Data Cleaning and Transformation

Image Credit: https://analyticsindiamag.com/10-best-data-cleaning-tools-get-data/

- ## Summary of Data Cleaning Steps

Step Count	Steps to perform EDA	Tools Used
Step 1	Importing the data	Python
Step 2	Selecting only the required columns	Python
Step 3	Replace blank/Nan values with "Not Answered" as it's a survey-based data and hence can't take mean value etc.	Python
Step 4	Meaningful renaming of the Columns	Python
Step 5	Exporting and saving "refined file" only with the required data for analysis, in xlsx format(Filtered_Data.xlsx)	Python
Step 6	Formatting Salary Column from Filtered_Data	Excel
Step 7	Converting "Job Satisfaction rating" to numeric value for better visualisation	Excel

- **Screenshots for the Steps with Python**

Importing data using Python

```
In [3]: # Step 1: Load the data
        import pandas as pd
        data_frame = pd.read_csv('D:\DA Course Material\DAB 103 Python etc\Project 1 Final Files\survey_results_public.csv',low_memory=Fa
        data_frame.head()
```

Out[3]:

	Respondent	Hobby	OpenSource	Country	Student	Employment	FormalEducation	UndergradMajor	CompanySize	DevType	...	Exercise	Gend
0	1	Yes	No	Kenya	No	Employed part-time	Bachelor's degree (BA, BS, B.Eng., etc.)	Mathematics or statistics	20 to 99 employees	Full-stack developer	...	3 - 4 times per week	Ma
1	3	Yes	Yes	United Kingdom	No	Employed full-time	Bachelor's degree (BA, BS, B.Eng., etc.)	A natural science (ex. biology, chemistry, phy...	10,000 or more employees	Database administrator;DevOps specialist;Full-...	...	Daily or almost every day	Ma
2	4	Yes	Yes	United States	No	Employed full-time	Associate degree	Computer science, computer engineering, or sof...	20 to 99 employees	Engineering manager;Full-stack developer	...	NaN	Na
3	5	No	No	United States	No	Employed full-time	Bachelor's degree (BA, BS, B.Eng., etc.)	Computer science, computer engineering, or sof...	100 to 499 employees	Full-stack developer	...	I don't typically exercise	Ma
4	7	Yes	No	South Africa	Yes, part-time	Employed full-time	Some college/university study without earning ...	Computer science, computer engineering, or sof...	10,000 or more employees	Data or business analyst;Desktop or enterprise...	...	3 - 4 times per week	Ma

- We used Pandas and read_csv command to read the raw csv dataset file. Since the file size is huge, hence we use "Low memory"= False

Selecting only the required columns for analysis

```
In [3]: #Step 2: Select the required columns
        filtered_data_1 = pd.read_csv('D:\DA Course Material\DAB 103 Python etc\Project 1 Final Files\survey_results_public.csv',
                                      low_memory=False,usecols=['Country','FormalEducation','CompanySize','DevType',
                                                                'JobSatisfaction','LastNewJob','ConvertedSalary'])
        filtered_data_1.head(8)
```

Out[3]:

	Country	FormalEducation	CompanySize	DevType	JobSatisfaction	LastNewJob	ConvertedSalary
0	Kenya	Bachelor's degree (BA, BS, B.Eng., etc.)	20 to 99 employees	Full-stack developer	Extremely satisfied	Less than a year ago	NaN
1	United Kingdom	Bachelor's degree (BA, BS, B.Eng., etc.)	10,000 or more employees	Database administrator;DevOps specialist;Full-...	Moderately dissatisfied	More than 4 years ago	70841.0
2	United States	Associate degree	20 to 99 employees	Engineering manager;Full-stack developer	Moderately satisfied	Less than a year ago	NaN
3	United States	Bachelor's degree (BA, BS, B.Eng., etc.)	100 to 499 employees	Full-stack developer	Neither satisfied nor dissatisfied	Less than a year ago	NaN
4	South Africa	Some college/university study without earning ...	10,000 or more employees	Data or business analyst;Desktop or enterprise...	Slightly satisfied	Between 1 and 2 years ago	21426.0
5	United Kingdom	Bachelor's degree (BA, BS, B.Eng., etc.)	10 to 19 employees	Back-end developer;Database administrator;Fron...	Moderately satisfied	Between 2 and 4 years ago	41671.0
6	United States	Some college/university study without earning ...	10,000 or more employees	Back-end developer;Front-end developer;Full-st...	Slightly satisfied	Less than a year ago	120000.0
7	Nigeria	Bachelor's degree (BA, BS, B.Eng., etc.)	10 to 19 employees	Designer;Front-end developer;QA or test developer	Slightly satisfied	Less than a year ago	NaN

- The Raw data of the survey is humongous and all of the data is not required for us to complete the analysis, hence, here, we are selecting only those columns which are required to complete our task, making our file size smaller and more tidier

Replacing the NA Values with customised phrase

```
In [5]: #Step 3: Replace Nan values with "Not Answered" as its a survey based data and hence cant take mean value etc.
filtered_data_2=filtered_data_1.fillna("Not Answered")
filtered_data_2
```

Out[5]:

	Country	FormalEducation	CompanySize	DevType	JobSatisfaction	LastNewJob	ConvertedSalary
0	Kenya	Bachelor's degree (BA, BS, B.Eng., etc.)	20 to 99 employees	Full-stack developer	Extremely satisfied	Less than a year ago	Not Answered
1	United Kingdom	Bachelor's degree (BA, BS, B.Eng., etc.)	10,000 or more employees	Database administrator;DevOps specialist;Full-...	Moderately dissatisfied	More than 4 years ago	70841
2	United States	Associate degree	20 to 99 employees	Engineering manager;Full-stack developer	Moderately satisfied	Less than a year ago	Not Answered
3	United States	Bachelor's degree (BA, BS, B.Eng., etc.)	100 to 499 employees	Full-stack developer	Neither satisfied nor dissatisfied	Less than a year ago	Not Answered
4	South Africa	Some college/university study without earning ...	10,000 or more employees	Data or business analyst;Desktop or enterprise...	Slightly satisfied	Between 1 and 2 years ago	21426
...	...	...	...	...	...	...	...
98850	United States	Not Answered	Not Answered	Not Answered	Not Answered	Not Answered	Not Answered
98851	Spain	Not Answered	Not Answered	Back-end developer;Front-end developer	Not Answered	Not Answered	Not Answered
98852	India	Bachelor's degree (BA, BS, B.Eng., etc.)	Not Answered	Not Answered	Not Answered	Not Answered	Not Answered
98853	Russian Federation	Some college/university study without earning ...	Not Answered	Not Answered	Not Answered	Not Answered	Not Answered
98854	Cambodia	Not Answered	Not Answered	Not Answered	Not Answered	Not Answered	Not Answered

98855 rows × 7 columns

- In the screenshot above, we've used fillna() function within pandas to replace all NAN values with "Not Answered" as its a survey based data and also to make our data look more meaningful

Renamed the columns to make it readable.

In [6]:
```
#Step 4: Rename Columns
final_data=filtered_data_2.rename(columns={'FormalEducation':"Education Completed",'CompanySize':'Size_of_Company',
                                           'DevType':'Technology_used','JobSatisfaction':'Job Satisfaction Rating',
                                           'LastNewJob':'Last Job Start Time','ConvertedSalary':'Salary in USD'})
final_data.head(20)
```

Out[6]:

	Country	Education Completed	Size_of_Company	Technology_used	Job Satisfaction Rating	Last Job Start Time	Salary in USD
0	Kenya	Bachelor's degree (BA, BS, B.Eng., etc.)	20 to 99 employees	Full-stack developer	Extremely satisfied	Less than a year ago	Not Answered
1	United Kingdom	Bachelor's degree (BA, BS, B.Eng., etc.)	10,000 or more employees	Database administrator;DevOps specialist;Full-...	Moderately dissatisfied	More than 4 years ago	70841
2	United States	Associate degree	20 to 99 employees	Engineering manager;Full-stack developer	Moderately satisfied	Less than a year ago	Not Answered
3	United States	Bachelor's degree (BA, BS, B.Eng., etc.)	100 to 499 employees	Full-stack developer	Neither satisfied nor dissatisfied	Less than a year ago	Not Answered
4	South Africa	Some college/university study without earning ...	10,000 or more employees	Data or business analyst;Desktop or enterprise...	Slightly satisfied	Between 1 and 2 years ago	21426
5	United Kingdom	Bachelor's degree (BA, BS, B.Eng., etc.)	10 to 19 employees	Back-end developer;Database administrator;Fron...	Moderately satisfied	Between 2 and 4 years ago	41671
6	United States	Some college/university study without earning ...	10,000 or more employees	Back-end developer;Front-end developer;Full-st...	Slightly satisfied	Less than a year ago	120000
7	Nigeria	Bachelor's degree (BA, BS, B.Eng., etc.)	10 to 19 employees	Designer;Front-end developer;QA or test developer	Slightly satisfied	Less than a year ago	Not Answered
8	United States	Some college/university study without earning ...	100 to 499 employees	Back-end developer;C-suite executive (CEO, CTO...	Moderately satisfied	Between 2 and 4 years ago	250000
9	India	Bachelor's degree (BA, BS, B.Eng., etc.)	500 to 999 employees	Designer	Not Answered	Not Answered	Not Answered
		Master's degree (MA, MS, M.Eng...	1,000 to 4,999 ...	Back-end developer;Database ...			Not ...

- The default column names have been changed to make the data more readable for visualization

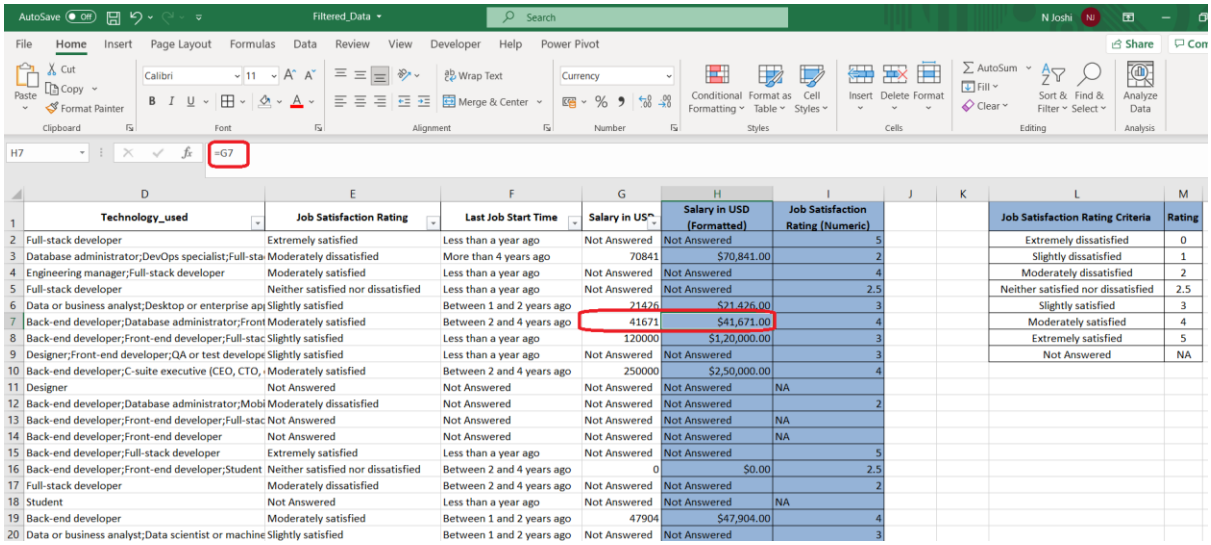
Saved the file to Personal Drive.

In [7]:
```
#Step 4: Exporting the data in CSV format
final_data.to_excel('D:\DA Course Material\DAB 103 Python etc\Project 1 Final Files\Filtered_Data.xlsx',
                    sheet_name='Filtered Data',index=False)
```

- Here, after making all the required changes in the file, we've exported and saved it to personal drive as a excel sheet. We've used to_excel function of pandas

- Screenshots for the Steps with MS-Excel

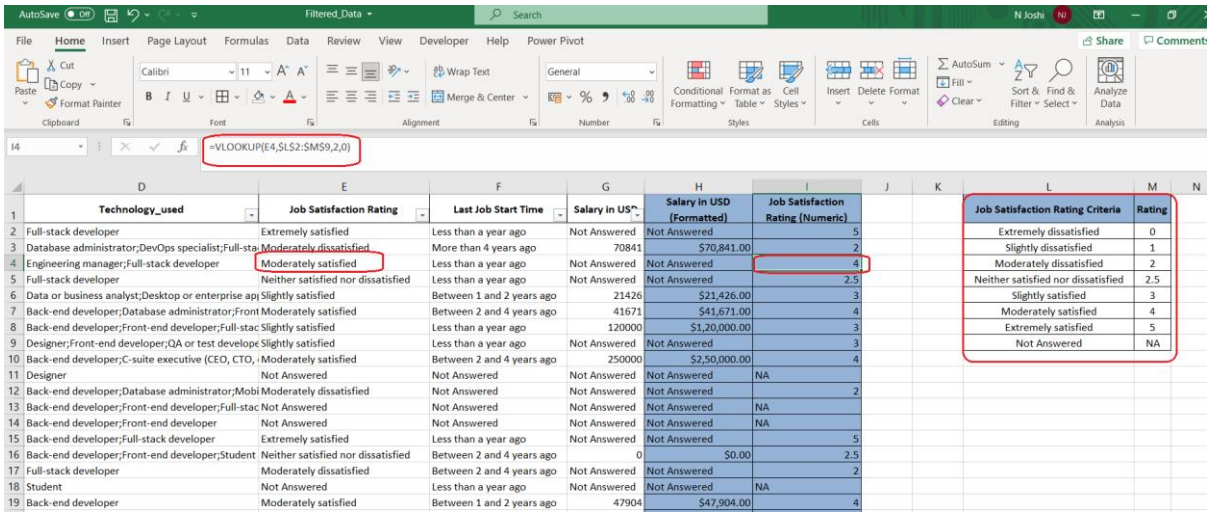
Format change for Salary



	D	E	F	G	H	I	J	K	L	M
1	Technology_used	Job Satisfaction Rating	Last Job Start Time	Salary in USD	Salary in USD (Formatted)	Job Satisfaction Rating (Numeric)			Job Satisfaction Rating Criteria	Rating
2	Full-stack developer	Extremely satisfied	Less than a year ago	Not Answered	Not Answered	5			Extremely dissatisfied	0
3	Database administrator;DevOps specialist;Full-sta	Moderately dissatisfied	More than 4 years ago	70841	$70,841.00	2			Slightly dissatisfied	1
4	Engineering manager;Full-stack developer	Moderately satisfied	Less than a year ago	Not Answered	Not Answered	4			Moderately dissatisfied	2
5	Full-stack developer	Neither satisfied nor dissatisfied	Less than a year ago	Not Answered	Not Answered	2.5			Neither satisfied nor dissatisfied	2.5
6	Data or business analyst;Desktop or enterprise ap	Slightly satisfied	Between 1 and 2 years ago	21426	$21,426.00	3			Slightly satisfied	3
7	Back-end developer;Database administrator;Front	Moderately satisfied	Between 2 and 4 years ago	41671	$41,671.00	4			Moderately satisfied	4
8	Back-end developer;Front-end developer;Full-stac	Slightly satisfied	Less than a year ago	120000	$1,20,000.00	3			Extremely satisfied	5
9	Designer;Front-end developer;QA or test develope	Slightly satisfied	Less than a year ago	Not Answered	Not Answered	3			Not Answered	NA
10	Back-end developer;C-suite executive (CEO, CTO,	Moderately satisfied	Between 2 and 4 years ago	250000	$2,50,000.00	4				
11	Designer	Not Answered	Not Answered	Not Answered	Not Answered	NA				
12	Back-end developer;Database administrator;Mobi	Moderately dissatisfied	Not Answered	Not Answered	Not Answered	2				
13	Back-end developer;Front-end developer;Full-stac	Not Answered	Not Answered	Not Answered	Not Answered	NA				
14	Back-end developer;Front-end developer	Not Answered	Not Answered	Not Answered	Not Answered	NA				
15	Back-end developer;Full-stack developer	Extremely satisfied	Less than a year ago	Not Answered	Not Answered	5				
16	Back-end developer;Front-end developer;Student	Neither satisfied nor dissatisfied	Between 2 and 4 years ago	0	$0.00	2.5				
17	Full-stack developer	Moderately dissatisfied	Between 2 and 4 years ago	Not Answered	Not Answered	2				
18	Student	Not Answered	Less than a year ago	Not Answered	Not Answered	NA				
19	Back-end developer	Moderately satisfied	Between 1 and 2 years ago	47904	$47,904.00	4				
20	Data or business analyst;Data scientist or machine	Slightly satisfied	Between 1 and 2 years ago	Not Answered	Not Answered	3				

- Here, we created a new column H where we want to format the salary column to "Currency format". We referenced the cell from the original column G and then converted it by going to *"Format cells>Number>Currency>USD> OK"*

Interpreting Column with Character values



	D	E	F	G	H	I	J	K	L	M	N
1	Technology_used	Job Satisfaction Rating	Last Job Start Time	Salary in USD	Salary in USD (Formatted)	Job Satisfaction Rating (Numeric)			Job Satisfaction Rating Criteria	Rating	
2	Full-stack developer	Extremely satisfied	Less than a year ago	Not Answered	Not Answered	5			Extremely dissatisfied	0	
3	Database administrator;DevOps specialist;Full-sta	Moderately dissatisfied	More than 4 years ago	70841	$70,841.00	2			Slightly dissatisfied	1	
4	Engineering manager;Full-stack developer	Moderately satisfied	Less than a year ago	Not Answered	Not Answered	4			Moderately dissatisfied	2	
5	Full-stack developer	Neither satisfied nor dissatisfied	Less than a year ago	Not Answered	Not Answered	2.5			Neither satisfied nor dissatisfied	2.5	
6	Data or business analyst;Desktop or enterprise ap	Slightly satisfied	Between 1 and 2 years ago	21426	$21,426.00	3			Slightly satisfied	3	
7	Back-end developer;Database administrator;Front	Moderately satisfied	Between 2 and 4 years ago	41671	$41,671.00	4			Moderately satisfied	4	
8	Back-end developer;Front-end developer;Full-stac	Slightly satisfied	Less than a year ago	120000	$1,20,000.00	3			Extremely satisfied	5	
9	Designer;Front-end developer;QA or test develope	Slightly satisfied	Less than a year ago	Not Answered	Not Answered	3			Not Answered	NA	
10	Back-end developer;C-suite executive (CEO, CTO,	Moderately satisfied	Between 2 and 4 years ago	250000	$2,50,000.00	4					
11	Designer	Not Answered	Not Answered	Not Answered	Not Answered	NA					
12	Back-end developer;Database administrator;Mobi	Moderately dissatisfied	Not Answered	Not Answered	Not Answered	2					
13	Back-end developer;Front-end developer;Full-stac	Not Answered	Not Answered	Not Answered	Not Answered	NA					
14	Back-end developer;Front-end developer	Not Answered	Not Answered	Not Answered	Not Answered	NA					
15	Back-end developer;Full-stack developer	Extremely satisfied	Less than a year ago	Not Answered	Not Answered	5					
16	Back-end developer;Front-end developer;Student	Neither satisfied nor dissatisfied	Between 2 and 4 years ago	0	$0.00	2.5					
17	Full-stack developer	Moderately dissatisfied	Between 2 and 4 years ago	Not Answered	Not Answered	2					
18	Student	Not Answered	Less than a year ago	Not Answered	Not Answered	NA					
19	Back-end developer	Moderately satisfied	Between 1 and 2 years ago	47904	$47,904.00	4					

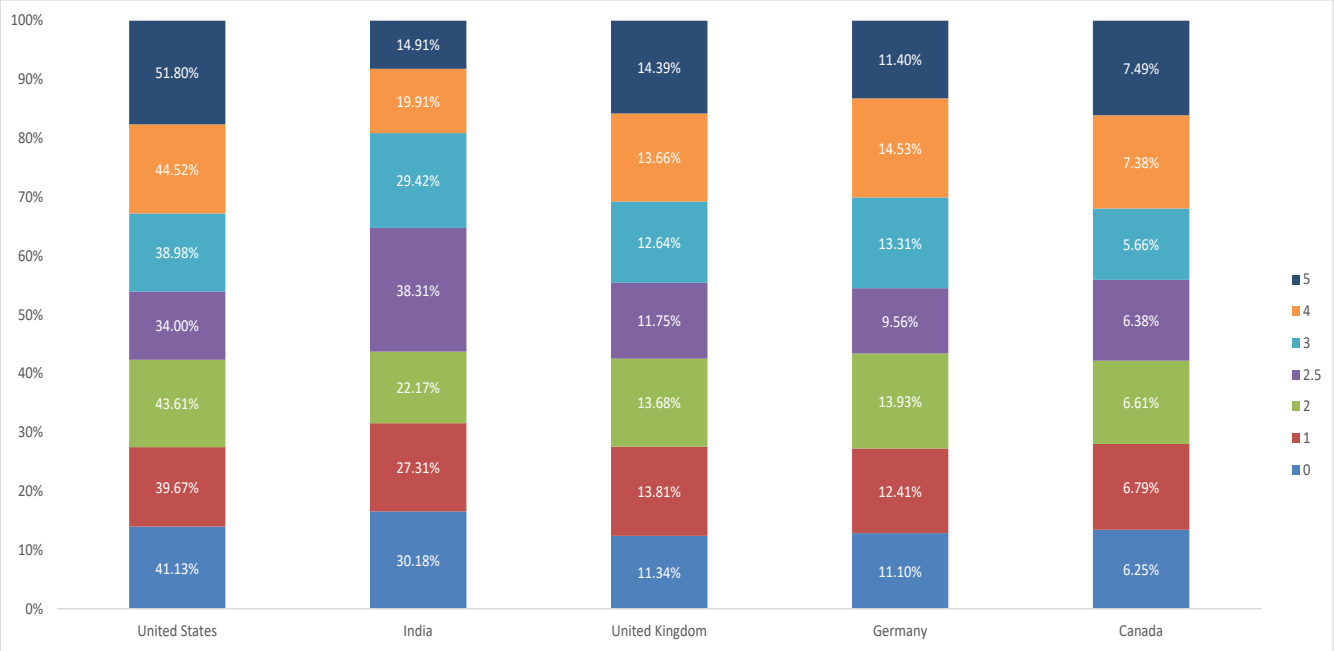
- Here, we wanted to convert the character type survey satisfaction rating (Column E) to numeric to make our analysis a little easier. Hence, we first assigned a numeric value to these ratings (Column L and M) and then we created a new column I to reflect these values using Vlookup function in Excel (kindly refer to the screenshot for the formula).

Data Analysis

Image Credit: https://www.techrepublic.com/article/your-big-data-reports-are-not-being-used-as-they-should-heres-how-to-change-that/

Some of the Key Patterns:

Proportion of Employee Satisfaction



The above plot shows us the proportion about employee satisfaction in the top 5 Countries where 0 being the worst and 5 being the best. (Kindly refer to the details given below).

Job Satisfaction Rating Criteria	Rating
Extremely dissatisfied	0
Slightly dissatisfied	1
Moderately dissatisfied	2
Neither satisfied nor dissatisfied	2.5
Slightly satisfied	3
Moderately satisfied	4
Extremely satisfied	5
Not Answered	NA

Following are some of the observations from the plot:

- US had both the worst and highest satisfaction rate.
- For India, comparison of bottom 3 and top 3 parameters (refer to the table above),i.e., 0/1/2 vs 3/4/5 shows us that majority of the employees are not satisfied with the job.

Top 5 Countries with highest average Salary in USD

Top 5 Countries	Average of Salary in USD
Andorra	$5,25,090
Liechtenstein	$2,84,028
Venezuela, Bolivarian Republic of...	$2,41,824
Ireland	$1,86,313
Botswana	$1,70,000

From the above plot, we clearly notice that although there are some countries offering highest salary packages across the world (as per available data), however, that isn't the sole criteria for employee satisfaction!

(Please note: The name of the 3rd country isn't complete in the Survey Data itself)

Final Analysis of the Survey

Image Credit: https://www.tlpnyc.com/blog/project-step-1-create-a-problem-statement

The 4 main highlights that lead us to choose the dataset are as listed below:

Relation between size of organization and Job satisfaction
High paying positions and technologies
Information on time frame about employees leaving their jobs
Relation between education and high salary

Please see the Visualizations below for the answers:

Visualization 1)

This analysis has been done using "Tableau"

a) In the above analysis, we have divided the survey responses into "Satisfied" or "Promoters" category (on the right as "Satisfaction Scores") and "Dissatisfied" or "Detractors" category (on the left side as Dissatisfaction Scores)
b) NPS or Net promoter Score has been calculated as below.
(Total percentage of all positive responses) - (Total percentage of all negative responses)*100
c) The data clearly reflects that employee satisfaction is not related to the size of an organization, i.e., a renowned organization doesn't guarantee satisfaction in the work environment and *hence, an individual shouldn't prioritise the size of an organization while looking for job satisfaction.*

Visualization 2)

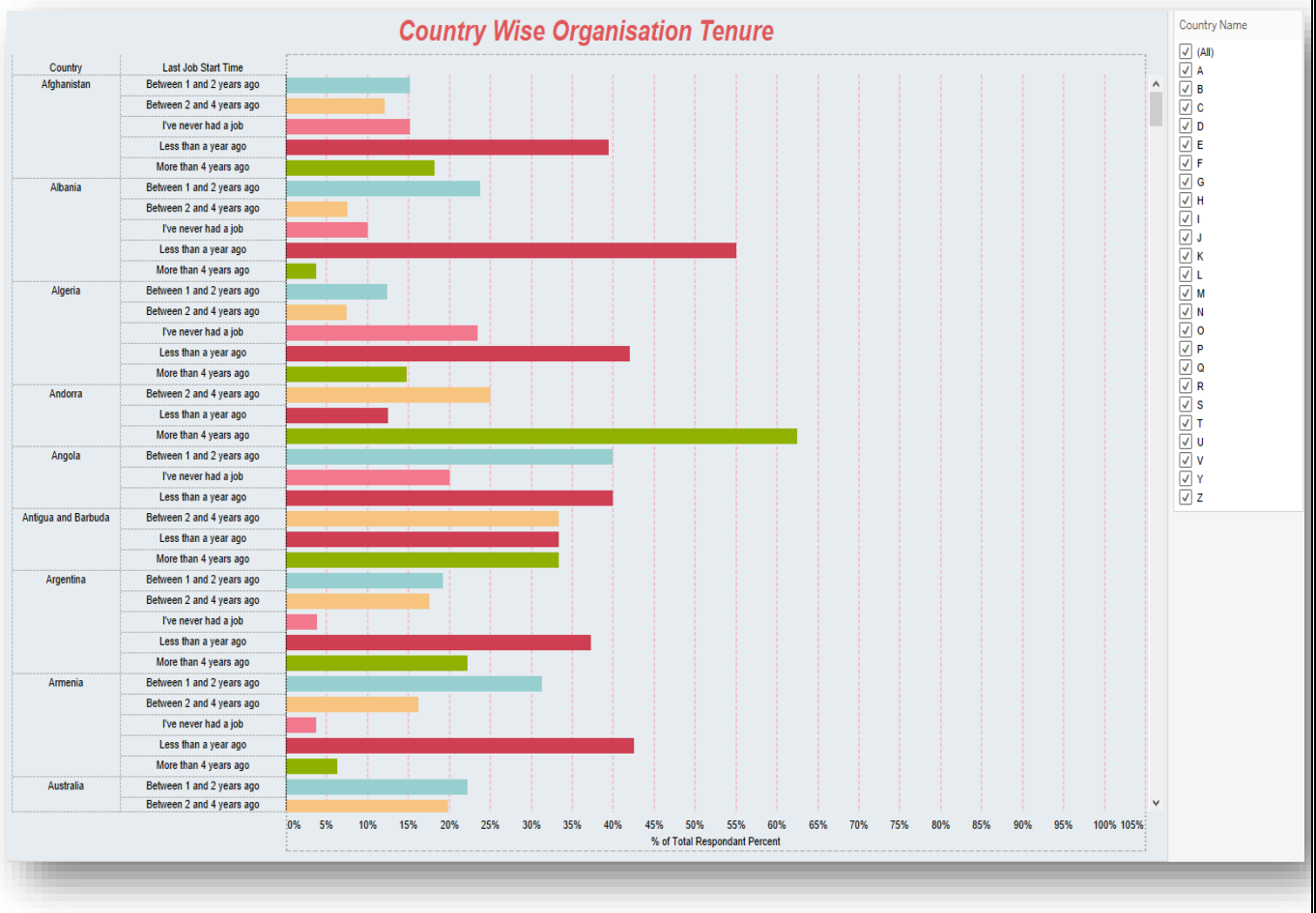
Technologies/Designations	Percentage of Individuals	Avg. Salary in USD
Engineering manager	6%	$1,29,926.44
C-suite executive (CEO, CTO, etc.)	4%	$1,17,667.22

DevOps specialist	10%	$1,15,029.31
Product manager	5%	$1,12,144.07
Marketing or sales professional	1%	$1,06,729.30
Data or business analyst	8%	$1,06,542.40
Data scientist or machine learning specialist	8%	$1,01,430.37
System administrator	11%	$1,01,142.15
Full-stack developer	48%	$1,00,156.39
Database administrator	14%	$97,767.51
Back-end developer	58%	$96,609.91
Desktop or enterprise applications developer	17%	$96,060.25
QA or test developer	7%	$95,877.93
Front-end developer	38%	$95,027.72
Embedded applications or devices developer	5%	$94,803.10
Designer	13%	$89,856.50
Mobile developer	20%	$84,073.32
Educator or academic researcher	4%	$83,182.16
Game or graphics developer	5%	$82,366.68
Student	17%	$43,759.39

This Visualization has been done using MS Excel.

a) Here, we clearly see that Engineering Manager/C-Suite Executive and DevOps Specialists are amongst the top 3 positions/designations with highest paying salary package.
b) The “Percentage of Individuals” column shows us what percentage of people have been on the given position or are using a specific technology as listed in “Technology/Designations” section.

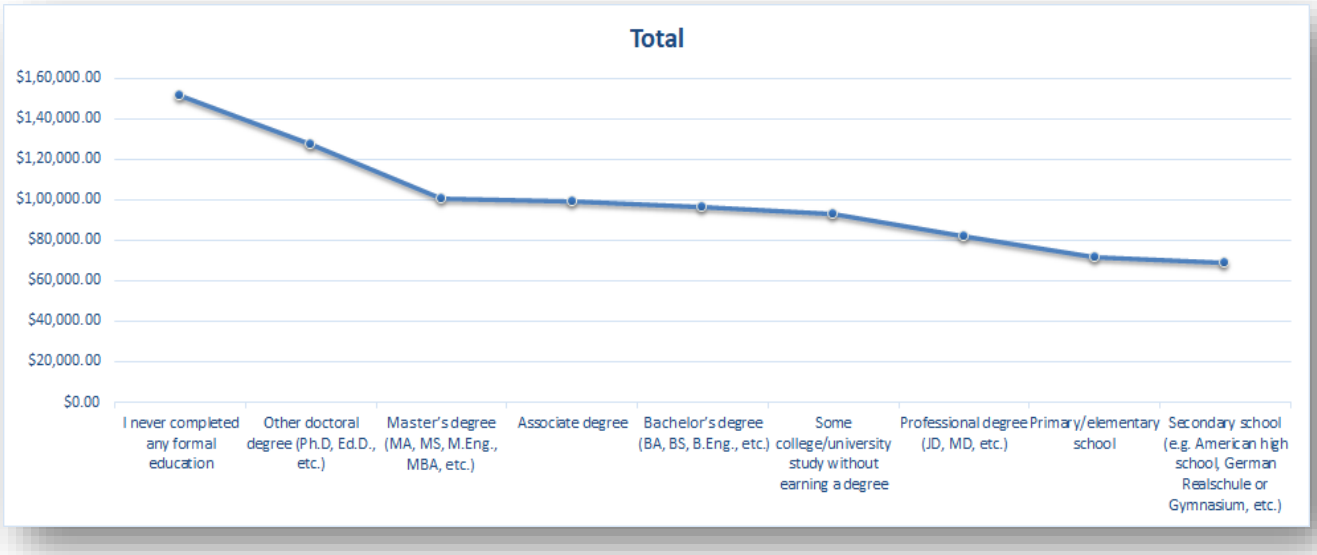
Visualization 3)



This analysis has been done using "Tableau"

a) This visualization is important for those who want to see what percentage of survey respondents have left their previous employment in recent time.
b) It can be helpful for those who want to perform research on the tenure of an individual depending on the country.
c) Since the data has been collected from around the world, hence, the "Check boxes" on the right, will help the user to select country by its first alphabet and hence analysing the data easily.

Visualization 4)



This analysis has been done using "MS Excel"

a) This analysis shows us that while higher education degree does directly increase the chances of getting a higher salary, however, the majority of respondents, who are being paid the most, haven't completed any formal education.
b) Followed by people with no formal education, others with PHD/Doctoral degree and then Master's Degree are being paid the most.

Conclusion:

To conclude the analysis/visualization completed above, we've the following observations:

- ∅ Job Satisfaction is not related to size of an organization and depending on the requirement of an individual, he/she should focus accordingly.
- ∅ From Visualization2, we see that at some point in their career, more than 50% of IT employees have been back-end developers, hence students can focus on similar technology to open more job opportunities.
- ∅ Least paid average salary is for game developers, hence if someone wants to pursue a new course as a student, he/she can take an informed decision.
- ∅ Although higher degrees tend to get someone high paying jobs, however, we do see an increasing trend of people getting paid more without a formal education, hence, an organization may choose to focus on hiring raw talent whilst the aspiring students who cannot complete formal education probably due to some personal reason, shouldn't lose hope.