Task 1-Data Exploration

1.The data types are text, whole number and percentage.

2.There is missing data in the url and shazam columns.

3.Yes, there are duplicates.

4. Some issues are misspelled names, song titles and missing text in the key column, but overall formatting is correct.

Task 2-Handling missing and incorrect data

1. How could you validate the values in streams? The way that the streams were able to be validated were through the Interquartile Range (IQR) Method to find outliers based on the spread of data between the Q1 and Q3.
2. What data/info would be needed to calculate the value to estimate if the data correct? The data that was needed to calculate the value to estimate if the data is correct was the Standard deviation, the Q1, Q3 to calculate the Lower Bound and Upper Bound. Once the Lower and Upper Bound was calculated any values in ‘Streams’ below the lower bound and above the upper bound were filtered through a conditional column as “Outlier” or “Normal”.

Task 4- Removing duplicates

1. How many duplicate records did you find?

Out of a total of 953 rows there were only 4 duplicates found based on artist-name, trackname and released year.

Task 5 – Documenting the process

1. What issues did you find in cleaning the data- which columns were changed? Issues that were found while cleaning the data were many things like missing values in the data set, duplicates and misspelled names/song titles. An outlier column was added to identify the suspicious streams column.
2. For each column that was changed, what did you change? Replacing the null and empty cells the value 0.
3. Were any duplicate records found? Yes, Duplicate records were found.
4. If so, what process did you use to confirm the record was a duplicate? He combined those columns and deleted the duplicates within those columns.
5. What assumptions were made when cleaning the data? All data types were going to be the same. Easy navigating through Power BI.