Data Cleaning

Brought dataset into Power BI, to be transformed (cleaned). The following was performed:

Empty Rows:

There are a bunch of completely empty columns/features, which were deleted. Specifically:

- "Browser"
- "OS"
- "City"
- "Country"
- "Referrer

Many types of "Other":

There are many different unique values in the "Q1 – Which Title…" column, specifically for "Other". Thus, all values that contain "Other", will be grouped together, labelled as "Other", to dramatically reduce the number of categories, and thus make this column much easier to graph and filter effectively. This also applies to the following columns:

- "Q4 – What Industry…"
- "Q5 – Favorite…"
- "Q8 – If you…"
- "Q11 – Which…"
- "Q13 – Ethnicity"

Note on Q5, Q11:
The "Q5 – Favourite…" column had several entries (proportionally speaking) for SQL, thus SQL entries were grouped under "SQL". This also applies to Q11, with several "Nigeria" entries.

Created Income Column:

Duplicated the salary column "Q3 – Current…", then converted it into integers (quantitative), instead of ranges (categorical), to make analysis later on, easier. Accomplished by substituting ranges, for their averages (so for example, "0-40k", was replaced with the integer 20,000, and "150k-225k", replaced with the integer 187,500). We can now also sort the "Q3" column properly, by using this new column.

Date Taken Format:

Changed format of the "Date Taken" column to Date, and restructured from Month/Day/Year, to Day/Month/Year, using Locale settings, (as I live in Australia, and likely anyone else who reads this). Additionally, "Time Spent" format changed to duration, and "Time Taken" column was deleted, as I could not think of a use for it, for the dashboard.

Column Header Lengths:

Many of the columns have long headings, especially the question columns. Thus, they have all been renamed, to be considerably shorter, as follows:

- Date Taken (America/New_York) => Date
- Q1 - Which Title Best Fits your Current Role? => Job Title
- Q2 - Did you switch careers into Data? => JobSwitch ToData?
- Q3 - Current Yearly Salary (in USD) => Salary (USD)
- Q4 - What Industry do you work in? => Current Industry
- Q5 - Favorite Programming Language => Fav Program Lang
- Q6 - How Happy are… (Salary) => Happy Salary?
- Q6 - How Happy are… (Work/Life Balance) => Happy WorkLifeBalance?
- Q6 - How Happy are.. (Coworkers) => Happy Coworkers?
- Q6 - How Happy are… (Management) => Happy Management?
- Q6 - How Happy are.. (Upward Mobility) => Happy UpwardsMobility?
- Q6 - How Happy are.. (Learning New Things) => Happy Learning?
- Q7 - How difficult… break into Data? => Difficult BreakIntoData?
- Q8 - If you were… => NewJob ImportantThing?
- Q9 - Male/Female? => Gender
- Q10 - Current Age => Age
- Q11 - Which Country do you live in? => Country
- Q12 - Highest Level of Education => Highest Education
- Q13 – Ethnicity => Ethnicity

Education ordering:

Created a new conditional column, “Order Highest Education”, based on Education, used solely to order education, as follows:

- High School => 1
- Associates => 2
- Bachelors => 3
- Masters => 4
- PhD => 5

Converting Difficulty column:

Difficulty column, was converted into integers, to allow for averaging of difficulty, amongst each Highest Education group (which was then used to build a line chart).