

## POWER BI PROJECT

### DATA CLEANING (FIRST PHASE)

The data source is completely row. We are not going to make changes in data source ie the Excel file.

The transformation will be done in Power BI Transform not in in Excel.

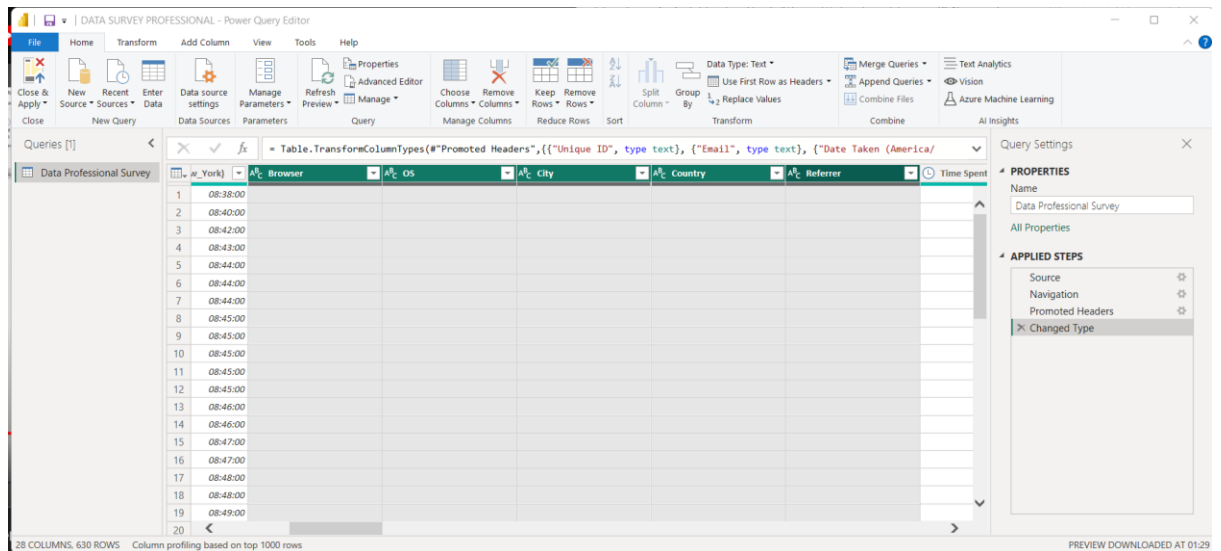
The column Q1 has Other attribute value has lot of data...that needs to be cleaning

The column Q2 has yearly salary in ranges. So that's needs to be cleaned.

Lets get started.

#### First Transform data.

- 1) We delete the columns as shown below.



- 2) The Q1 columns. We need to remove those extra values attached to value=Other. So we create split column based on delimiter '('.

File Home Transform Add Column View Tools Help

Close & Apply New Source Recent Sources Enter Data Data source settings Manage Parameters Refresh Preview Advanced Editor Choose Remove Columns Manage Columns Keep Remove Rows Sort Split Column Group By Use First Row as Headers Replace Values Data Type: Text Merge Queries Append Queries Combine Files Text Analytics Vision Azure Machine Learning

Queries [1]

Table.RemoveColumns("#'Changed Type'",{"'Browsed...")

Data Professional Survey

	W_York)	Time Spent	A <sup>B</sup> Q1 - Which Title Best Fits your
1	08:38:00	00:00:44	Data Analyst
2	08:40:00	00:01:30	Data Analyst
3	08:42:00	00:02:18	Data Engineer
4	08:43:00	00:02:10	Other (Please Specify):Analytics Co
5	08:44:00	00:01:51	Data Analyst
6	08:44:00	00:02:34	Data Analyst
7	08:44:00	00:01:15	Data Scientist
8	08:45:00	00:01:25	Data Engineer
9	08:45:00	00:02:10	Data Analyst
10	08:45:00	00:01:27	Data Analyst
11	08:45:00	00:01:29	Data Analyst
12	08:45:00	00:02:31	Data Analyst
13	08:46:00	00:03:20	Data Analyst
14	08:46:00	00:00:55	Data Scientist
15	08:47:00	00:01:24	Data Analyst
16	08:47:00	00:00:47	Data Analyst
17	08:48:00	00:01:06	Data Analyst
18	08:48:00	00:01:04	Student/Looking/None
19	08:49:00	00:01:05	Student/Looking/None
20			

Split Column by Delimiter

Specify the delimiter used to split the text column.

Select or enter delimiter

--Custom--

(

Split at

☒ Left-most delimiter

☐ Right-most delimiter

☐ Each occurrence of the delimiter

Advanced options

Quote Character

"

☐ Split using special characters

Insert special character

OK Cancel

File Home Transform Add Column View Tools Help

Close & Apply New Source Recent Sources Enter Data Data source settings Manage Parameters Refresh Preview Advanced Editor Choose Remove Columns Manage Columns Keep Remove Rows Sort Split Column Group By Use First Row as Headers Replace Values Data Type: Text Merge Queries Append Queries Combine Files Text Analytics Vision Azure

Queries [1]

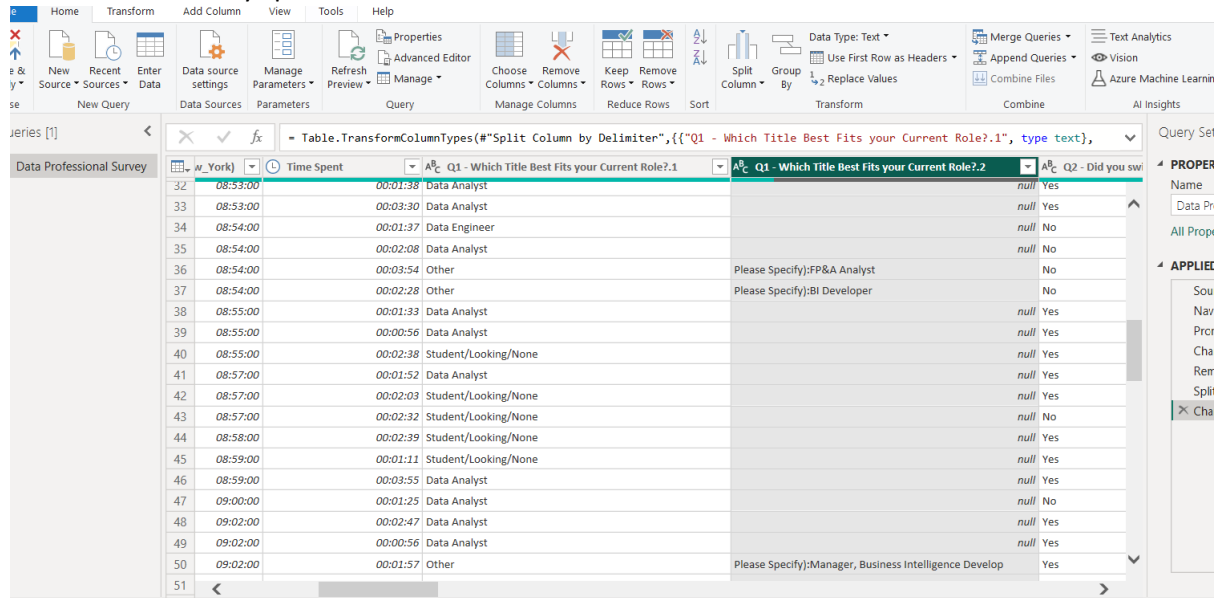
Table.TransformColumnTypes("#'Split Column by Delimiter'",{{"Q1 - Which Title Best Fits your Current Role?1", type text},

Data Professional Survey

	W_York)	Time Spent	A <sup>B</sup> Q1 - Which Title Best Fits your Current Role?1	A <sup>B</sup> Q1 - Which Title Best Fits your Current Role?2	A <sup>B</sup> Q2 - Did you sv
32	08:53:00	00:01:38	Data Analyst	null	Yes
33	08:53:00	00:03:30	Data Analyst	null	Yes
34	08:54:00	00:01:37	Data Engineer	null	No
35	08:54:00	00:02:08	Data Analyst	null	No
36	08:54:00	00:03:54	Other	Please Specify):FP&A Analyst	No
37	08:54:00	00:02:28	Other	Please Specify):BI Developer	No
38	08:55:00	00:01:33	Data Analyst	null	Yes
39	08:55:00	00:00:56	Data Analyst	null	Yes
40	08:55:00	00:02:38	Student/Looking/None	null	Yes
41	08:57:00	00:01:52	Data Analyst	null	Yes
42	08:57:00	00:02:03	Student/Looking/None	null	Yes
43	08:57:00	00:02:32	Student/Looking/None	null	No
44	08:58:00	00:02:39	Student/Looking/None	null	Yes
45	08:59:00	00:01:11	Student/Looking/None	null	Yes
46	08:59:00	00:03:55	Data Analyst	null	Yes
47	09:00:00	00:01:25	Data Analyst	null	No
48	09:02:00	00:02:47	Data Analyst	null	Yes
49	09:02:00	00:00:56	Data Analyst	null	Yes
50	09:02:00	00:01:57	Other	Please Specify):Manager, Business Intelligence Develop	Yes
51					

14 COLUMNS. 630 ROWS Column profiling based on top 1000 rows

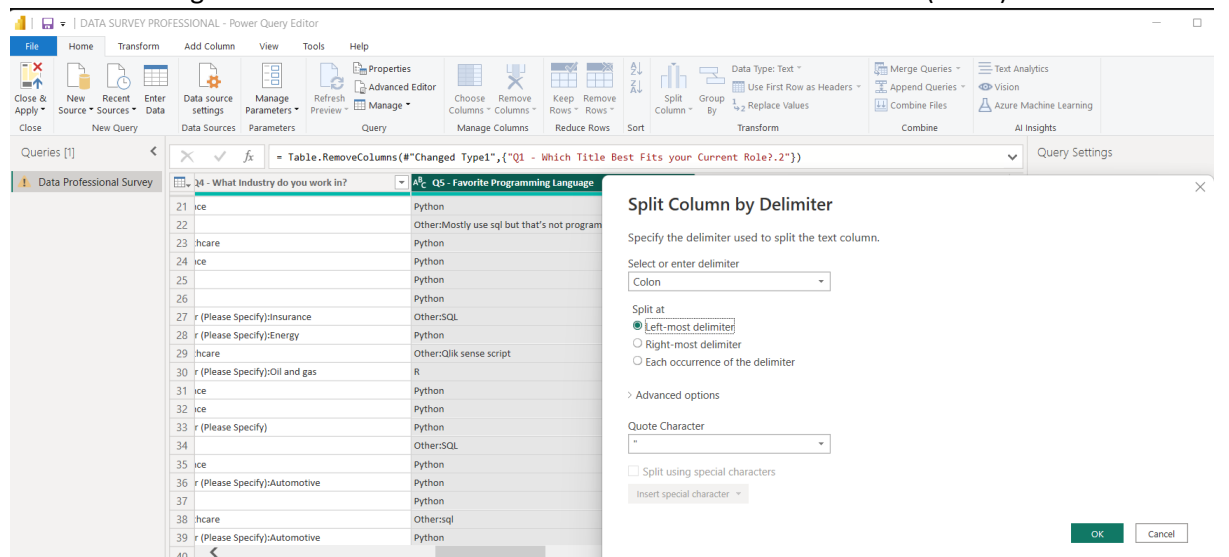
### 3) We delete the newly splitted column.



The screenshot shows the Power Query Editor interface. The table has columns: Data Professional Survey, Time Spent, Q1 - Which Title Best Fits your Current Role?, and Q2 - Did you switch roles?. The Q1 column has been split by delimiter, and the new column is being deleted.

Data Professional Survey	Time Spent	Q1 - Which Title Best Fits your Current Role?	Q2 - Did you switch roles?
32	08:53:00	Data Analyst	null Yes
33	08:53:00	Data Analyst	null Yes
34	08:54:00	Data Engineer	null No
35	08:54:00	Data Analyst	null No
36	08:54:00	Other	Please Specify:FP&A Analyst No
37	08:54:00	Other	Please Specify:BI Developer No
38	08:55:00	Data Analyst	null Yes
39	08:55:00	Data Analyst	null Yes
40	08:55:00	Student/Looking/None	null Yes
41	08:57:00	Data Analyst	null Yes
42	08:57:00	Student/Looking/None	null Yes
43	08:57:00	Student/Looking/None	null No
44	08:58:00	Student/Looking/None	null Yes
45	08:59:00	Student/Looking/None	null Yes
46	08:59:00	Data Analyst	null Yes
47	09:00:00	Data Analyst	null No
48	09:02:00	Data Analyst	null Yes
49	09:02:00	Data Analyst	null Yes
50	09:02:00	Other	Please Specify:Manager, Business Intelligence Develop Yes
51			

### 4) The above two steps is applied for column Q5 Favourite Programming Language. We split the others from original column. And delete the new column. Note the delimiter (colon).



The screenshot shows the Power Query Editor interface. The table has columns: Data Professional Survey, Q4 - What Industry do you work in?, and Q5 - Favorite Programming Language. The Q5 column has been split by delimiter, and the new column is being deleted.

Data Professional Survey	Q4 - What Industry do you work in?	Q5 - Favorite Programming Language
21	ice	Python
22		Other:Mostly use sql but that's not program
23	hicare	Python
24	ice	Python
25		Python
26		Python
27	r (Please Specify):Insurance	Other:SQL
28	r (Please Specify):Energy	Python
29	hicare	Other:Qlik sense script
30	r (Please Specify):Oil and gas	R
31	ice	Python
32	ice	Python
33	r (Please Specify)	Python
34		Other:SQL
35	ice	Python
36	r (Please Specify):Automotive	Python
37		Python
38	hicare	Other:sql
39	r (Please Specify):Automotive	Python
40		

**Split Column by Delimiter**

Specify the delimiter used to split the text column.

Select or enter delimiter: Colon

Split at:
 

- ☒ Left-most delimiter
- ☐ Right-most delimiter
- ☐ Each occurrence of the delimiter

Advanced options

Quote Character: "

☐ Split using special characters

Insert special character

OK Cancel

- 5) Now we need clean the Q3 salary. We create duplicate copy of that column which appears at end. And we split that copied column by Digit to Non-Digit.

Queries [1] fx = Table.DuplicateColumn(#"Removed Columns2", "Q3 - Current Yearly Salary (in USD)", "Q3 - Current Yearly Salary (in USD) - Copy"

	Q12 - Highest Level of Education	Q13 - Ethnicity	Q3 - Current Yearly Salary (in USD) - Copy
1		White or Caucasian	106k-125k
2		Asian or Asian American	41k-65k
3		Black or African American	0-40k
4		White or Caucasian	150k-225k
5		Black or African American	41k-65k
6	democratic of Congo	Black or African American	0-40k
7		Black or African American	0-40k
8		Asian or Asian American	125k-150k
9		Asian or Asian American	86k-105k
10		Hispanic or Latino	41k-65k
11		White or Caucasian	66k-85k
12		White or Caucasian	0-40k
13	a	Other (Please Specify):Latino with Italian roots	0-40k
14		White or Caucasian	0-40k
15		White or Caucasian	41k-65k
16		White or Caucasian	41k-65k
17		Black or African American	0-40k
18		Black or African American	0-40k
19		Asian or Asian American	41k-65k
20			

4 COLUMNS: 630 ROWS Column profiles based on top 1000 rows

Queries [1] fx = Table.SplitColumn(#"Duplicated Column", "Q3 - Current Yearly Salary (in USD) - Copy",

	Q3 - Current Yearly Salary (in USD) - Copy.1	Q3 - Current Yearly Salary (in USD) - Copy.2	Q3 - Current Yearly Salary (in USD) - Copy.3
1	106	k-125	k
2	41	k-65	k
3	0	-40	k
4	150	k-225	k
5	41	k-65	k
6	0	-40	k
7	0	-40	k
8	125	k-150	k
9	86	k-105	k
10	41	k-65	k
11	66	k-85	k
12	0	-40	k
13	0	-40	k
14	0	-40	k
15	41	k-65	k
16	41	k-65	k
17	0	-40	k
18	0	-40	k
19	41	k-65	k
20			

We remove the column that has k's.

Then we look at Q3..Copy 2. Here we need to remove k , then remove '-' and some have only + which needs to be by corresponding value of that row.

Table.RemoveColumns("#Split Column by Character Transition",{"Q3 - Current Yearly Salary (in USD) - Copy.3"})

Location	Q13 - Ethnicity	Q3 - Current Yearly Salary (in USD) - Copy.1	Q3 - Current Yearly Salary (in USD) - Copy.2
1	White or Caucasian	106	k-125
2	Asian or Asian American	41	k-65
3	Black or African American	0	-40
4	White or Caucasian	150	k-225
5	Black or African American	41	k-65
6	Black or African American	0	-40
7	Black or African American	0	
8	Asian or Asian American	125	
9	Asian or Asian American	86	
10	Hispanic or Latino	41	
11	White or Caucasian	66	
12	White or Caucasian	0	
13	Other (Please Specify):Latino with Italian roots	0	
14	White or Caucasian	0	
15	White or Caucasian	41	
16	White or Caucasian	41	
17	Black or African American	0	
18	Black or African American	0	
19	Asian or Asian American	41	
20			

Replace Values

Replace one value with another in the selected columns.

Value To Find: k

Replace With:

Advanced options

OK Cancel

Table.ReplaceValue("#Removed Columns3", "k", "", Replacer.ReplaceText, {"Q3 - Current Yearly Salary (in USD) - Copy.2"})

Location	Q13 - Ethnicity	Q3 - Current Yearly Salary (in USD) - Copy.1	Q3 - Current Yearly Salary (in USD) - Copy.2
1	White or Caucasian	106	-125
2	Asian or Asian American	41	-65
3	Black or African American	0	-40
4	White or Caucasian	150	-225
5	Black or African American	41	-65
6	Black or African American	0	-40
7	Black or African American	0	
8	Asian or Asian American	125	
9	Asian or Asian American	86	
10	Hispanic or Latino	41	
11	White or Caucasian	66	
12	White or Caucasian	0	
13	Other (Please Specify):Latino with Italian roots	0	
14	White or Caucasian	0	
15	White or Caucasian	41	
16	White or Caucasian	41	
17	Black or African American	0	
18	Black or African American	0	
19	Asian or Asian American	41	
20			

Replace Values

Replace one value with another in the selected columns.

Value To Find: -

Replace With:

Advanced options

OK Cancel

6) As you see two rows have + sign so replace with the value from column Copy1.

Table.SelectRows("#Replaced Value1", each ([#Q3 - Current Yearly Salary (in USD) - Copy.2] = "+"))

Location	Q13 - Ethnicity	Q3 - Current Yearly Salary (in USD) - Copy.1	Q3 - Current Yearly Salary (in USD) - Copy.2
1	White or Caucasian	225	+
2	Asian or Asian American	225	+

7) Convert the type of Copy1 and Copy2 column into Whole Numbers. Then we take average of both in new column Average Salary. And delete the Copy1 and Copy2 columns.

DATA SURVEY PROFESSIONAL - Power Query Editor

File Home Transform Add Column View Tools Help

Column From Custom Examples - Column Function

General

From Text From Number From Date & Time

Trigonometry - Statistics Standard Scientific Rounding - Information - Date Time Duration Text Vision Azure Machine Learning

Queries [1]

Data Professional Survey

Q13 - Ethnicity

Q3 - Current Yearly Salary (in USD) - Copy.1

Q3 - Current Yearly Salary (in USD) - Copy.2

106 125

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

Available columns

Q9 - Male/Female? Q10 - Current Age Q11 - Which Country do you I... Q12 - Highest Level of Educati... Q13 - Ethnicity Q3 - Current Yearly Salary (in... Q3 - Current Yearly Salary (in...

<< Insert

Learn about Power Query formulas

No syntax errors have been detected.

OK Cancel

PROPERTIES

Name

Data Professional Survey

CHANGED STEPS

Changed Type removed Columns split Column by Delimiter changed Type1 removed Columns1 split Column by Delimiter1 changed Type2 removed Columns2 duplicated Column split Column by Character ... removed Columns3 replaced Value replaced Value1 replaced Value2 changed Type3

PREVIEW DOWNLOADED AT 01:

25 COLUMNS, 630 ROWS Column profiling based on top 1000 rows

Table.AddColumn("#Changed Type3", "Average Salary", each ([#Q3 - Current Yearly Salary (in USD) - Copy.1]+[#Q3 - Current Yearly Salary (in USD) - Copy.2])/2)

	Q3 - Current Yearly Salary (in USD) - Copy.1	Q3 - Current Yearly Salary (in USD) - Copy.2	Average Salary
n	106	125	115.5
erican	41	65	53
merican	0	40	20
n	150	225	187.5
merican	41	65	53
merican	0	40	20
merican	0	40	20
erican	125	150	137.5
erican	86	105	95.5
	41	65	53
n	66	85	75.5
n	0	40	20
cify):Latino with Italian roots	0	40	20
n	0	40	20
n	41	65	53
n	41	65	53
merican	0	40	20
merican	0	40	20
erican	41	65	53

Make sure that Average Salary is in Whole Number type.

8) Similary try to remove the Others from The column Q11 and Q4.

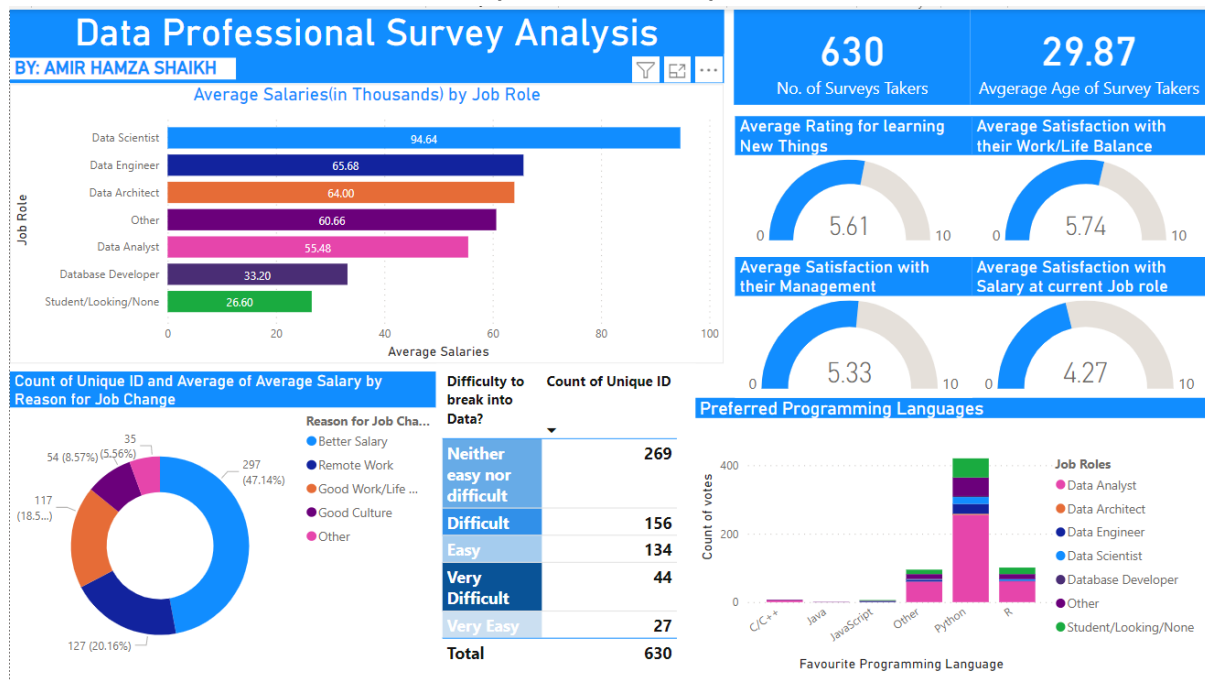
The first screenshot shows a data table with columns: 'Yearly Salary (in USD)', 'Q4 - What Industry do you work in?', and 'Pyth'. The 'Split Column by Delimiter' dialog is open for the 'Q4' column. The delimiter is set to '(' and 'Split at' is set to 'Each occurrence of the delimiter'.

The second screenshot shows a data table with columns: 'Male/Female?', 'Q10 - Current Age', 'Q11 - Which Country do you live in?', 'Q12 - Highest Level of Education', and 'Q13 - Ethnicity'. The 'Split Column by Delimiter' dialog is open for the 'Q11' column. The delimiter is set to '(' and 'Split at' is set to 'Left-most delimiter'.

This enough for Data Cleaning. The way taking insights from Salary range might not be so accurate in taking averages. But we this method/ way as simple.

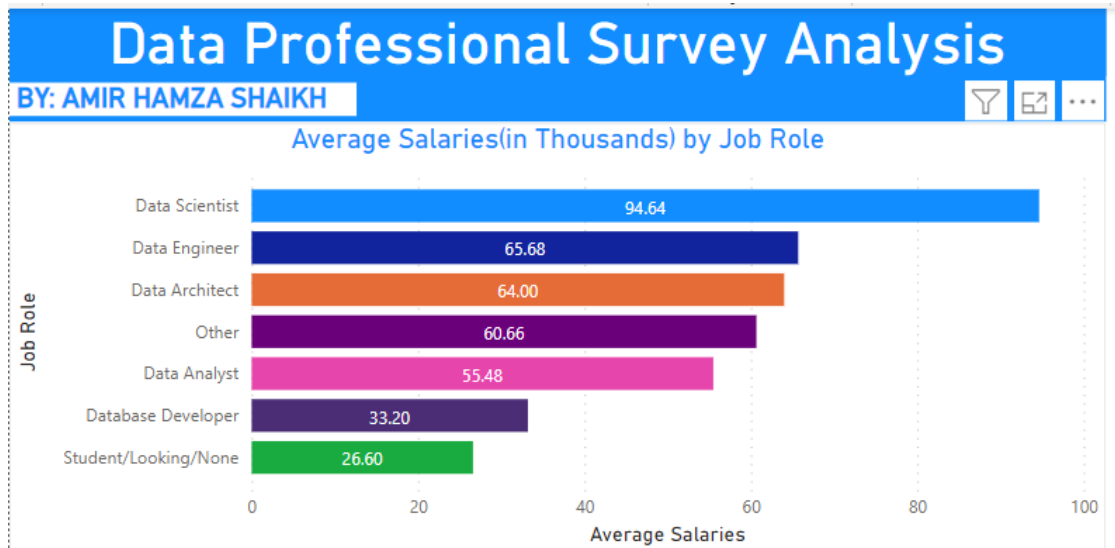
Then click on Close&Apply button.

## Second INSIGHTS data visualization .(SECOND PHASE)



## Observations:

1)

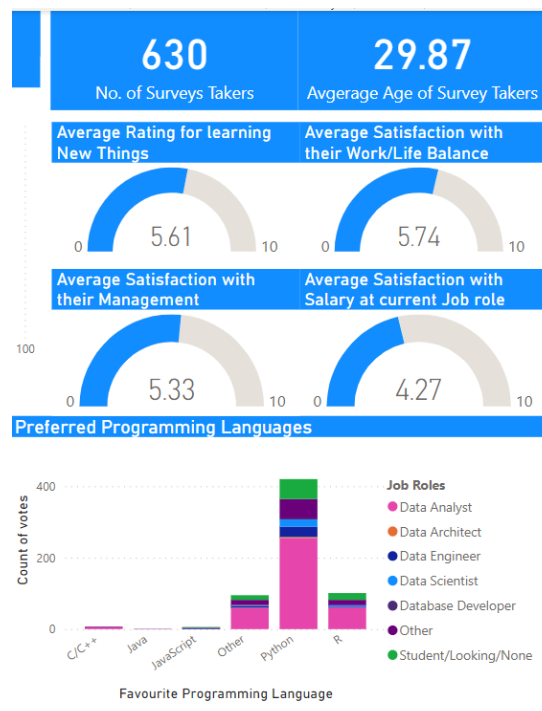


**Count of Unique ID and Average of Average Salary by** **Difficulty to** **Count of Unique ID**

Here we see that we have highest Average Salary for Data Scientist and Database developer as lowest while ignoring the Student/Looking/None. The Data Analyst earn 55.48k on an average.

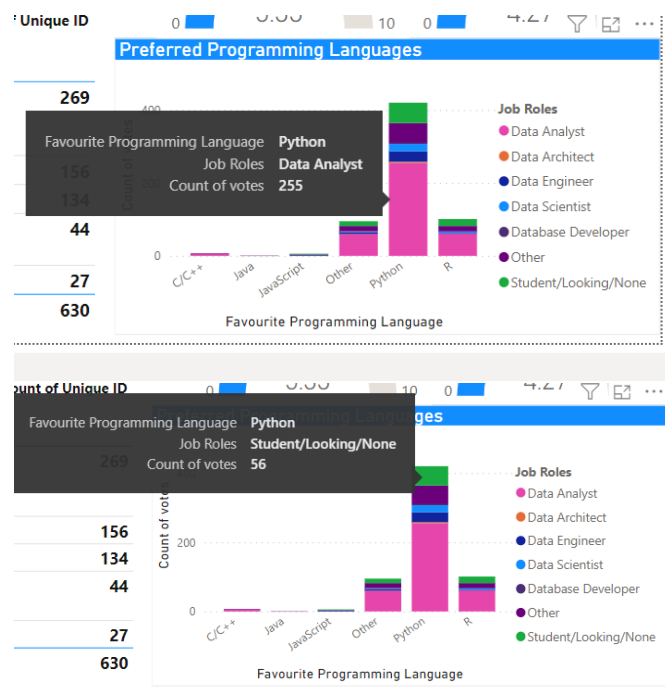


2)

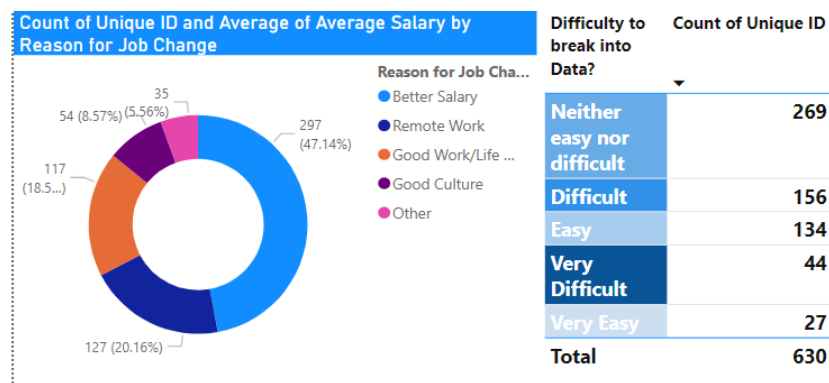


We have four Gauge Charts which says the following for example The people learning new things have average rating of 5.6/10 irrespective of their the Job role. One important thing to know that most of people are not happy with their Job work/satisfaction and with their current role and salary.

The Most preferred language preferred for various Roles is Python. If I just hover it. On python pink section. Python is mainly used by Data Analyst as compared to R. For the students/freshers they highly prefer Python as their Main Language.



3)



We have two observations for this.

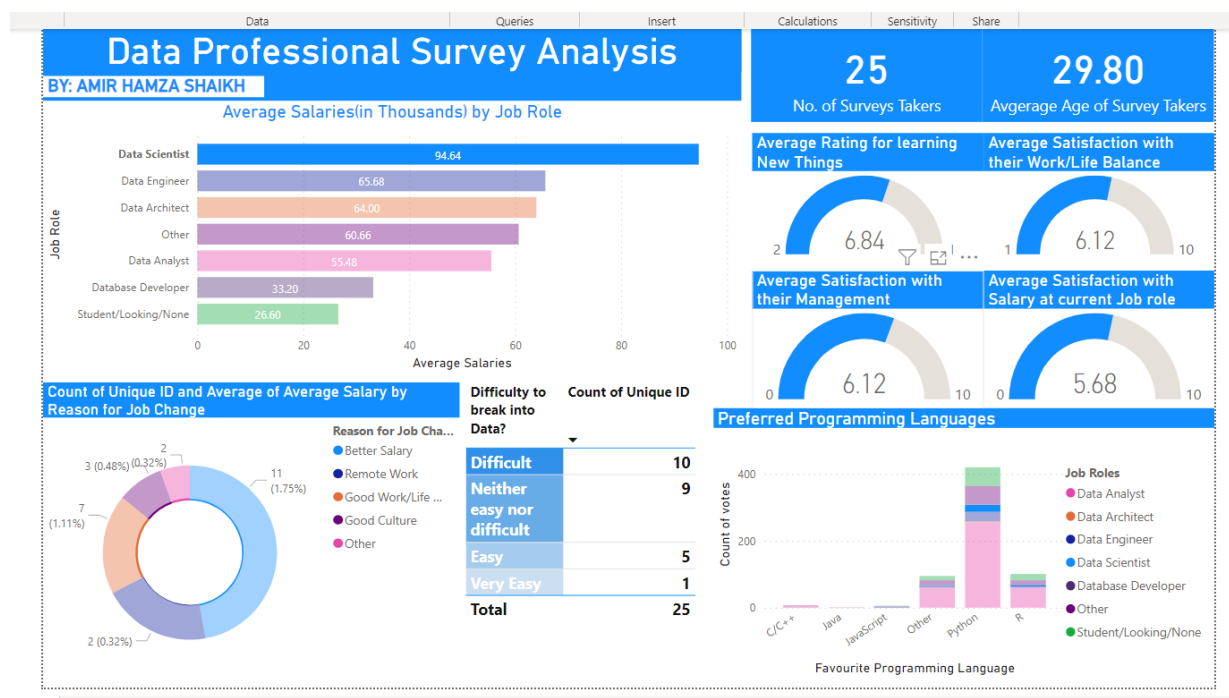
**On right side** is self explainable. Most of people find that it is average/medium level to break into Data related Jobs. Only 134+27 found that getting Job is Easy!

**On Left side** we found that people switch to the next Job for two Main reasons:

- To have better Salary as compared to current job position and salary
- Most people prefer to work at home instead of working in office. As they have become adaptable working at home during pandemic.

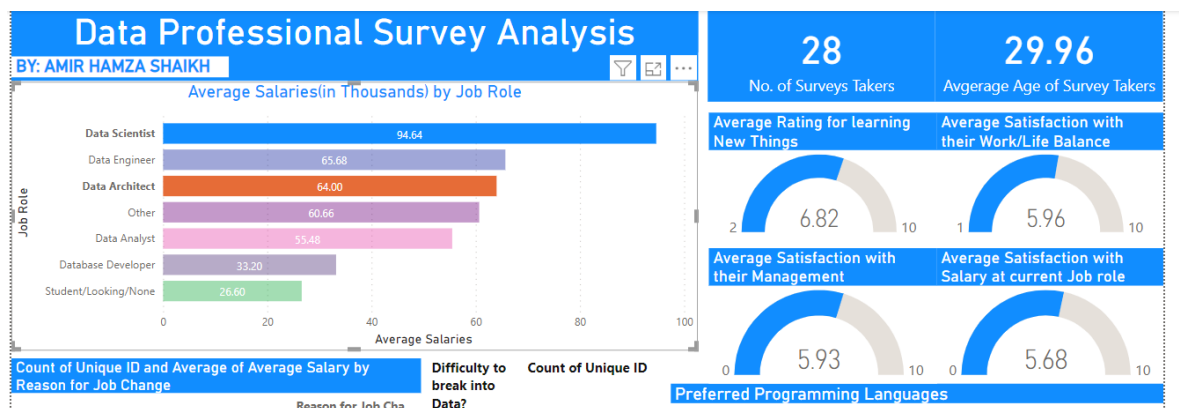
Deeper Observation:

1)



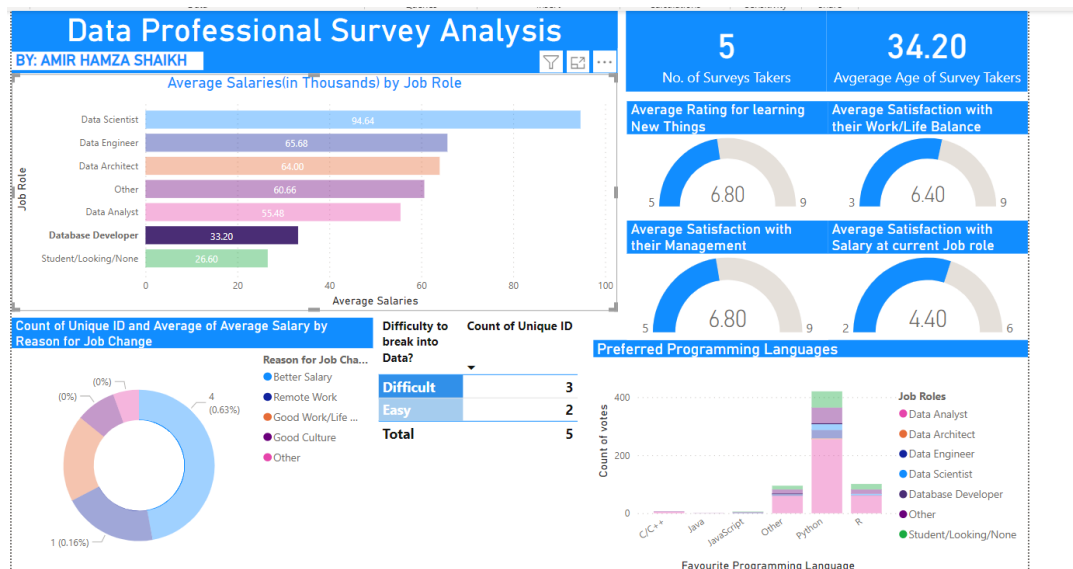
The Data scientists have the highest rating for learning New things.

2)



Both Data Scientists and Data architects have highest Job Satisfaction with current role and salary.

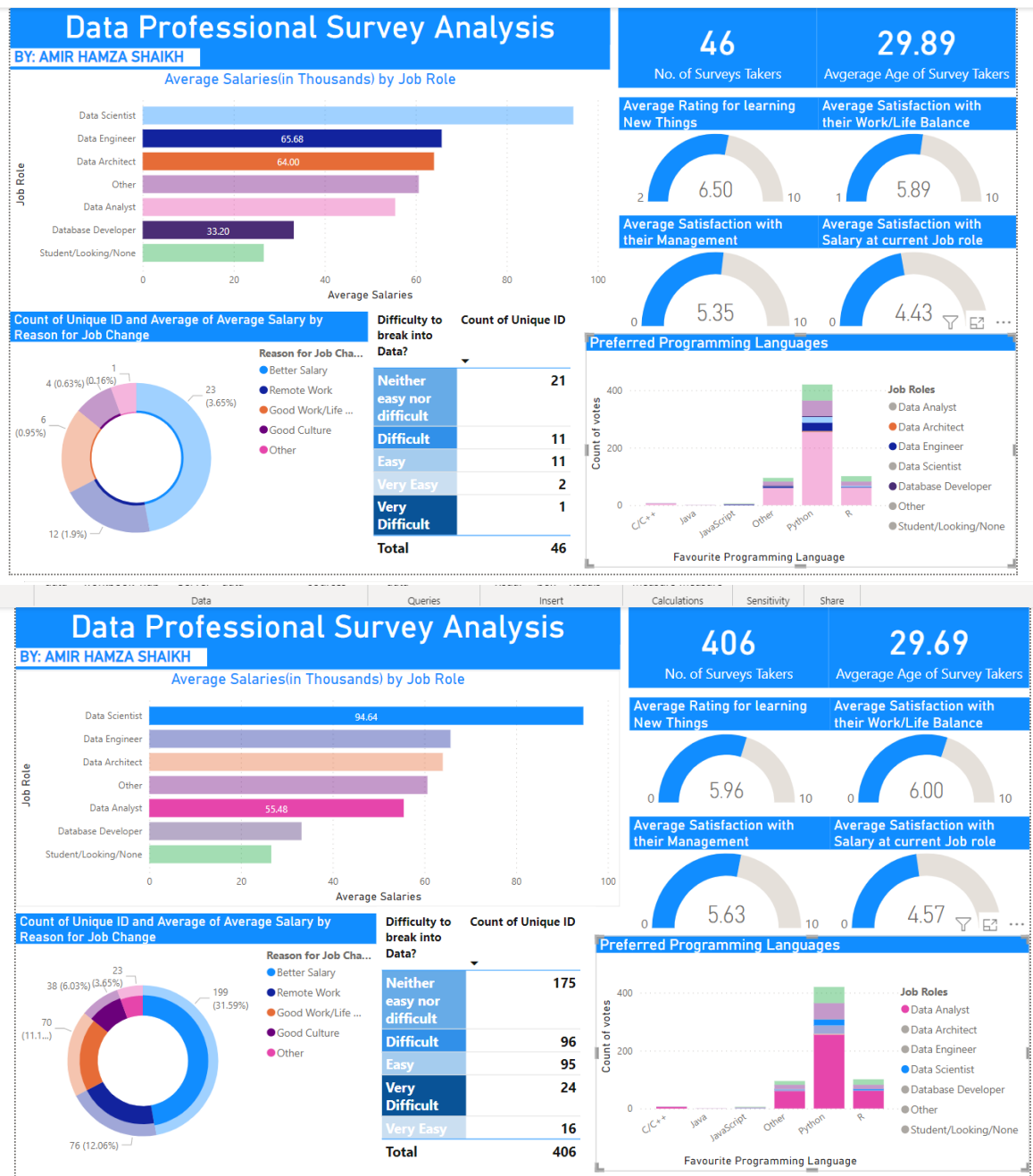
3)



The Database developer have no tension with their Management but they not happy with their current role and income.

4)

4)



The people working with their current role as Data analyst or scientist are open to get better company as they gain experiences. They are highly demanding jobs.