Cleaning to Viz

The cleaning process for the project was done using Excel and then the visualisation project was done using Power BI. It is my first using Power BI hence why I did the heavy cleaning with Excel. This let me transform the data the way I wanted.

The first step was deleting the columns that were completely empty ().

Then used the filters to find and remove blanks.

Find and replace was used to replace blank educational level with “Not Specified.” I did not see a reason to delete that much data.

The email column was anonymous; I deleted them since they had no use.

The “Current Yearly Salary” column was in a range i.e 10k-50k. This is not usable, so it was split with the formulas, ‘=LEFT(G2,SEARCH("-",G2)-1)’ and ‘=RIGHT(G2,LEN(G2)-SEARCH("-",G2))’

After filling out the newly created cells it was necessary to convert values in thousands that appeared with ‘k’, i.e 1k to 1000. The formula ‘=SUBSTITUTE(SUBSTITUTE(J2,"k",""),"k","")\*IF(RIGHT(J2)="k",1000,1)’

Copy and paste values and delete helper columns. The values are in the right data type to work with

A simple find and replace worked like magic for the column ‘Which Title Best fits Your Current Role’. ‘Other (Please Specify):’ was replaced with blank to leave user descriptions only. This broke down the category called ‘Other’ and gave a broader perspective of the job titles of the survey participants. The same was done for the ‘What Industry do you work in’ column and the ‘Favourite Programming Language Column’ (‘Other:’ replaced with blank). The same was done for the column ‘If you were to look for a new job today, what would be the most important thing to you’? The same was done for the column ‘Which Country do you live in’? Giant of Africa in ‘Other’? And finally, for the ‘Ethnicity’ column.

The data is now quite clean for easy work in Power BI. The data types are right too.

In Power BI a little transformation is done. A custom column for average salary is created.

Both the dirty data from the source and my cleaned data had a major issue in the date column. It was not able to parse the data in the format it was in. This was fixed by splitting the column by characters and then splitting by delimiter (/). These steps left numbers that were changed to whole numbers. The three cells were merged and the type was set to date. The error was fixed.

The transformed data were used to set up a visualisation. See here and give feedback on possible areas of improvement. Cheers!