



Creating a Scatter Chart

QlikView Technical Brief

18 Mar 2013, HIC

www.qlikview.com

Contents

Contents	2
Introduction	3
The data	3
The chart	3
Loading the data	4
Creating a basic chart	5
The first steps	6
Correcting the errors	7
Making it beautiful	8
Summary	8
More advanced scatter charts	9
Bubbles	9
Arrows	10
Grouping	11
Animations	12

Introduction

This Technical Brief is an instruction of how to create a scatter chart or bubble chart based on a couple of files with country data fetched from United Nations.

The data

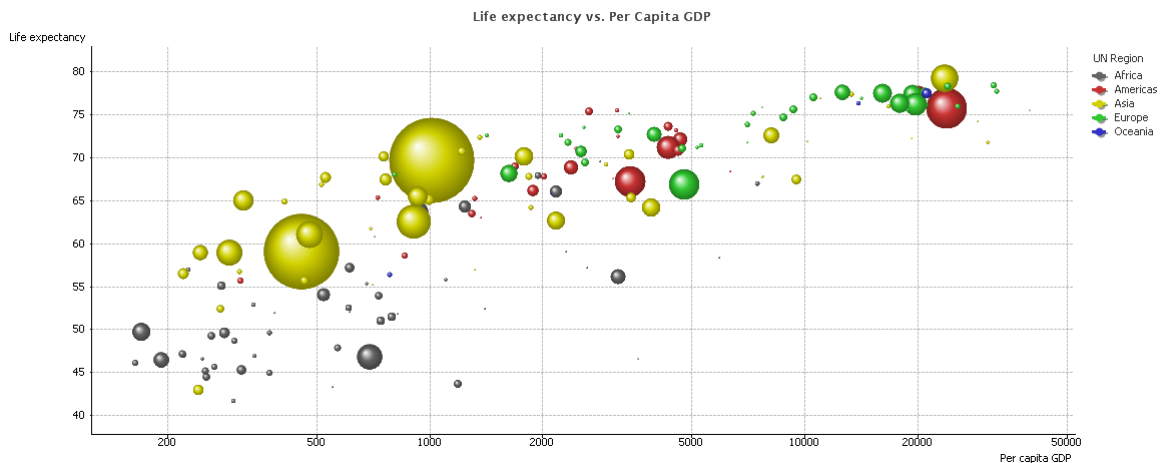
There are two tables:

- UNData.txt
A table listing the population, the life expectancy, and the per capita GDP of the world's countries over a number of years. The numbers are all stored with decimal points (and not decimal commas). There is one record per combination of year and country.
- UNCountries.txt
A table listing the countries and which region the country belongs to. There is one record per country.

Both files are comma delimited text files and can thus be opened with QlikView as well as a number of other programs, including Microsoft Excel.

The chart

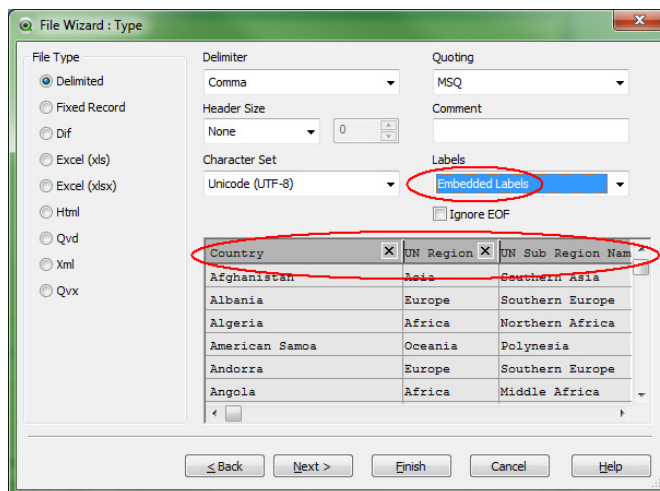
A scatter chart or bubble chart is a chart showing the correlation between two numbers for a set of entities. The picture below shows a bubble chart where each country is a bubble and the two numbers are per capita GDP and life expectancy for the country.



Loading the data

It should be straightforward to load the two files, but I will nevertheless give the instructions from scratch.

1. Download the two data files. Save them in a new folder called e.g. *ScatterChartExample*.
2. Open QlikView.
3. Create a new document. [*New* button in toolbar or *File – New*]
4. If a wizard labeled “*Step 1 – Select data source*” opens, press *Cancel*.
5. Save the QlikView document in the *ScatterChartExample* folder. [*Save* button in toolbar or *File – Save*]
6. Open the script editor. [*Edit Script* button in toolbar or *File – Edit Script*]
7. Tick the “*Relative Paths*” check box.
8. Click the “*Table Files...*” button.
9. Navigate to your *ScatterChartExample* folder. Mark both source files (using <ctrl>-click) and press *Open*.
10. Check that the pre-view of the file looks OK. You may need to change *Labels* to *Embedded Labels* so that you get the field names in the top row of the preview. Click *Finish*.



11. Do the same for the second file.
12. Check that your two load statements look like the script below. (You probably have a “*Directory;*” statement also, and here you can specify a working directory. But since the statement lacks parameter, it is just a placeholder. Delete it, if you want to.)

```

LOAD [UN Region],
     [UN Sub Region Name],
     Country
FROM UNCountries.txt (txt, utf8, embedded labels, delimiter is ',', msq);
LOAD Country,
     Year,
     Population,
     [Life expectancy],
     [Per capita GDP]
FROM UNData.txt (txt, utf8, embedded labels, delimiter is ',', msq);

```

13. Run the script. [*Reload* button in the toolbar]
14. Add the fields *Country*, “*UN Region*”, “*UN Sub Region Name*” and *Year* to the list of displayed fields.
15. Press *OK*.
16. Save the file

You are now done loading data.

Creating a basic chart

Before you start to create your scatter chart, there are a couple of things you need to think about. The first thing is the “Dimension”. Many think that this is identical to one of the axes of the chart, and for most chart types it is. **But not for a scatter chart.** Here, the logical dimension is **not** the same as the graphical.

Instead, you should visualize your graph and ask yourself: “What should each bubble represent?”

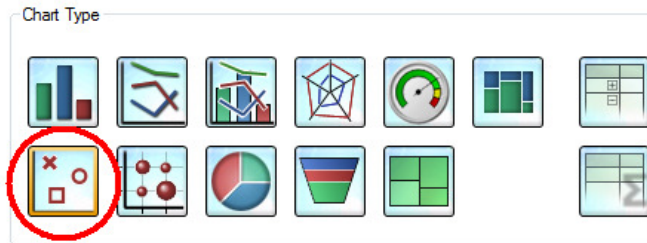
In the graph above, each bubble represents a country. **This** is your dimension. For other data, it could perhaps be one bubble per customer, supplier or product.

The next question is: “Where should the bubble or the point be positioned?” In the graph above, the x-coordinate is *per capita GDP* and the y-coordinate is the *life expectancy*. **These** are your measures – your expressions. For other data, it could e.g. be total order value, gross margin, net cost or some other numbers.

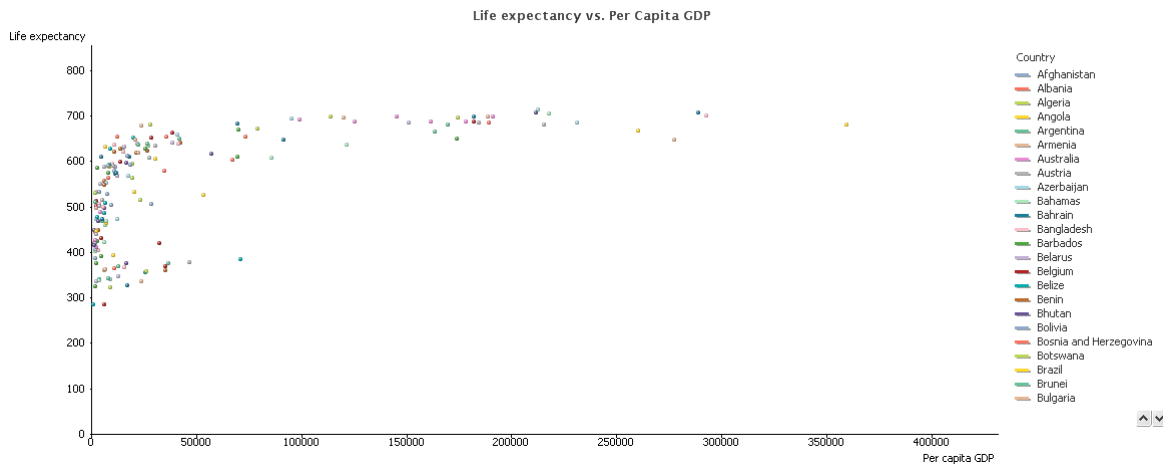
Now that we know which dimension and which two expressions to use, we can start making the chart.

The first steps

1. Create a chart. [Right-click on the background, choose “New sheet object” – “Chart”]
2. On the *General* sheet, choose scatter chart. Click *Next*.
3. On the *Dimensions* sheet, add the field *Country* to the “Used Dimensions”. The reason why you should choose country as dimension is that you want **one point or bubble per country**. Click *Next*.
4. On the Expressions sheet, choose “Per capita GDP” as x-coordinate for the data point (the X control). Then choose “Life expectancy” as y-coordinate for the data point (the Y control).
5. Click Finish.



Now you will have made a very basic scatter chart. It is not yet beautiful and it contains a couple of errors – but it *is* a scatter chart.



Correcting the errors

The above approach is the simple one and it often (unfortunately) creates errors. One error here is that the numbers are wrong: The Life Expectancies are around 400 – 700 years, which obviously is incorrect.

Further, the GDP per capita should for the richest countries be around 30 – 100 kUSD, and here they are much higher. So, what has happened?

The reason for the incorrect numbers is that each country has several lines in the data – one line per year. And QlikView has **summed** over all the years, so the two numbers are much larger than they should be. Instead QlikView should **average** over the years.

Hence, back to the chart properties:

6. Open the chart properties. [Right-click the chart and choose *Properties*]
7. Click the *Expressions* tab
8. Tick the “Advanced Mode” check box.
9. Mark an expression and look in the box to the right labeled “Definition”.

Here we can see the aggregation functions used to calculate the graph coordinates for each country. Both definitions use the Sum() function.

The fact that you need an aggregation function might be confusing: A country has only one GDP, but QlikView still wants you to use an aggregation function, e.g. Sum() or Avg(). The reason is that QlikView cannot “know” if your data has one or several records for each dimensional value. So, you need to use an aggregation function.

In this specific case there are several records possible per country so we need to decide how to aggregate them. The average function Avg() is in this case the appropriate one, but QlikView has chosen the Sum() function. Another possible function could be $\text{Sum}([\text{Per capita GDP}] / \text{Count}(\text{distinct Years}))$.

If there really is only one record per dimensional value, then it doesn't matter if you use Sum() or Avg(). Any of the two will work fine.

In any case, here we need to change the functions.

10. Change the Sum([Per capita GDP]) into Avg([Per capita GDP])
11. Change the Sum([Life expectancy]) into Avg([Life expectancy])
12. Click OK.

Now the chart numbers should be OK.

Making it beautiful

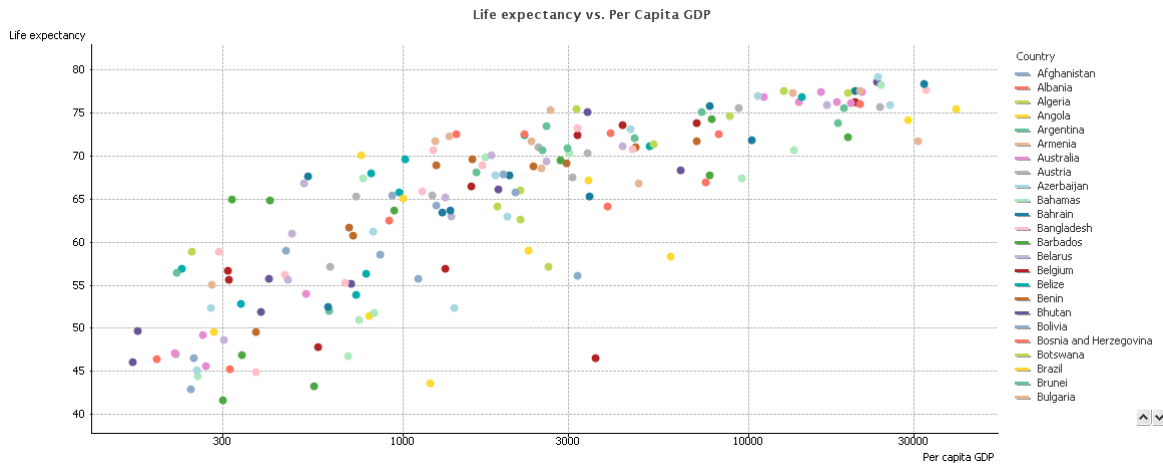
A couple of small additional changes can make the chart a lot more beautiful. First of all, choose a symbol of your liking and make sure that it is big enough so that it is clearly visible:

13. Choose a style with small dots as symbols. [*Style – Look*]
14. Increase the size of the symbols. [*Presentation – Representation – Symbol Size*]

Both axes start at zero, but often you want to look at a more limited range that not necessarily starts at zero. Further, the GDP per capita is several magnitudes larger for the rich countries than for the poor. In such a case it is often useful with a logarithmic axis:

15. Remove the force 0 option for both the axis. [*Axes – X-axis/Y-axis – Forced 0*]
16. Use a logarithmic scale on the x-axis. [*Axes – X-axis – Log Scale*]
17. Turn on grid lines. [*Axes – X-axis/Y-axis – Show Grid*]

Now you have made a scatter chart that shows the right numbers and looks good.



Summary

- Start with one dimension only.
- Ask yourself “What should each dot represent?” Your answer indicates what you should choose as dimension.
- Start with two expressions. They should be aggregations of numeric fields corresponding to the x and y-coordinates in the graph.
- If you have data where you have several records per dimensional value, then it is important that you choose the right aggregation function. Do you want to sum the records? Or do you want the average?

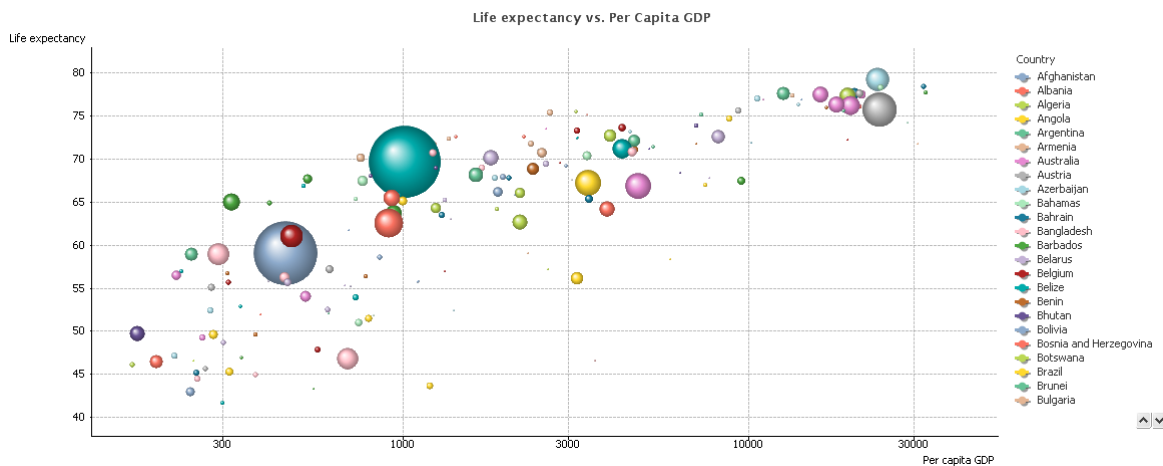
More advanced scatter charts

Bubbles

In the above charts, all dots in the scatter chart have the same size. But it is also possible to use the size of the dot to represent an additional piece of information. With the data used in this case, it would make sense to have the bubble size represent the size of the country's population.

1. Add a third expression: Avg(Population). [*Expressions*]
2. Make sure you use a look that supports differently sized symbols, e.g. the one to the bottom right. [*Style – Look*]
3. Turn the symbol autosizing off. [*Presentation – Representation – Autosize Symbols*]
4. Increase the size of the bubbles. 24 pt gives decent size bubbles. [*Presentation – Representation – Max Bubble Size*]

You should now have a chart with bubbles instead of dots.

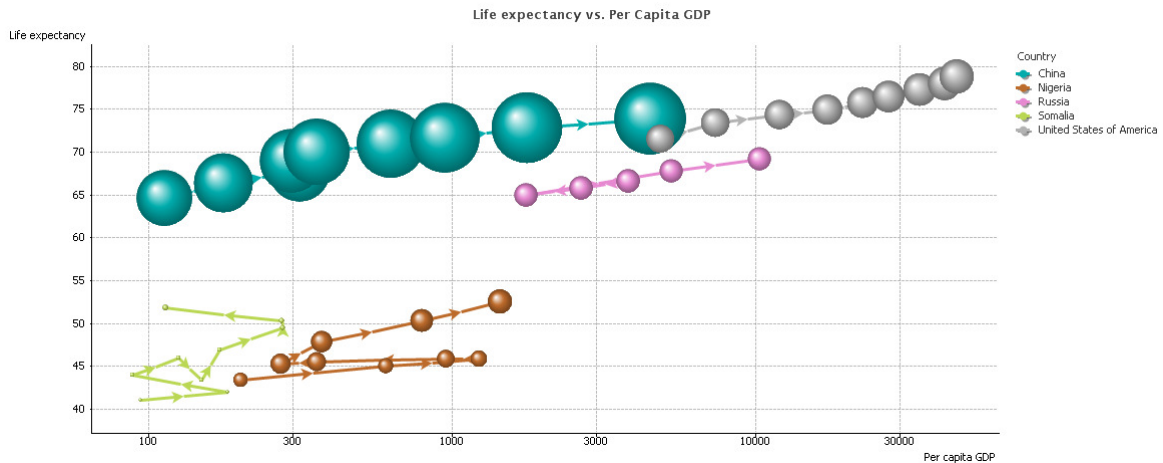


Arrows

Sometimes you want to show the evolution of something and connect the different bubbles with arrows. With the data used in this case, it would make sense to have several dots for each country – one dot per year – and connect them with arrows.

1. Add “Year” as first dimension. I.e. you need to add it and promote it so that it is the dimension in the top of the list. [*Dimensions – Used Dimensions*]
2. Make sure that you have “Both Lines and Symbols” as representation. [*Presentation – Representation*]
3. Show arrows. [*Presentation – Representation – Show Arrows*]

You should now have a chart with arrows between the bubbles. Note that this chart does not make sense if you have too many countries possible. But with just a few, it can be very informative.

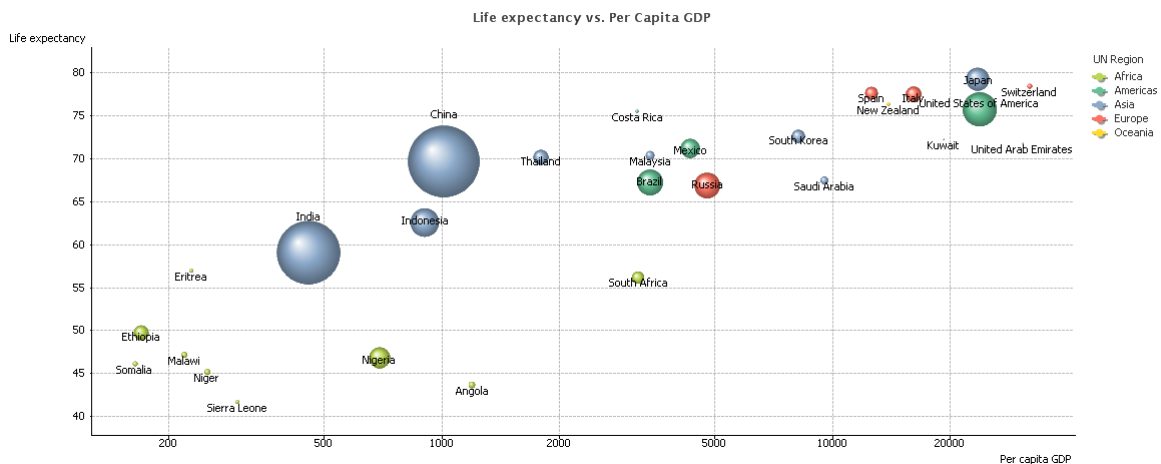


Grouping

In some cases, the dimensional values can be grouped together, products in product groups, customers in regions etc. With the data used in this case, it would make sense to group several countries together in regions. This information can be used in the scatter chart e.g. to color the dots.

1. Add “UN Region” as second dimension. I.e. you need to make sure that “Country” is the first dimension (that it is the dimension in the top of the list) and that “UN Region” is the second. [*Dimensions – Used Dimensions*]
2. Make sure that you have “Only Symbols” as representation. [*Presentation – Representation*]
3. Check “Labels on Datapoints”. [*Presentation – Labels in Chart*]

You should now have a chart where the bubbles have the same color if the countries belong to the same region, and the legend lists the regions and not the countries. Note that the labels on the bubbles will not be displayed if there are too many possible countries.



An alternative way to color the bubbles is to use a color function. In the data sample, there is a Color field that can be used.

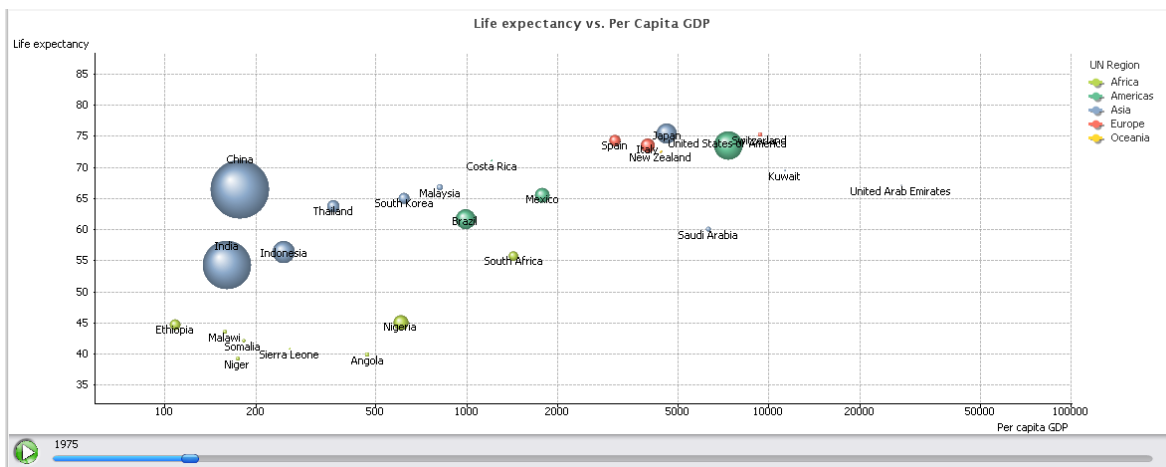
1. Select the background color for the first expression. [*Expressions – The plus sign next to “Per capita CDP” – Background Color*]
2. Set the background color to “=Color”. [*Expressions – Definitions*]

Animations

All graphs can be animated and for a scatter chart this can be very informative. With the data used in this case, it would make sense to have one dot per country and animate this over the different years in the data.

1. Add “Year” as first dimension. I.e. you need to add it and promote it so that it is the dimension in the top of the list. [*Dimensions – Used Dimensions*]
2. Make sure that you have “Only Symbols” as representation. [*Presentation – Representation*]
3. Press the “Animate” button. [*Dimensions – Bottom left*]
4. Tick “Animate first dimension”. Press OK.

You should now have a chart that will be animated when you press the Play button.



HIC