WIE3007 2023/2024 S1

Tutorial 6

1. Data Profiling with Talend Prep

Question: Given a sample dataset of customer information, use Talend Prep to profile the data. Identify and list the following:

1) Number of missing values for each column

	CustomerID	Name	Address	Country	Email	OrderDate	ProductName	ProductID
1)	20	5	12	2	6	0	0	18

2) Any columns that have inconsistent data types
   OrderDate
3) The top 5 most frequent values in the "Country" column
   1) UK – 28
   2) USA – 25
   3) Australia – 23
   4) Canada – 22
   5) Empty - 2

2. Data Cleaning in Talend Prep

Question: You've been provided a dataset with known errors, such as misspellings in the product names and duplicate customer entries. Using Talend Prep:

Correct the misspellings in the product names using transformation suggestions.

By using ‘ Replace the cells that match ‘

Smartt Phones -> Smart Phones

Earphons -> Earphones

Laptoop -> Laptop

Identify and remove duplicate customer records.**

3. Data Standardization using Talend

Question: The date formats in the 'OrderDate' column of a sales dataset are inconsistent, with some dates in MM/DD/YYYY and others in DD-MM-YYYY format. Using Talend Prep:

Standardize all dates to the DD-MM-YYYY format.

Use change date format -> d-M-yyyy

4. Data Validation with Talend

Question: In a given e-commerce transaction dataset, every transaction should have a corresponding customer ID and product ID. Using Talend Prep:

Set up validation rules to ensure that no transaction record is missing either a customer ID or a product ID.

Generate a report highlighting any records that fail this validation.