Multivariate Descriptive Analytics

Descriptive Analytics: Data Visualization and Storytelling with Data

# Interactive Visual Multivariate Descriptive Analytics

For all but the simplest datasets, complex analytics requires a multivariate understanding of the data being studied. A data set is said to be *multivariate* when it consists of multiple observations involving attributes, characteristics, or properties that may or may not be related. You can think of multivariate analytics as the simultaneous analysis of multiple univariate features. Accordingly, the primary objective of *multivariate descriptive analytics* is to examine how variables change over time from one observation to the next and look for relationships and structures among those variables that show which ones are *independent* of each other and which ones are *dependent* on each other.

Ultimately, we seek to answer the question, “How did \_\_\_\_ happen?” by considering the interactions and relationships among the variables in the dataset in the context of the statistical, mathematical, and heuristic structures that underpin their connections. Visual interactivity with the data is a key component of multivariate analytics and makes finding higher dimensional relationships in complex datasets more intuitive than standard static graphical approaches or those purely focused on computation.

## Scaling and Data Transformations Hi

Before you can compare or contrast values between variables, you need to look at the variable data types. As you recall from the univariate analytics module, data variables can be discrete or continuous. In univariate analytics, the inherent scales of the raw data may not pose any problems as we are often concerned with the unfiltered values.

When two or more continuous variables lie on very different scales, it may be necessary to perform a transformation on one or more of them to allow a comparison. Even comparing univariate data to itself might require a transformation—such as when comparing different time periods or slices partitioned by a discrete variable. These activities are all part of the data preparation process and constitute a necessary step for developing models that involve normalized dependent variables.

In general, these kinds of data transformations fall under the umbrella of *feature engineering* and include *data normalization methods* such as:

* Min-max normalization
* Mean normalization
* Z-score standardization
* Scaling to unit length
* Log transforms
* Differencing

Following are several examples of situations in which a transformation may be necessary.

Suppose you want to compare the economic output of one country to the output of another by comparing the countries’ gross domestic products (GDPs). A direct comparison of these quantities may not be appropriate, but perhaps the *per capita* GDP is more appropriate because it takes into consideration the difference in the two countries’ populations.

Now suppose you want to compare GDP fluctuations from one period to the next for two very different countries—a wealthy country and a not-so-wealthy country. Since the starting points are likely to be very different, it may be more appropriate to look at differences over time in terms of percentages rather than absolute values.

Finally, you could index the two countries’ GDP variables to the same starting date and calculate the cumulative change for each variable since that date. As you can see, there are numerous ways to adjust the variables. The main point is to ensure that the variables are on *the same scale* and are thus *comparable*.

Transformations play a crucial role in avoiding problems with visual fallacies such as that of the dual axis (Figure 1). By altering the scales of the axes, it’s possible to plot two variables in a way that either shows or obfuscates any relationship that may be present. With careful adjustments, it’s possible to identify spurious correlations between completely unrelated variables. (See Tyler Vigen’s [running blog of spurious correlations](http://www.tylervigen.com/spurious-correlations) for a variety of interesting and amusing examples.)

Two curves drawn on different y-axis scales illustrating the ability to manipulate scales so that any desired conclusion can be reached


*Figure 1: Two curves drawn on different y-axis scales illustrating the ability to manipulate scales so that any desired conclusion can be reached*

As you’ve seen, the first question to ask when approaching the comparison of two variables is whether they can be compared directly or if one of them first needs to be transformed into a scale that allows comparison with the other. While this step is also part of univariate analysis, it’s absolutely critical for multivariate analysis.

## Comparisons

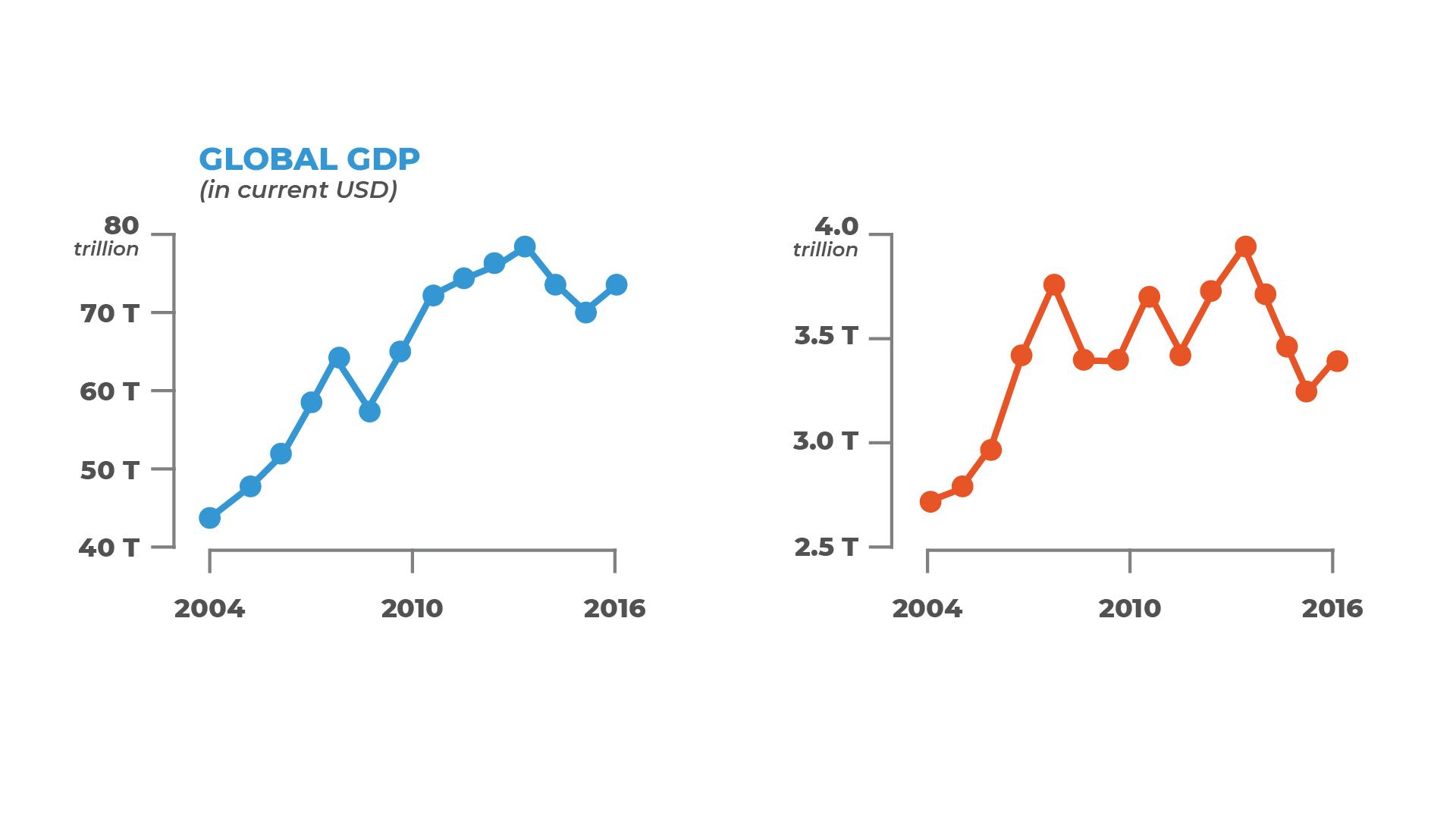
In multivariate analysis, we seek to reveal:

* Differences
* Associations
* Relationships

Because of the visual cues associated with various chart types, certain visualizations are better at expressing these comparisons than others.

### Facets

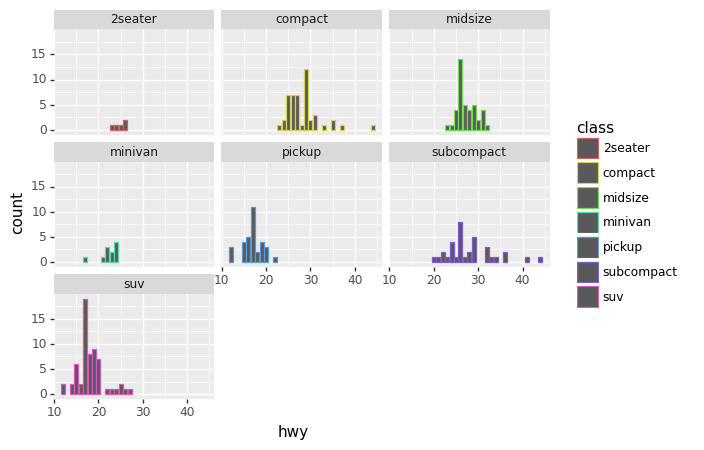
The simplest means by which to compare variables is through side-by-side comparisons.



*Figure 2: Hypothetical side-by-side comparison of global GDP vs. country GDP*

Variables that are left in an untransformed, raw data state can be quickly compared in this fashion although they are still sensitive to axis scaling issues.

*Faceting*can be used to divide the data into subsets conditioned on a discrete variable or on partitions of a continuous variable. This is useful when exploring whether patterns are the same or different across a set of given conditions. Much like side-by-side plots, faceting allows subsets of data to be visualized adjacent to one another (in what is otherwise known as conditioned or trellis plots). The faceting specification identifies the variables that should be used to split up the data as well as how the facets should be arranged.

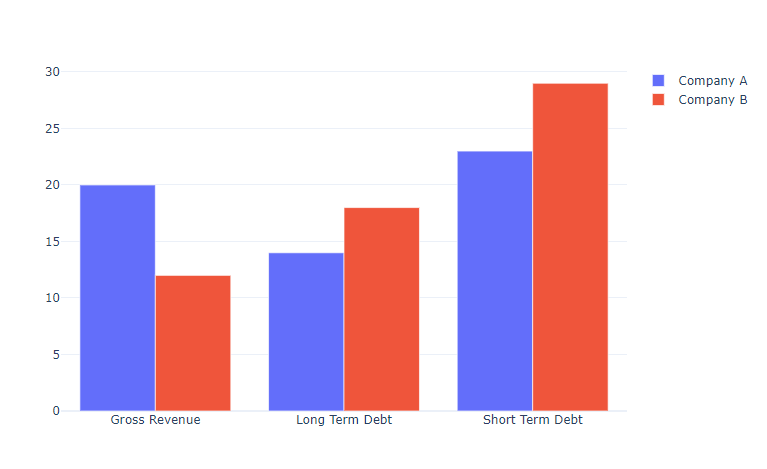


*Figure 3: An example facet showing histograms of highway miles per gallon faceted by the class of vehicle*

The faceted view of the data is a quick way to compare variable measures divided by variable dimensions.

### Layered Univariate Visualizations

Faceting is essentially an application of the grammar of graphics using the additional variable to spatially separate individual graphs. In general, you begin with a univariate visualization and map additional aesthetics to incorporate variables for comparison.



*Figure 4: Comparison of two companies, A and B, using a layered grammar of graphics approach; bars represent absolute level in $10 million (USD) increments with color and position used to distinguish between company A and B values*

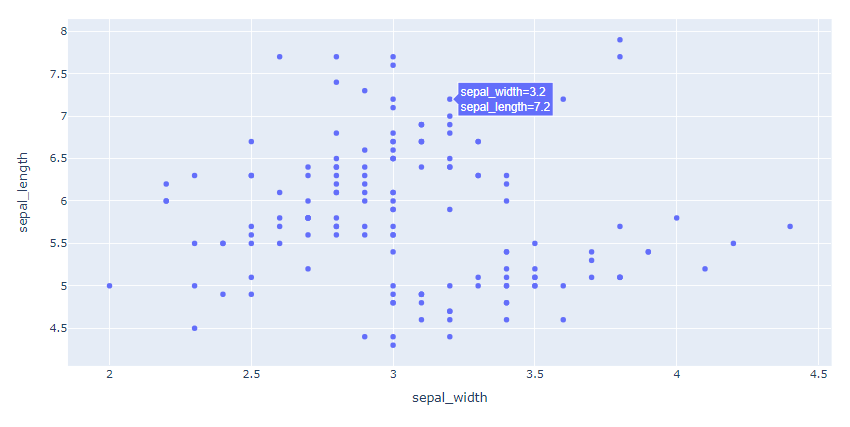
Univariate visualizations such as bar charts and distribution plots can illuminate the differences and similarities among multivariate data when mapped to aesthetics like color.

### Scatter Plots

Comparison views are useful for quickly identifying differences and similarities among data sets. However, gaining an understanding of how two variables are interrelated requires a look at pairwise relationships. For multivariate data, you can use *scatter plots* to probe these relationships visually. A scatter plot graphs two variables in the Cartesian coordinate plane with the x-axis and y-axis representing the ranges of each variable and each point on the graph represents a single observation.

Let’s explore scatter plots in depth using data on iris plants as our example. (This [dataset](https://archive.ics.uci.edu/ml/datasets/iris) is available from the UCI machine learning repository.) The dataset contains 50 samples from each of three iris species (Iris setosa, Iris virginica, and Iris versicolor). Four plant characteristics for each sample were measured (all in centimeters): sepal length, sepal width, petal length, and petal width.

Let’s create our first scatter plot by specifying the x-axis variable as the sepal width and the y-axis variable as the sepal length, and then mapping each data point (xi,yi) onto this plane (Figure 5).

*Figure 5: Scatterplot of iris plants width and length*

Note that:

* The x-axis is mapped to a numeric variable, which is often assumed to be the *independent* variable.
* The y-axis is also mapped to a numeric variable, which is often assumed to be the *dependent* variable.
* As stated earlier, a point is plotted at the (xi,yi) position in the coordinate space for each observation.

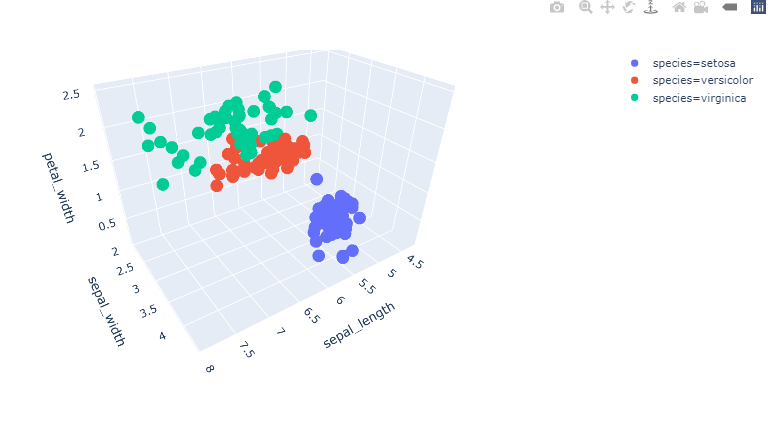
We can layer additional variables on top using:

* **Aesthetics:** Shapes and colors can be used for each point to indicate the values of additional variables.
* **Facets:** Different variable combinations can be viewed simultaneously.
* **Coordinates:** A third axis can be added to show additional observed values in three-dimensional space.

In the original scatter plot above, the addition of color and size aesthetics mapped to species and petal lengths, respectively, distinguishes the species setosa much more clearly (Figure 6).

*Figure 6: Color and size aesthetics mapped to species and petal length*

If we add another dimension to the coordinate system by mapping petal width to the z-axis, we arrive at a 3D scatter plot (Figure 7).



*Figure 7: 3D scatter plot for the iris data*

The 3D plot clearly shows how the species can be identified based on a combination of the three variables. Additional layers for point geometry, color, size, and shape can be added to the 3D plots, as well.

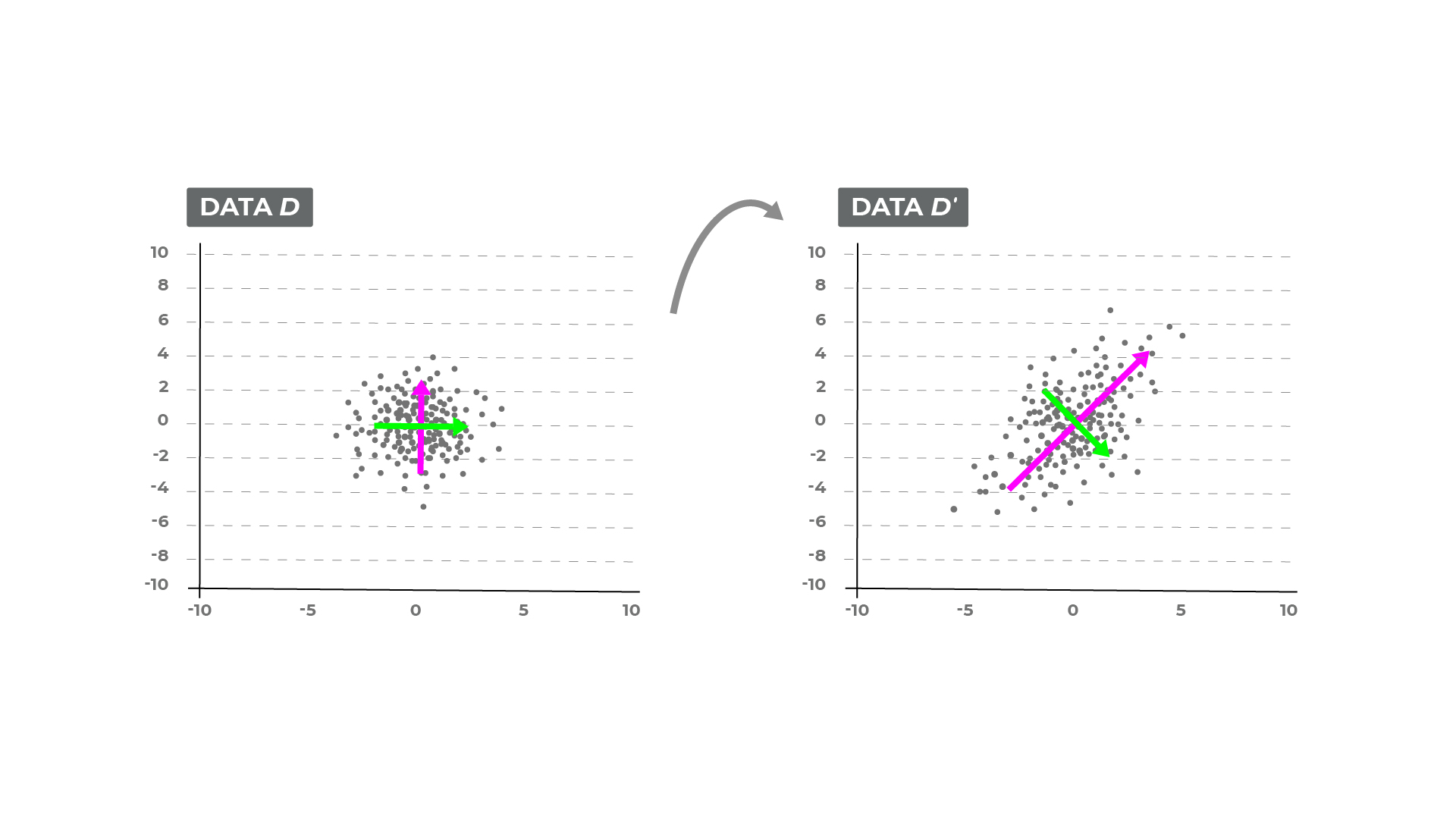
While there are obviously many benefits to using a variety of aesthetics along with 3D plotting, there is a point of diminishing returns when trying to understand multidimensional datasets purely through naive visualizations. In fact, the use of 3D visualizations for other than exploratory analysis is not recommended, particularly in publications or presentations. However, the ability to interact with higher dimensional datasets in 3D (especially in conjunction with dimensionality reduction methods like PCA or t-SNE) makes it an invaluable tool for exploring multiple variables.

Scatter plots have many uses, including:

* Estimating relationships through:
  + Linear regression
  + Curve fitting
  + Classification
  + Model formulation
* Preparing data by:
  + Finding and eliminating outliers
  + Transforming data
  + Filtering data

### Scatter Plot Matrix

When there are many variables in a dataset, creating pairwise scatter plots for every combination of variables is useful as a starting point. This leads to a *scatter plot matrix,* which consists of a grid of individual scatter plots as shown in Figure 8. Note that the diagonal scatter plots have row and column variables that are identical. In those cells, the distribution of the variables is shown as either a histogram or density plot.

*Figure 8: Scatterplot matrix for the iris dataset*

In the scatterplot matrix, we can add layers in the same way as we added layers to individual plots. Here, we apply the species categorization to the hue to show the differentiation. Notice also that the diagonals are now overplotted density curves for each of the species distributions. The off-diagonal cells comprise a symmetric ordering of each pairwise scatter plot.

Scatterplot matrices are used in the same manner as regular scatterplots. Because the grid view considers all combinations of the variables in the dataset at once, this visualization in some sense provides a flattened multidimensional view, which is useful when modeling relationships that have simultaneous dependencies on more than two variables. Of course, the main limitation to scatter plot matrices arises when the number of variables is exceedingly large.

## Statistical Relationships

Visual analytics in conjunction with robust statistical or mathematical methods can help tease out numerical relationships.

### Heatmaps

#### Covariance Heatmaps

The simplest way to explore the relationships among a large number of numeric variables is to calculate pairwise bivariate *covariances* and list them in a *covariance matrix*. Covariance is a measure of the linear association between two variables and is expressed by the standard formula:

for variables X and Y.

* A positive value indicates a direct or increasing linear relationship.
* A negative value indicates a decreasing relationship.
* Values only indicate the direction for the covariance matrix.

The calculation of the covariance has a nice geometric interpretation, which can be understood as a linear transformation of data (see Figure 9).

*Figure 9: Linear transformation of data*

With this concept in mind, the covariance visualization of a large set of variables provides a quick way to view the relationships among the variables in lieu of a scatterplot matrix.

A tabular heatmap of the covariance matrix with the range of covariance values mapped to a color scale


*Figure 10: Covariance heatmap in Seaborn*

Figure 10 depicts a tabular heatmap of the covariance matrix with the range of covariance values mapped to a color scale. The diagonals provide the variance of the random variable (analogous to the scatterplot matrix, which shows the distributions in the diagonals) and the covariance values appear in the off-diagonal cells.

The utility of this view of the covariance matrix is similar to that of a scatterplot matrix but is used particularly when the direction of the relationships is important. It is typically viewed during:

* Feature selection
* Relationship estimation
* Model validation

#### Correlation Heatmaps

The *correlation* between two variables X and Y is defined as the covariance between the variables normalized by their standard deviation. The *correlation matrix* is thus derived from the covariance matrix as follows:

This means that the diagonals of the matrix are necessarily 1.0 (the variable’s variance normalized by its variance). The advantage of a correlation heatmap is that the correlation metric now has meaning for both the direction and the strength of the linear relationship (assuming that the correlation is specifically the Pearson correlation). See Figure 11.



*Figure 11: Correlation heatmap with Pearson correlation*

The correlation matrix and its visualizations are used broadly

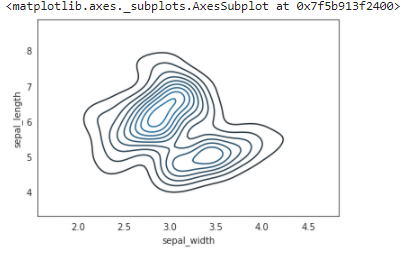
* to summarize large datasets with many variables;
* to view and investigate potential patterns and relationships;
* as inputs into other analyses (correlation matrices are often inputs for exploratory factor analysis, confirmatory factor analysis, structural equation models, optimization, and linear regression); and
* as a diagnostic tool (one example of which occurs in linear regression, where a high degree of correlation between variables suggests that the linear regression estimates will be unreliable).

### Distributions

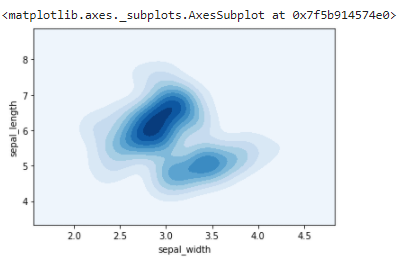
The covariance and correlation matrices provide a narrow view of the full joint distribution among variables in a manner similar to the summary statistics table for univariate analyses. Underlying the covariance calculation is an assumption of linearity—which may not, in fact, be the case. In more advanced analytics, assuming joint distributions for modeling or simulations requires an understanding of the nature of the dependencies among the random variables. This includes using copulas in risk modeling for applications such as quantitative finance, weather and climate modeling, reliability engineering, and many more. Visualizing the joint distribution is one method by which to check modeling and simulation assumptions.

#### Contour plot

Joint distributions are typically viewed as contour plots (Figures 12 and 13) created through kernel density estimates (much as in the univariate case).



*Figure 12: Contour plot with iris data*



*Figure 13: Contour plot of iris data with added color*

Much as with univariate distributions, we utilize the joint distribution plots to:

* visualize the shape of the distribution;
* look at areas of central tendency;
* estimate the dispersion; and
* validate the quality of the data.

# Interaction and Action Intents

The real power of visual analytics lies in the ability to quickly manipulate and interact with data at an abstract level. Interactive visual analytics allows us to manipulate graphical data in real-time. Recall the patterns of interaction and the action intents from the last module (Table 1).

## Interaction Intents

| **Intent** | **Description** |
| --- | --- |
| **Mark as interesting** | Identify and tag interesting artifacts in a representation. |
| See another view, arrangement, or representation | For large and complex data, seeing all the information in a single static visualization would be difficult. Analysts must be able to explore different parts of the data and experiment with various combinations of visual mappings, perspectives, and representations. |
| Show more or less detail | The level of detail needs to be constantly adjusted in order to zoom in on subtleties of the data while also keeping in mind the larger picture. |
| **Show information conditioned on certain criteria** | Interactively filtering out or attenuating data helps limit the view to data that adhere to certain conditions relevant to the task at hand. |
| **Show related information** | Visually linking information helps build a conceptual map of how data in one part of a dataset might logically link to other parts of the dataset. |

*Table 1: Intents and their corresponding descriptions*

## Action Patterns

*Action patterns* capture the actions users will actually take to gain insight during visual data analyses (Table 2).

| **Pattern** | **Description** |
| --- | --- |
| Arranging | Changes ordering spatially, temporally, alphabetically, numerically, etc. |
| Assigning | Binds features or values to be encoded |
| Blending | Fuses visual representations together to form one entity |
| Comparing | Determines relative similarities or differences |
| Drilling down/ Drilling up | Allows changing the resolution of detail of the data |
| **Filtering** | Displays subsets based on certain criteria |
| Navigating | Moves in, around, and through data |
| **Selecting** | Focuses on or chooses individuals or sets of data |
| Collapsing/Expanding | Folds in or compresses visual representations |
| **Composing/Decomposing** | Assembles and groups together representations |
| **Linking/Unlinking** | Creates temporary associations or relationships |
| Storing/Retrieving | Temporarily stores or retrieves representations for quick recall |

*Table 2: Action patterns*

These basic functions, which are needed to interact with univariate data, are similarly applicable to multivariate data. The bold-faced interaction and action intents fall under the idea of *brushing and linking* visual graphics to two or more data variables such that we can physically see structural connections.

## Brushing and Linked Plots

*Brushing* is the action of highlighting regions of a plot. The process of brushing can:

* isolate data points;
* zoom in or out the graph;
* filter in/out points; or
* carry out any of the interaction patterns and intents described above.

Brushing essentially creates a collection of points or a subset region of the data point space.

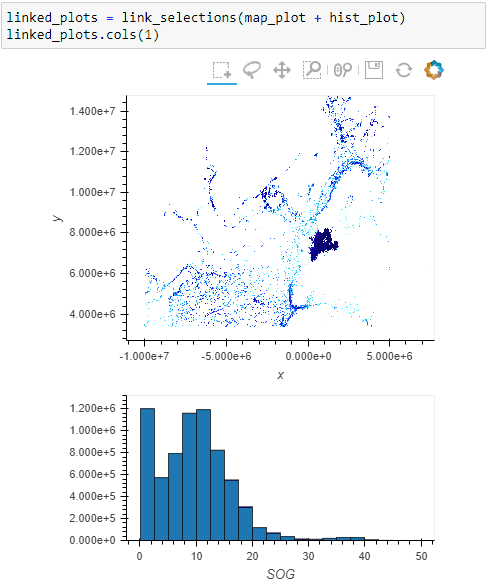
When plots are linked, changes to one plot automatically appear in all the others. Possible actions include:

* zooming in or out the graph;
* drilling down or up;
* highlighting areas;
* isolating data points;
* filtering in/out points; or
* carrying out any of the interaction patterns and intents described above.

A linked action essentially performs a complementary visual analytic action on another view of the data (see Figure 14).

Brushing and linking allow users to investigate relationships in an exploratory fashion. The primary benefits are:

* Local exploration
* Feature localization and isolation



*Figure 14: Brushing and linking in action*

Interactive visual analysis applies brushing and linking to multiple multivariate plots and allows you to view data simultaneously from multiple perspectives, at varying levels of detail, and through different filters. This allows for an assumption-free approach to building mental models of the data you’re analyzing.

# Key Points

As a review of this module, here are several of the key points:

* Scale transformations are often required to compare continuous variables that are not inherently scale-free.
* Multivariate visualizations are primary created to highlight:
  + Differences
  + Associations
  + Relationships
* Multivariate visualizations include the following types:
  + Layered univariate visualizations
  + Faceted graphs
  + Scatterplot and scatterplot matrices
  + Joint distribution plots
  + Statistical heatmaps
* Brushing and linking allow for multivariate exploratory analysis.