**Data Science Jobs: Data Cleaning Process Documentation**

**Purpose**: To ensure data accuracy and consistency

Process Overview:

1. Duplicates: Checked for duplicates using the “Remove Duplicates” feature, there were no duplicates in the dataset.
2. Job titles were converted to proper casing using the “PROPER ()” function for uniformity.
3. The Salary Estimate column had its values as follow: $137K-$171K (Glassdoor est.), It was split into ”Min Salary Estimate” and “Max Salary Estimate” using the “Text to columns” feature.
4. The company name column contained names and ratings combined e.g. (Healthfirst  
   3.1), the 2 were split so that the company names only include names. This was achieved by using the “LEFT ()” AND “FIND ()” functions.
5. Company name column values were standardised using the “PROPER()” function.
6. The Size column values which had value of -1 were replaced with “Unknown” in order to indicate missing data.
7. Founded column cells which had value of -1 were replace with 0, for unknown year.
8. Names in the type of ownership were cleaned to remove “Company-” prefix of the names using the “Replace” feature.
9. The competitors column had cells which had a value of -1 and they changed to “No competitor” for clarity.
10. The Revenue and Description columns were deleted because they were not required for analysis.

Conclusion: The dataset was cleaned and it now have standardised names and values which makes it ready for analysis.