**Data Description:** The Metropolitan Museum of Art presents over 5,000 years of art from around the world for everyone to experience and enjoy. The Museum lives in three iconic sites in New York City—The Met Fifth Avenue, The Met Breuer, and The Met Cloisters. Millions of people also take part in The Met experience online.

Since it was founded in 1870, The Met has always aspired to be more than a treasury of rare and beautiful objects. Every day, art comes alive in the Museum's galleries and through its exhibitions and events, revealing both new ideas and unexpected connections across time and across cultures.

The Metropolitan Museum of Art provides select datasets of information on more than 470,000 artworks in its Collection for unrestricted commercial and noncommercial use.

**Critical Details:**

i. Included with these directions should be a .csv file (**MetObjects\_American.csv**) that consists of only a small subset of objects (~18k) in the museum. You should use this file as a basis for all instructions that follow.

ii. You must use either a .ipynb notebook with separate cells per problem or a .py file with separate functions per problem in your submission.

iii. You should maintain the modified data-frame from one problem to the next (ex: the resulting data-frame from #3 should be used in #4, etc.)  **However, it is critically important that your intermediate results (i.e. between each problem) are available and easily accessible during grading.**

iv. With the exception of problem 1 (which is trivial) you should include a few comments in your code that make it clear what your thought process and/or code does to address each problem.

1. Load the .csv file into a pandas data-frame (DF) with appropriate rows and columns.

2. Many columns of this data are missing *entirely* (i.e. no entries are present), because this subset of the data has no pieces with relevant information. Use Python to determine which columns have data that are entirely missing, and remove them from the data-frame.  
Hint 1: remember that ***drop*** is a simple way to remove a column or columns, but requires specific syntax.  
Hint 2: Any column that consists entirely of missing data will be treated as “numeric” in construction.

3. Find all columns in the data-frame that consist of a single, uniform value, and remove it or them from the DF.

4. The column *Artist Role* indicates the relationship of the entity cited in *Artist Display Name* to the object. Suppose we want to populate this column with integers succinctly summarizing each role as follows:  
**Artist** or **Maker** or **Manufacturer** should be replaced by 1.  
**Designer** or **Decorator** or **Modeler** should be replaced by 2.  
**Publisher** or **Patron** or **Retailer** should be replaced by 3.  
All Others or Missing should be replaced by 4.  
Assuming entries that include multiple roles should use the **smallest** relevant value, reformat the Artist Role column so that each entry contains only a single value from the above (1 to 4).

5. Locate at least 3 columns in your modified data-frame that have *some* missing data. For each of these:   
i. Compute the proportion of the column that is missing as a percentage.  
ii. For the missing values, substitute the most common value (if more than one are equally as frequent, choose one arbitrarily). In a sentence or two, explain whether the replacements are reasonable or not for each column.

6. Produce simple summary statistics (mean, quartile information, etc.) in tabular format, as well as histogram representations of all numeric columns (*if you follow the approach we have taken to this process in class, this should be trivial*). Can you glean anything meaningful or notable about the data from these depictions? Explain why or why not in about 1 paragraph (using comments)