Data Through Time

Descriptive Analytics: Data Visualization and Storytelling with Data

# Time Series

A descriptive *time series* analysis seeks to answer the question, “What happened over a certain period of time,” and explore any structures and patterns that may be present in the data involving either previous values of a univariate series data or multivariate relationships among several time series. Analyzing time series data is notoriously difficult, but fortunately there are numerous techniques for visualizing the data by transforming, decomposing, and segmenting it. Aside from the challenge of making forecasts, there are several characteristics of time series that introduce difficulties:

* Strict ordering of data
* Temporal relationships
* Multivariate relationships
* Seasonality/cyclicality
* Frequency mismatch
* Historical data revisions
* Noisy data or missing data
* Windowing
* Structural gaps like weekends, holidays
* Exogenous one-time events that impact the data
* ...and many more

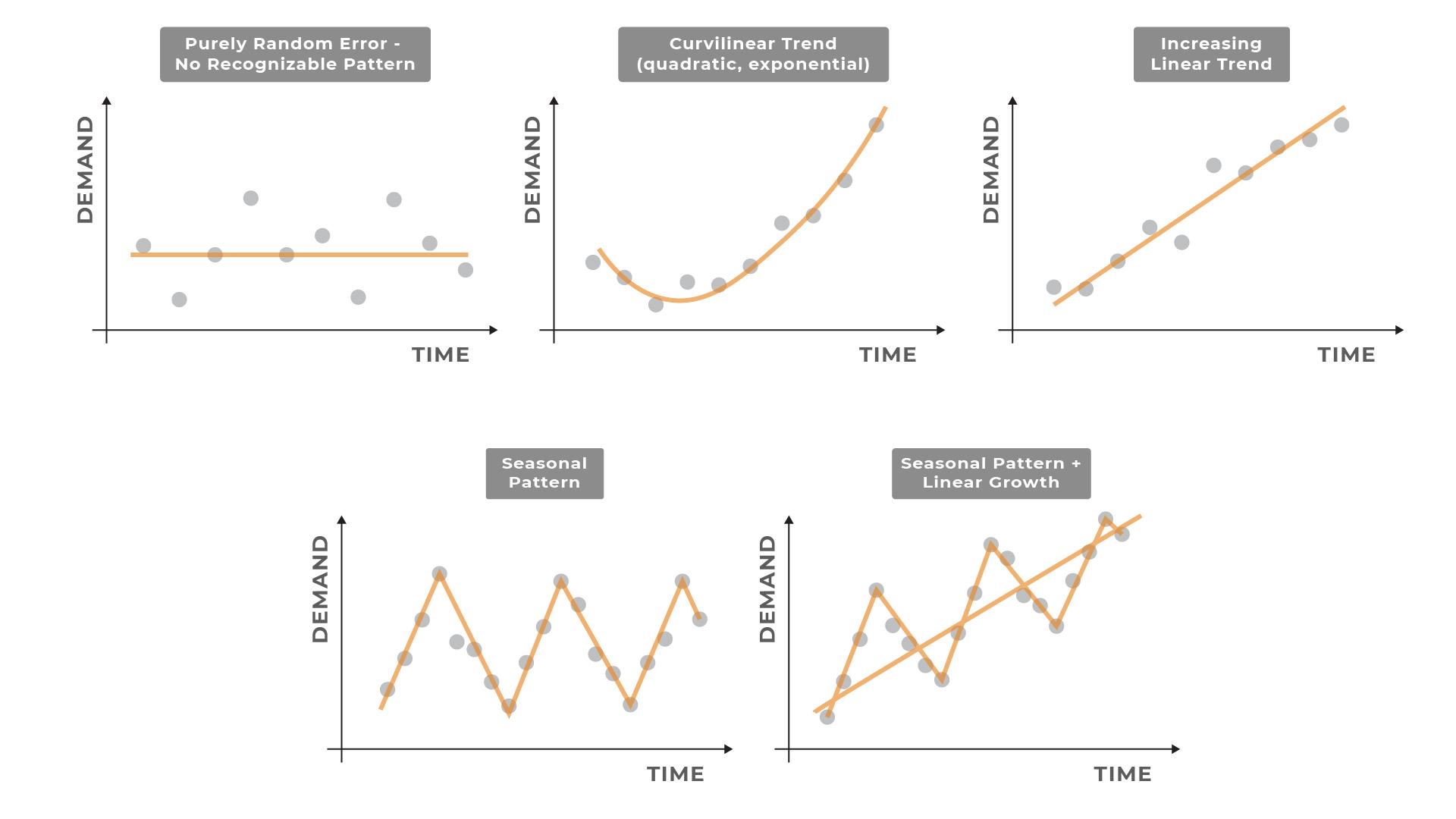
Consider a univariate time series, the simplest case. This is a time-ordered data set in which each value has an associated time or date. We can certainly ignore the time component and use any of the univariate methods we’ve studied previously to look at distributions, maximum/minimum values, etc., but this would ignore a large source of information. Perhaps there are relationships or dependencies among consecutive values in time. Perhaps there are relationships or patterns that repeat with a certain frequency.

If we take a more complex situation in which we compare several sets of univariate time series data, we may find multivariate relationships among the data. Once again, we can ignore the time component and view scatterplots of the data in search of statistical relationships, but as before there may be important information in the time component that could impact the different series simultaneously.

Consequently, the treatment of time series data requires a particular focus on addressing time dependence. The time component acts as an independent variable that can be treated as both a dimension and as a measure. This allows a good deal of freedom in developing descriptive visualizations since the time component can be sliced, diced, and manipulated much like any other variable. There may be structural relationships that can be observed as a function of time or among prior values. The potential dependence on both absolute time and prior values within the same dataset make analyzing a time series rather challenging.

## Line Graphs

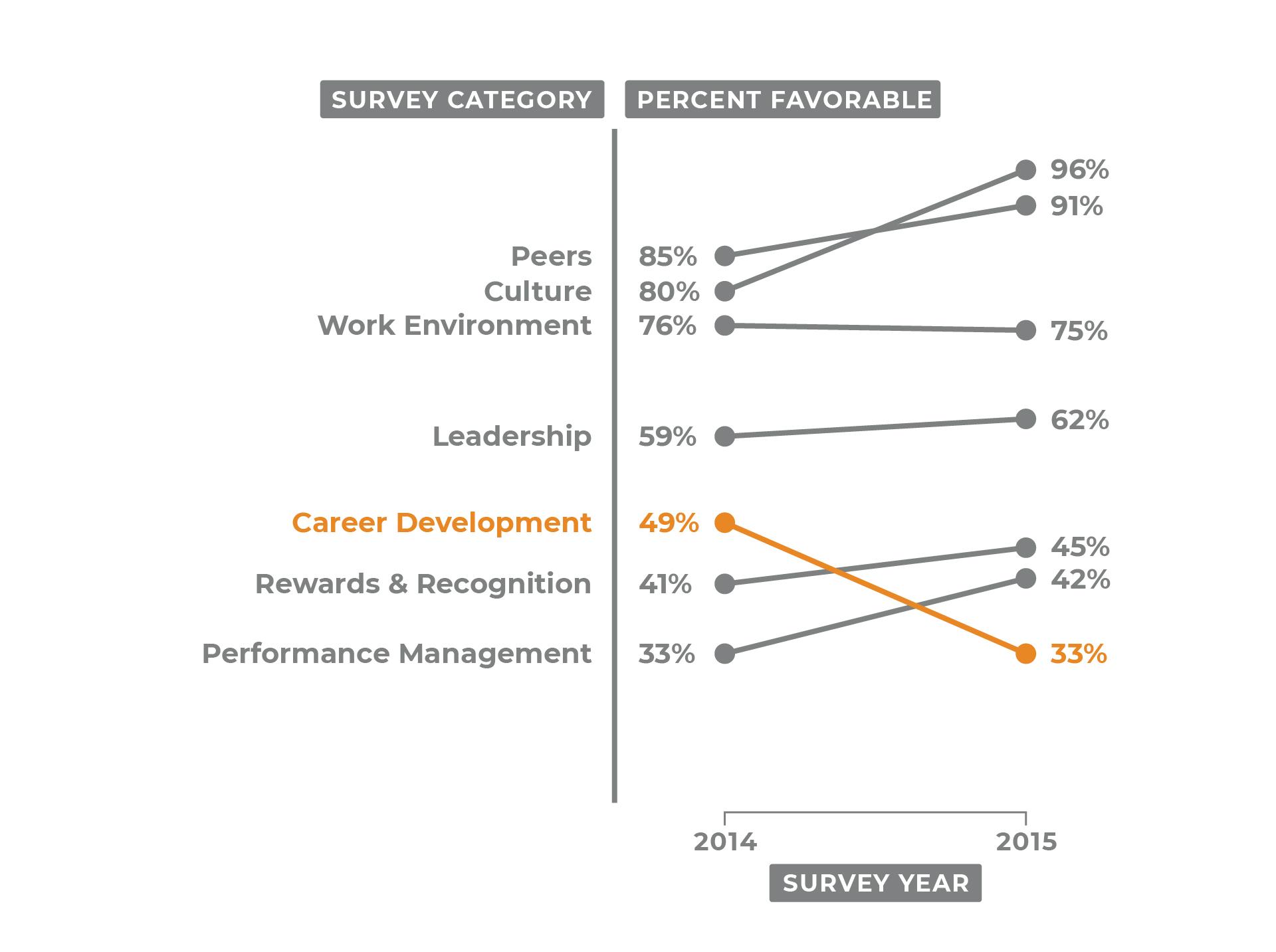
Despite the analytical challenges associated with time series, visualization is rather intuitive. A *line chart graph* of a time series always shows linear time on the x-axis increasing to the right with the y-axis value representing the data at particular points in time. The line graphs of the five time series shown in Figure 1 clearly show the evolution of the data over time and quickly provide visual cues for trends, noise, patterns, cyclicality, etc.

*Figure 1: Visual cues in time series charts: purely random (no recognizable pattern), curvilinear trend (quadratic or exponential), increasing linear trend, seasonal pattern, and seasonal pattern combined with linear growth*

## Slope Graphs

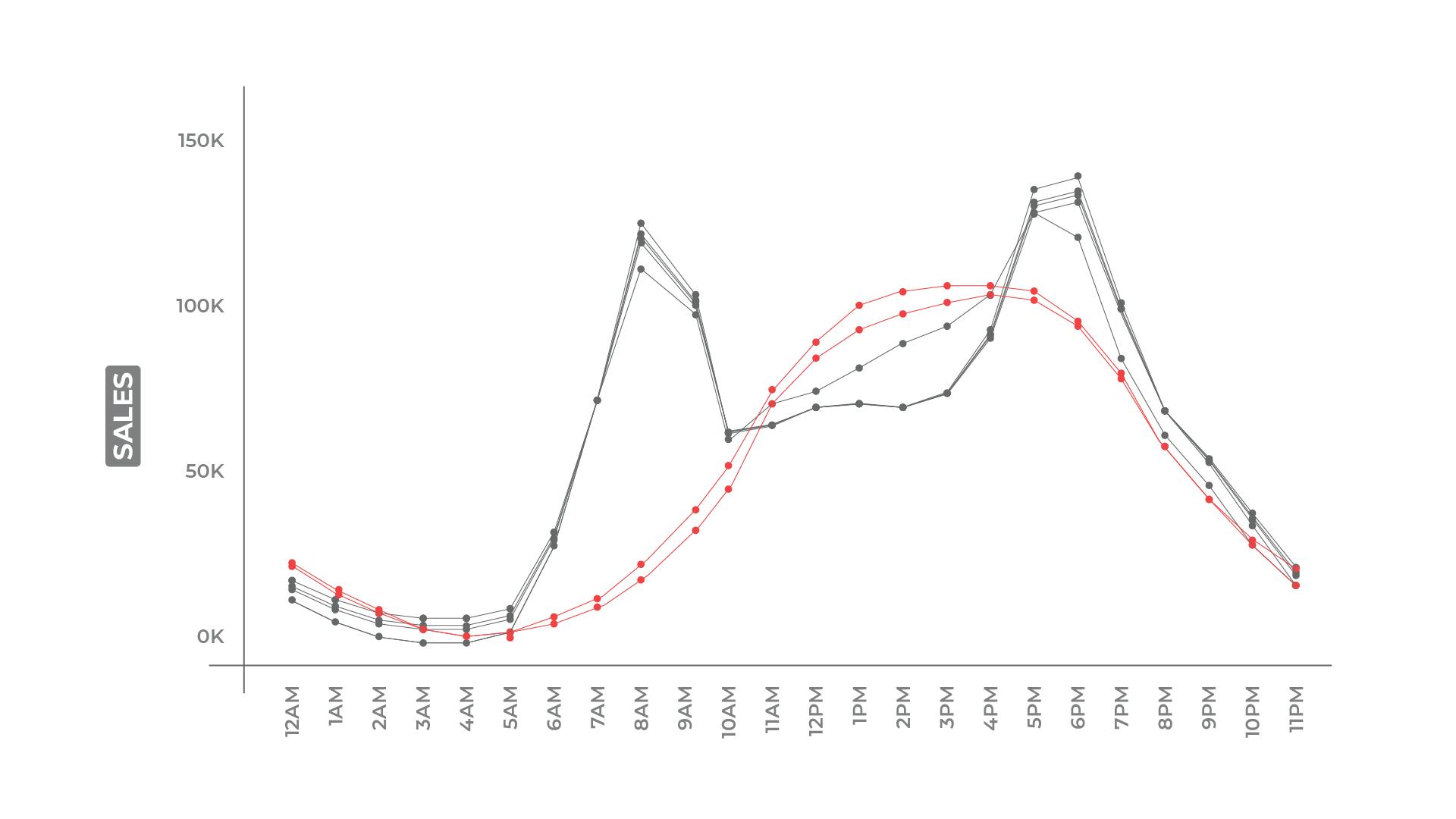
In some cases, the details of how a variable changes over time is less important than the starting and ending points. Such a case might arise when looking at annual cash balances for a company, net user subscriptions before and after a marketing campaign, polling survey results, etc., where the path between the points is not important. The driving question in these cases might be, “What is the ending value and how quickly did it get there?”

*Slopegraphs* (Figure 2) are useful for showing relative increases and decreases for a set of categorical variables. The slopes of these lines reflect the relative rates of change for each variable and are simple to use because in-between data are omitted.

*Figure 2: Example of a slope graph*

## Cycle plots

A particularly useful variation on the line chart is called a *cycle plot*. This type of plot reveals patterns that repeat over time. In Figure 3, you can see how cycle plots treat certain time dimensions as categorical groupings and align the data accordingly. For example, sales data can be grouped by the hour and the data points overplotted to show cyclical trends by time of day. This provides a visual method by which to find trends and cyclical patterns from stacked data.

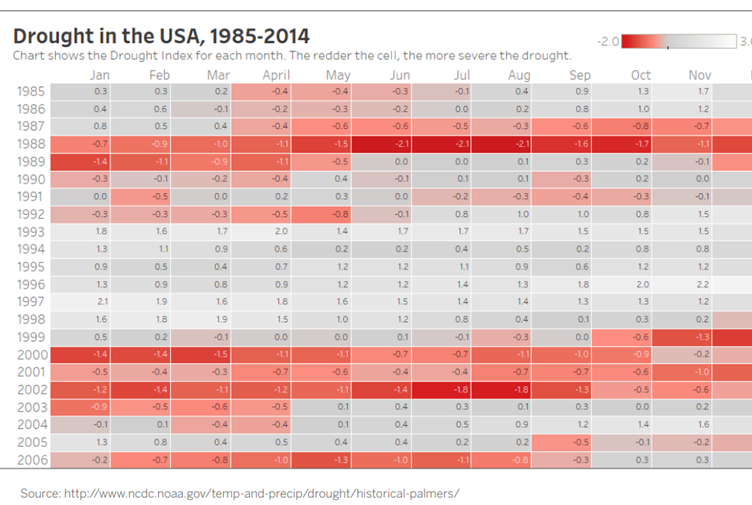


*Figure 3: Sales by the hour during the week (grey) and weekend (red). Note how weekend sales are significantly higher in the morning and late afternoon hours than on weekdays.*

Aligning a time series by the start of an event is another common way to compare the evolution of a set of variables when starting dates or times differ. This is useful for showing relative changes in the rates of increase and decrease as well as trends or cyclicality. While visual analysis of a time series in this manner does not supplant a formal time series analysis, it does inform the modeling of time series methods if prediction is even possible. For descriptive analyses, cycle plots provide a different contextual view of the data without implicit assumptions. The website *Our World in Data* has a great example of this with their [COVID-19 daily new confirmed cases chart.](https://ourworldindata.org/coronavirus-data-explorer?yScale=log&zoomToSelection=true&time=earliest..latest&country=BRA~IND~MEX~NGA~PAK~QAT~ZAF~SWE~GBR~USA&region=World&casesMetric=true&interval=smoothed&aligned=true&hideControls=true&smoothing=7&pickerMetric=total_deaths&pickerSort=desc)

## Highlight Tables

Another way to show cyclicality or seasonality is with a *highlight table* (sometimes called a *heat map*)*,* which uses an aesthetic color layer to indicate another dimension (Figure 4). In this case, the dimension is simply the value of the cell in the table where the cell’s color is determined from a range of values mapped to a color scale.



*Figure 4: Drought in the USA, 1985-2014 (Source:* [*NCDC*](https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00675)*)*

## Decomposition

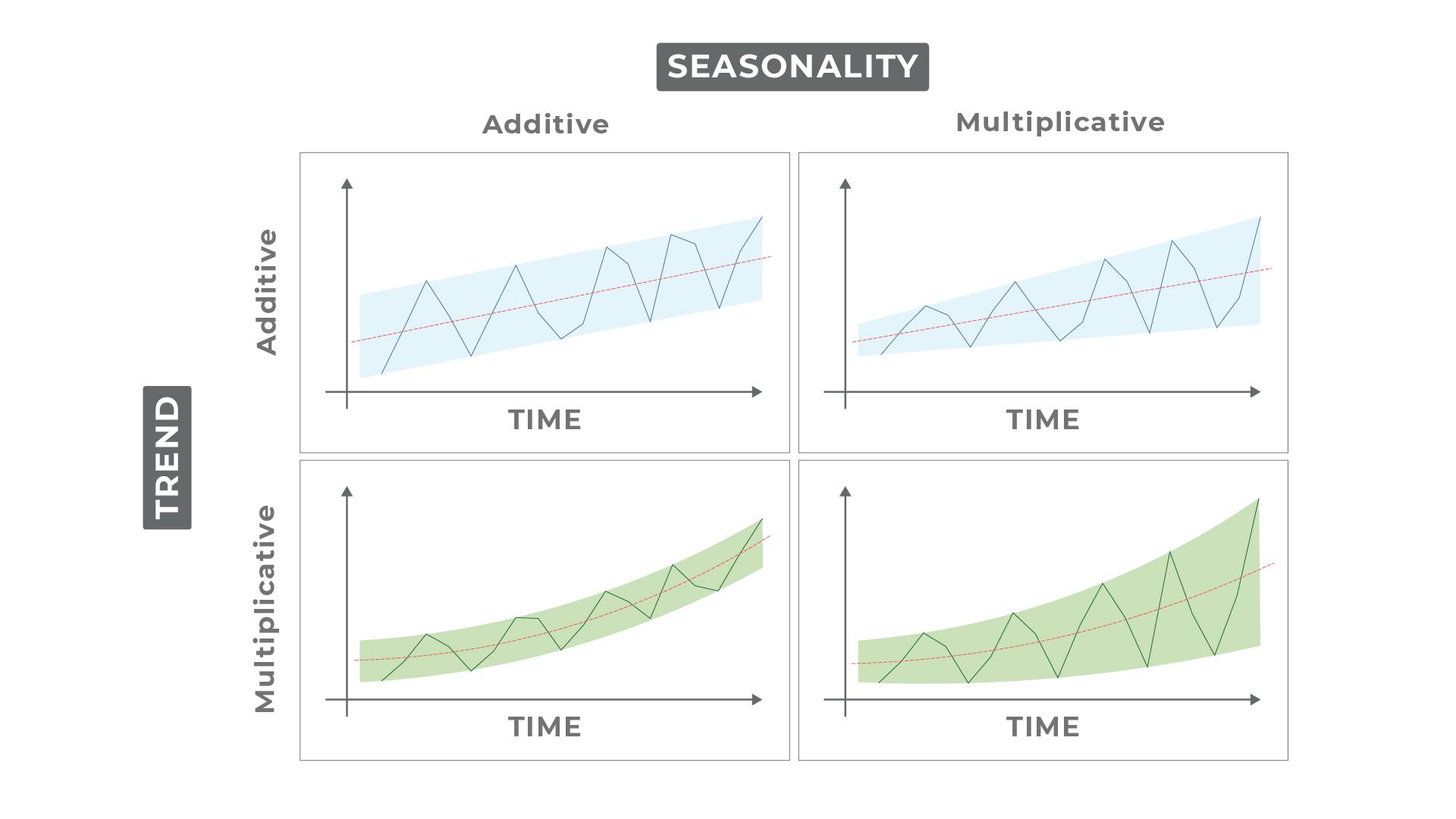
Carrying out a time series analysis is fiendishly difficult—especially during the descriptive analytics stage. The time series charts mentioned above attempt to answer questions like:

* What is the trend of the time series?
* What are the regular fluctuations of the time series?
* What are the long-running departures from the trend during that cycle?
* What aspects of the time series are attributed to noise?

Often, practitioners transform and plot the data using:

* Smoothing methods to reduce noise
* Differencing or log differencing to look at the variations in change
* Linear regressions through time to estimate trends
* Groupings like the highlight table to calculate seasonal/cyclical lifts

The trend and seasonality components (Figure 5) underlying a time series can be multiplicative or additive, which complicates attempts to plot these components independently without time series analysis.

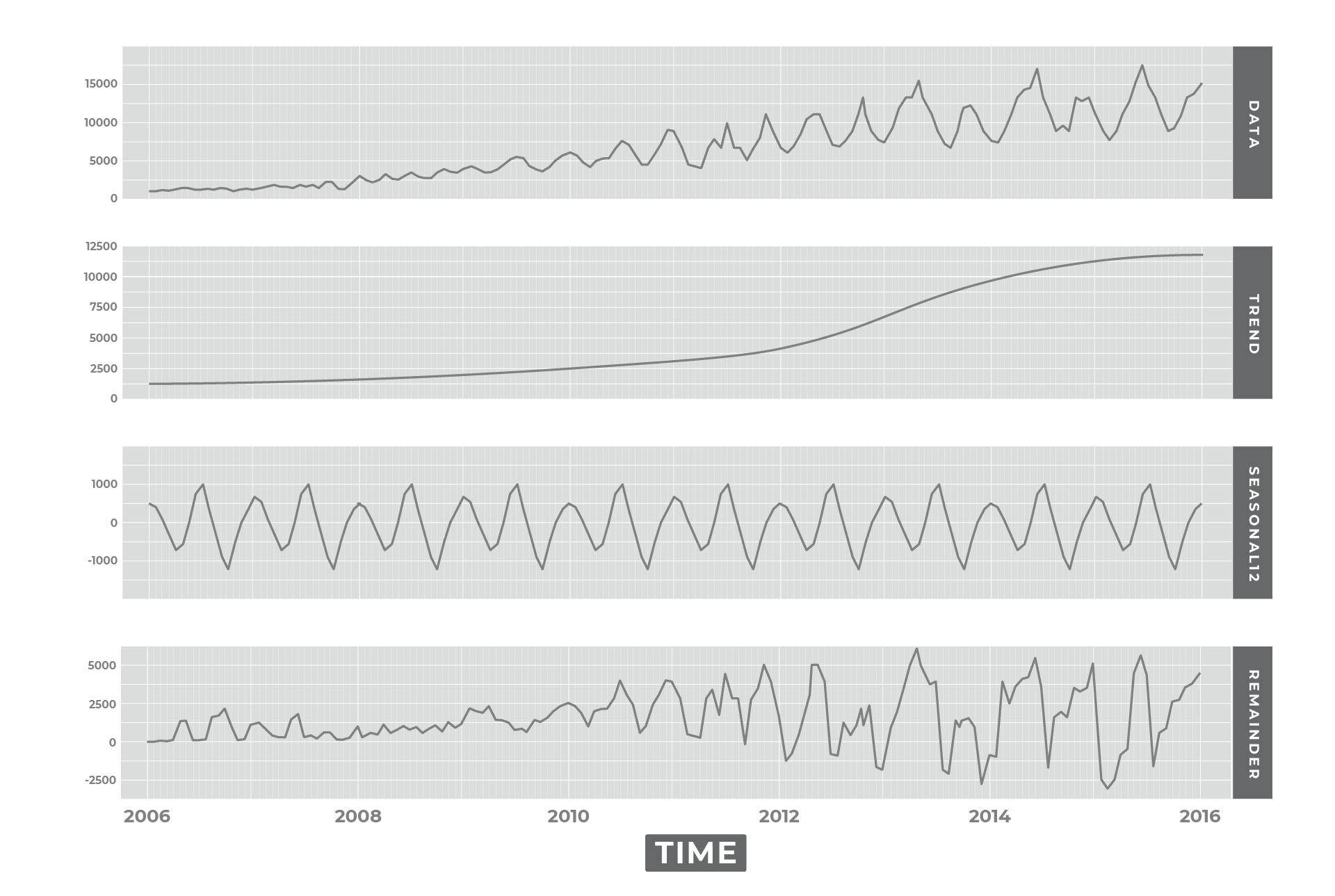


*Figure 5: Seasonality and trend in data*

These questions map naturally to the classical time series decomposition which you may recall as:

where

* *t* is the **time** component.
* *Ut* is the **value** of the time series at time *t.*
* *Tt* represents a **trend**, which is the tendency of a series to rise or fall and exhibit upward or downward movement (both short- and long-term).
* *St* represents **seasonality**, which is the regular fluctuation of a time series within a certain period. These fluctuations form a predictable pattern that tends to repeat from one seasonal period to the next.
* *Ct* represents **cycles**, which are long departures from a trend that occur along longer time intervals than seasonality. The length of time between successive peaks or troughs of a cycle are not necessarily the same.
* *Rt* represents **noise**, which is movement in the series after trend, seasonality, and cyclical movements are removed. Noise is usually random in a time series.



*Figure 6: Seasonal and trend decomposition plot for hotel bookings*

The mathematics are less important at this stage but the conceptual framework provided by the classical decomposition can help you determine how to approach the analysis. For example, if you are analyzing the seasonal impact on sales (Figure 6), you can subtract the long-running trend from the data to isolate the seasonality component. This will show the inherent impact holiday shopping might have on sales. Conversely, you can subtract the seasonal component to view the true trend line of sales to determine whether or not last quarter’s sales increase was due to strategic changes or simply seasonal impacts.

So, instead of struggling with various transformations of the data in a piecemeal fashion, it is often much more expedient to perform the decomposition in one shot and plot the components individually.

Many time series software packages allow you to isolate the various components giving you the freedom to use any plotting package or function. One particularly robust package for modeling and isolating time series components is [Facebook’s Prophet Time Series Library](https://facebook.github.io/prophet/docs/quick_start.html#python-api).

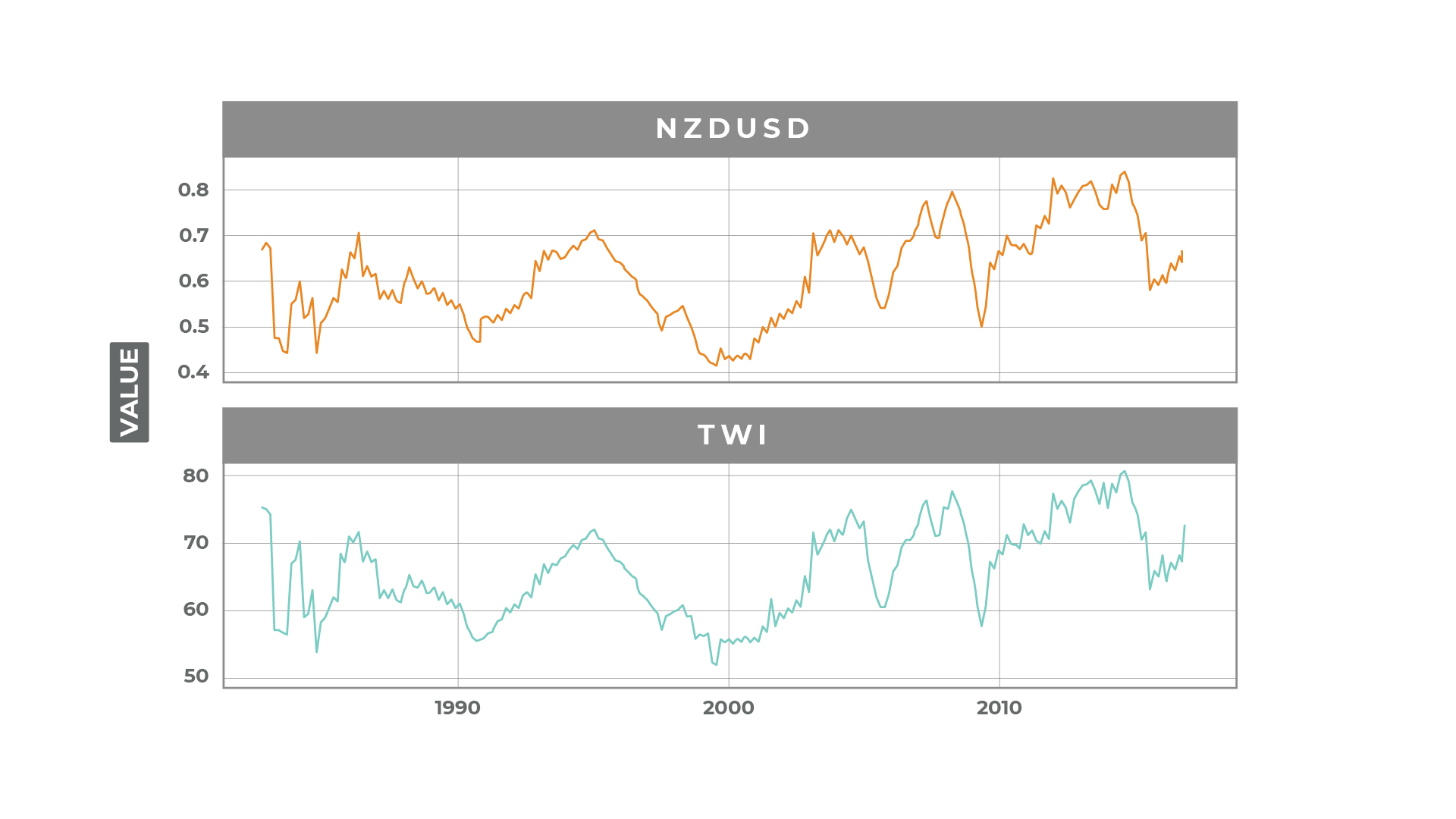
The univariate time series problem is hard, but the multivariate time series problem is even harder because of the dependencies that may exist both within the series and with the time component. If the time order of the values is unimportant, the relationship among the time series values can be visualized as a scatterplot ignoring the time dimension altogether. In most cases, however, the timing and ordering of events are important especially if the aim is to predict future values.

In formal modeling, a univariate time series problem is modeled as a relationship among the series’ prior values (such as the previous time step value, cyclical values, etc.), some noise or randomness, and a dependency on other variables’ values (which may be lagged values, cyclical values, etc.). The dependency among different time series could be bidirectional as well, with one time series influencing others at different moments in time.

## Multiple Line Plots

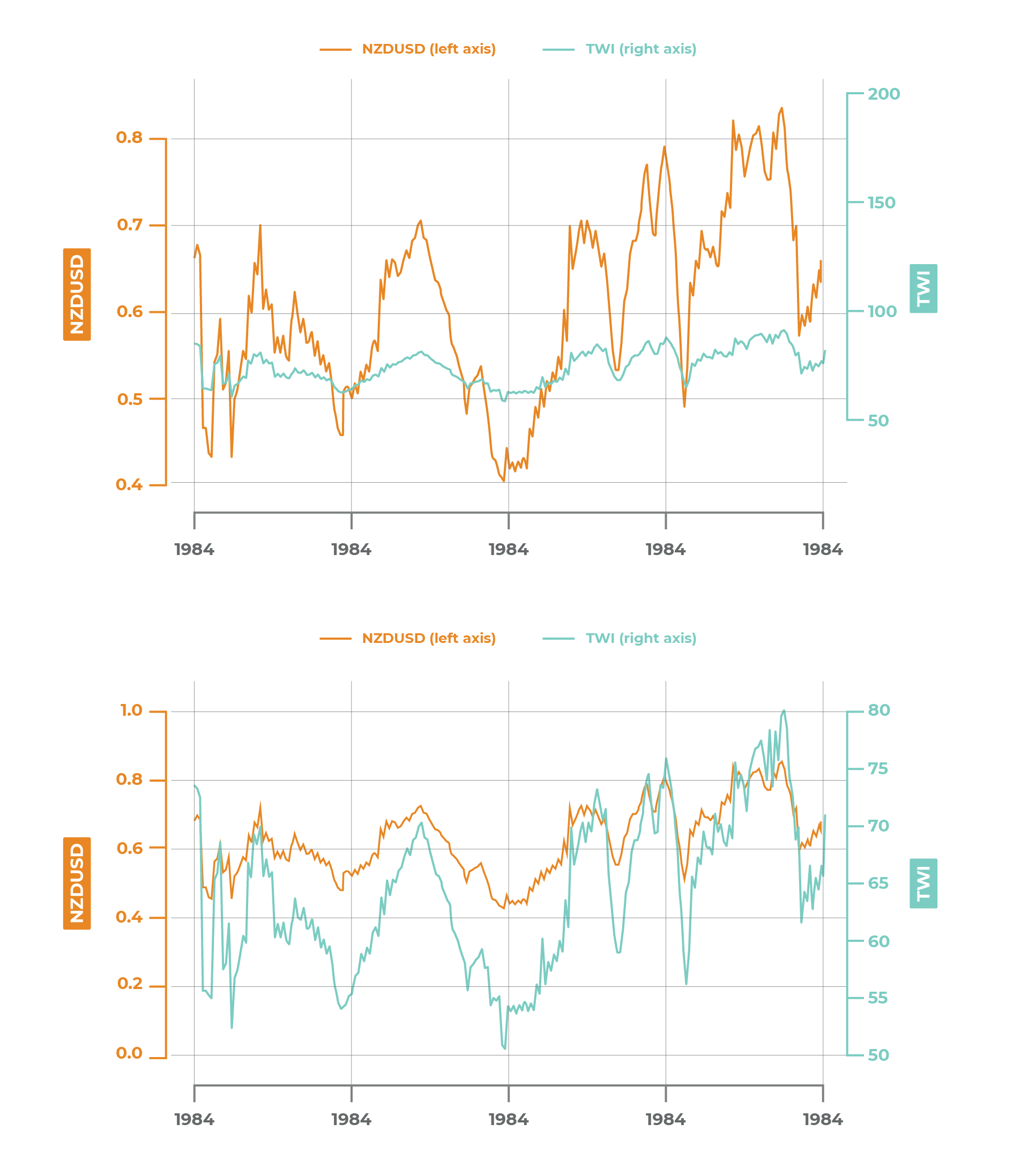
When comparing two or more time series, a number of issues can arise if care is not taken to avoid introducing potentially fallacious comparisons.

In Figure 7, using facets to examine two time series may actually reduce comparability. Superficially, they appear to have a great deal of similarity.



*Figure 7: Comparison of two time series line charts*

However, using arbitrary scales on the axes can lead to widely different conclusions (Figure 8). Changing the scales can either overstate or understate the relationships.

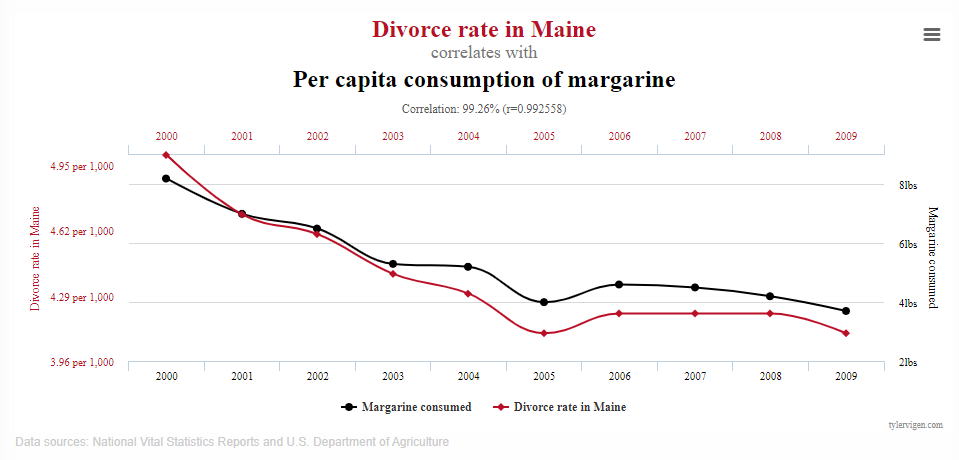


*Figure 8: Arbitrary scales on line plot comparison charts*

When evaluating graphs such as the ones shown above, it’s important to keep the following points in mind:

* The choice of scales can have a big impact on the viewer’s interpretation of the data.
* “Cross-over points” (where one series crosses another) are the result of design choices and are not intrinsic to the data. Viewers (particularly unsophisticated viewers) may attribute more significance to crossover events than they actually merit.
* Artificial visual similarities may cause the viewer to confuse correlation with causation and fail to take into account autocorrelation and other time series issues.

Creating visualizations with seemingly strong correlations (Figure 9) is a rather easy trap to fall into:



*Figure 9: Example visualization with misleading correlation (Source:* [*Spurious Correlations)*](http://tylervigen.com/spurious-correlations)

To avoid these issues when creating multiple line time series plots, consider these practices:

* Transform the series to be measurable on the same scale and plot the series using a single axis. Use color or line aesthetics to distinguish among multiple series.
* If the series cannot be put onto the same scale, use a consistent process to limit and scale both axes (such as using an automated process). Don’t manually adjust the axes until it appears that there is some relationship between the series.
* Use conceptually similar scales on both axes. If growth rate is shown along one axis and raw data are shown along the other, it’s certainly likely to be an “apples to oranges” comparison.

# Key Points

* Time series data can be analyzed with any of the univariate and multivariate analytics methods.
* When looking at temporal questions such as “how did the data change over time,” line plots are the most common form of visualization.
* Slope graphs show a simplified view of data and trend change over time by eliminating interim path values.
* Cycle plots and highlight tables are a few of the common ways to observe seasonality patterns in time series.
* Classical time series decomposition helps break down time series into components: trends, seasonality and cycles, and noise. This may be helpful for drawing attention to one or more of the patterns or structures in the series.
* Multiple line plots with dual axes can pose many interpretation risks. It is best to use these with caution.