PROG 8460

ASSIGNMENT #1

Problem

Business Intelligence project. Your initial task will be pre-processing a dataset containing their collected works of art. A local museum has contracted you, the Waterloo Museum of Art, to help implement a Business Intelligence project. Your initial task will be pre-processing a dataset containing their collected works of art.

Task 1 – Data Cleaning [13 marks]:

You are required to clean the dataset using any tool you prefer (Excel, Python, R, etc.) subject to the following conditions:

* Each data record must, at minimum, contain the fields listed below.

|  |  |
| --- | --- |
| Field Name | Required |
| Object ID | X |
| Department | X |
| Object Name | X |
| Title | X |
| Culture | X |
| Artist Nationality | X |
| Object Begin Date | X |
| Object End Date | X |
| Medium | X |
| Credit Line | X |
| Country | X |

* Any field that is not explicitly listed above can be considered irrelevant in this context and can be removed
* Any field designated as "Required" must contain a valid data value.
* The field named "Object ID" represents a unique identifier.
* For any given field, the data values stored within it must be represented consistently
* All data values must make logical sense and correspond to actual possible values

**Apply at least five different data cleansing concepts we’ve covered. For full marks, you must document your entire cleansing process. Maintain a log file recording your actions and reasoning.**

1. REMOVING THE EXTRA COLUMNS:

* Remove
  + Object Number,
  + Period, Dynasty,
  + Reign, Portfolio,
  + Artist Role,
  + Artist Prefix,
  + Artist Display Name,
  + Artist Display Bio,
  + Artist Suffix,
  + Artist Begin Date
  + Artist End Date
  + Object Date
  + Dimensions
  + City
  + State
  + County
  + Region
  + Subregion
  + Locale
  + Locus
  + Excavation
  + River
  + Classification
  + Rights and Reproduction
  + Link Resource
  + Metadata Data
  + Repository
  + Tags

Columns, since they are not mandatory fields, and the values don’t make sense ultimately.

1. CHECKING VALID DATA:

Removed the special characters and made the content relevant to the field. Also, fill the blank space with the most related field values.

1. MAKING OBJECT ID UNIQUE:

Remove the duplicate values in the object ID field.

Credit: The data used for this assessment belongs to the Metropolitan Museum of Art.