

A DATA ANALYSIS PROJECT

BASED ON A SURVEY OF DATA PROFESSIONALS ACROSS COUNTRIES

PREPARED BY: CHINEDU ELEKWA. P

BRIEF INFORMATION

A survey was given to several data professionals across different countries to create a data with which analyst can help explain the work life among them. The data analysis would answer certain puzzling questions that gives insights to Business Owners and Human Resource Managers. The survey data was collected and stored in an excel spreadsheet.

- DATA CLEANING TOOL: Power Query
- DATA VISUALIZATION TOOL: PowerBi
- NUMBER OF COLUMNS: 27
- NUMBER OF ROWS: 631

PROBLEM STATEMENT

1. Favorite programming language of data professionals
2. Average Salary of data professionals
3. Average of data professionals happy learning through work
4. Average of data professionals happy with salary

COLUMN HEADINGS OF THE DATA

- Unique ID, Email Date Taken (America/New_York), Time Taken (America/New_York), Browser, OS, City, Country, Referrer, Time Spent, Q1 - Which Title Best Fits your Current Role?, Q2 - Did you switch careers into Data?, Q3 - Current Yearly Salary (in USD), Q4 - What Industry do you work in?, Q5 - Favorite Programming Language, Q6 - How Happy are you in your Current Position with the following? (Salary), Q6 - How Happy are you in your Current Position with the following? (Work/Life Balance), Q6 - How Happy are you in your Current Position with the following? (Coworkers), Q6 - How Happy are you in your Current Position with the following? (Management), Q6 - How Happy are you in your Current Position with the following? (Upward Mobility), Q6 - How Happy are you in your Current Position with the following? (Learning New Things), Q7 - How difficult was it for you to break into Data?, Q8 - If you were to look for a new job today, what would be the most important thing to you?, Q9 - Male/Female?, Q10 - Current Age, Q11 - Which Country do you live in?, Q12 - Highest Level of Education, Q13 - Ethnicity

I imported the excel file into PowerBi and clicked the transform button so as to clean the data using Power Query

Email	Date Taken (America/New_York)	Time Taken (America/New_York)	Browser	OS	City	Country	Referrer	Time Spent	Q1 - Which Title
anonymous	6/10/2022	8:38						0:00:44	Data Analyst
anonymous	6/10/2022	8:40						0:01:30	Data Analyst
anonymous	6/10/2022	8:42						0:02:18	Data Engineer
anonymous	6/10/2022	8:43						0:02:10	Other (Please Specify)
anonymous	6/10/2022	8:44						0:01:51	Data Analyst
anonymous	6/10/2022	8:44						0:02:34	Data Analyst
anonymous	6/10/2022	8:44						0:01:15	Data Scientist
anonymous	6/10/2022	8:45						0:01:25	Data Engineer
anonymous	6/10/2022	8:45						0:02:10	Data Analyst
anonymous	6/10/2022	8:45						0:01:27	Data Analyst
anonymous	6/10/2022	8:45						0:01:29	Data Analyst
anonymous	6/10/2022	8:45						0:02:31	Data Analyst
anonymous	6/10/2022	8:46						0:03:20	Data Analyst
anonymous	6/10/2022	8:46						0:00:55	Data Scientist
anonymous	6/10/2022	8:47						0:01:24	Data Analyst
anonymous	6/10/2022	8:47						0:00:47	Data Analyst
anonymous	6/10/2022	8:48						0:01:06	Data Analyst
anonymous	6/10/2022	8:48						0:01:04	Student/Looking for a job
anonymous	6/10/2022	8:49						0:01:05	Student/Looking for a job
anonymous	6/10/2022	8:49						0:01:21	Data Analyst
anonymous	6/10/2022	8:49						0:01:23	Data Analyst
anonymous	6/10/2022	8:50						0:01:35	Data Analyst
anonymous	6/10/2022	8:51						0:01:25	Data Analyst

Navigator

Display Options

Power BI - Final Project (1).xlsx [1]

☒ Data Professional Survey

Data Professional Survey

Unique ID	Email	Date Taken (America/New_York)	Time Taken
62a33b3db4da29969c62df3d	anonymous	6/10/2022	
62a33ba1bae91e4b8b82e35c	anonymous	6/10/2022	
62a33c2cbc6861bf3176bec1	anonymous	6/10/2022	
62a33c8624a26260273822f9	anonymous	6/10/2022	
62a33c91f3072dd892621e03	anonymous	6/10/2022	
62a33cb6cf25554317300177	anonymous	6/10/2022	
62a33cb72e54c9003e531c65	anonymous	6/10/2022	
62a33cd30f8c8599d5af0f8f	anonymous	6/10/2022	
62a33cd3cf255543173001d9	anonymous	6/10/2022	
62a33cd8bc6861bf3176c05f	anonymous	6/10/2022	
62a33ce918134ddc75ce8c30	anonymous	6/10/2022	
62a33cf30a77c1a77f65baa2	anonymous	6/10/2022	
62a33d15f408bae018ed370d	anonymous	6/10/2022	
62a33d1ebae91e4b8b82e707	anonymous	6/10/2022	
62a33d4624a26260273824c4	anonymous	6/10/2022	
62a33d5c0f8c8599d5af107c	anonymous	6/10/2022	
62a33d850f8c8599d5af10cb	anonymous	6/10/2022	

The data in the preview has been truncated due to size limits.

Load Transform Data Cancel

I deleted irrelevant columns (Browser – Referrer) by clicking on the “browser” header, held “shift”, then “referrer” and finally removed the columns.

The screenshot displays the Microsoft Power Query Editor interface. The main window shows a data table with the following columns: Browser, OS, City, Country, and Referrer. The 'Transform' tab is active, and the 'Remove Columns' option is selected. The 'Query Settings' pane on the right shows the 'APPLIED STEPS' list with 'Changed Type' selected.

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

PREVIEW DOWNLOADED AT 12:33 PM

Q1 includes texts in brackets and after it that are not relevant, to deal with this, I split Q1 column (using Split column). “Split column By Delimiter” using the “c” delimiter, “Split at left-most delimiter”. This creates a new column – Q1.2 of text within () and proceeding it.. I deleted the new column. The same for column Q4.

The screenshot displays the Power Query Editor interface. The main area shows a table with 23 rows and 3 columns. The first column contains timestamps, the second column contains job titles, and the third column contains 'Yes' or 'No' responses. The 'Applied Steps' pane on the right lists the transformations: Source, Navigation, Promoted Headers, Changed Type, Split Column by Delimiter, and Changed Type1.

	A ^B _C Q1 - Which Title Best Fits your Current Role?.1	A ^B _C Q1 - Which Title Best Fits your Current Role?.2	A ^B _C Q2 - Did you switch careers into Da
1	12:00:44 AM Data Analyst		null Yes
2	12:01:30 AM Data Analyst		null No
3	12:02:18 AM Data Engineer		null No
4	12:02:10 AM Other	Please Specify):Analytics Consultant	Yes
5	12:01:51 AM Data Analyst		null Yes
6	12:02:34 AM Data Analyst		null Yes
7	12:01:15 AM Data Scientist		null Yes
8	12:01:25 AM Data Engineer		null Yes
9	12:02:10 AM Data Analyst		null Yes
10	12:01:27 AM Data Analyst		null Yes
11	12:01:29 AM Data Analyst		null Yes
12	12:02:31 AM Data Analyst		null Yes
13	12:03:20 AM Data Analyst		null Yes
14	12:00:55 AM Data Scientist		null Yes
15	12:01:24 AM Data Analyst		null No
16	12:00:47 AM Data Analyst		null Yes
17	12:01:06 AM Data Analyst		null Yes
18	12:01:04 AM Student/Looking/None		null Yes
19	12:01:05 AM Student/Looking/None		null No
20	12:01:21 AM Data Analyst		null No
21	12:01:23 AM Data Analyst		null Yes
22	12:01:35 AM Data Analyst		null Yes
23	12:01:25 AM Data Analyst		null Yes

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

PREVIEW DOWNLOADED AT 12:33 PM

Q5 includes texts after colon which are not relevant. To deal with this, I repeated the same process for Q5. I splitted the columns by delimiter, using the “colon” as delimiter, “Split at left-most delimiter”. This creates a new column Q5.2. I then deleted the new column. The same for column Q13

The screenshot displays the Power Query Editor interface. The ribbon at the top includes tabs for File, Home, Transform, Add Column, View, Tools, and Help. The Transform tab is active, showing options like Split Column, Group By, and Replace Values. The main area shows a data table with the following columns: Q5 - Favorite Programming Language.1, Q5 - Favorite Programming Language.2, and Q6 - How Happy are you in your Current Position. The data rows show various programming languages (Python, R, SQL) and happiness ratings (null, 1, 2, 3). The right-hand pane shows the Query Settings for 'Data Professional Survey', including the Name, All Properties, and Applied Steps. The Applied Steps list includes: Source, Navigation, Promoted Headers, Changed Type, Split Column by Delimiter, Changed Type1, Split Column by Delimiter1, and Changed Type2.

	A ^B C Q5 - Favorite Programming Language.1	A ^B C Q5 - Favorite Programming Language.2	1 ² 3 Q6 - How Happy are you in your Current Position
1	Python		null
2	R		null
3	Python		null
4	R		null
5	R		null
6	Python		null
7	Python		null
8	Other	SQL	
9	R		null
10	Python		null
11	Python		null
12	Python		null
13	R		null
14	Python		null
15	R		null
16	Python		null
17	Python		null
18	Python		null
19	R		null
20	Python		null
21	Python		null
22	Other	Mostly use sql but that's not programming language..	
23	Python		null

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

PREVIEW DOWNLOADED AT 12:33 PM

Q3 is in ranges which is not good enough. It should have definite values. First, I duplicated the column, splitted the duplicate (By digits to non-digits). I deleted the column with “k” values (copy 3) Then for copy 2, using “Find and Replace”, I replaced “+” with 225, “-” with nothing and “k” with nothing.

transformColumnTypes("#Split Column by De

A ^B C	Q3 - Current Yearly Salary (in USD)
	106k-125k
	41k-65k
	0-40k
	150k-225k
	41k-65k
	0-40k
	0-40k
	125k-150k
	86k-105k
	41k-65k
	66k-85k
	0-40k
	0-40k
	0-40k
	41k-65k
	41k-65k
	0-40k
	0-40k
	41k-65k
	0-40k
	41k-65k
	106k-125k
	0-40k

I converted the two duplicate columns data type to “whole numbers”. Furthermore, to create the needed salary column, I clicked on “Add Column”, selected “Custom Column” and named it “Average Salary”. In the Custom Column formula, I inserted “Copy 1” + “Copy 2”/2

([#”Q3-current yearly salary (in USD)-copy.1”]+[#”Q3-current yearly salary (in USD)-copy.2”]/2

Then I deleted copy1 and copy2, left with column Average Salary.

ie.RenameColumns(#[“Added Custom1”,{“Avderage Salary”, “Average Salary”}})

3 Q3 - Current Yearly Salary (in USD) - Copy.1	123 Q3 - Current Yearly Salary (in USD) - Copy.2	ABC 123 Average Salary
106	125	115.5
41	65	53
0	40	20
150	225	187.5
41	65	53
0	40	20
0	40	20
125	150	137.5
86	105	95.5
41	65	53
66	85	75.5
0	40	20
0	40	20
0	40	20
41	65	53
41	65	53
0	40	20
0	40	20
41	65	53
0	40	20
41	65	53
106	125	115.5
0	40	20

With the cleaning done so far, the data is finally ready for the analysis.

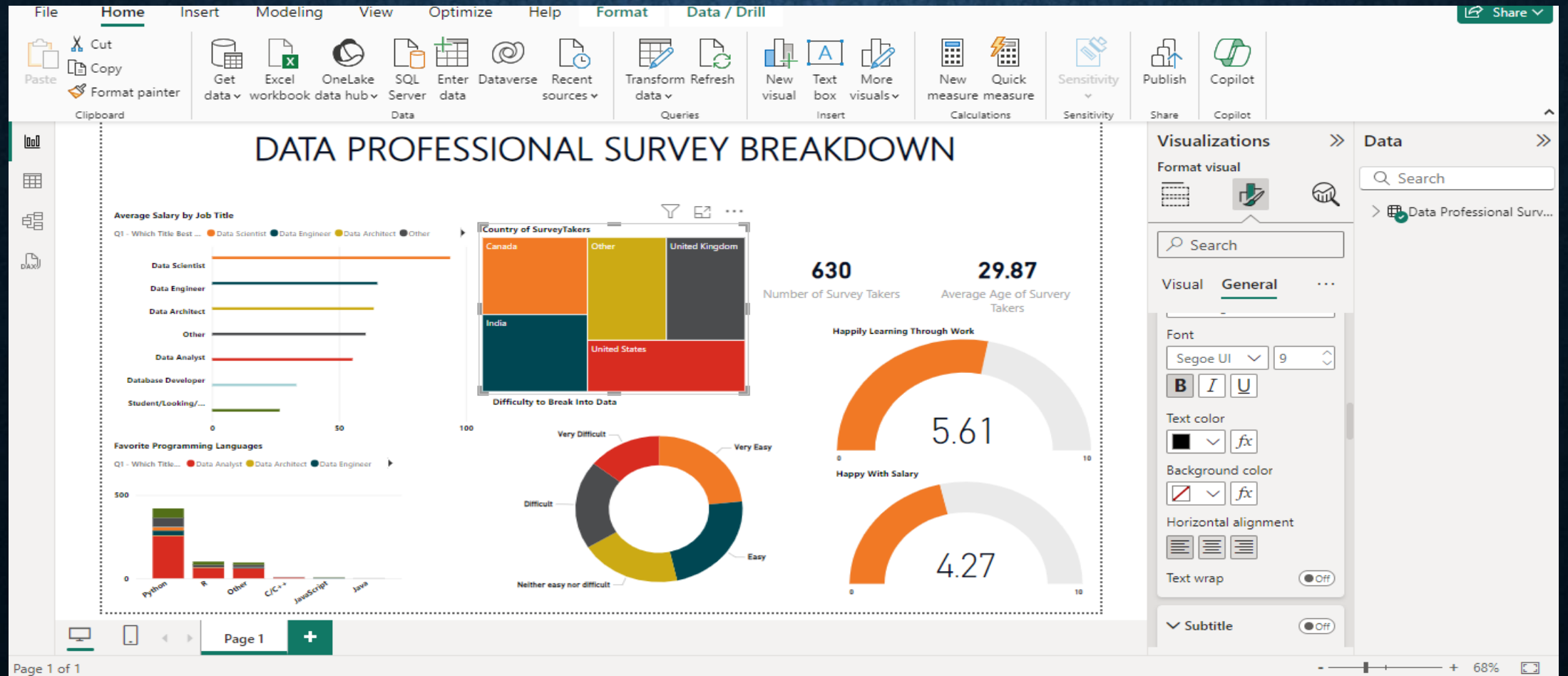
I then clicked “CLOSE AND APPLY to load the ready cleaned data into the visualization pane chart creations

The screenshot shows the Microsoft Power BI Desktop interface. The ribbon at the top includes the following groups: Clipboard (Cut, Copy, Paste, Format painter), Data (Get data, Excel workbook, OneLake data hub, SQL Server, Enter data, Dataverse, Recent sources), Queries (Transform data, Refresh), Insert (New visual, Text box, More visuals), Calculations (New measure, Quick measure), Sensitivity, Share (Publish), and Copilot. The main canvas area displays the text "Build visuals with your data" and "Select or drag fields from the Data pane onto the report canvas." Below this text is a graphic showing a dashed box representing a report canvas and a list of fields with a checkmark and an arrow pointing to the canvas. The right-hand pane is divided into two sections: "Visualizations" and "Data". The "Visualizations" section has a "Build visual" dropdown and a grid of visualization icons. The "Data" section has a search bar and a list of fields under the heading "Data Professional Su...". The fields listed are: Average Salary, Browser, City, Country, Date Taken (A..., Email, OS, Q1 - Which Ti..., Q1 - Which Ti..., Q10 - Current..., Q11 - Which ..., Q12 - Highest..., Q13 - Ethnicity, Q2 - Did you ..., Q3 - Current Y..., Q3 - Current Y..., Q3 - Current Y..., and Q4 - What Ind....

RESULTS

- The total number of survey takers is 630
- Country of Survey Takers are: Canada, India, United Kingdom, United States and Others
- Average Age of survey Takers is 30 years
- Favorite programming language is Python
- Difficulty to break into Data: Easy
- There are more data professionals happy learning through work
- There are less data professionals happy with their salary
- Average Salary by job roles (in USD): DATA SCIENTIST – \$93.78K, DATA ENGINEER – \$65.09K, DATA ARCHITECT – \$63.69K, OTHER – \$60.49K, DATA ANALYST – \$55.30K, DATABASE DEVELOPER – \$33.20K

SCREENSHOT OF THE DASHBOARD



LINKS

- PORTFOLIO LINK: <https://www.datascienceportfol.io/ElekwaChinedu>
- GITHUB LINK: <https://github.com/Nedupelekwa>
- LINKEDIN LINK: <https://www.linkedin.com/in/chinedu-elekwa-7656a91b0/>
- All documents for this project can be downloaded and viewed from my github account specifically at: <https://github.com/Nedupelekwa/DATA-PROFESSIONALS-SURVEY.git>