Data Analyst Survey Project

For this project , where I am doing an analysis of a data analyst survey, I downloaded a dataset and performed transformations within the Power Query Editor in Power BI.

To do this I opened the csv file in Power BI and chose to transform the data.

Data Cleaning

The first thing carried out was removing empty columns that existed on the dataset. This strips unnecessary columns that will be distracting when making visualisations.

After this, I looked at the column 'Q1 – Which Title Best Fits your Current Role?' When selecting the drop down of this column it displays all the unique field names of that column and we can see several that fall under the 'Other' category with a specific profession after it. As this varies the data too much, it would be more appropriate to just rename this to fall under just the category 'Other' without any specifics.

To do this I first Split the Column by a Delimiter being the '(' which was used for the input suggestion in the survey as '(Please Specify). This way anything including and after the open bracket will be taken from the rows of the original column and put in a separate column leaving only the term other 'Other' in the original rows. We can then remove the column that was created from the split.

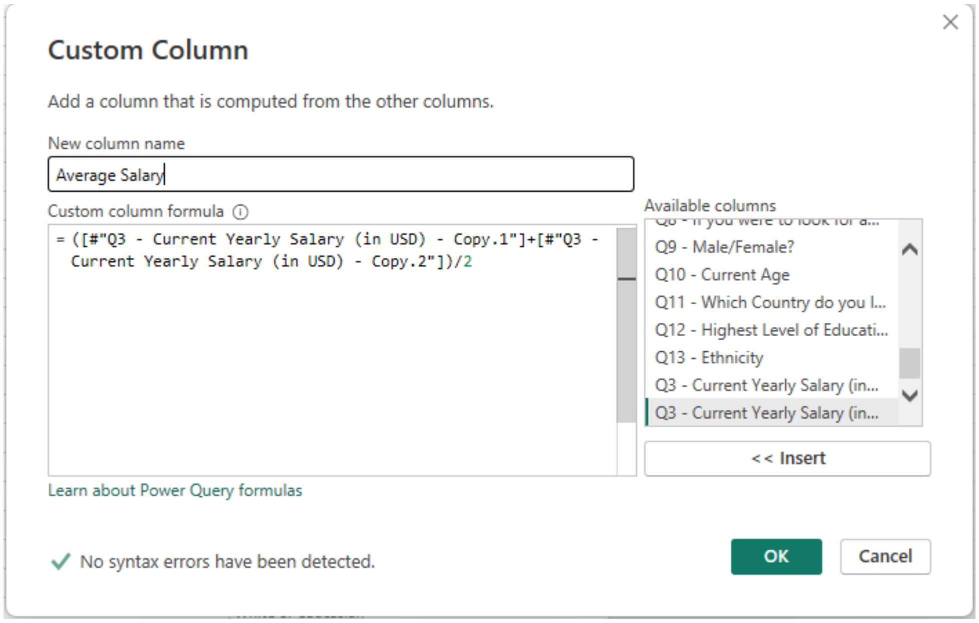
This process is repeated on the column 'Q11- Which Country do you live in?' .

We also repeat this process on the column 'Q4 – What Industry do you work in?' and then we proceed to the next column 'Q5 – Favourite Programming Language' however when we do the Delimiter split, we use a colon as the delimiter.

On the column 'Q3 – Current Yearly Salary' we duplicate the column so that we don't alter the original column. We then split the column using the option 'By Digit to Non-digit' to put the first salary number of the range in the column in one column and the last salary number of the range in another column. In the second column that was generated after the split we had to clean even further to remove the 'k' that represents the thousands and the '-' that was used to represent the range. To do this, we simply use the replace values features twice, once to replace any instance of 'k' and then again to remove the hyphen. We now have two columns with the low end and high end of the salary range for everyone in the survey dataset.

Additionally on the high-end column we must also replace + signs that were used to indicate whether a person made a salary above 225k. As we will be calculating the average salary it would be appropriate to replace the + signs with 225.

With these constructed columns we will create a custom column using a formula. The column will be called 'Average Salary' where we add both constructed columns and divide them by 2.



Custom Column

Add a column that is computed from the other columns.

New column name: Average Salary

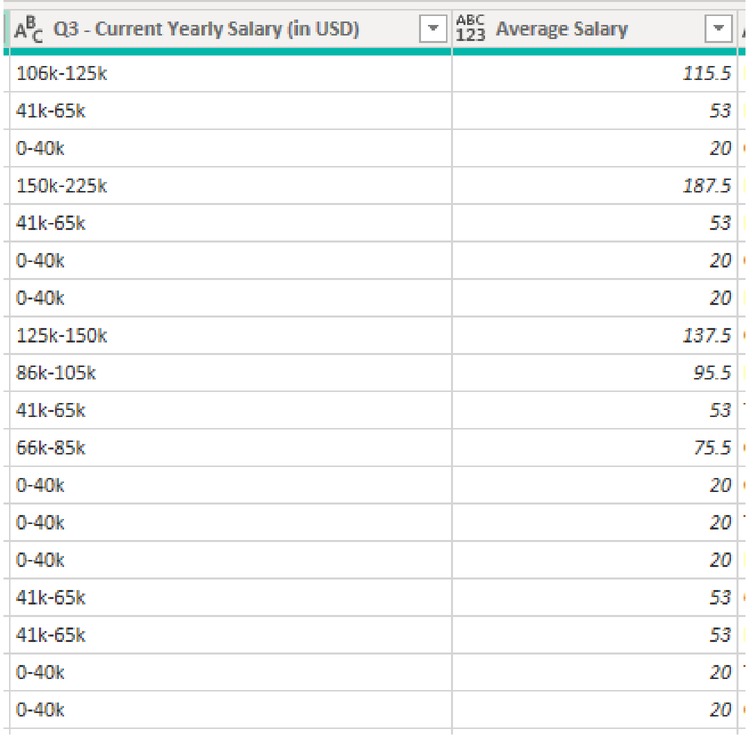
Custom column formula:

```
= ([#"Q3 - Current Yearly Salary (in USD) - Copy.1"]+[#"Q3 -
  Current Yearly Salary (in USD) - Copy.2"])/2
```

No syntax errors have been detected.

Available columns
Q9 - Male/Female?
Q10 - Current Age
Q11 - Which Country do you l...
Q12 - Highest Level of Educati...
Q13 - Ethnicity
Q3 - Current Yearly Salary (in...
Q3 - Current Yearly Salary (in...
<< Insert
Learn about Power Query formulas
OK
Cancel

As result we have a column providing an Average Salary of each member of the survey. This can be used for better visuals when creating the dashboard.



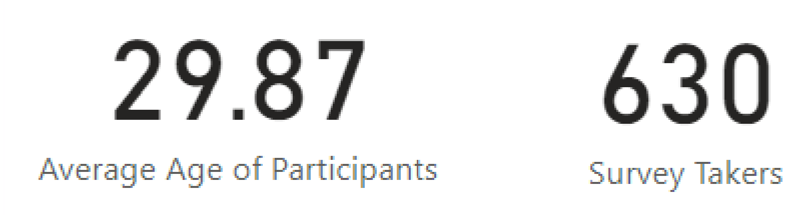
Q3 - Current Yearly Salary (in USD)	Average Salary
106k-125k	115.5
41k-65k	53
0-40k	20
150k-225k	187.5
41k-65k	53
0-40k	20
0-40k	20
125k-150k	137.5
86k-105k	95.5
41k-65k	53
66k-85k	75.5
0-40k	20
0-40k	20
0-40k	20
41k-65k	53
41k-65k	53
0-40k	20
0-40k	20

After carrying out the changes we can select close and apply the power query editor to apply the changes to the loaded data.

Creating the Dashboard and Analysis of Visuals

After creating a title and adjusting it to cover the length of the board. I will start off with a simple card that will display a count of the unique Id and rename it to be the 'Count of Survey Takers'.

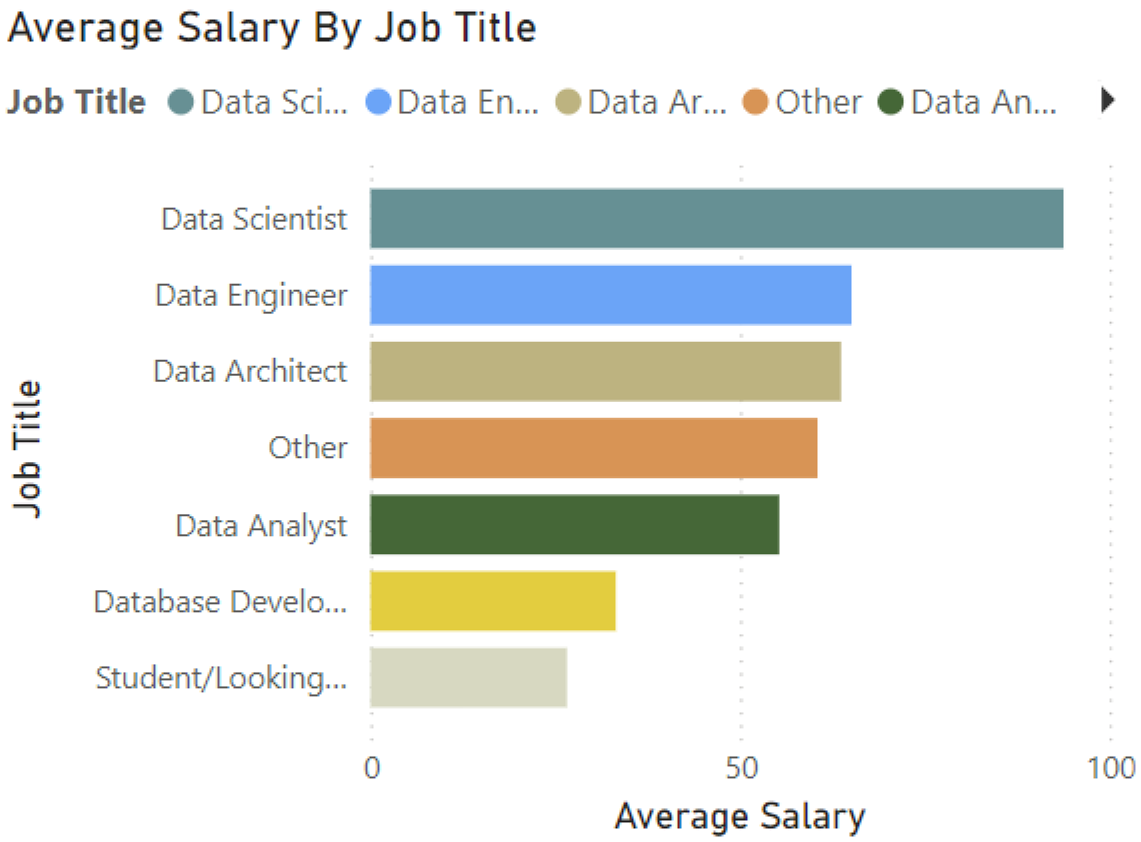
We will also create another card, that will display an average aggregate of the survey takers' age.



29.87
Average Age of Participants

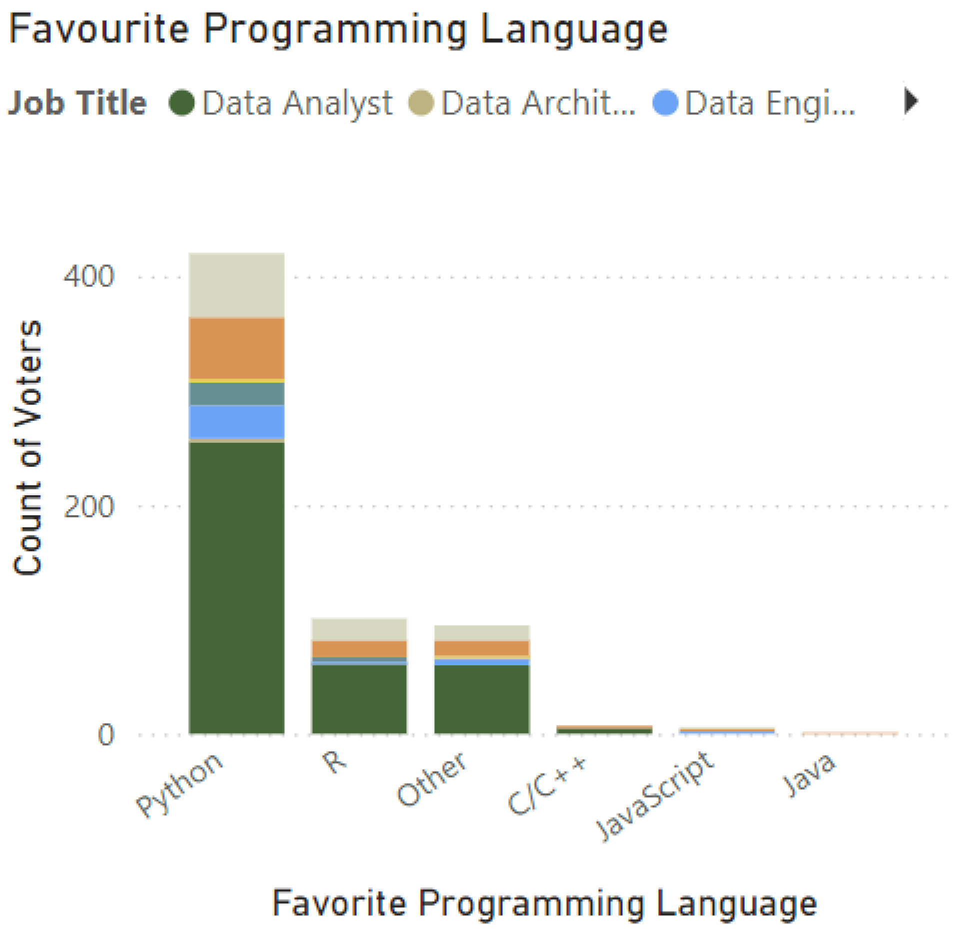
630
Survey Takers

Next, I create a stacked bar chart that will discover the average salary by job title. To do this, I will use the custom column we created 'Average Salary' as the x axis and the 'Job Title' being the Y-axis, so it breaks the salary down by job titles.



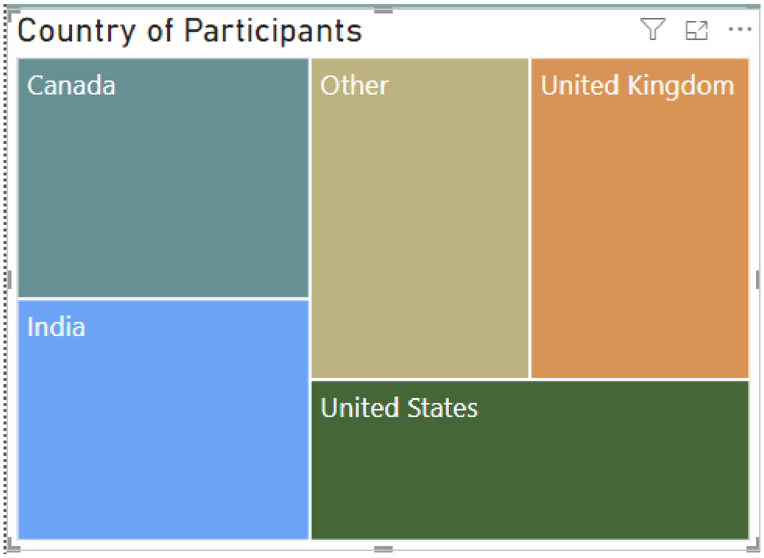
630 people took part in this survey. From the visual we can see that the highest paying profession amongst the survey respondents are Data Scientists making an average salary of 93.78k. The job title with the lowest salary is 'Students/Looking for Jobs'. This makes sense as these respondents are students who may be working part-time whilst looking for a job related to their education.

The next visual focuses on what was the survey takers favourite programming language in conjunction with their current job title. To display this, I created a stacked column chart where the X-axis represented each programming language, the Y-axis representing the number of voters (using a count aggregate on the id) and each portion of a bar was coloured in relation to the legend created based on the Job title.



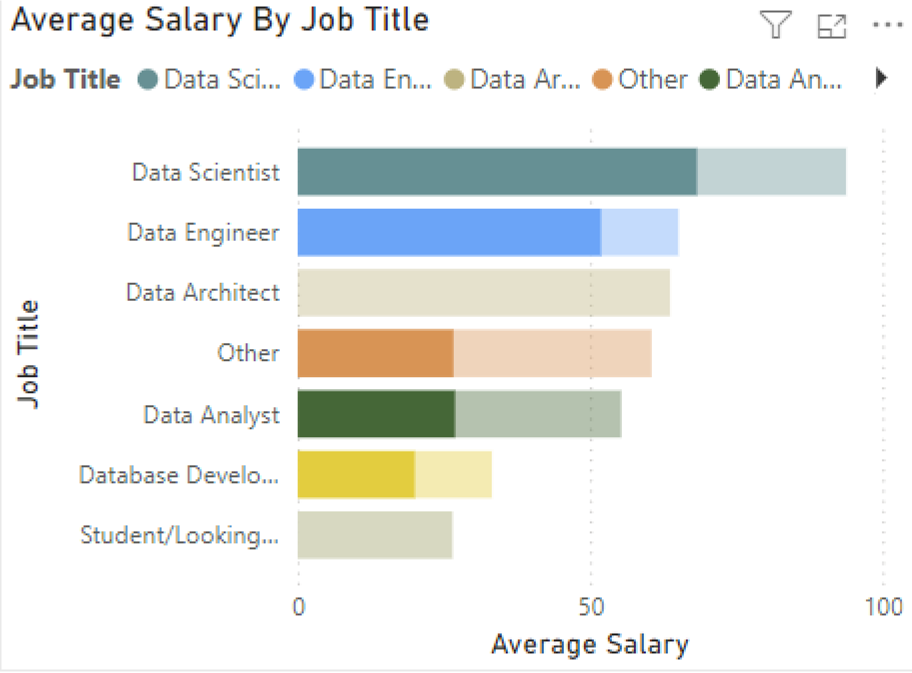
From this visual we can see that Python is vastly more popular amongst the respondents particularly those who are currently Data Analysts as it is comprised of 255 respondents. Python is favored by each profession significantly more than other programming languages and this can be understood as the user-friendly aspects of the language are assumed to be better in comparison to others.

The next visual is being created to show the various countries the respondents are from to show the diversity of the respondents. This was done by using a tree map and setting the values to show the distinct countries of the survey.



With each country displayed on the tree map the user can click on each country to filter the dashboard visual to show the visuals as a view of that data filtered by the country.

For example, by selecting India on the tree map we can see an altered visual of the Average Salary by Job Title.



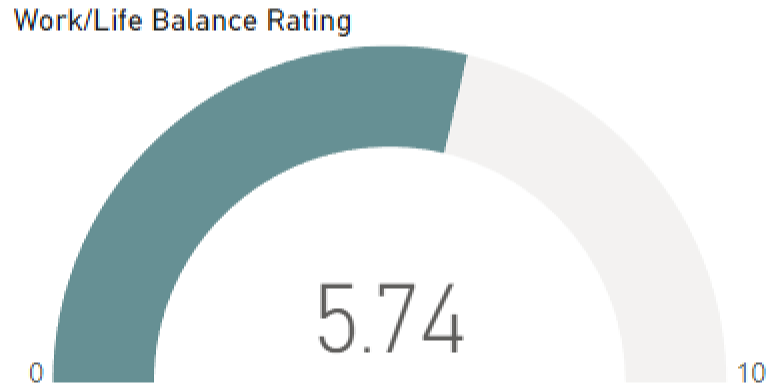
In India Data Scientist is the highest paying job with an average salary of 68.25k whereas there is no salary for Data Architect. With no salary for Data Architects in India, we can assume that there is no Data Architect Role in India.

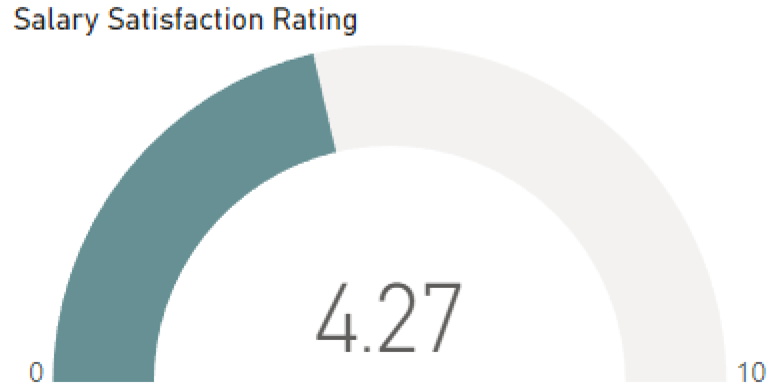
We can also see that there were 73 participants of the survey from India who had an average age of 27.

The next visuals that were implemented were gauges. These are useful to gain insights into work/life balance and salary satisfaction amongst respondents.

So to create the Work/Life Balance Rating gauge we set the Minimum value of the gauge to be the minimum aggregate of the column ‘How Happy are you in your current position with your Work/Life Balance?’. The maximum value was set to be the maximum aggregate of the same column. Finally, the value that was to be read from the gauge was set to be the average aggregate of the same column.

We repeat the aggregates on another gauge but use the column ‘How Happy are you in your current position with your salary?’

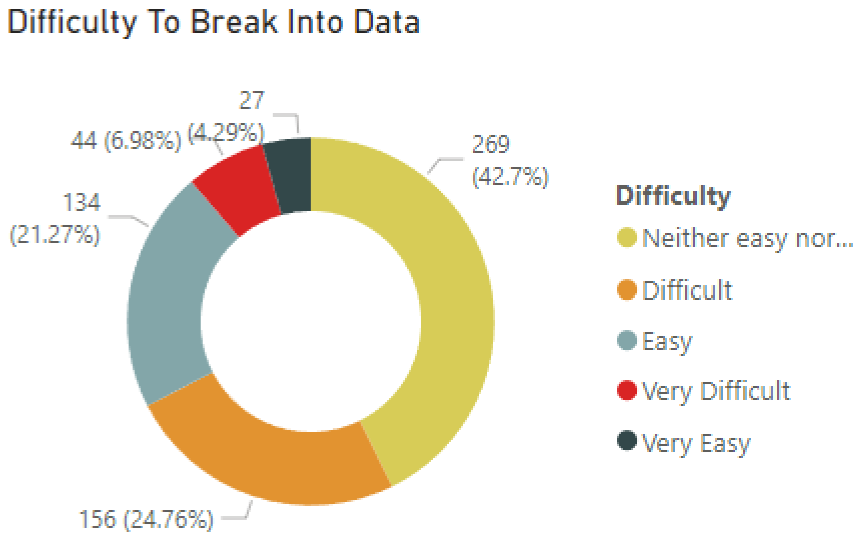




As you can see, from the first gauge the average Work/Life Balance Rating for the respondents of this survey is 5.74. This means that the consensus is that the Work/Life Balance is manageable and positive, but it still requires a lot of improvements. This could be on fair deadlines, more holidays and less load per individual in a company.

The second gauge we can see that the average Salary Satisfaction Rating is 4.27. This is below half so we can assume that most respondents are not satisfied with salaries in the related job fields. So, companies could provide benefits, incentives and rewards to improve satisfaction with salaries.

The final visualization is a donut chart that focuses on the question ‘How difficult was it for you to break into Data?’ the respondents were split by their responses, and it shows us the following:



We can see that the highest proportion of respondents (42.7%) found it was neither easy nor difficult to get into data. This was closely followed by respondents finding it difficult (consisting of 24.76%). From this we can assume that it is typically moderately challenging to enter the data field. This suggests that there is a need for structured learning resources, mentorship programs and useful career guidance in order to help ease the transition into data focused roles.