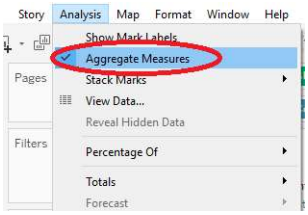
Practice task 1

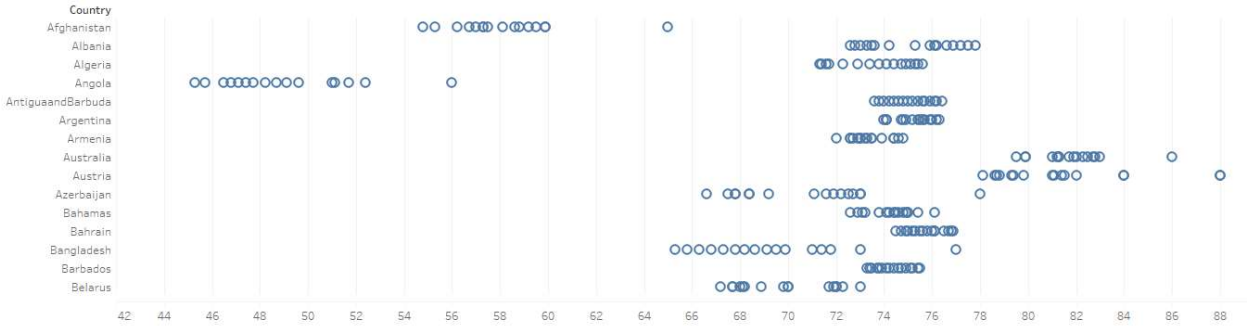
Download the file WHO_health_data.csv provided and load it into Tableau. The csv file contains data collected from WHO (World Health Organisation) and the United Nations website, and contains data by year and country on various health metrics.

The original data file was downloaded from https://www.kaggle.com/augustus0498/life-expectancy-who/downloads/life-expectancy-who.zip/1. There are more details about the different variables on this website.

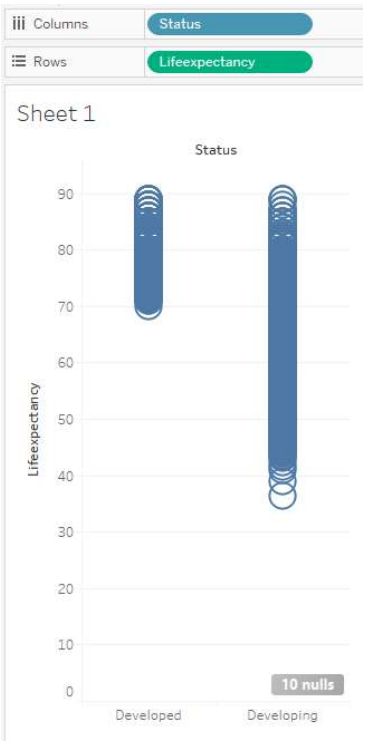
Let's start by making a plot of Life expectancy data by country. Make sure to de-select *Aggregate Measures* from *Analysis* in the ribbon at the top.



The following excerpt has been included for your reference.

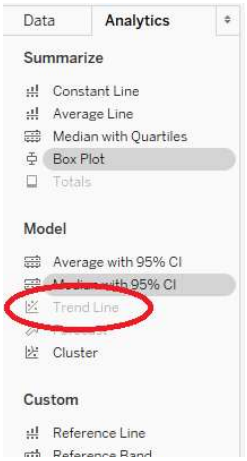


Just from this simple plot, we can get a decent idea of which countries have a relatively higher life expectancy than others. Let's look a bit deeper though. One of our *Dimensions* is "Status", which says whether a country is

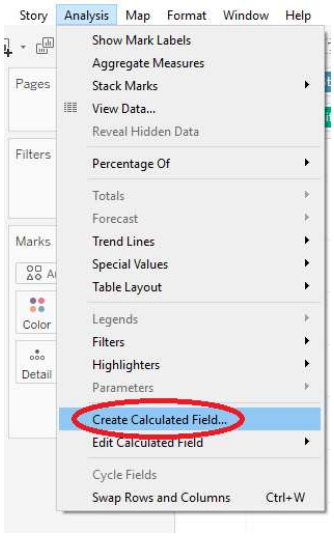


'developing' or 'developed'. Instead of "country", let's plot life expectancy against "Status". You should have the following plot:

As we would expect, our data is telling us that developing countries overall have a lower life expectancy than developed countries. But we would like to get a better idea of the level of this difference. If we go to the Analytics tab, we see that we cannot add a *Trend Line*.



This is because the data in "Status" is being regarded as a string. In order to add a trend line, we will need to convert "Status" into a numeric value. We will do this using a calculated field. Click on *Analysis* and select *Create Calculated Field*.

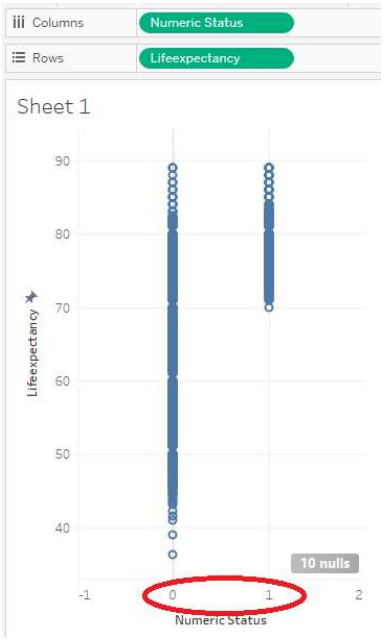


We shall call this calculation 'Numeric Status'. Assign a value of 0 to developing countries and a value of 1 to developed countries using the following code:

(you could just as easily assign the numbers the other way around). Hopefully this calculation is fairly self-explanatory. Press OK.

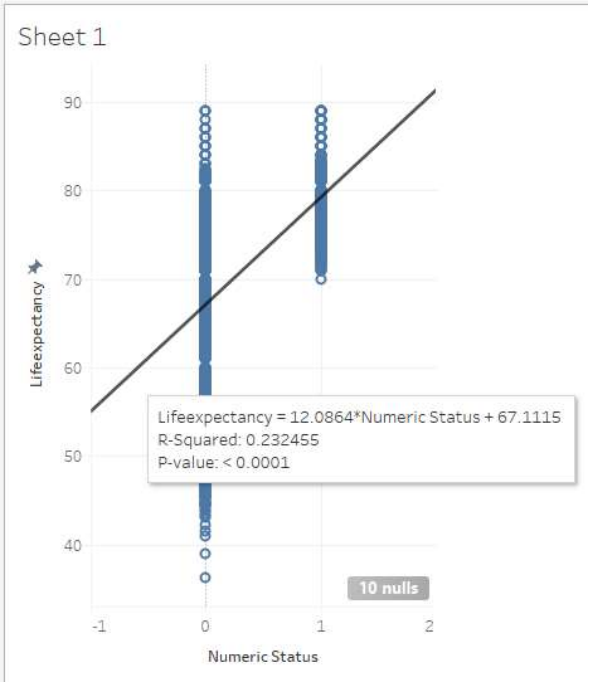
You should see a new *Measure* appear. Leave “Numeric Status” as a *Measure*.

Replace “Status” with our newly created “Numeric Status” in *Columns*. You will see essentially the same plot, but you will notice that instead of the words ‘Developing’ and ‘Developed’ on the x-axis, you will instead have the numbers 0 and 1.



It seems like a small change, but now that “Status” has been converted to a number, we can apply regression. Go back to the *Analytics* tab. You will now find that *Trend Line* has become available. Click-and-drag *Trend Line* onto *Linear*.

A line of best fit has appeared in our graph. Mouse-over the line of best fit for a more detailed description.



What is the description telling us? Well, the formula for the line of best fit is

$$\text{Life Expectancy} = 12.0864 \times \text{Numeric Status} + 67.1115.$$

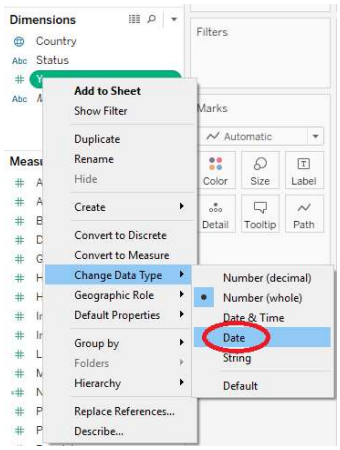
This means that the average Life expectancy across developing countries for all years' data (we will discuss this point in more detail in a moment) is 67.1115, and developed countries have an extra 12.0864 years extra life expectancy (again, across all developed countries and all years).

$R^2 = 0.232455$, which tells us that there is still a lot of fluctuation that is not explained by whether a country is developed or not.

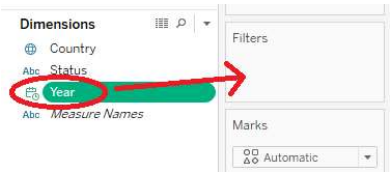
Finally, the p-value is very small, which tells us that whether a country is developed or not does indeed have a significant impact on life expectancy.

Let's look further into this analysis. We remarked that we have combined data across all available years. Since, as a general rule, life expectancy increases over time, it makes sense to try and remove this separate effect when comparing developed and developing countries. To do this, we may wish to only look at data from a given year.

Currently, "Year" is being interpreted as a general numeric value (we can tell because it has an adjacent pound/hash symbol). Let's first convert it to a time measurement. Right-click "Year" under *Dimensions*, go to *Change Data Type*, and select *Date*.

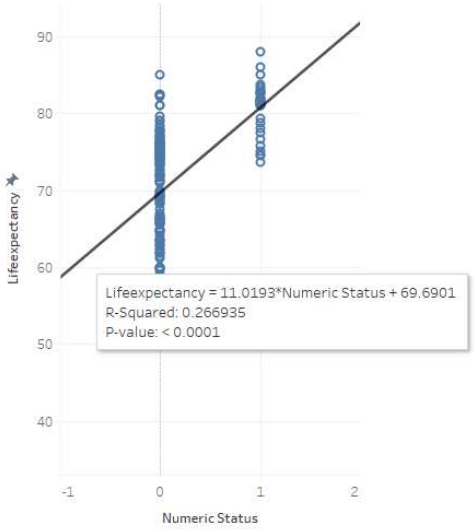


Click-and-drag "Year" from *Dimensions* into *Filters*.



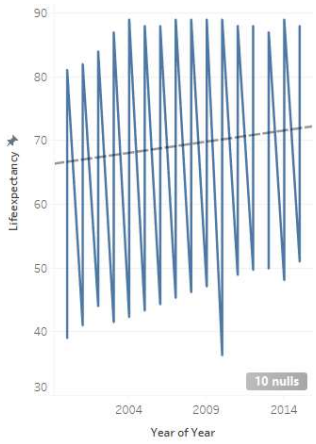
In the new dialog, select *Years* and press *Next*. Let's choose only data for 2015 (you can experiment and see what happens in other years). Select the checkbox for 2015 and press OK.

You should notice that there are fewer data points included in our plot (we have just filtered out all years except 2015).



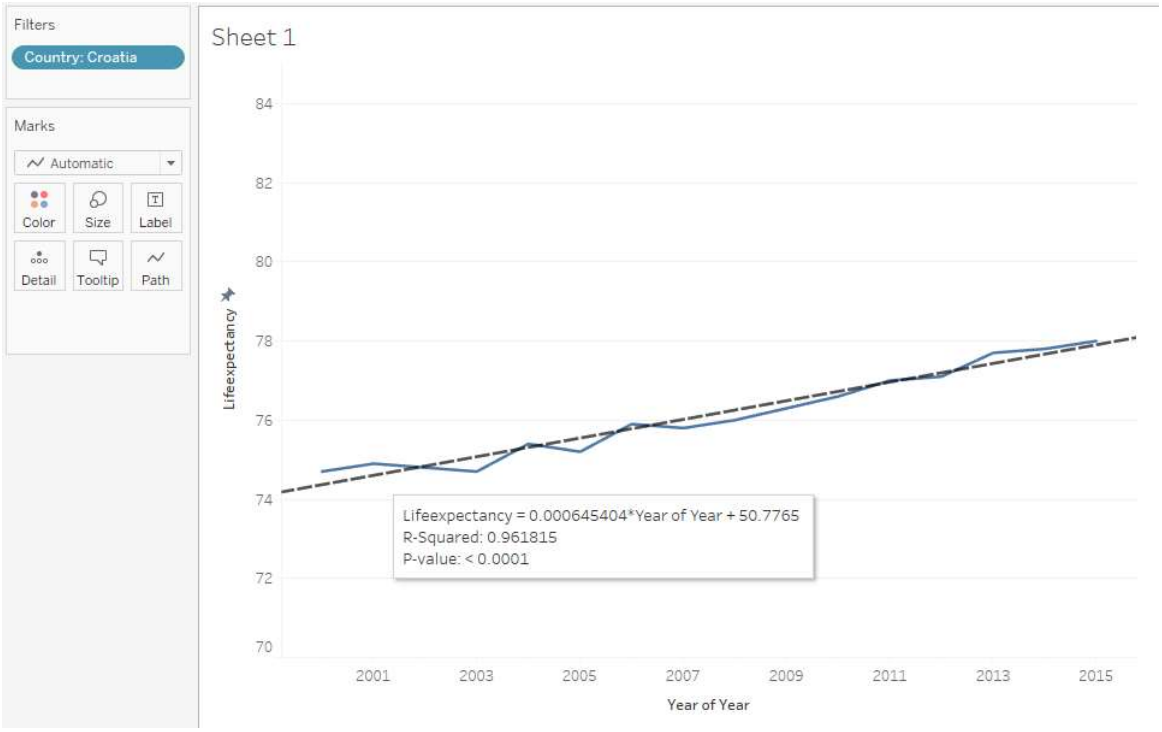
Looking only at 2015 data, we see that the average life expectancy for developing countries is 69.6901 (as compared with 67.1115 when we include all years, supporting our intuition that life expectancy increases over time), and developed countries have an extra 11.0193 years extra life expectancy (we may argue that this means the gap in life expectancy between developed and developing countries is gradually closing).

Since we've mentioned it, let's also look at live expectancy over time. Replace "Numeric Status" in *Columns* with "Year". You should get a graph that doesn't make much sense:



What's happened is that within each year, Tableau has plotted the life expectancies for each country, then tried to connect all of them with a line.

Let's only look at a single country. Click-and-drag "Country" into *Filters*. For our exercise, we shall choose "Croatia" (again, feel free to choose a different country and see what happens in the resulting graph). Press OK.



From the graph, we see that the observed life expectancy over time fairly closely matches the line of best fit. Moreover, the coefficient in our regression is 0.0006454. This is actually on a daily scale (as the time scale is technically in days). Multiplying this number by 365 (we are ignoring leap years for this approximate calculation), we get 0.235571.

This tells us that each year, the life expectancy in Croatia goes up by about 0.236 years. As always, we will leave it to you to play around with this data set and see what other insights can be drawn from this data.

(End of Task)