PROG 8460

Assignment 1
Due Date: Feb 11th, 2021

Problem

You have been contracted by a local museum, the Waterloo Museum of Art, to help implement a Business Intelligence project. Your initial task will be to pre-process a dataset containing their collected works of art and then to visualize insights into their data.

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 Number	X
Object ID	X
Department	X
Object Name	X
Title	X
Culture	
Period	
Dynasty	
Reign	
Portfolio	
Artist Role	
Artist Prefix	
Artist Display Name	
Artist Suffix	
Artist Nationality	
Object Begin Date	X
Object End Date	X
Medium	X
Dimensions	
Credit Line	
Geography	
City	
State	
Country	
Region	
Subregion	
Locale	
Locus	
Excavation	
River	
Classification	
Rights and Reproduction	
Link Resource	X
Metadata Date	
Repository	
Tags	

- Any field that is not explicitly listed above can be considered irrelevant in this context.
- The header line must be updated to reflect any changes made to the fields.
- Any field listed above that is designated "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.
- If you are unable to correct a data record, then you are permitted to remove it.

For full marks, you must document your entire cleansing process. Maintain a log file recording your actions and reasoning. Be sure to include any records that you have manually removed. **[9 marks]**

Short Answer

Once your dataset has been pre-processed, answer the following questions.

1. How many data records are in your final dataset?   **[1 mark]**
2. How many data records did you manually remove during the cleansing process?   **[1 mark]**
3. How many countries of origin are contained in your final dataset? List them.   **[2 marks]**

Task 2 - BI and Data related Questions [17 marks]

1. What are the 3 Vs of Data? **[1 marks]**

2. List capabilities of Business Intelligence systems. **[5 marks]**

3. Different types of data with example for each type. **[3 marks]**

4. Define data visualization. **[2 marks]**

5. What is BI system? **[1 marks]**

6. What are the 5 C's of Data for data preparation and the purpose of each? **[5 marks]**

Submissions

In your submission, include the following items:

- the cleaned dataset in CSV format,
- a log file containing your notes or Python script/notebook,
- a document with your answers to the short questions listed above.

Archive your work in ZIP or RAR format and submit the archive to eConestoga under Assignments section. Multiple submissions are allowed before the deadline. If multiple submissions are made before the deadline, I will grade the last one to be submitted.

You are responsible for ensuring that the submission can be extracted and that the archive contains the correct documents. A good way to check is to extract your work from the archive in a different location on your computer and making sure that everything still works. There will be no credit given if I cannot see your work. If you must resubmit to correct any errors, the standard late submission deductions will still apply.

Credit: The data used for this assessment belongs to the Metropolitan Museum of Art. I have modified parts of the dataset and, therefore, it should not be used as a catalog for their work.