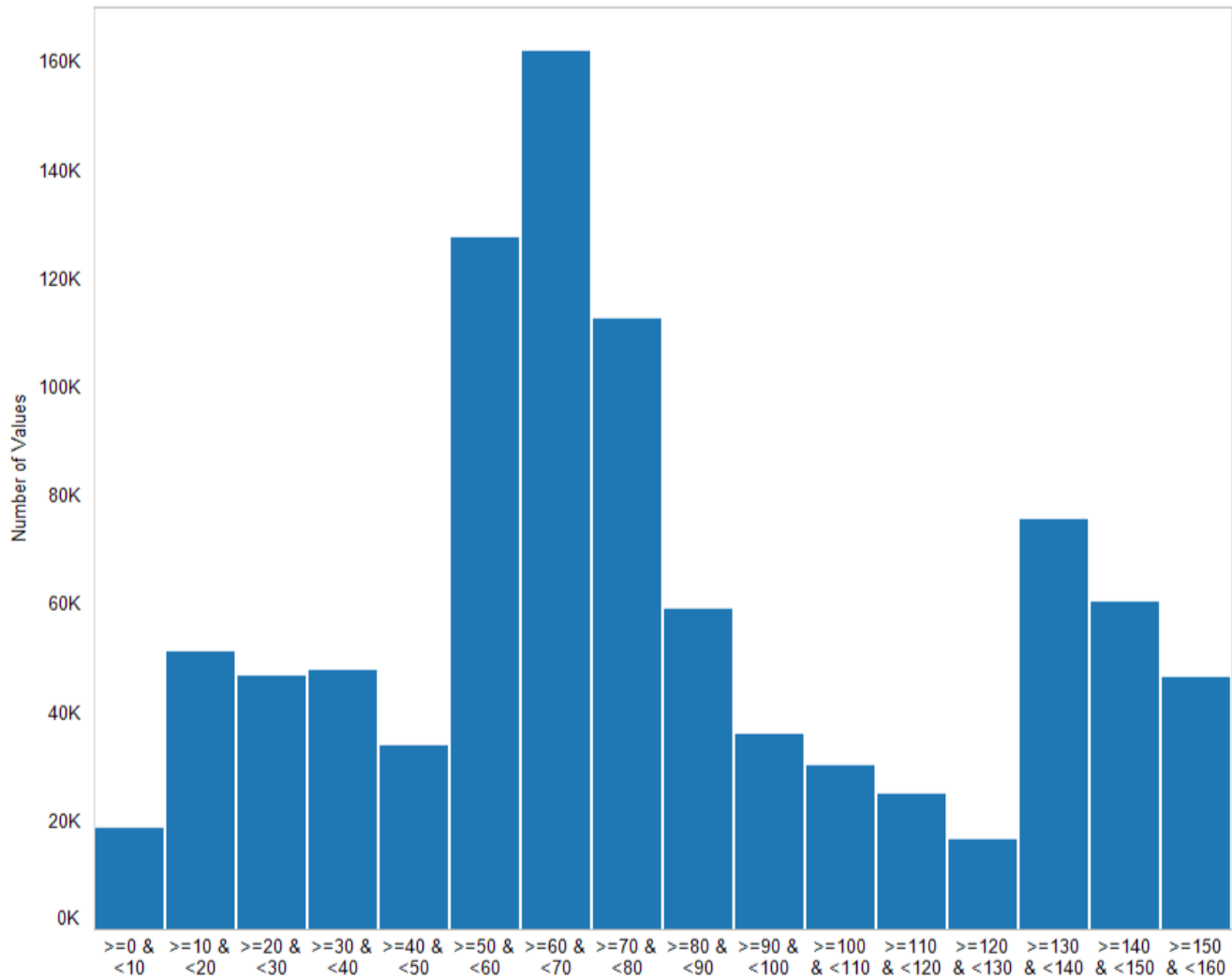


This is a standard display in SAS JMP for examining distributions.

The first thing we notice in the histogram is the one really long bar, which tells us that most revenue values fall within the low range of \$0 to \$250. This distribution of revenues is extremely skewed. Relatively few values are greater than \$2,000. The peak is at the low end, and then the values continue through a long tail that extends upwards past \$7,000.

According to the Quantile list, the maximum value is precisely \$7,539.74. While still looking at the list, let's find the other values that are of interest for getting an overview. First, opposed to the maximum value of \$7,539.74 is the minimum value of \$0.54, which tells us that the full spread of values is \$7,539.20. Both the mean of \$120.21 and median of \$41.25 provide measures of the distribution's center, but the median does a better job of telling us what's typical of revenues. The mean value of \$120.21 is much higher than what's typical, because it has been pulled upwards by extremely high, atypical revenues. We can tell from reading the 75th

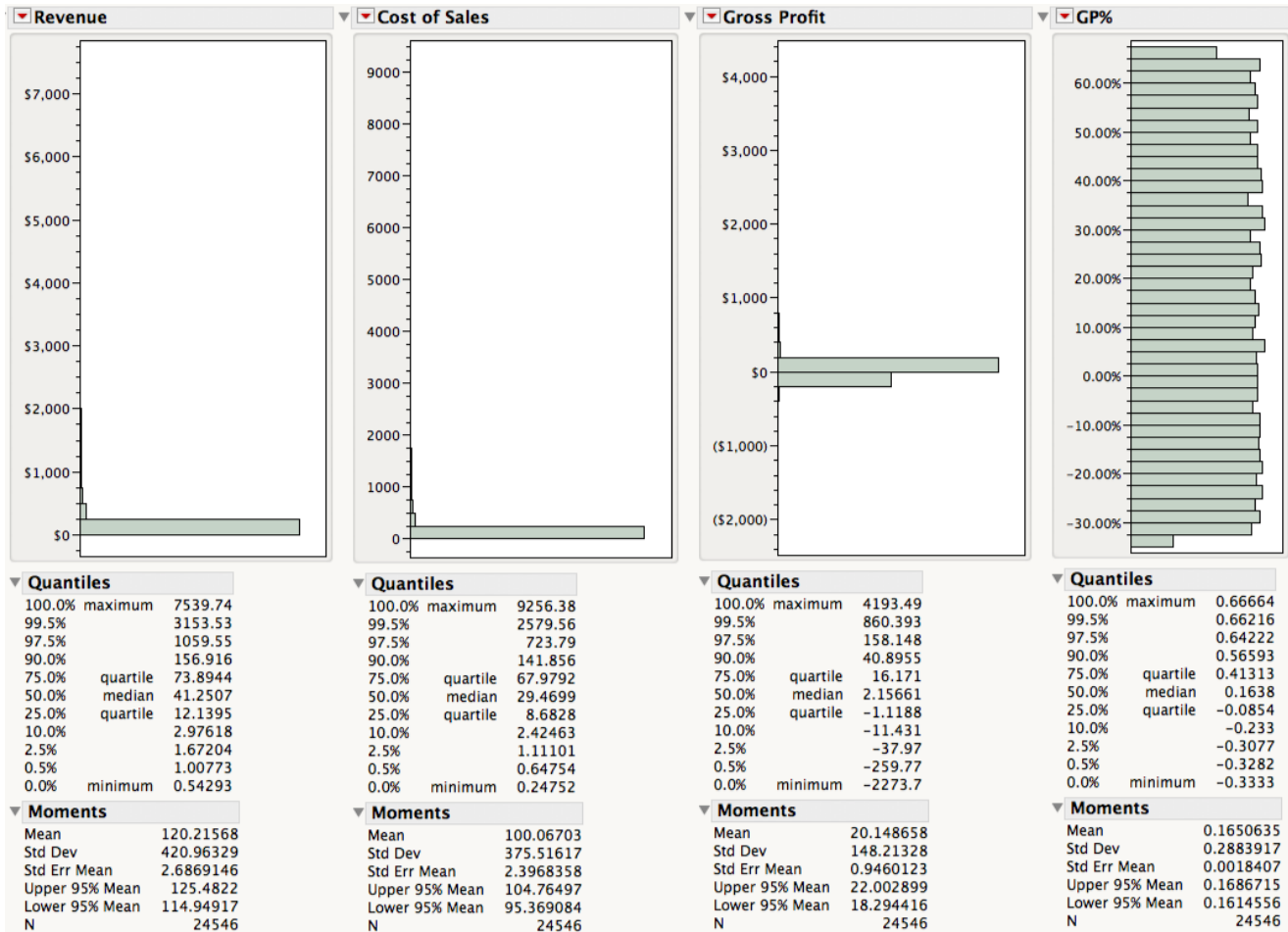
percentile value in the list that 75% of revenues are less than \$74. In fact, 90% are less than \$157. Because this distribution is so skewed, it would be useful to remove the outliers—perhaps every value about the 90th percentile of \$157—to better see how most values are distributed. In the next example, which I’ve reoriented by placing the intervals along the X axis so the bars are vertical, all values of \$160 and above have been filtered out. (Note: I chose \$160 as the upper limit because I’m using intervals of \$10 each, and I didn’t want to under represent the number of values in the final bin.)



This histogram was created using Tableau.

Now we see that there is a peak in revenues ranging from \$50 to just under \$80, with two lesser peaks ranging from \$130 to just under \$160 and from \$10 to just under \$40. Perhaps we should come back later and dig into the details to see if we can figure out what causes the peaks in these particular ranges. It would be interesting to see if particular product types or customers contribute to this effect, but we’re sticking with an overview for now.

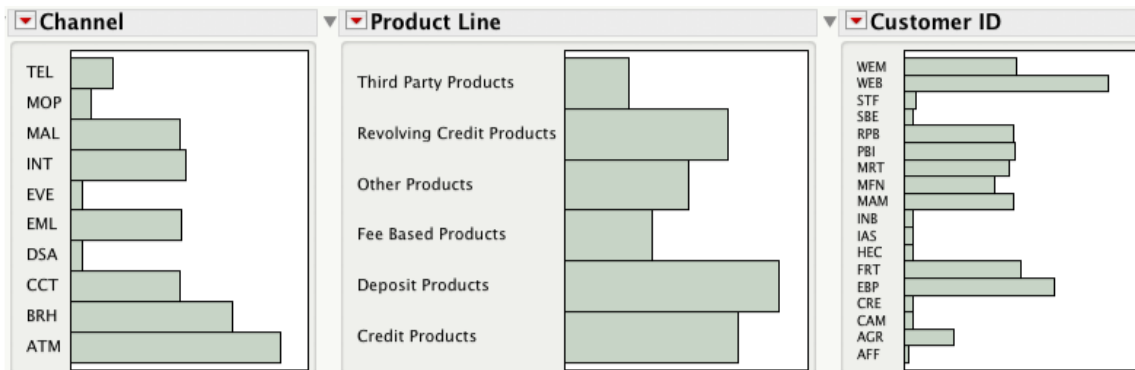
Rather than looking too closely at the distribution of a particular variable, let's stick to high-level overviews for now. In fact, let's look at the distributions of all four quantitative variables at the same time.



This example was created using the distribution feature of the product JMP, which make it easy to view several distributions at once.

At a glance, we can tell that Cost of Sales is similar to Revenue: its values are predominantly at the low end. Profits also bunch together as relatively small amounts that range only slightly above and below zero (80% are between \$41 and -\$12). Gross Profit % (GP%) is unlike the other variables: its distribution appears fairly uniform, ranging from 67% at the high end to -33% at the low end, with only minor variation in between.

To get acquainted with the categorical variables, we can construct a similar view that also uses bar graphs. Below, you see a series of bar graphs that count the number of rows in the data set that are associated with the members of each categorical variable, except for period, because time can be examined in better ways, which we'll get to later.

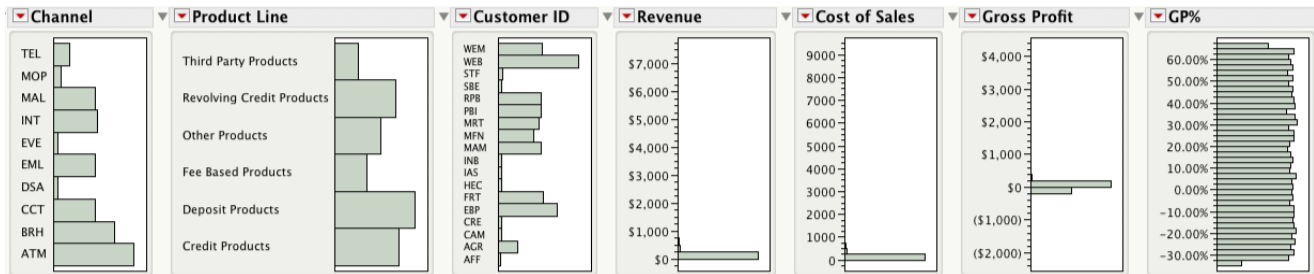


Several facts are immediately available. We can easily see if particular channels, product lines, or customers are dominant (e.g., more purchases are made through ATMs than any other channel and customers seem to prefer deposit, credit, and revolving credit products). We can also see if some contribute little (e.g., eight customers are much less active than the others). (Note: This view would work better if the bars in each of the graphs were sorted by value from large to small or vice versa, but this feature is not currently supported by the software.)

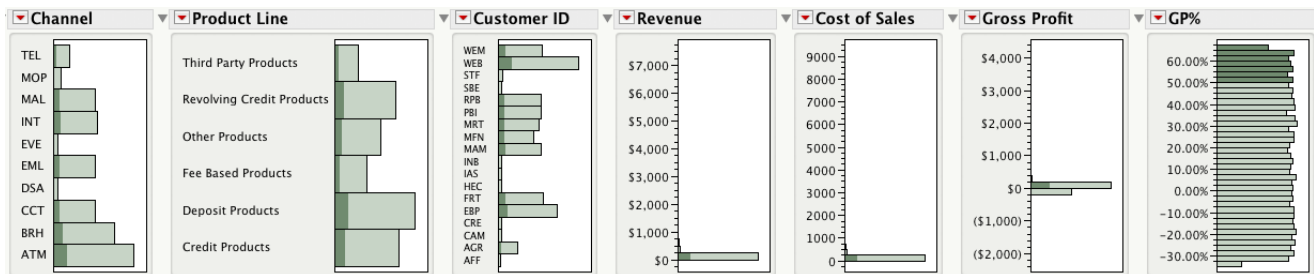
This is a good start, but it would be interesting to see how all of the variables interconnect. Our next topic, the data interaction technique known as brushing and linking, will help us see explore these interconnections.

### Brushing and Linking for Viewing Interconnections Among Variables

Let's begin by constructing a single display that includes all of the graphs that we've examined up until now. To see interconnections between these variables, we must view them at the same time.

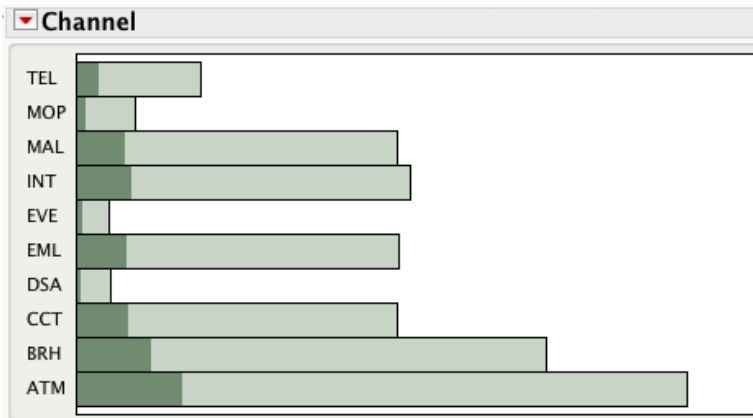


Let's start with a question that might be asked by an actual analyst of sales data: "What contributes to the highest profit margins?" To pursue this, let's select all of the bars in the GP% histogram on the right that represent values of 50% and above. Here's the result:

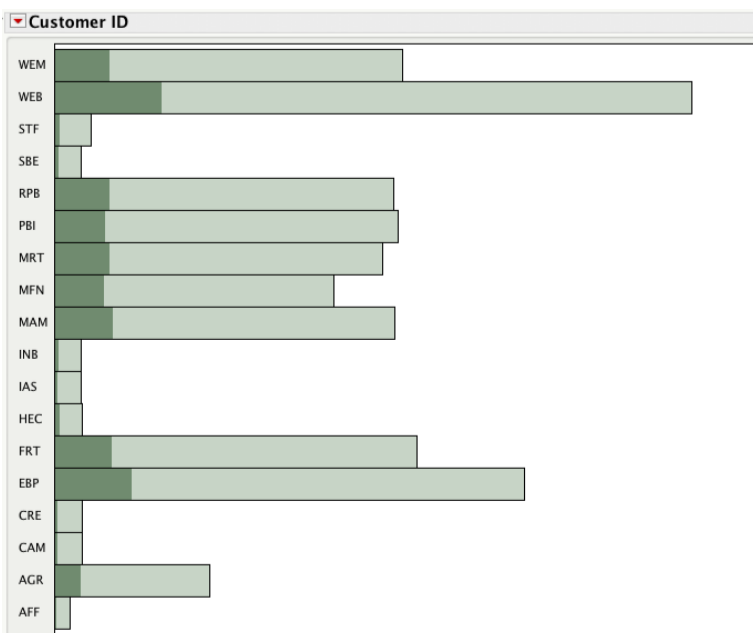


Notice that, not only are bars with values of 50% and above highlighted in dark green, but dark green portions of bars appear in every other graph as well. The act of selecting one subset of data (called brushing) has caused not only that subset to be highlighted but also the records that are associated with those high gross profit percentages in every other graph as well (called linking). Now we can see the interconnections between high profit margins and the other variables.

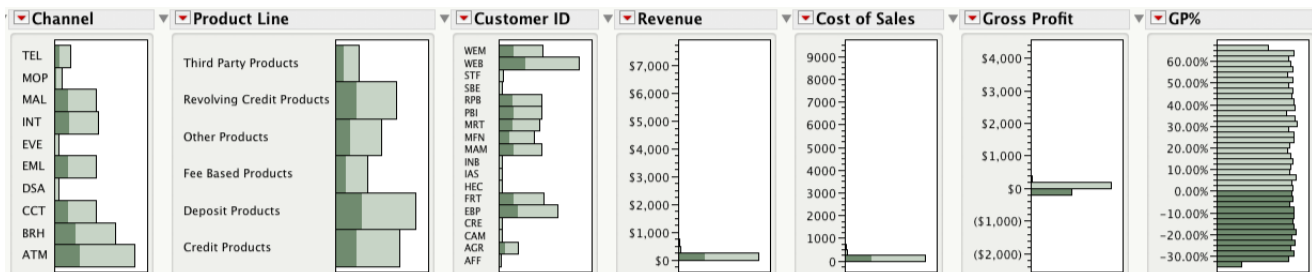
It isn't a surprise that high profit margins are not associated with gross profits under zero, as we can see in the Gross Profit histogram. Let's see if we can find an interconnection that's more interesting. Can you see any channels, product lines, or customers that are associated to a greater degree than others with high profit margins? Keep in mind that we're not just comparing the amounts of green highlighting in each of the bars, but the amounts in proportion to their bars. For example, the fact that the ATM channel has the largest highlighted section is to be expected, because its bar is the longest. The percentage of its bar that is highlighted, however, does not appear to be greater than the highlighted portions in the other bars. It looks like the channels MOP, EVE, and DSA, however, might not participate in high profit margins at all, but if we enlarge the graph a little, we can tell that this isn't the case as you can see on the following page.



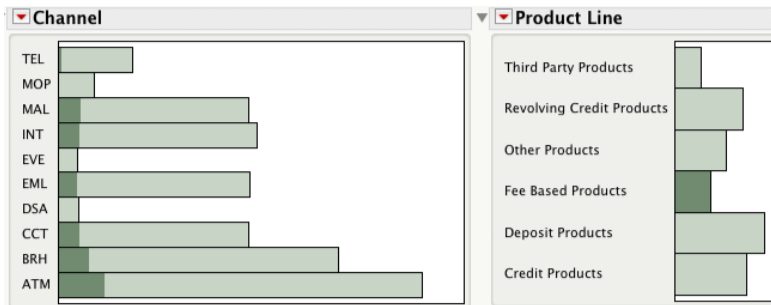
After enlarging the Customer ID graph as well, we can see that, even though some customers contribute slightly more or less than others to high profit margins, none do so to a significant degree.



Don't mistake what we've just noticed as meaningless. We've actually made a meaningful discovery: no channels, product lines, or customers contribute a substantially greater or lesser degree than others to high profit margins. That's useful to know. The same is true of negative profit margins as well, which we can see below.



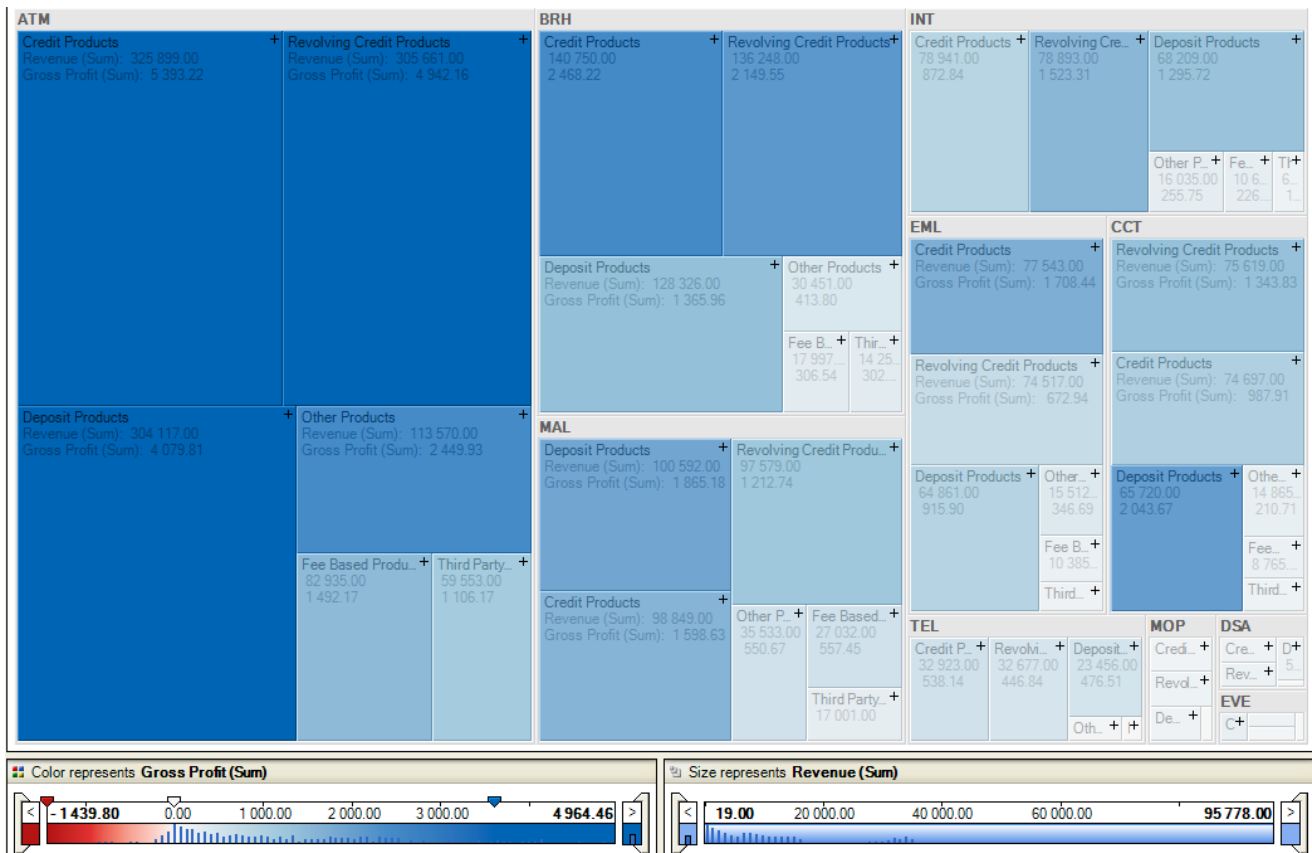
I went on to brush each of the channels to see if there were any interesting interconnections to other variables, but found none. When I began brushing product lines, however, I learned something. Can you see anything interesting in the following display?



Fee based products are not sold through three of the sales channels—MOP, EVE, and DSA—and they’re rarely sold through the TEL channel. As I continued to explore, I also learned that none of the product lines are loss free (i.e., they all participate in negative profits), and that some customers don’t purchase products through some channels. All of these facts contribute to a useful overview of the data. Although brushing and linking produced no dramatic insights in this particular data set, it often does.

### Treemaps and Bar Graphs for Viewing Rankings and Parts of the Whole

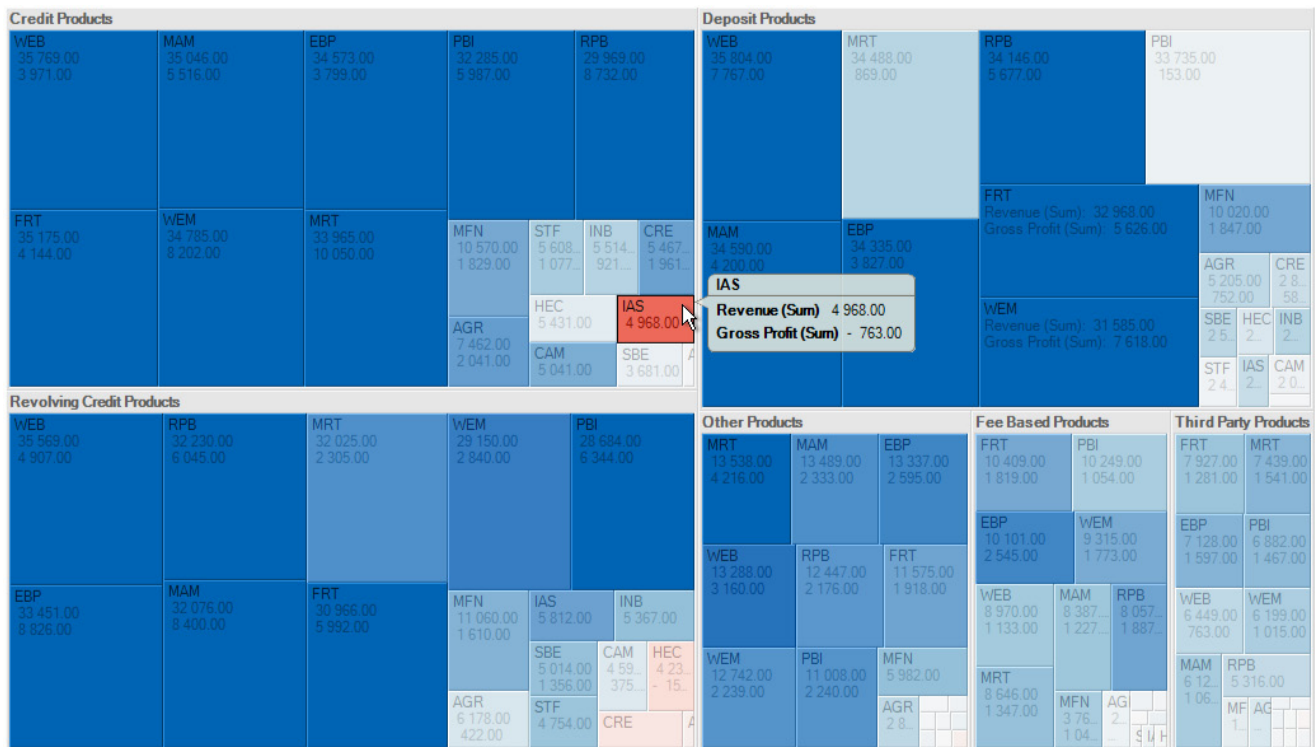
So far, we learned how records in the data set are distributed across ranges in quantitative variables and items in categorical variables and how they interconnect, but we haven’t looked at any of the quantitative amounts that are associated with categorical items. For instance, we don’t know yet how much revenue is associated with various channels, product lines, or customers, but only the number of revenue records of various amounts that are associated with them. Let’s shift our focus now to overviews of the actual amounts, first by looking at how the parts of something (e.g., revenue per channel) contribute to the whole (e.g., total revenue). We’ll begin by using a treemap, because it’s designed to handle a great deal of data, much more than we could easily display in a bar graph. The following treemap displays revenues as the sizes of rectangles and profits as their colors (blue for profits and red for losses, the darker the greater).



This treemap was created using Panopticon.

Revenues and profits are organized by channels (the large rectangles with gray borders that serve as containers for the smaller rectangles) and product lines (the smaller rectangles within the larger ones). We can easily see that ATM is the channel through which the greatest amounts of revenue are earned (the largest gray-bordered rectangle on the left) and that no product lines that were sold through this channel lost money overall (i.e., none of the rectangles are red). In fact, products sold through the ATM channel clearly earned the highest profits, revealed by the prevalence of dark blue rectangles. The greatest revenues are earned by credit products, revolving credit products, and deposit products that are sold through the ATM channel (i.e., they are the largest product rectangles of all).

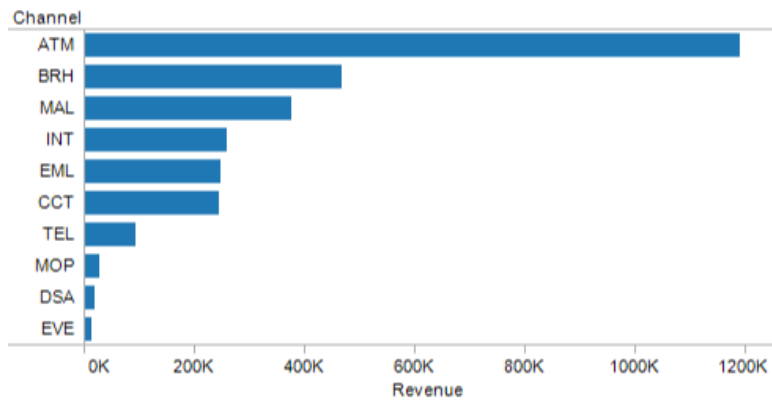
To explore the ATM channel more closely, I drilled down to view it alone (see below).



Noticing the one dark red rectangle, I hovered over it with my mouse to reveal the revenue and profit details for IAS, the one customer associated with a loss. We could continue exploring various sales channels at the customer level using the treemap, which would allow us to fill in the overview that we're trying to build a bit more.

Let's shift our focus for now to graphs that allow us to compare values more precisely. I've displayed the channel's relative rankings using the lengths of bars, which visual perception is well-tuned to handle, rather than the sizes and colors of rectangles, which we can only compare in approximate ways. We can put up with the lesser precision of treemaps when we need to view lots of data at once, but when we narrow the list of items, it's better to shift to bar graphs.

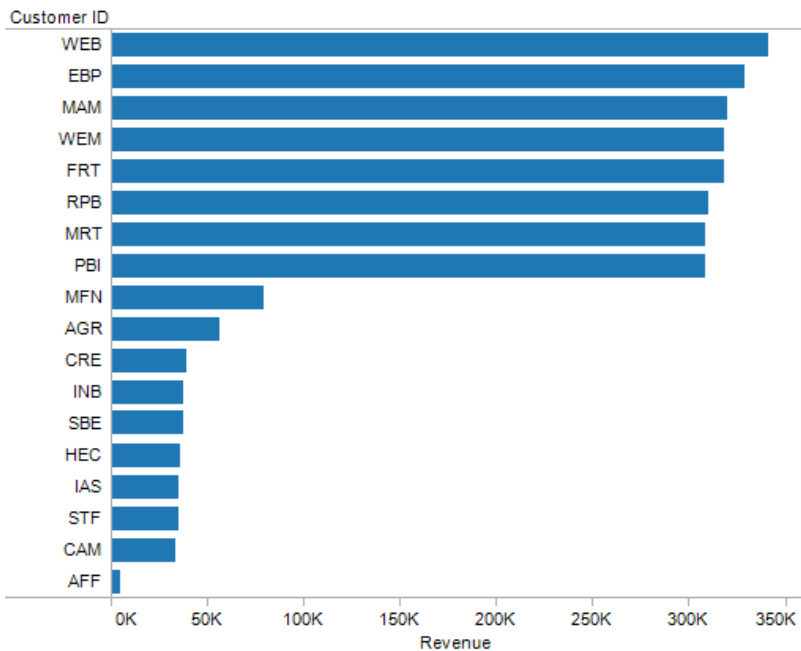
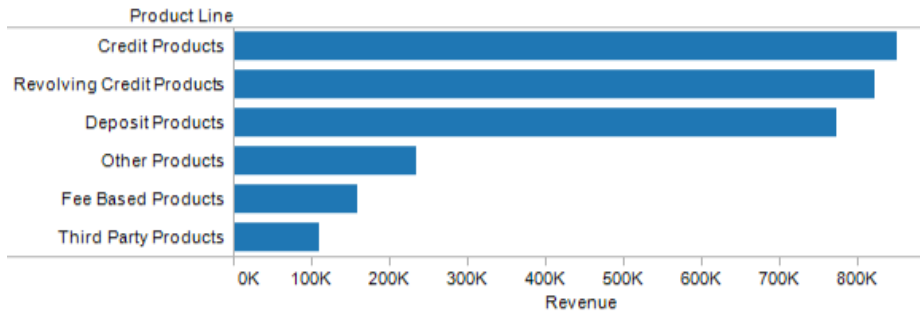
In the following bar graph, we can now compare the revenues associated with various channels with ease and precision:



Created with Tableau.

We can now see that revenues earned through the ATM channel are more than twice as great as those sold through the next highest channel: BRH. We can also see slight differences between revenues earned through the MOP, DSA, and EVE channels, which weren't discernable before.

I continued to explore these ranking and part-to-whole relationships using bar graphs and found that product lines and customers both exhibit a similar feature. Take a look at the following two graphs to see if anything pops out as similar.

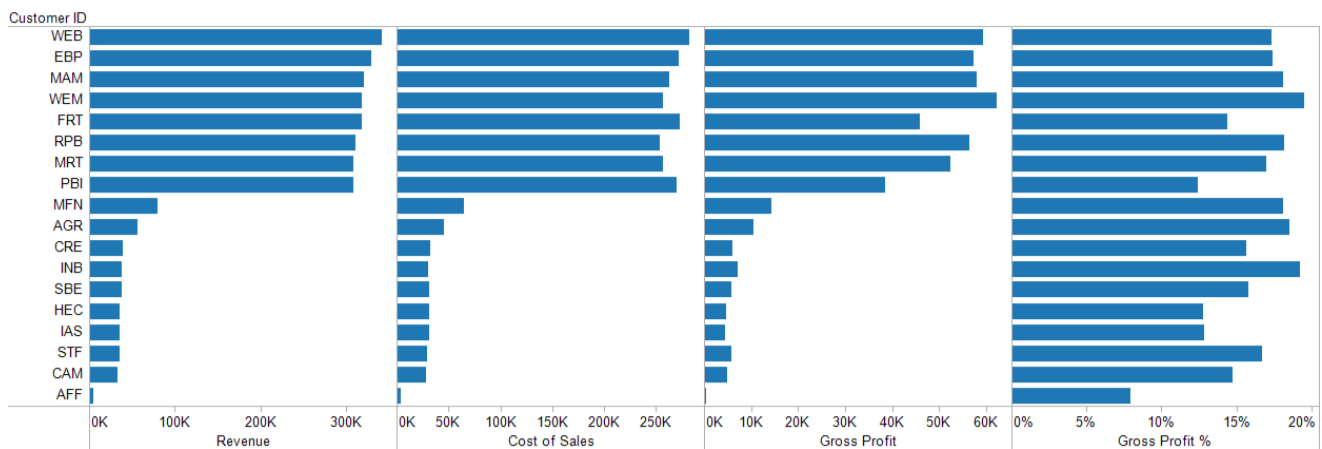


Both product lines and customers appear to be divided into two groups of items: high and low revenue producers. The top three product lines are high revenue producers and the others are relatively low; the top eight customers are high revenue producers and the other ten are relatively low. I explored further and found that this pattern is true of profits as well.

In these ranking and part-to-whole views, we've mostly focused on one categorical variable in association with one quantitative variable at a time (e.g., revenues by product line). Let's now shift to visualizations that will reveal how quantitative variables relate to one another—that is, how they correlate.

### Table Lenses for Viewing Rankings and Correlations

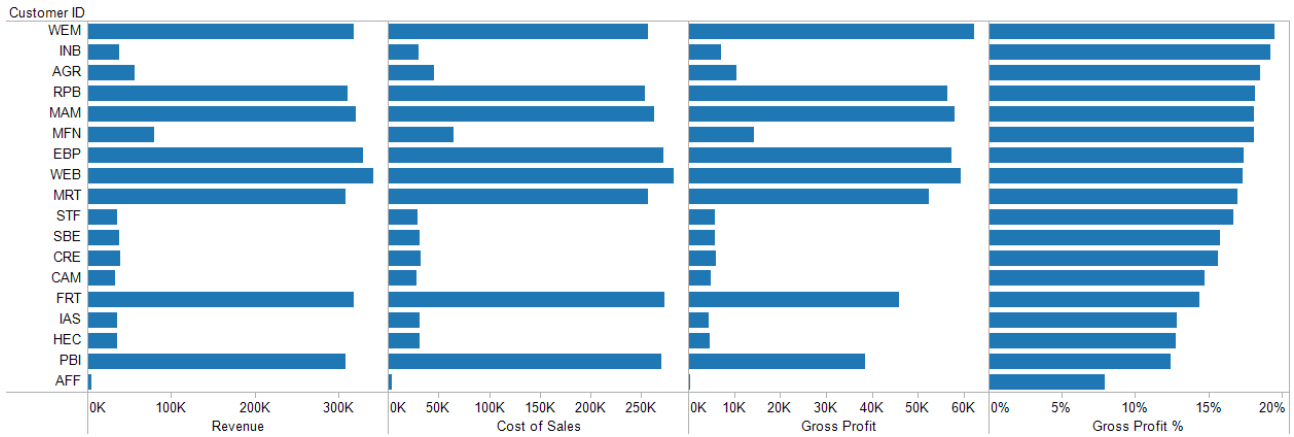
The best graph for viewing potential correlations between two quantitative variables is a scatterplot, which allows us to do so with great precision, but for now we're looking for ways to get a quick overview. A form of display that goes by the name *table lens* will serve our purpose well. Here's a table lens display of all four quantitative variables:



Created using Tableau.

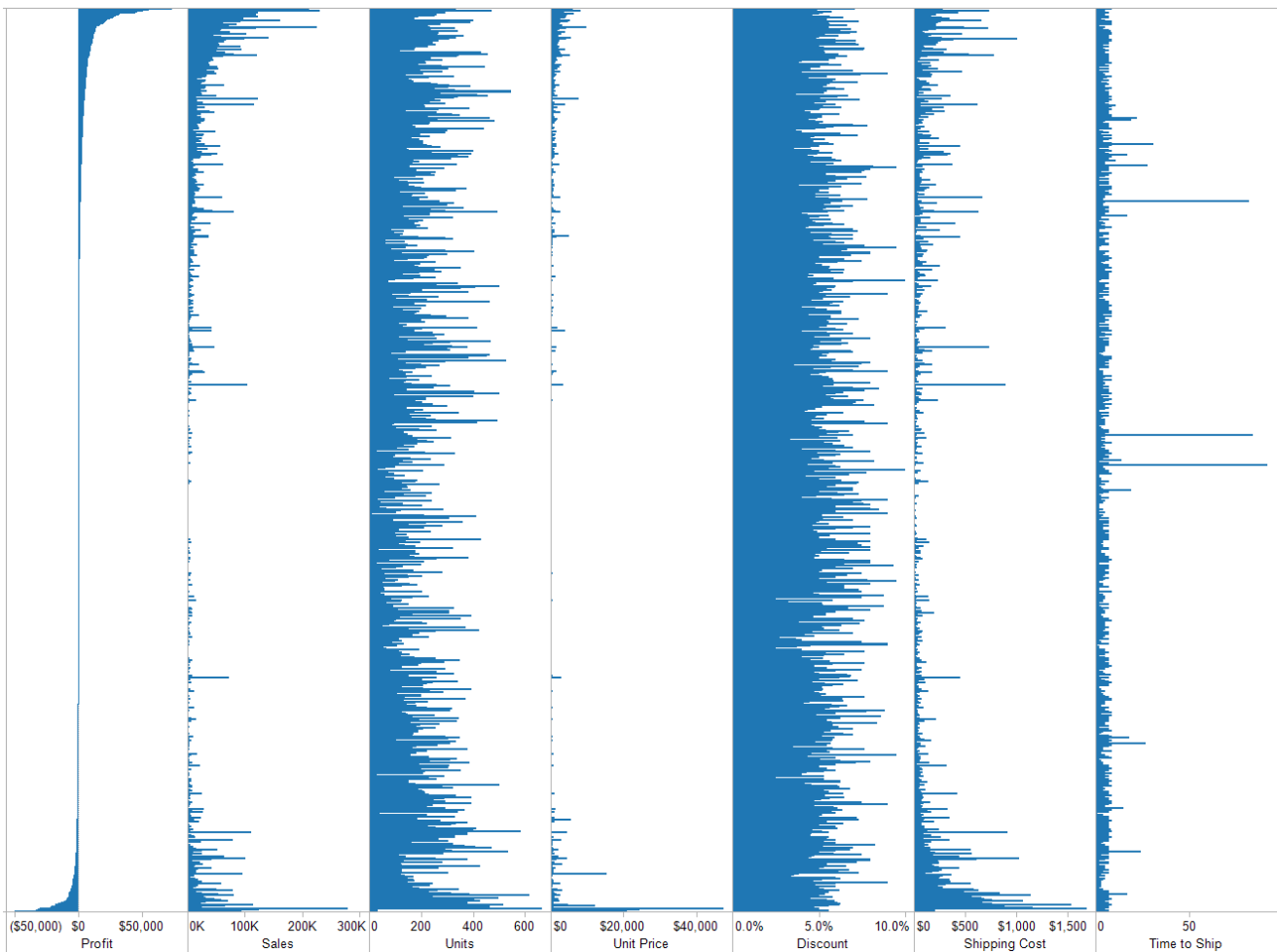
Essentially, a table lens is a series of side-by-side horizontal bars graphs that all share the same categorical variable (in this case customer ID) but each graph displays a different quantitative variable. By sorting the values associated with one of the variables, as I've done above with revenue (the leftmost graph), we can now look at the other variables to see if their graphs exhibit a corresponding pattern. We can see that the cost of sales graph exhibits a high value at the top to low value at the bottom pattern that is very similar to revenues, which reveals that these two variables are correlated in a positive way: the greater the revenue the greater the cost of sales tend to be also. The correlation isn't perfect (e.g., notice that the fifth bar down—FRT—in the cost of sales graph is longer than you would expect based on revenues), but it's fairly strong. There also appears to be a correlation between revenue and gross profit, although it is a bit weaker. If one of the graphs were to exhibit a small value at the top to low value at the bottom pattern, this would indicate that it had a negative correlation to revenue (i.e., the greater the revenue the smaller the value of this other variable).

Notice that the gross profit % graph does not seem to indicate a correlation between it and any of the other variables. To look at this closer, let's sort the tables lens based on the gross profit % values, shown in the example below.



This view indeed seems to confirm the lack of correlation between gross profit % and any of the other variables based on the fact that none of them exhibit a similar high-to-low value or reverse low-to-high value pattern.

The real power of a table lens display is hard to illustrate using a data set with only four quantitative variables. Its benefits as an overview of potential correlations becomes more apparent, however, with more variables and more rows of data. To give you a sense of this, here's a table lens that displays 1,263 products (a bar for each) across seven quantitative variables:



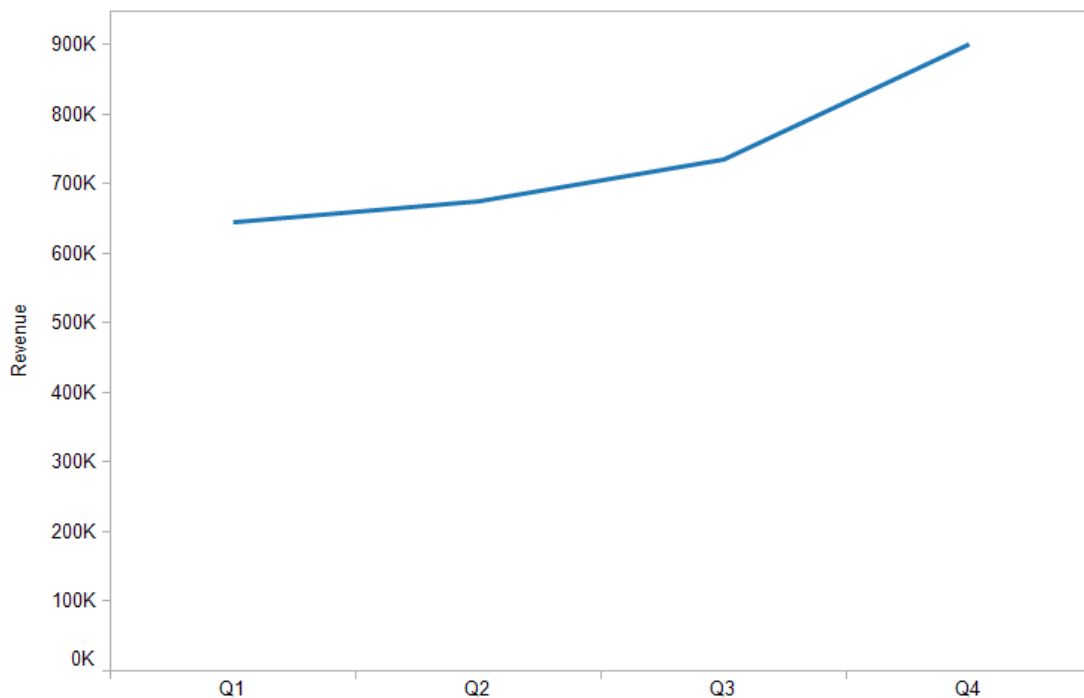
By sorting on profits from high to low in the leftmost graph, we can easily look at other variables for potential correlations. No obvious positive or negative correlations appear, but an interesting pattern emerges that would be worth investigating. Notice how sales, unit price, and shipping cost all display most of their highest values near the top and near the bottom of the graph, with lower values in the middle. Perhaps both the highest profits and the greatest losses tend for some reason to correspond to high sales, unit prices, and shipping costs.

We've yet to look at one of the most important variables in most data sets: time. An overview is incomplete and can be quite misleading if you haven't taken change through time into account. For that, we'll need to view the data differently.

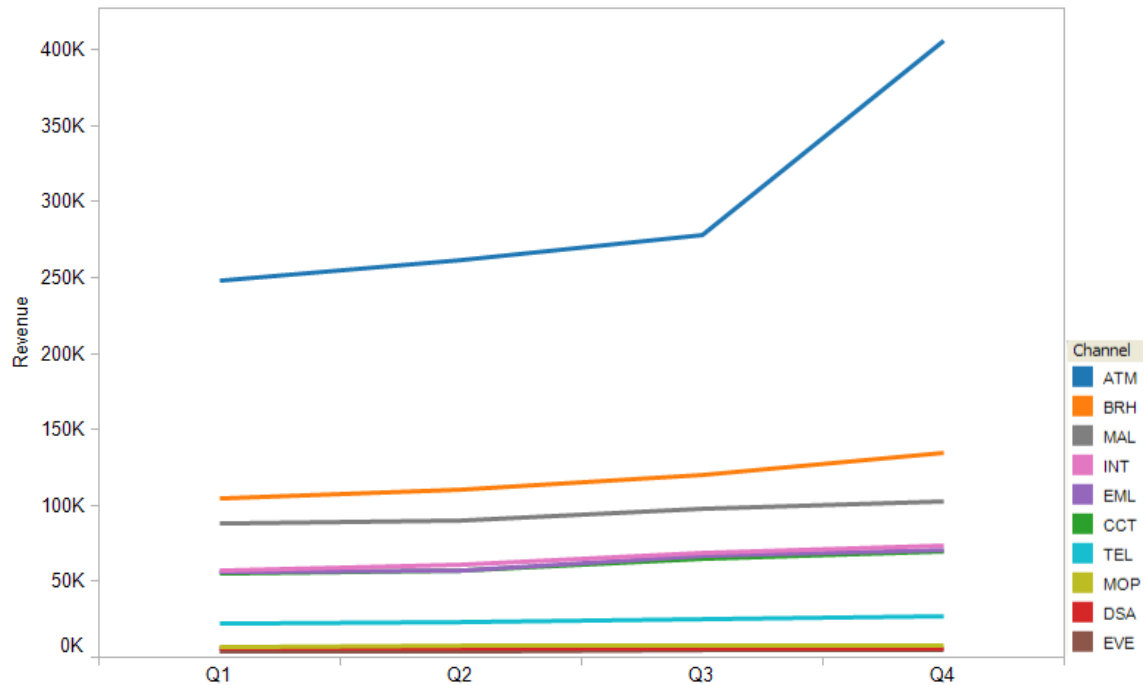
### Line Graphs for Viewing Change through Time

The data set that we've been using isn't rich in time-series information—it contains one year's worth of data at the quarterly level only—so we'll need to switch to another to fully explore overviews of change through time. Before doing so, however, I can illustrate a few things with the current data set.

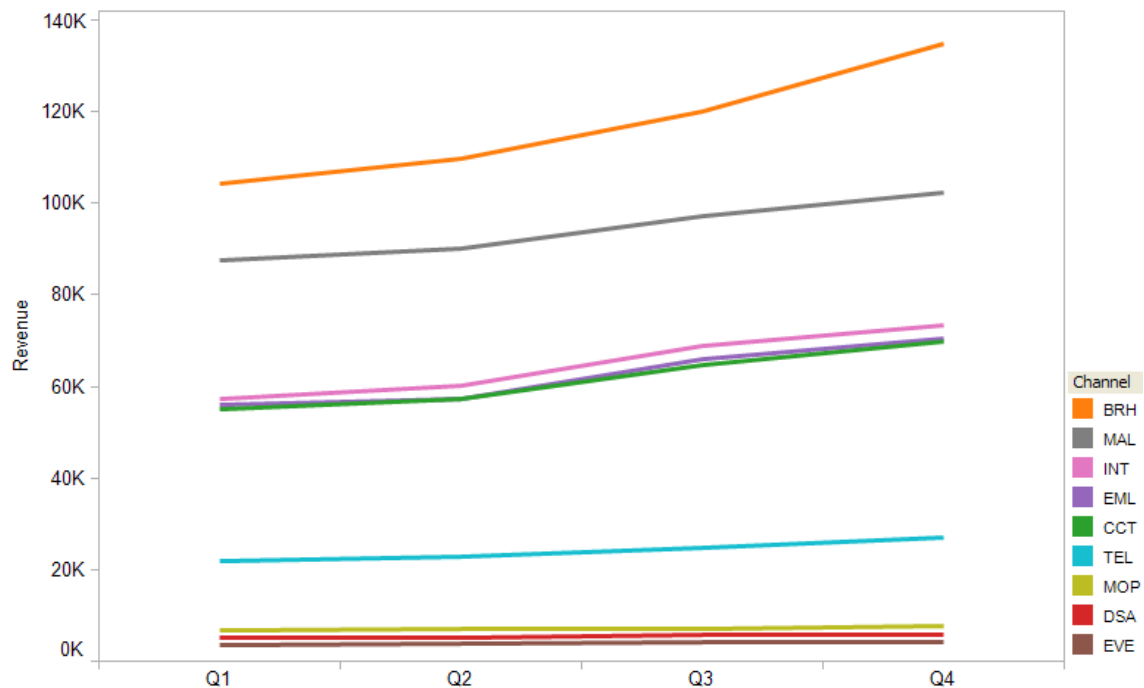
In general, nothing does a better job of displaying change through time than a simple line graph. Let's begin by looking at each quantitative variable, one at a time, as its values changed through time, both in total and at the level of its various parts (per channel, etc.).



Overall, revenues have been going up. In fact, they have increased each quarter, with the greatest increase in the last. Let's see how revenues changed per channel.

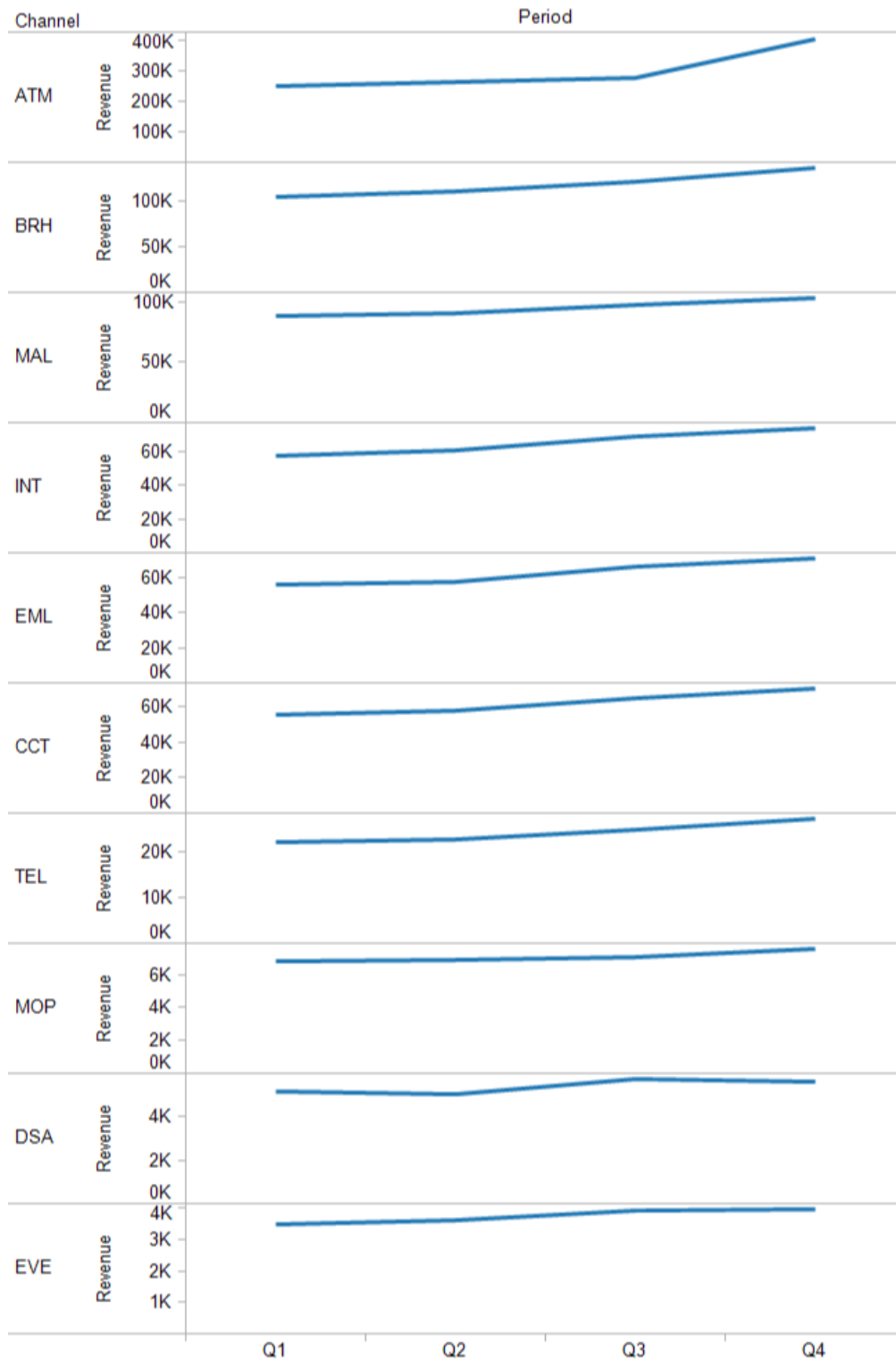


Sales through the ATM channel clearly increased much more in the last quarter than the others. This view, however, suffers from two potential problems: 1) Because many of the channels produce much smaller revenues than ATM, they appear as relatively flat lines near the bottom, which makes it difficult to see their pattern of change through time clearly; 2) Even though this graph makes it easy to compare amounts of change through time, it does not make it easy to compare rates of change (i.e., percentage change). The first problem can potentially be solved in two ways. One way is to remove ATM from the graph and allow the graph to rescale, thus spreading the other lines across much more space, which unflattens them a bit, making it easier to see patterns of change. Here's how that looks:



We've now solved the problem for most of the lines, but patterns in the bottom three are still undetectable. We could further remove all but the bottom three lines to solve this remaining problem, or we could use the next method.

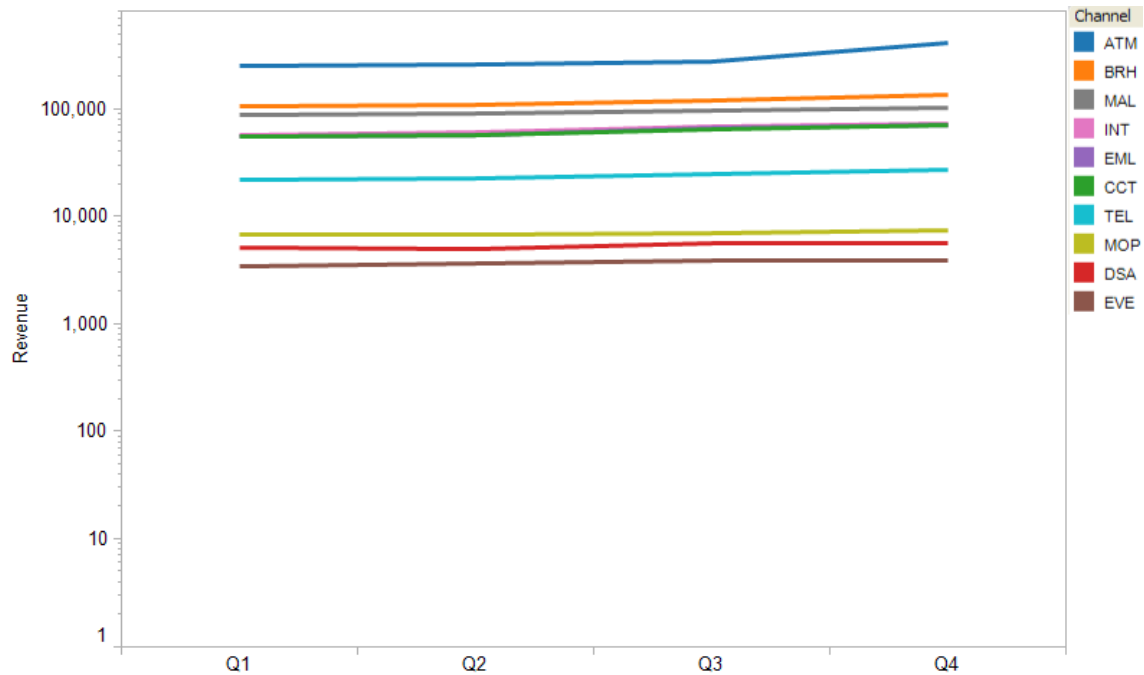
Rather than displaying each line in the same graph, let's display each in its own graph and allow the graphs to scale independently, as follows:



An independently scaled series of graphs like this can be easily created in Tableau.

We can't use this display to compare magnitudes of change among the various channels, because the quantitative scale of each is different, but we can use it quite easily to see and compare patterns of change. What is now clear is that revenues in every channel have been consistently increasing by quarter, except for the slight decrease in DSA revenues from Q1 to Q2.

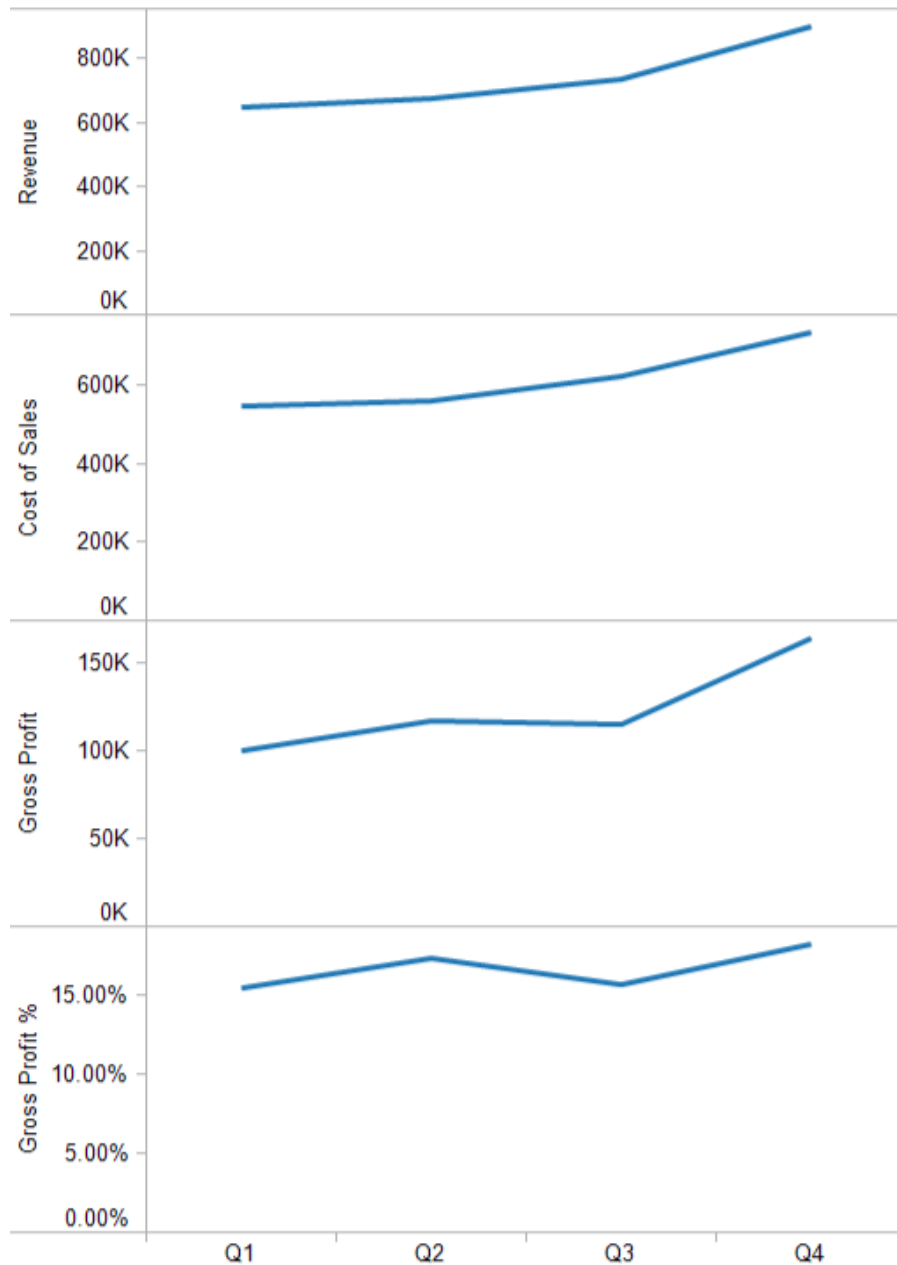
We still haven't compared rates of change. One simple way to do that is to switch the scale along the quantitative axis from linear to logarithmic (log). I did this to the original graph in this section, which displays a separate line for each channel in a single graph, and this is the result:



A useful characteristic of a log scale is the fact that, when using a line graph to display change through time, lines with equal slopes represent equal rates of change. For example, if you compare the second and third lines from the top you can see that in the third quarter, the orange line increased by a slightly greater rate (it slopes up a bit more) than the gray line. What is very clear now is the fact that ATM (the blue line at the top) increased at a faster rate than all others in the 4th quarter. Even if you don't understand log scales and their other uses, this is one simple use that doesn't require further knowledge.

So far, we've examined how a single quantitative variable—revenue—changed through time, in total and by channel. To get a full overview, we would want to break revenue down by the other two categorical variables as well (product type and customer), and we would want to examine the other quantitative variables individually, as we did with revenue.

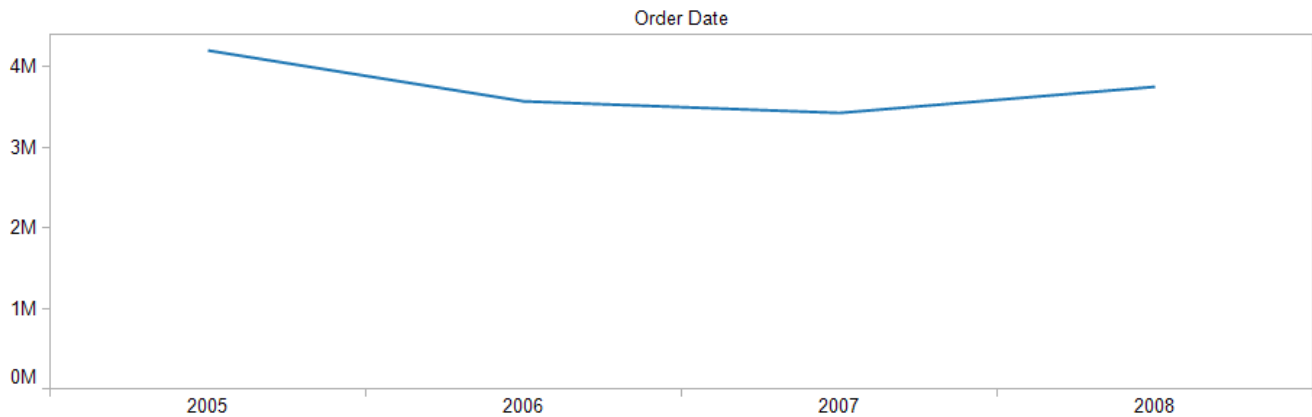
What next? There's one more overview of change through time that I can illustrate with this data set: the comparison of multiple quantitative variables. In the following display, we can compare patterns of change among all four quantitative variables at once:



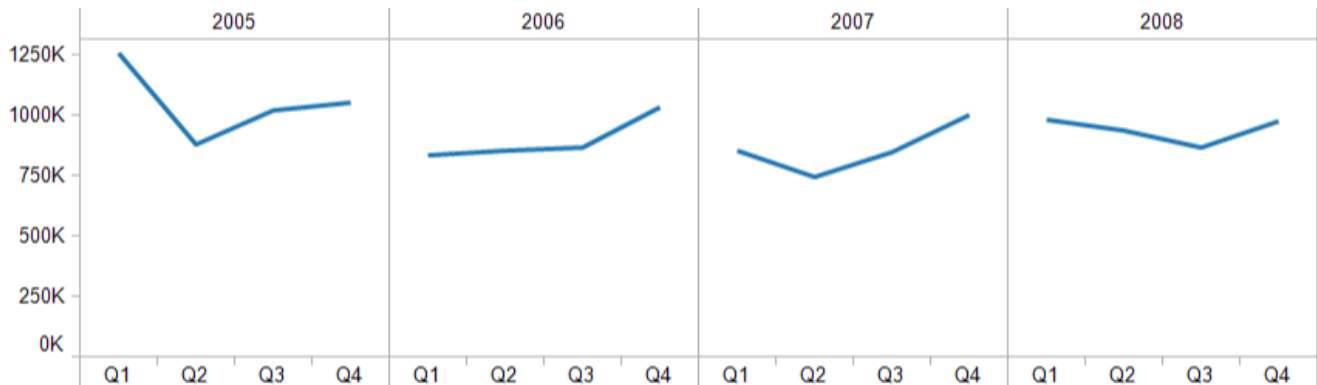
These graphs are independently scaled, so we can't use them to compare magnitudes of change among different variables, but that's not what we're after anyway—we simply want to compare patterns of change. Using this display we can easily see that, whereas both revenue and cost of sales increased each quarter, both gross profit and gross profit % decreased from Q2 to Q3. Something happened during that period that we should investigate, but for now we're sticking with an overview.

I can't go on without interrupting to make an incredibly important point. In most cases, to properly understand change through time, we need more than a single year's worth of data, and we need to see the data expressed at different intervals of time (i.e., not just by quarter). For example, because we only have one year's worth of data in this data set, we have no way of telling if the changes that we're seeing are typical quarterly patterns or are unique to this year. Also, by seeing change expressed at the quarterly level only, we have no way of seeing monthly or weekly patterns that might also be important.

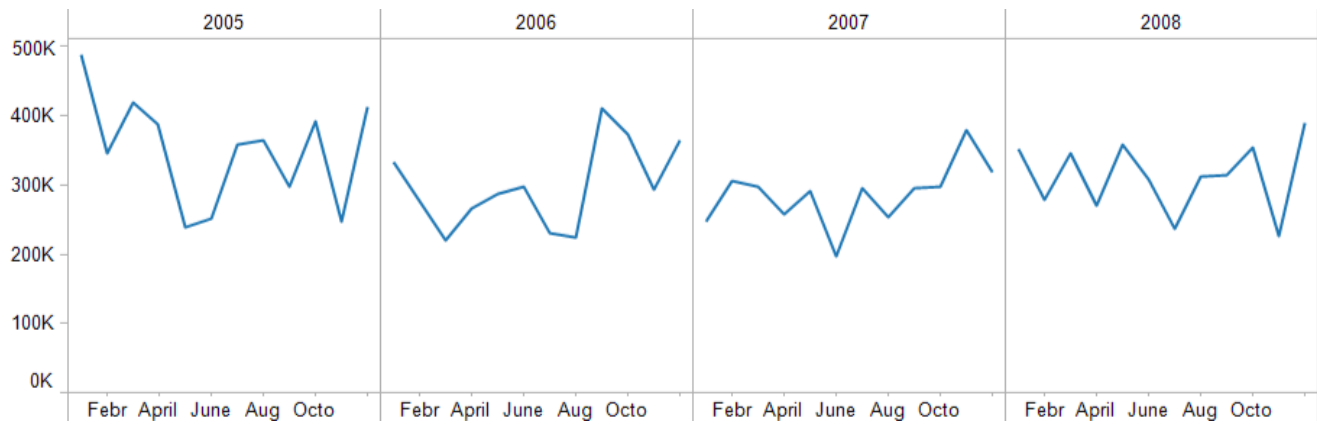
Using another set of sales data that contains more information about time, here's a high-level view of revenues by year:



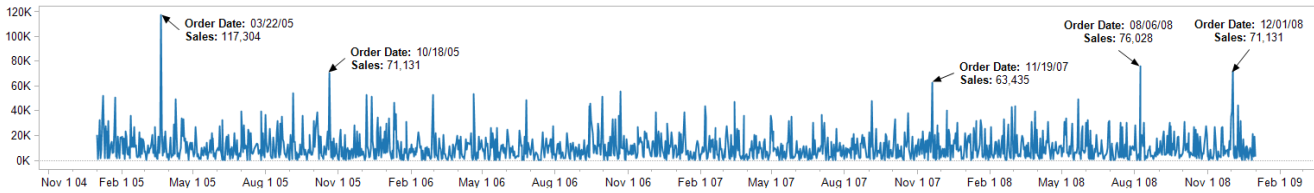
Revenues in 2008, although slightly higher than 2007, are roughly equal to 2006 and actually lower than 2005. Seeing the sum of revenues per year tells one story, but other useful stories live in this time series at different levels as well. Here's the same stretch of time by quarter:



The story of this display is that a great deal of variation is occurring by quarter that wasn't visible before. For instance, in the year 2005, which produced the highest overall revenues, Q1 was largely responsible for this. Despite other differences, each year exhibited an increase in revenues from Q3 to Q4, and three of the four years exhibited decreases from Q1 to Q2. (Note: This quarterly view and the monthly view below would work better if the line continued across the years without interruption, but this is not currently supported in a practical way by the software.) Let's see what story a monthly view reveals.



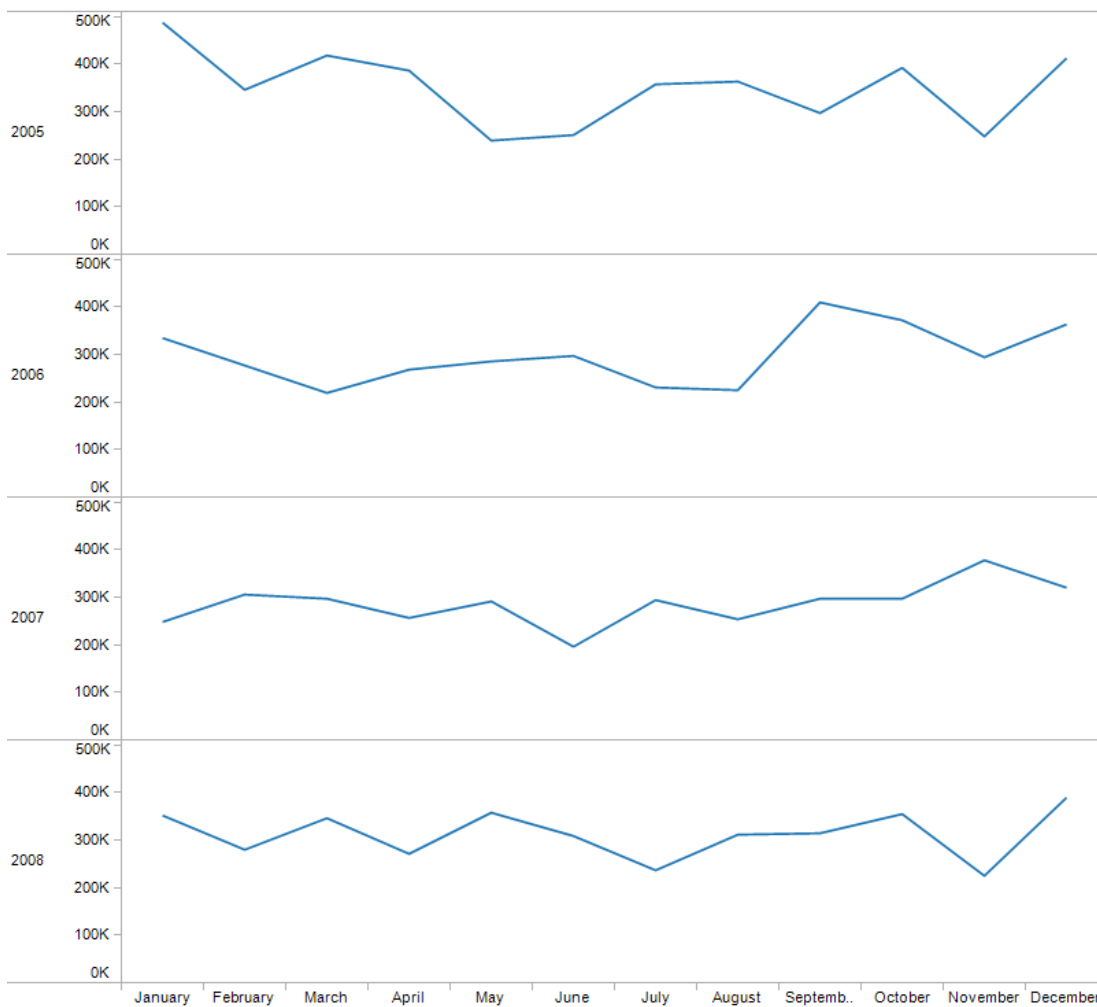
What we now see is that there is a great deal of variability by month. In 2005, despite the great performance in January, February took a dive, but March headed up again. In that same year, notice how much lower May, June, and November were compared to other months. Here's the same period of time by day:



I've added annotations to point out that the days of extraordinarily high sales each occurred during a different month of the year.

By expressing the same period in different intervals of time, we've gained a much richer understanding of change. Always, always, always examine time at different intervals if you truly wish to understand it.

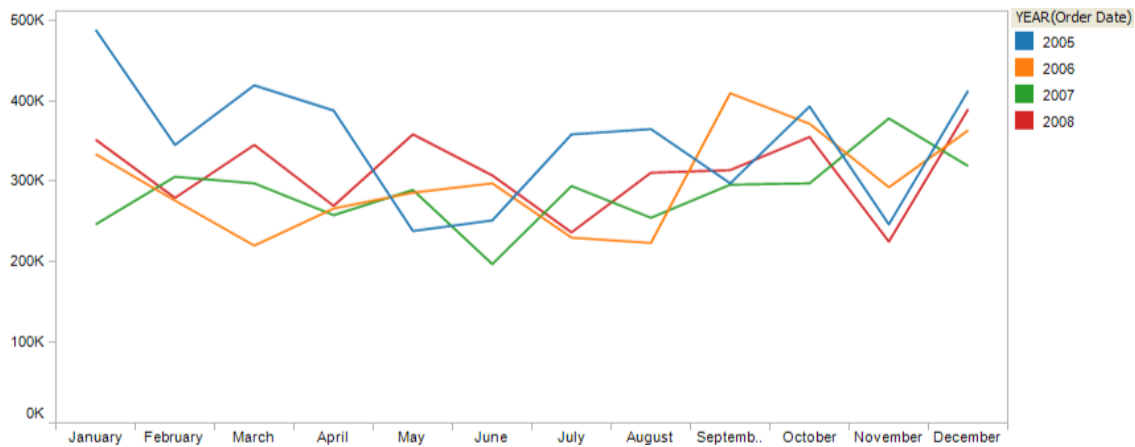
Another characteristic of time that we might want to add to our overview is the cyclical nature of change. For example, we might want to know if there are typical patterns across the months that remain fairly consistent from year to year. Even though we could already tell that the monthly patterns varied from year to year in the monthly view above, a better way to compare monthly patterns—not always the best, but the easiest—involves a display such as this:



By creating a separate line for each year and aligning the months in this manner, we can compare the monthly patterns of these four years more easily than we could in the earlier view that arranged the years from left

to right. We can now see that revenues are not exhibiting a consistent monthly pattern overall, nor is any particular month behaving in a particular way.

If we want to compare magnitude differences in a given month among the four years, combining the four lines into a single graph makes this a little easier.

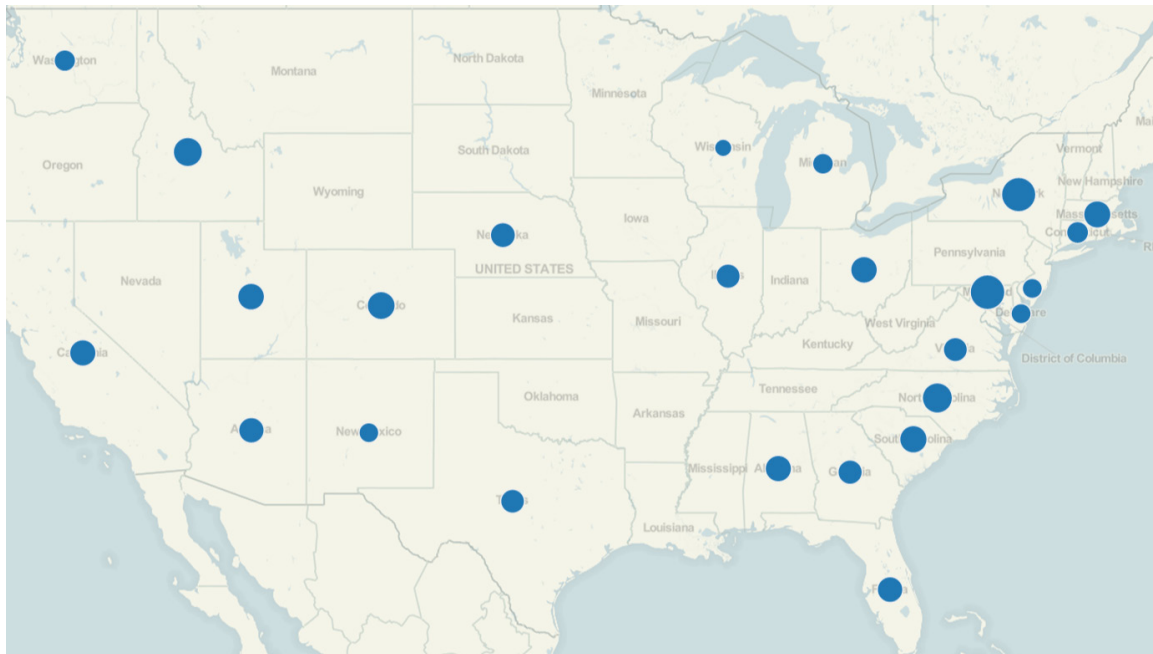


The biggest magnitude differences in particular months among the years occurred in January and March.

There is obviously a great deal more that we could do to examine change through time, but the views that we've examined have provided a good starting overview, so we'll move on.

### Geospatial Maps for Viewing Locations

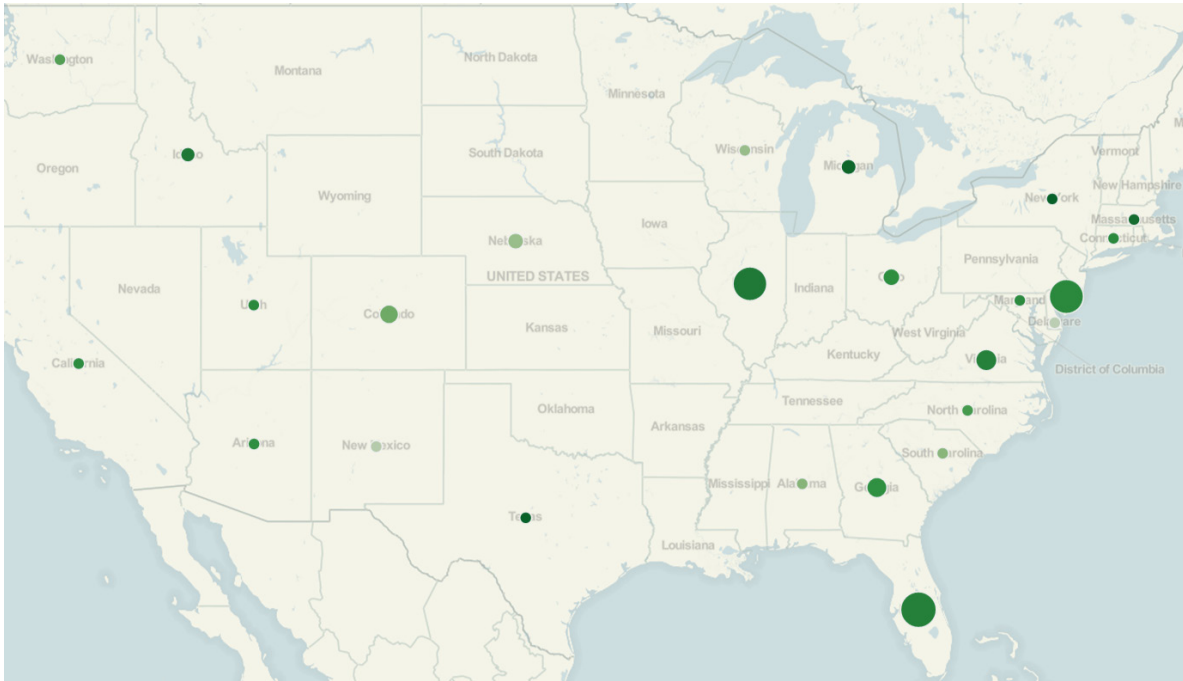
One overview visualization remains, but only when the location of things plays a role in the data set's story. For instance, geospatial displays would not be useful for understanding employee information if all employees work in a single location. In most cases, though, location is part of the story. Sticking with the same sales data set that we just used to explore time, here's a geospatial view of profits by state based on the sizes of the bubbles:



Created using Tableau

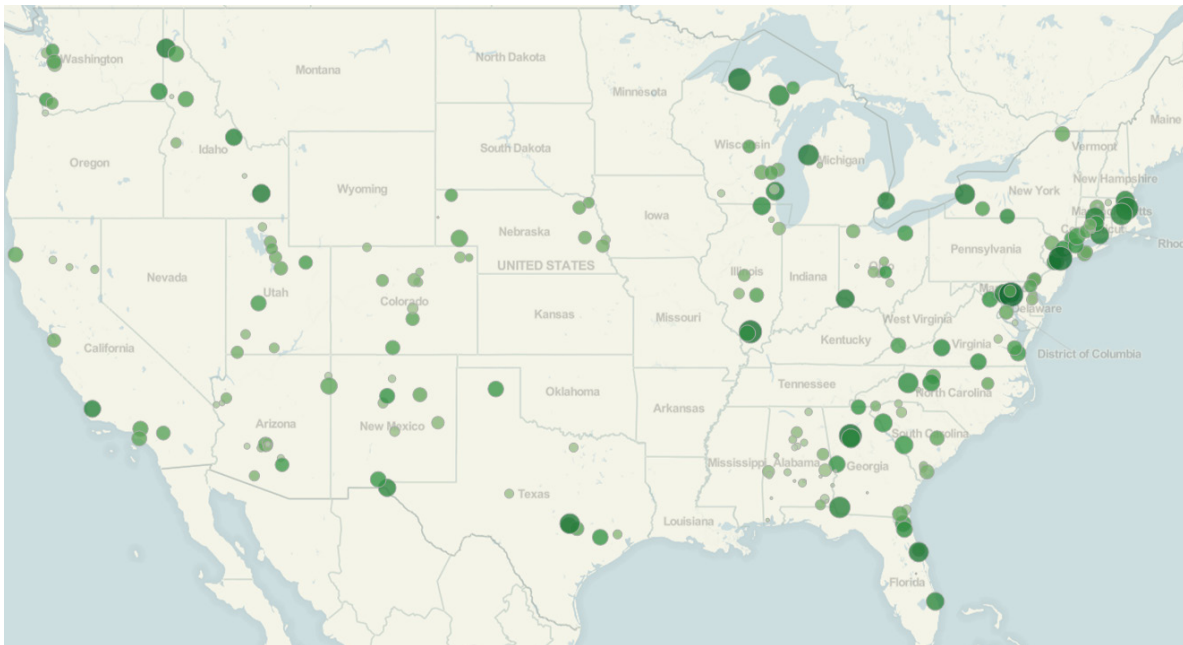
Even before comparing profit amounts, we can see that sales do not occur in several states and that the two states with the highest profits are Maryland and New York, both in the upper half of the east coast.

The number of days that it takes to ship orders varies quite a bit by location, as seen in the following:



In addition to presenting the number of days that it takes to ship orders by the sizes of the bubbles, I've used color intensity to represent the number of orders handled by each state to see if differences in order volume explain differences in shipping performance. Although Illinois handles a lot of orders and takes a while to ship them, several other states with similar order volumes manage to ship them quickly, so order volume alone probably doesn't explain the problem.

Just as it's important to view time series data at various intervals of time, it is also important to view geospatial data at various levels of geography. In the following display, we can see the same days to ship (bubble size) and order volume (color intensity) information as before, but this time at the zip code level.



We can see that Illinois struggles more to ship orders quickly in places where they handle the most orders.

An important characteristic of useful geospatial displays is a map that's designed to feature quantitative data on top of it. Displaying data on a map that was designed to provide driving directions, such as Google Maps, is perhaps better than nothing, but ghastly compared to a display such as the one above. Notice how well the bubbles stand out and how the map provides geographical context in the background to support the data without competing with it.

### All Together Now

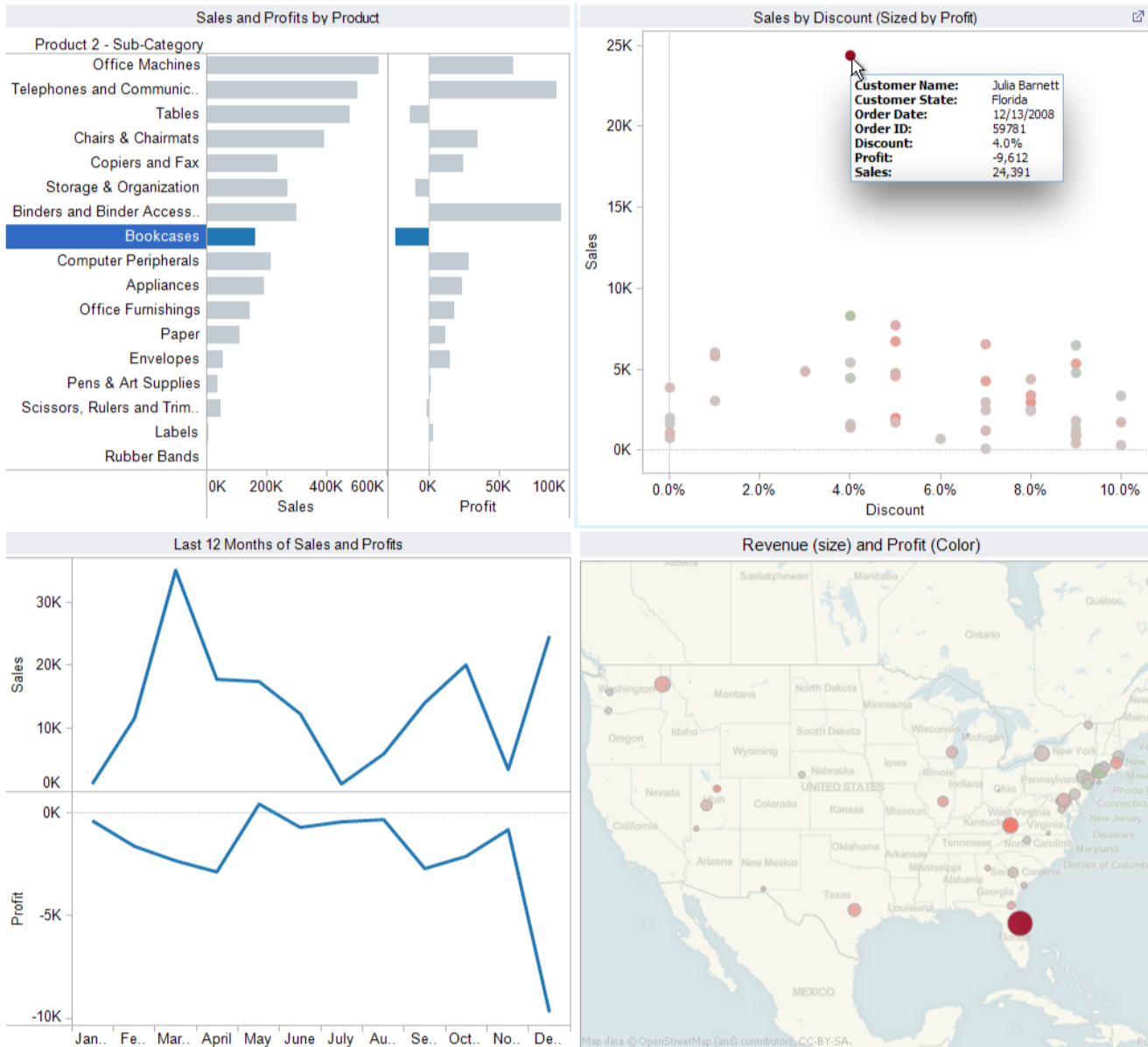
A thorough understanding of the data set cannot be constructed by using the five sets of visualizations in isolation from one another. Using them together will make an even richer data landscape available to our eyes, making it possible for us to see how the facts that we've learned relate to one another. The visualizations that we've examined will continue to serve as the core of our toolset as we venture from the initial overlook down into the depths of the data set. To illustrate, with the data set that we used in the last two sections, I quickly constructed the following combined display:



Created using Tableau.

In this one multi-chart display we can view rankings, parts of wholes, changes through time, correlations, and locations related to sales and profits for the past 12 months. Viewing this data set from multiple perspectives

all at once allows us to discover connections that viewing them independently would never produce, especially if the multi-perspective display is enriched through coordinated interaction, such as brushing and linking. For instance, noticing in the upper left-hand bar graph that bookcases are losing money, we might decide to focus exclusively on that for a moment by brushing bookcases to select them and using this action to filter out all but bookcase sales data in the other charts to produce the display below:



Noticing the one large order at the top of the scatterplot that produced the greatest loss of all, I hovered over it to get details to appear, which revealed what I had already guessed from looking at the line graphs and map—that this order was placed by a Florida customer in December. Glancing at the line graph tells us that only in one month—May—did the sale of bookcases not produce a loss.

## A Good Beginning Shapes the Journey

Getting the lay of the land at the start of the journey using these five sets of visualizations and one interaction technique will keep you oriented as you become immersed in any new territory. It will provide contextual awareness that will reduce wild goose chases and false conclusions that plague many analytical journeys. Always take a good look around you before proceeding, and especially before taking off at a run. Data analysis requires the skills of an explorer. Knowing how to begin the journey is fundamental to those skills.

---

## Discuss this Article

Share your thoughts about this article by visiting the [Exploratory Vistas](#) thread in our discussion forum.

---

## About the Author

Stephen Few has worked for over 25 years as an IT innovator, consultant, and teacher. Today, as Principal of the consultancy Perceptual Edge, Stephen focuses on data visualization for analyzing and communicating quantitative business information. He provides training and consulting services, writes the quarterly [Visual Business Intelligence Newsletter](#), and speaks frequently at conferences. He is the author of three books: *Show Me the Numbers: Designing Tables and Graphs to Enlighten*, *Information Dashboard Design: The Effective Visual Communication of Data*, and *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. You can learn more about Stephen's work and access an entire [library](#) of articles at [www.perceptualedge.com](http://www.perceptualedge.com). Between articles, you can read Stephen's thoughts on the industry in his [blog](#).