**Data Professional Survey Dataset**

Project Overview:

The data professional survey dataset analysis project aims to analyze a comprehensive survey dataset of data professionals. The dataset contains information such as participants' demographics, current roles, career paths, salary details, industry affiliations, programming language preferences, and job satisfaction ratings. The objective of the analysis is to gain insights into the data professional landscape, understand career transitions into data, evaluate job satisfaction factors, and identify key trends and patterns in the industry. The analysis will provide valuable information for professionals, organizations, and stakeholders in the data field, assisting them in making informed decisions, improving work environments, and enhancing career opportunities.

Data Description:

The dataset contains 28 columns and 630 rows and contains the following columns:

* Unique ID: An identifier for each survey response.
* Email: The email address of the participant.
* Date Taken (America/New\_York): The date the survey was taken in the America/New\_York time zone.
* Time Taken (America/New\_York): The time the survey was taken in the America/New\_York time zone.
* Browser: The web browser used to take the survey.
* OS: The operating system used to take the survey.
* City: The city where the participant is located.
* Country: The country where the participant is located.
* Referrer: The source that referred the participant to the survey.
* Time Spent: The duration of time spent on the survey
* Which Title Best Fits your Current Role?: The participant's current job title.
* Did you switch careers into Data?: Whether the participant switched careers into the field of data.
* Current Yearly Salary (in USD): The participant's current annual salary in USD.
* What Industry do you work in?: The industry in which the participant is employed.
* Favorite Programming Language: The participant's preferred programming language.
* How Happy are you in your Current Position with the following? (Salary): Happiness level with salary in the current position.
* How Happy are you in your Current Position with the following? (Work/Life Balance): Happiness level with work/life balance in the current position.
* How Happy are you in your Current Position with the following? (Coworkers): Happiness level with coworkers in the current position.
* How Happy are you in your Current Position with the following? (Management): Happiness level with management in the current position.
* How Happy are you in your Current Position with the following? (Upward Mobility): Happiness level with upward mobility opportunities in the current position.
* How Happy are you in your Current Position with the following? (Learning New Things): Happiness level with learning new things in the current position.
* How difficult was it for you to break into Data?: The perceived difficulty of entering the field of data.
* If you were to look for a new job today, what would be the most important thing to you?: The most important factor in considering a new job.
* Male/Female?: The gender of the participant.
* Current Age: The age of the participant.
* Which Country do you live in?: The country of residence of the participant.
* Highest Level of Education: The highest level of education attained by the participant.
* Ethnicity: The ethnicity of the participant.

The link to the dataset can be found here

Tool Used: Power BI

Observation with the data

The following observations were made in the data

1. Null/missing values
2. Wrong Formatting
3. Wrong spellings and abbreviations of words

Data Cleaning and Preprocessing:

For the data cleaning and preprocessing step in the data professional dataset, I followed the following steps:

Firstly, I loaded the data into Power BI and accessed the Power Query Editor by clicking on the transform tab. This allowed me to initiate the data cleaning process.

I addressed the issue with the "date taken" column by formatting it to the date data type. However, I encountered errors in some rows. To resolve this, I investigated and identified that changing the regional settings for the file from Nigeria to the United States resolved the problem effectively.

To ensure clarity and descriptive labeling, I renamed all the columns appropriately, providing a clear understanding of the values they hold.

Since the columns for email, browser, OS, city, country, and referrer were all empty and wouldn't contribute to my analysis, I removed them from the dataset.

Next, I focused on the "current yearly salary" column (in USD). As the values were entered in ranges, I used a delimiter to split the column into minimum and maximum values. Additionally, I replaced the "k" symbol (representing thousand) in those columns with three zeros.

After splitting the column, I noticed that the maximum value column contained null values, indicating that respondents had only provided one value. To address this, I replaced the null values with zero. I then utilized the custom column feature to calculate the average of the maximum and minimum values for each row, adding a new column named "average salary" to the dataset.

Moving on, I encountered null values in columns such as "How Happy are you in your Current Position with the following? (Salary)," "How Happy are you in your Current Position with the following? (Work/Life Balance)," "How Happy are you in your Current Position with the following? (Coworkers)," "How Happy are you in your Current Position with the following? (Management)," "How Happy are you in your Current Position with the following? (Upward Mobility)," and "How Happy are you in your Current Position with the following? (Learning New Things)." I replaced these null values with the respective medians of each entire column.

In the "job role" column, I replaced the value "Other (specify)" with a blank, leaving only the specified values in the column. If no specified value was present, I input "Not Stated" as a replacement.

To provide a rating for the columns mentioned in step 7, I employed a conditional column approach. For columns like "How Happy are you in your Current Position with the following? (Salary)," "How Happy are you in your Current Position with the following? (Work/Life Balance)," "How Happy are you in your Current Position with the following? (Coworkers)," "How Happy are you in your Current Position with the following? (Management)," "How Happy are you in your Current Position with the following? (Upward Mobility)," and "How Happy are you in your Current Position with the following? (Learning New Things)," I used the following rating scale: If the selected column value is less than or equal to 2, it is considered "Very Dissatisfied." If the value is less than or equal to 4, it is labeled "Dissatisfied." If the value is less than or equal to 6, it is categorized as "Neutral." If the value is less than or equal to 8, it is regarded as "Satisfied." Otherwise, if the value is greater than 8, it is labeled as "Very Satisfied."

Lastly, I addressed the "age" column. Observing that the maximum age was 92, I utilized the conditional column approach once again to establish age ranges. The age ranges were defined as follows: Ages 18-34, Ages 35-49, Ages 50-64, Ages 65-79, and Ages 80-92.

I then went through my ethnicity column to clean it and get rid of all errors. After which I finally went through each column again one after the other.

By following these systematic steps, I ensured that the data was properly cleaned and preprocessed, setting the stage for meaningful analysis and insights.

Data Analysis Techniques:

To gain valuable insights from the dataset. Some of the used data analysis techniques include:

1. Descriptive Statistics
2. Data Visualization
3. Data Aggregation
4. Data Cleaning and Preprocessing
5. Data Interpretation and Communication

By applying these data analysis techniques, I was able extract valuable insights, make data-driven decisions, and derive actionable recommendations from the dataset at hand.

Findings and Insights:

Visualizations and Reports**:**

Limitations and Assumptions:

Recommendations and Conclusions: