# Scenario Worksheet

## The X-Files – UFO Sightings Data Investigation

**Background:**  
The *Department of Unexplained Phenomena (DUP)* has hired your data team to review recent UFO sighting reports. The data has been collected by field agents, but some of it is messy and inconsistent. Your task is to clean, blend, and analyse the data to help identify credible patterns.

You will work with two linked files:

* **AgentProfiles.csv** – details of investigation agents
* **SightingReports.csv** – UFO sightings recorded by those agents

**Step-by-Step Task Guide**

**Step 1 – Import both datasets**  
Open Excel and import the two CSV files. Save the workbook as **XFiles\_Analysis.xlsx**.  
Check that column headers are visible and that both tables have matching **AgentID** fields.

**Step 2 – Inspect the data for problems**  
Notice any inconsistencies in capitalisation (e.g., “london” vs “London”), extra spaces, or blank rows.  
Use **Filter** and **Sort** to check for:

* Duplicate Agent IDs
* Missing Regions or Credibility Scores
* Unnecessary spaces in names

**Step 3 – Clean the text data**  
Use functions to tidy the data:

* =TRIM() → remove extra spaces
* =PROPER() → capitalise names and locations
* =UPPER() → make codes consistent  
  Delete any duplicate rows using **Remove Duplicates**.

**Step 4 – Blend the data**  
Use **XLOOKUP** or **VLOOKUP** to bring Agent details (Region, ExperienceYears) into the SightingReports table using **AgentID** as the key.  
Check that all lookups return a valid match, if not say why.

*(Stretch: do this using Power Query > Merge Queries.)*

**Step 5 – Add calculated columns**  
Add two new columns:

* **“Classification”** – use an IF() formula to mark “High Credibility” if Score ≥ 80, “Medium Credibility” otherwise
* **“Adjusted Duration”** – divide Duration by ExperienceYears to see time spent per experience level.
* Format **“Adjusted Duration”** to two decimal places

**Step 6 – Apply conditional formatting**  
Highlight cells where:

* Credibility Score > 90 (green fill)
* Duration > 120 minutes (orange fill)

This helps spot the most interesting or time-intensive cases.

**Step 7 – Calculate regional insights**  
Use SUMIF and AVERAGEIF to summarise:

* Average credibility score per Region
* Sum of duration for confirmed cases

**Step 8 – Create a Pivot Table and Chart**  
Insert a Pivot Table showing:

* Rows = Region
* Values = Average Credibility, Count of Sightings  
  Then create a Pivot Chart (Column or Bar).  
  Rename it **Regional\_Credibility\_Chart**.

**Step 9 – Interpret your findings**  
Write a short paragraph summarising what you discovered.  
For example:

* Which region reported the most credible sightings?
* Are more experienced agents linked to higher credibility reports?

**Step 10 – Reflect on your process**  
Answer briefly:

1. What cleaning steps improved your results?
2. How did using functions like SUMIF and AVERAGEIF make your analysis easier?
3. If you repeated this task, what would you automate using Power Query or formulas?