Chapter No. 6
"Correlation Analysis"
In this package, you will find:

A Biography of the author of the book

A preview chapter from the book, Chapter NO.6 "Correlation Analysis"

A synopsis of the book’s content

Information on where to buy this book

About the Author

Karl Pover is co-owner of Evolution Consulting (http://www.evolcon.com), which provides QlikView consulting services throughout Mexico. Since 2006, he has been dedicated to providing QlikView pre-sales, implementation, training, and expert services. He has worked in more than 50 companies and government agencies, and set up QlikView competence centers that expand the globe. Most importantly, he has formed a team of highly capable consultants that together have done far more than him.

Recently, he has started a blog (http://www.poverconsulting.com) that will continue to share his experiences in the world of data discovery.

For More Information:
I couldn't have written this book without the loving support and patience of my wife, Pamela.

I would also like to thank the consulting team at Evolution Consulting, especially my business partner, José Angel, and the founding consultants, Carlos and Julian, for their excellent work day in, day out.

Thanks to my old boss, John, for introducing me to QlikView back in 2006.

Finally, thanks to my family, and my friend, Eric, for giving me a shot of confidence and my dog, Axel, for keeping me company during those long nights of writing and revising.
Learning QlikView Data Visualization

Data visualization is a powerful analytical technique and an exciting form of communication. Only in the past few decades has the advent of the personal computer helped it become a more widely used method to explain events, investigate cause-effect relationships, and search for opportunities among growing amounts data.

Data visualization has come a long way since being used solely as eye-candy on top of tabular spreadsheets. Slowly, but surely, we are coming to realize that flashy 3-dimensional charts are neither evidence of a particular software's effectiveness nor of a well-executed analysis. Instead of glitzy graphs, we are now looking for ways to quickly and easily create insightful data visualizations in software that complements our thought processes.

QlikView has been that software for thousands of empowered business users. The ability to rapidly produce powerful analysis and data visualization on top of a data model developed to imitate human thought has placed QlikView at the vanguard of data discovery tools.

QlikView lends the responsibility of choosing the most suitable data visualizations to us. In real world implementations, the freedom to choose how data is visualized has minimized the resistance to the change brought on by migrating old, static corporate reports, and spreadsheets to a new platform. However, we need to grow and learn how to create the best data visualization that allows us to fully benefit from QlikView's dynamism.

In the following chapters, we propose a data visualization style guide and apply it to various forms of analysis in QlikView. We will not cover all of the software's available visualization options, but rather we review a selected mix of both basic and advanced functions that covers the most valuable analytical techniques.

We pack as much content as possible into as few pages as possible to give you a quick return on your investment of time and money. The book is written to be read from start to finish, and then, used as a reference book for your own data discovery experience. We hope this book is part of a continual learning process to create great data visualization with QlikView.

For More Information:
What This Book Covers

*Chapter 1, First Things First*, explains how finding the right people, data, and tools is key to creating great data visualizations. Finally, we start our first exercise in data discovery.

*Chapter 2, Rank Analysis*, explains how to use bar charts to create analysis that ranks values. We introduce the data visualization style guide.

*Chapter 3, Trend Analysis*, helps us discover how line charts show us how our company has changed over time.

*Chapter 4, Multivariate Analysis*, explains ways to analyze a large amount of variables using straight and pivot tables along with heat maps.

*Chapter 5, Distribution Analysis and Statistics*, takes our analysis further by adding more sophisticated statistical analysis. We review the histogram, frequency polygon, and box plot chart.

*Chapter 6, Correlation Analysis*, looks for relationships between variables using scatterplot charts.

*Chapter 7, Geographical Analysis*, brings to light how location adds insightful information with a geographical chart. We introduce the use of extensions.

*Chapter 8, What-if Analysis*, explains how to include variables that we can change to create possible future scenarios.

*Chapter 9, Dashboard and Navigation*, brings everything together to communicate the results of our analysis. In the process, we propose a solution and create a way to monitor its execution.

For More Information:
Correlation Analysis

When analyzing data, we are constantly looking for the causes and effects of certain events. Discovering a causal relationship between events is significant because we can take action upon it. If the effect is beneficial, we will promote what causes it; however, if the effect is detrimental, we will discourage the cause. Correlation analysis will help us discover the relationships between metrics. Then, it is up to us to apply human intuition to determine if they are causal relationships.

What is correlation analysis?

Samantha notes that several projects have incurred losses. What other events coincide with those losses? Could one of the other events be the cause? What is the effect of these losses on QDataViz, Inc.'s financial health?

The losses might have been caused by some descriptive value such as the poor selling practices of a particular salesperson. We could have formed this hypothesis after viewing a simple bar chart that displays profit margins by salesperson.

However, what if Samantha wants to test if losses are related to, and possibly caused by, another metric? If project losses are correlated to the number of external consultants involved in the project, the number of days budgeted for the project, or the number of sunny days during the project then we can recognize this perfectly using a scatterplot chart.

Samantha uses a scatterplot chart to discover a possible relationship, but it is up to her to decide if the relationship is either causal, a related effect of the same cause or purely coincidental. For example, the number of external consultants could be a cause of project losses that could be avoided if more internal consultants were used. Project losses and the number of days budgeted could be a shared symptom of poor project scoping methods. Finally, a relationship between the number of sunny days during the project and project losses is complete coincidence since no outdoor activities are involved.

For More Information:

Correlation Analysis

Samantha will use both visual cues and mathematical methods to prove or disprove relationships between metrics.

**Scatterplot chart**

The scatterplot chart involves plotting a symbol that represents two different metrics along an x-axis and y-axis.

To start, we open our QlikView application, copy the Distribution Analysis sheet and rename the new sheet Correlation Analysis. Finally, we delete the distribution analysis charts from the Correlation Analysis sheet.

We want to analyse the relationship between the number of days worked for a project and project margin, so let's create a chart with the following characteristics:

<table>
<thead>
<tr>
<th>Type of chart</th>
<th>Scatter chart (shown in the previous image)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension</td>
<td>[Project Code]</td>
</tr>
<tr>
<td>First metric</td>
<td>sum([Days Worked])</td>
</tr>
<tr>
<td>Second metric</td>
<td>sum([Profit Margin])</td>
</tr>
</tbody>
</table>

Note that the Expressions window is different for scatterplot charts. The chart wizard reminds us that a scatterplot must have at least two metrics: one metric along the x-axis and one metric along the y-axis. Optionally, we can include a third metric along a z-axis that refers to the size of the dot and not an actual third axis.

In the scatterplot's Expressions window, we create the chart defined in the previous table by selecting Days Worked in the X drop-down box and Profit Margin in the Y drop-down box. QlikView automatically assumes that the values will be summed.

Select the Advanced Mode checkbox to view the normal Expressions window. The first metric corresponds to the x-axis and the second to the y-axis.

For More Information:
We now have the following scatterplot chart:

Data visualization style guide for scatterplot charts
The scatterplot chart contains several new items to consider in our data visualization style guide. We take the time to review a few previous rules and add some new ones.

Rule 1: use adequate labeling
All components of a chart should be made coherent by adequate labeling. We will use **Tahoma** as the font of all labels and use font size 8 to 10.

Define the following labels as we’ve discussed in previous chapters:

<table>
<thead>
<tr>
<th>Label</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chart</td>
<td>Days worked vs. project margin by project</td>
</tr>
<tr>
<td>Dimension</td>
<td>Project Code</td>
</tr>
<tr>
<td>Metric</td>
<td>Profit Margin and Days Worked</td>
</tr>
<tr>
<td>Axes label</td>
<td>$K$ as the Thousand Symbol</td>
</tr>
</tbody>
</table>

Finally, verify that the number format of each expression is **Integer** in the **Numbers** tab and add a reference line at 0 in the **Presentation** tab.

For More Information:  
Trendlines

Trendlines play an important role in visualizing the strength of the relationship between two metrics in a scatterplot. A trendline, or a linear regression, is basically the average of all the plotted dots. The closer all the dots are to the average, the stronger the relationship is between the two metrics.

The correlation can be positive or negative, weak or strong. Often, no correlation exists between two metrics. Rarely are real correlations so simple as the following example correlations, so we draw a linear regression to help determine how two metrics are related, if at all. The following figure shows a few example correlations:

![Trendlines Diagram]

For More Information:
We add a linear regression to a scatterplot chart by the following steps:

1. Right-click on the scatterplot chart and select Properties…
2. In the Expression tab, select the Advanced Mode checkbox.
3. Select the second expression, Profit Margin, and select the Linear checkbox in the Trendlines section.
4. Additionally, select the Show Equation and Show R2 checkboxes in the same Trendlines section.
5. Click on OK.

After adding adequate labeling and a linear regression, we obtain the following chart:

The equation represents the line. If the relationship between the two metrics was strong, then that equation would help us estimate the profit margin based solely on the number of days worked.

The independent variable or possible cause is placed on the x-axis and the dependent variable or possible effect is placed on the y-axis.

The strength of the relationship between the two metrics is represented by the value of $R^2$. An $R^2$ of 1 means that the relationship between the two metrics is well-defined while an $R^2$ of 0 means that no relation exists between the two metrics. Usually, the $R^2$ value falls between 0 and 1, so the closer the value is to 1, the more well-defined the relationship is.

For More Information:
Correlation Analysis

According to the previous chart, no correlation exists between the number of days worked for a project and the project's profit margin. This is not what Samantha expected. Ideally, as the number of days worked increases, so should the project margin. Let's see how we can add more data to our scatterplot chart to solve this mystery.

Rule 2 – convert color into data

We use color to group the plotted dots in the chart and analyze whether stronger relationships exist within those groups. Groups that have a correlation greater than the whole are called clusters.

Before we begin, clone the Days worked vs. project margin by project chart and rename the new chart Days worked vs. profit margin by project (Grouped by division).

Perform the following steps in the Chart Properties window of the cloned chart:

1. In the Expressions tab, expand the Days Worked expression.
2. Select Background Color and click on ... in the Definition textbox.
3. Type the following code in the Edit Expression window:
   ```qlik
   =pick(wildmatch(Division, 'Government','LargeEnterprise','Small and Medium Businesses','*'),Red(),LightBlue(),Black(),LightGray(),)
   ```
4. Click OK.

Since the dots are solid spheres, some dots could hide others that are located below them. We can minimize this defect if we change the dots into crosses in the Chart Properties window.

1. In the Style tab, select the option that includes crosses in the Look section.
2. In the Presentation tab, set the Symbol Size to 3pt.
3. Click on OK.

An alternative solution to prevent dots from hiding under others is to make their colors transparent.

For More Information:
After removing the linear regression, we can analyze whether the projects grouped by their corresponding divisions create clusters within the scatterplot chart as shown in the following figure:

![Scatterplot Chart](image)

**Important function**

Let's take a closer look at the powerful function we recently used in the previous exercise before moving on to the next rule.

- pick(n, expr1[, expr2,...exprN]): The `pick()` function saves time and space. The following if statement works perfectly: `if(x=1,'a', if(x=2, 'b'))`. However, it is much easier and more elegant to write `pick(x,'a','b')`. This is especially the case when the if-statement is extensive. The `wildmatch()` function used in the exercise works well with the `pick()` function because it returns the number that corresponds with the position of the string that matches the field value. For example, if the **Division** value being evaluated is Large Enterprise then `wildmatch(Division, 'Government','LargeEnterprise','Small and Medium Businesses','*')` returns 2. The `pick()` function then uses that 2 to return its second expression.
Rule 3 – add more detail

Like previous charts, a scatterplot chart is a great candidate to convert into a trellis chart. We can also add more detail using the following methods:

- Z axis
- Trails
- Animation

Z axis

The Z axis in QlikView corresponds to the size of the dot instead of a third axis.

Before we add the Z axis, clone the Days worked vs. profit margin by project (Grouped by division) chart and rename the new chart as Days worked vs. profit margin vs. days worked by external consultants by project (Grouped by division).

Perform the following steps in the Chart Properties window of the cloned chart:

1. In the Expression tab, click on Add.
2. Type \text{sum}([^\text{Employee Type} = \text{External}]) [Days Worked] in the Edit Expression window.
3. Label the new metric as Days Worked by External Consultants in the Expressions tab and define it as an Integer in the Number tab.
4. In the Style tab, change the style to hollow circles of varying sizes.
5. Click on OK.

If the circle grows bigger as the values on the x axis and y axis increase or decrease, then we could be visualizing some kind of relationship between the three metrics. We can confirm this by creating two other scatterplots that compare the first and third metrics and then the second and third metrics. The result of our previous exercise is shown in the following figure:

For More Information:
Samantha notices that as the profit margin decreases, the number of days worked by external consultants increases. She creates an additional scatterplot chart to confirm her observation and discovers a negative correlation between days worked by external consultants and the profit margin. Her scatterplot chart displayed in the following chart is filtered to show only the Large Enterprise division:
**Correlation Analysis**

**Trails**

Surprisingly, it is possible to use a scatterplot to show the changes of a correlation over time. The first way to display change is to show the path a dot has traveled over time.

Before we add trails, clone the *Days worked vs. profit margin vs. days worked by external consultants by project (Grouped by division)* chart and rename the new chart *Days worked vs. profit margin vs. days worked by external consultants by project over time trails (Grouped by division)*.

Perform the following steps in the **Chart Properties** window of the cloned chart:

1. In the **Dimension** tab, remove *Project Code* from the **Used Dimensions List** and add *Year* and *Division*. *Year* should be first in the list of **Used Dimensions**.
2. In the **Presentation** tab, select the **Show Arrows** checkbox and increase the **Arrow Size** to 3pt.
3. Click on **OK**.

We change the dimension from *Project Code* to *Division* because too many trails makes the scatterplot chart unintelligible. Now we can see in the following figure how the relationship by division has evolved over the past three years:

---

**For More Information:**

Animation

Hans Rosling is famous for his animated presentations at Technology, Entertainment, Design (TED) conferences where he tells stories with animated scatterplot charts. Each frame of the animated scatterplot chart is a photo of that chart at a certain moment in time.

Before we add animation, clone the Days worked vs. profit margin vs. days worked by external consultants by project over time trails (Grouped by division) chart and rename the new chart Days worked vs. profit margin vs. days worked by external consultants by project over animated time (Grouped by division).

Perform the following steps in the Chart Properties window of the cloned chart:

1. In the Dimensions tab, click on Animate… in the lower, left-hand corner.
2. Select the Animate the First Dimension and the Show Animation Dimension Value checkboxes.
3. In the Alignment section, select Centered in the Horizontal drop-down box.
4. Click on Font… and change the font to Tahoma and the size to 24. Click on OK.
5. Click on OK twice.

Our scatterplot now includes a play button that animates the chart. Animation creates less clutter than trails, and as such, can handle more dots. The scatterplot is shown in the following figure:

For More Information:

Correlation Analysis

Rule 4 – throw away chartjunk
Almost all the ink in the scatterplot chart is used for data or usability. Like previous charts, we recommend removing the print icon from the caption.

If the dots plotted in the scatterplot chart number more than fifteen, displaying the legend is useless and wasteful. We can remove the legend in the Presentation tab of the Chart Properties window.

Rule 5 – respect usability
Like the previous chart, verify the usability within the following areas of the chart object:

<table>
<thead>
<tr>
<th>Area</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caption</td>
<td>Add the shortcut that copies the chart's image to the clipboard</td>
</tr>
</tbody>
</table>

Rule 6 – be honest
The scatterplot is better viewed when both axes have the same length. Redistribute the area used by the chart and other components by holding down Ctrl+Shift and adjusting the red lines that appear within the chart object.

We save our QlikView application before going on to the next chapter.

Summary
Correlation analysis can be used to detect relationships between different metrics. The type of relationship (cause and effect, common symptom, or coincidence) is up to Samantha to decide given her knowledge of QDataViz, Inc.

After conducting her analysis, she has come to the conclusion that the number of external consultants assigned to a project greatly affects its profitability. Before moving on to finding a possible solution, let's look at how we can analyze our data geographically and discover where we are using external consultants.

For More Information:
Where to buy this book


Free shipping to the US, UK, Europe and selected Asian countries. For more information, please read our shipping policy.

Alternatively, you can buy the book from Amazon, BN.com, Computer Manuals and most internet book retailers.