Contents

Summary

- Source some data and shape/format it for use in your own analysis.
- Practise using data quality tools to assess data quality and its impact on analysis requirements, while also identifying solutions for any quality issues.
- You'll also cleanse and transform data using relevant tools and techniques, assess its integrity and reliability
- Record all activities, issues, and resolutions.
- Produce a **document** where you describe your research findings, and then produce the necessary analysis, plans, and evaluations to demonstrate your knowledge use of tools.
- Evidence can be gathered from your usual workplace activities and also from a simulated data task that would be similar to any you carry out in your role.
- Practise stage datasets that are available to you via Mimeo. If you haven't already done so, download the Practise stage materials. ???

Overview - Data collation, formatting, and transformation

- **Understanding datasets**:
    - Load all datasets and get an overview of their structures and content.
      *Load into PowerBi or Jupyter Notebook
    - Identify the primary attributes, data types, and preliminary quality issues.
      *Can you see anything to use as a primary key or unique identification, what datatypes look suitable, text(varchar), date, time.
      *Is there any duplicated data, anything out of place, any different data types in a column, missing data
- **Collation and formatting**:
    - Standardise individual datasets: ensure you have uniform naming conventions, date formats, and unit measures.
      * Be consistent i.e. if you use this date format 12/09/24, then use this in date values, camel case employeeID, employee.employeeID
    - Address any preliminary data defects like outliers or erroneous values.
      *Are any values much smaller or larger than the average values, run a =avg() then =max()
    - Prepare datasets for amalgamation by adding complementary attributes.
      * You may need to add some columns to make relationships between multiple tables. For Example add primary key.

- **Data merging and combination**:
    - Identify and use primary keys or identifiers for combining data.
      * You can do this in PowerBi or Power Query in Excel. For example create a key for each record and use this to connect the 2 tables.
    - Handle missing values: Decide whether to impute, drop, or flag them.
      * Impute usually means replacing blank values with mean(average) values, you might also just pull those records out into another table
    - Address any inconsistencies or conflicts resulting from the combination, like same attributes with different values.
      *Check all the matching columns have the same values i.e. Primary Key to Foreign Key (these should have the same values/data types)
- **Reshaping and restructuring**:
    - Pivot or unpivot data tables, as necessary.
    - Reorder, rename, or recategorize columns or variables to fit the desired final structure.
    - Transform data values where necessary, e.g., converting units or encoding categories.
- **Documentation and summary**:
    - Record the steps taken in the collation, combination, and transformation processes.
    - Document challenges encountered and how they were addressed.
    - Reflect briefly on the importance of these processes in the broader data analysis pipeline.

- **Formatted datasets**:
    - Individual datasets after initial formatting, ready for combination.
      *Possibly take screengrabs of the cleaned datasets
- **Final merged dataset**:
    - The resultant dataset post-combination, reshaping, and restructuring.

      *Possibly take screengrabs of the cleaned dataset
- **Process documentation**:
    - The methodologies used, decisions made, and challenges encountered during the tasks.
      * for example did you replace blank values with mean values, how did you choose a primary key to relate tables.

Task 1 Demonstrating knowledge on assessing data quality

In this task, you will demonstrate your ability to use data quality toolsets to assess data for missing values, input issues, inconsistencies, and any patterns indicating misrepresentation.

To complete this activity, you will need to work with a dataset known to have data quality issues. *Try and find a dataset with, missing values, duplicates, wrongly placed datatypes etc.

Please note: You must use data quality toolsets to identify and document these issues.
* These could be PowerBi, Excel, Python (Jupyter Notebook)

The purpose of this task is for you to display proficiency in using data quality toolsets,

- **Familiarisation with dataset:**
    - Load the dataset into the chosen data quality toolset.
    - Perform an initial visual examination to understand the nature and structure of the data.
      * Have a scan, maybe look at the data using the Column Quality tools in PowerBi or Excel
- **Missing data identification:**
    - Use the toolset to identify and quantify missing data across columns/variables.
    - Document the severity and potential implications of these missing values.
      * Missing or incorrect values will create problems for loading into a RDMS if this is the intention.
- **Spotting input issues:**
    - Check for unusual input values (e.g., negative numbers were not expected textual data in numeric fields).
    - Record such anomalies and speculate on their causes.

- **Inconsistency detection:**
    - Compare similar data points to find inconsistencies (e.g., differing date formats, mislabelled categories).

      *Lots of ways for this in Excel:

      1. Quick conditional formatting to compare two columns of data

      This method might be the quickest and most simple method. It will allow you to highlight a cell or range of cells based upon defined criteria. A Duplicate Values setting box is available in the Conditional Formatting drop down list, where you can define the formatting and selection of Duplicate or Unique values. Formatting of values identified can then be defined for both Duplicate or Unique values for datasets in both lists.

      2. Match Data using Row Difference Technique

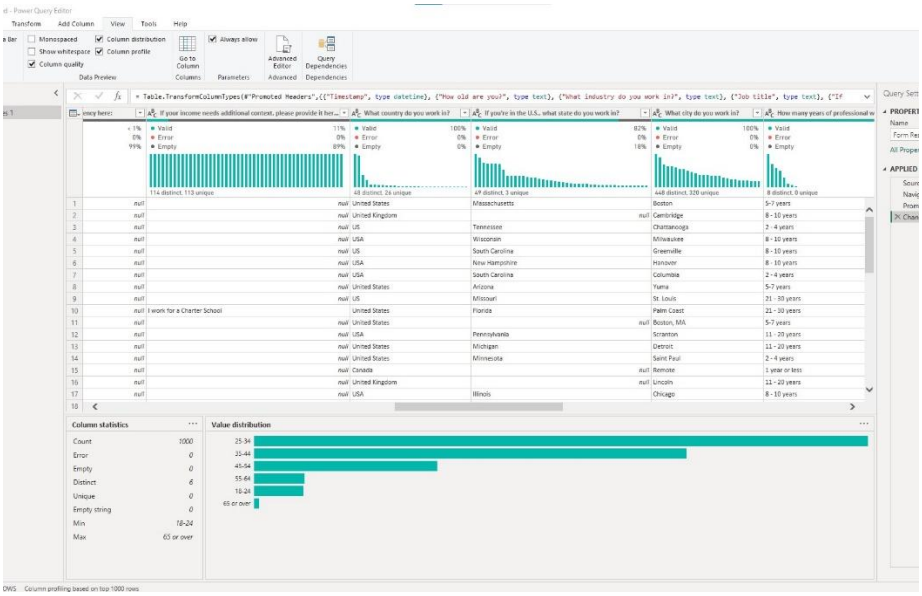
      When comparing two lists of data, select both columns of data, press F5 key on the keyboard, select the "Go to special" dialog box. Then select "Row difference" from the options. Matching cells of data across the rows in the columns are in white color and unmatched cells appear in grey color.

      3. Row Difference using IF Condition

      The IF Condition formula states if there is a match in the row when comparing two lists of data. If there is a match, the result of the formula will be "Matching" and if not then "Not Matching". The formula would look something like this: =IF(A2=B2, "Matching", "Not Matching"). The formula would need to be copied down the row of cells.

    - Document the nature and extent of these inconsistencies.

- **Pattern analysis for misrepresentation:**
    - Use visualisation or statistical tools within the toolset to identify patterns that may indicate data misrepresentation.
      * Use the Column Quality tool



    - For example, look for unusually uniform data distributions, spikes, or unexpected gaps in time series.

- **Summary and documentation:**
    - Create a brief report of all identified issues, including missing data, input problems, inconsistencies, and patterns of misrepresentation.

Task 1 Evidence required for assessment:

1. **Data quality report:**
    - A brief report highlighting all the data quality issues discovered, organised by type.
2. **Screenshots/exports:**
    - Visual evidence from the toolset, such as screenshots or exported graphs, which show the detected issues.
3. **Reflective note:**
    - A brief note reflecting on the task, the challenges faced, and the implications of the data quality issues identified for potential downstream analysis.

Task 2 Impact assessment and data quality remediation

In this task, you will demonstrate your ability to assess the impact of data quality levels on analysis requirements, and identify solutions to rectify data quality issues.

1. **Dataset examination**:
    - Load and familiarise yourself with the dataset.
    - Perform an initial overview to identify obvious quality issues.
2. **Assessing impact on analysis**:
    - Determine the data analysis requirements, e.g., what kind of insights or patterns one aims to derive.
      *What does the data show, for example is there a column of data that shows imbalance of values (Negative or positive skew), will you use a particular column to look for a trend.
    - For each identified data quality issue, hypothesise its potential impact on the analysis. This could be in terms of accuracy, reliability, or the scope of insights that can be derived.
      * Will outliers effect the analysis
    - Document each issue along with its potential consequences.
3. **Identifying remediation options**:
    - For each identified quality issue, brainstorm workable solutions.
    - Consider a range of remediation strategies, from data cleaning techniques to sourcing additional or alternative data.
    - Document the proposed solutions and, if possible, the pros and cons of each.
4. **Summary and reflection**:
    - Write a concise summary of the impacts and proposed solutions.
    - Reflect on the relationship between data quality and analysis, emphasising the importance of addressing quality issues.

Task 2 Evidence required for assessment:

1. **Impact assessment document**:
    - A structured document detailing each identified data quality issue, its potential impact on the analysis requirements, and the proposed remediation strategies.
2. **Demonstrative examples**:
    - Examples or mini-case scenarios demonstrating the direct effect of a quality issue on an analytical result.
3. **Reflective note**:
    - A brief note confirming your understanding of the significance of data quality in analysis and the challenges of remediation.

Task 3 Data cleansing and transformation

In this task, you will demonstrate your ability to cleanse data by identifying and rectifying common data defects. You will also transform data into formats suitable for subsequent analysis using relevant tools and techniques.

K&Ps: P6 and P7

To effectively accomplish this task, follow the steps below.

1. **Initial dataset review**:
    - Load the dataset into the chosen data processing tool.
    - Conduct a preliminary examination to grasp its structure, content, and obvious quality issues.
2. **Data cleansing**:
    - **Duplicates**: Use the tool's features to identify and eliminate duplicate rows or entries.
    - **Typos**: Check for and rectify typographical errors in textual data fields.
3. **Outdated data**:
    - Filter and remove records that are no longer relevant or that fall outside a specified date range or criteria.
4. **Other defects**:
    - Scan for other anomalies such as inconsistent entries, improbable values, etc., and take appropriate corrective measures.
5. **Data transformation**:
    - Understand the desired format for analysis.
    - Transform categorical data, if necessary (e.g., one-hot encoding).
    - Normalise or standardise numerical data, especially if they are on different scales.
    - Convert data types as needed, like changing strings to datetime objects.
    - Use appropriate functions or methods in the tool to reformat, pivot, or restructure data tables.

6. **Documentation and summary**:
    - Write a concise note detailing the cleansing and transformation processes undertaken, the rationale behind certain choices, and any challenges encountered.

Task 3 Evidence required for assessment:

1. **Cleansed dataset**:
    - The dataset after all cleansing operations, showing the absence of previously present defects.
2. **Transformed dataset**:
    - The dataset in its transformed state, ready for analysis.
3. **Process documentation**:
    - A brief report outlining the steps taken, the decisions made, challenges faced, and how they were addressed.

Task 4 Data integrity testing and quality documentation

In this task, you will test and assess the integrity and reliability of a given dataset and document all data quality-related activities, issues, and your resolutions.

To effectively accomplish this task, follow the steps below.

1. **Initial dataset inspection:**
    - Load the dataset and briefly review its structure and contents.
    - Make initial notes of any glaring issues or anomalies.
2. **Data integrity testing:**
    - Reliability tests: Check the dataset for consistency. This might involve comparing the data against known benchmarks or external data sources to verify accuracy.
    - Validity checks: Confirm that the data is within expected ranges and formats. For example, checking if dates are in the correct format or if numerical values fall within plausible ranges.
    - Uniqueness tests: Ensure there are no unnecessary duplicates that could affect the dataset's integrity.
    - Referential integrity: If the dataset has multiple tables or is relational, ensure that relationships between tables (e.g., primary, and foreign keys) are maintained without any orphaned records.
3. **Documenting quality activities:**
    - Note down all tests performed, their outcomes, and any challenges faced.
    - Highlight any data quality issues found, from minor to major.
4. **Documenting resolutions:**
    - Describe briefly the actions taken to address any identified issues. If certain issues were not addressed, provide reasoning.
    - Reflect briefly on the implications of these issues had they not been addressed.

Task 4 Evidence required for assessment:

1. **Data quality activity log**:
    - A chronological record of all tests and checks performed on the dataset, along with the outcomes.
2. **Issue and resolution documentation**:
    - A structured list or table detailing the identified data quality issues, their potential impacts, and the steps taken (or recommended) to resolve them.