## Goal

The objective of this exercise is to gain further practice in using PowerBI Query Editor features to cleanse and reformat data. It focuses on merge and append operations as well as text reformatting.

You will take the two datafiles (UFO\_Details and UFO\_Comments) and join them on the ID column in PowerBI. You will then use a range of PowerBI transform utilities to cleanse the data.

Steps

1. Open a new PowerBI workbook

2. Import UFO\_Details.txt and UFO\_Comments.txt as text files. Take care with the date formats as they are US dates. For this we need to change the regional settings in PowerBI. In Query Editor go to File-> Options and Settings -> Options-> Regional Settings and change the Locale to English (United States)

3. There are no headings in the data file. Add column headings to your tables following import

• UFO\_Details

• ID

• SIGHTING\_DATE

• CITY\_STATE

• COUNTRY\_CODE

• SHAPE

• DURATION

• DATE\_POSTED

• UFO\_Comments

• ID

• COMMENTS

4. Merge the two datasets on the ID field by creating a Merge query (Merge as New). Make sure the resulting query shows only the unique columns across the two datasets.

5. Rename the new merge query as “UFO\_Merge”. You’ll notice the COMMENTS column at the end is prefixed with the previous table name. Remove the table prefix on this column to “COMMENTS”

6. Some new sightings have just come through on another file (UFO\_latest.txt) – this time on a single file showing the *details and comments together*. Import the data into PowerBI and ensure this new table has the same column headings (check case sensitive!) as the merged table (“UFO\_Merge”) from step 5. Label this new table “UFO\_Latest”

7. Create a *new* query to hold latest sighting data (“UFO\_Latest”) *appended* to the bottom of your existing merged data (“UFO\_Merge”). Label this new query “UFO\_Append”

8. Check the append has worked successfully by reviewing the column headings in “UFO\_Append”. There should be the same number as in “UFO\_Merge” and “UFO\_Latest”. If

not, check the column headings in the “UFO\_Merge” and “UFO\_Latest” tables. They need to be exact otherwise additional columns will be created in the appended result.

9. We now have a successfully merged and appended UFO data file. We now need to ensure the data quality in our columns is fit for purpose and the data types correct. First, we’ll look at our data types.

• Both SIGHTING\_DATE and DATE\_POSTED are set as text fields. These need to be changed to DATETIME and DATE respectively

• Ensure ID and DURATION are both set to Whole Numbers

10. Now to cleansing. Reviewing the data we can see issues with:

• SHAPE:

• Replace the missing values with “Not Provided”

• Amend the casing to capitalise each word. Go to Transform ribbon -> Format->Capitalise Each Word

• COMMENTS – nonprinting characters of:

&#44 (,)

&#39 (‘)

&#33 (!)

&#8220(“)

&#8230 (horizontal ellipsis …)

Use the REPLACE function on the Transform ribbon to fix these issues. The horizontal ellipsis you can just remove entirely.

11. There are also issues with the CITY\_STATE column. We can see embedded county and country names held in brackets after the city. What we would like is to have the CITY\_STATE to just hold the city value.

- Use the SPLIT function to split out the city from the text held in brackets in the CITY\_STATE field. This will place the data in brackets into a new field. Relabel CITY\_STATE to just “CITY”

- Use the SPLIT function again to parse out the country name (eg “England”, “wales”, “Scotland”….) from the rest of the text in the new field. Once you’ve got it extracted label this field “COUNTRY”

- Use the REPLACE function on the new COUNTRY field to remove the final “)” character.

- Any intermediate columns as a result of the SPLIT can be removed

- We’re almost there but not quite. If you review the data in the newly generated COUNTRY field you’ll see there are still issues with some misspellings and non-standard values. We will try and correct that by generating a lookup table to fix the values.

- Before we generate the lookup table we can quickly fix some casing issues in the CITY and COUNTRY\_CODE data.

- Highlight the CITY column and go to Transform ribbon -> Format->Capitalise Each Word

- Highlight the COUNTRY\_CODE column and Format -> Uppercase

- We will fix the COUNTRY casing as part of the lookup cleanse

- Highlight the new COUNTRY column and right-click “Remove Duplicates”. Copy the values. This will give us a unique list of values for our lookup table. We will go back and remove this step once we have created our new lookup table.

- On the Home ribbon click the “Enter Data” button. This will bring up a new empty table. Paste the copied values into it



- Click in the empty column to the right and add a new column header for COUNTRY\_CLEAN

- Manually add the correct values for each of the unique values similar to below. We don’t want the null value, so remove that row entirely



- Save the table and label to COUNTRY\_LOOKUP

- Back in Query Editor go to the “UFO\_Append” table and ensure you delete the step “Removed Duplicates” in the Applied Steps window. Delete the step by clicking on the cross on the LHS of the step. This will reset our data to include all the rows that we previously filtered – which we did just to create a set of unique values for our lookup table.



- Now use the COUNTRY\_LOOKUP table to extract a standardised country name for the UFO\_Append dataset. You can do this by merging the two datasets on a common column (hint: have a think which column you can use as join key between these two tables). Add the clean country name as a new column and remove the old unstandardized version.

- We are now finished in the Query Editor. Configure the tables so that only the UFO\_Append table gets loaded into the model

12. In Data view configure the data so that we can create a map visualisation using the CITY\_STATE column. Ensure you configure the column such that only those cities in the UK will be shown on the map.

13. Create 2 visualisations to show your results:

* **BI Card to show the total number of sightings (1911 sightings)**
* **Map visualisation which can be filtered by Shape**