Museum Collection Analysis

Yan Gao & Tianxin Deng

Introduction

Dataset - Metropolitan Museum of Art

- Open-access dataset been waived all copyright for data mining
- Rich amount of detailed artworks' information

Motivation

- This dataset has lots of data cleaning problems, which is suitable for practicing data cleaning.
- Most of the data are string rather than numeric type, which could contain some NLP and data extracting works.
- We are interested in the distribution of the artworks by different groups.

Attributes

- Object Name
- Is Highlight
- Is Public Domain
- Department
- Credit Line
- Etc.

	Is_Public_Domain	Department	Object_Name	Credit_Line
448325	False	Photographs	Photograph	Gift of Weston J. Naef, 1974
275017	False	Drawings and Prints	Baseball card, photograph	The Jefferson R. Burdick Collection, Gift of J...
91385	False	Costume Institute	Blouse	Brooklyn Museum Costume Collection at The Metr...
308586	False	Islamic Art	Coin weight	Gift of Joseph W. Drexel, 1889
29353	True	Arms and Armor	Armet	Bashford Dean Memorial Collection, Funds from ...

Research Questions

Distribution of departments

Input: Department column

Output: Visualize the departments that are most worth visiting.

Distribution of artworks' types

Input: Object_Name column

Output: Visualize the top amount of types in collection.

Distribution of Credit lines by timeline

Input: Credit_Line column

Output: Visualize the timeline of amount of new coming artworks per year.

Distribution of Artist roles

Input: Artist_Role column

Output: Visualize the most popular artist roles in the collection

* Some works we already implemented, like checking the typos in artists' names, checking the typos in artist begin date and artist end date, are not included.

Data Cleaning, analyzing and visualizing

The total artwork amount is 492450.

Missing values in each column:

Object_Number	0
Is_Highlight	0
Is_Public_Domain	0
Object_ID	0
Department	0
Object_Name	4393
Title	31069
Culture	283517
Period	402975
Dynasty	469164
Reign	481245
Portfolio	470439
Artist_Role	208640
Artist_Prefix	394396
Artist_Display_Name	206550
Artist_Display_Bio	255362
Artist_Suffix	480785
Artist_Alpha_Sort	206585
Artist_Nationality	299004
Artist_Begin_Date	252956
Artist_End_Date	255734
Object_Date	15034

Object_Begin_Date	0
Object_End_Date	0
Medium	7655
Dimensions	76600
Credit_Line	722
Geography_Type	432029
City	460280
State	489533
County	483884
Country	415578
Region	460473
Subregion	470284
Locale	476885
Locus	485118
Excavation	476482
River	490352
Classification	56481
Rights_and_Reproduction	467731
Link_Resource	0
Metadata_Date	0
Repository	0
Tags	213750

Columns with fewer missing values:

- more likely to get general results.
- Analyzed with prior

Columns with considerable missing values

- dropped as less contribution
- filled by existing contents

Distribution of departments

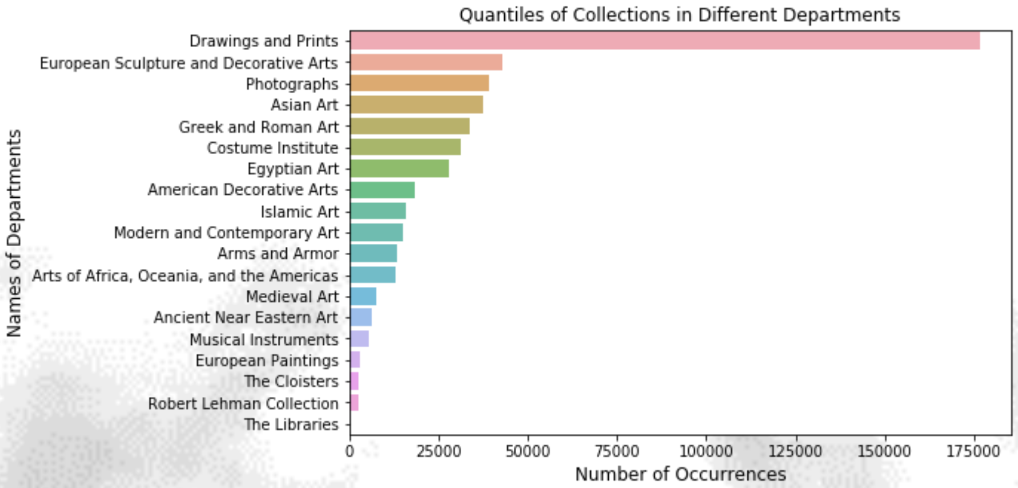
Cleaning: None. This columns is well formatted.

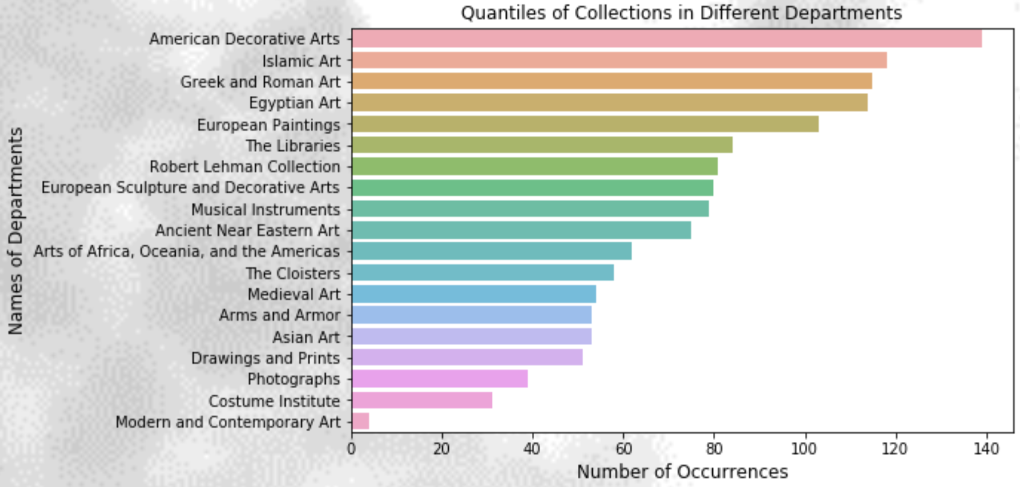
Analyzing:

1. We want to see the value counts
2. The amount of artworks in each department does not indicate which is worth visiting. We want to find the department with most highlighted artworks which are on display

Results: American Decorative Arts department is the most worth visiting

Visualizing:





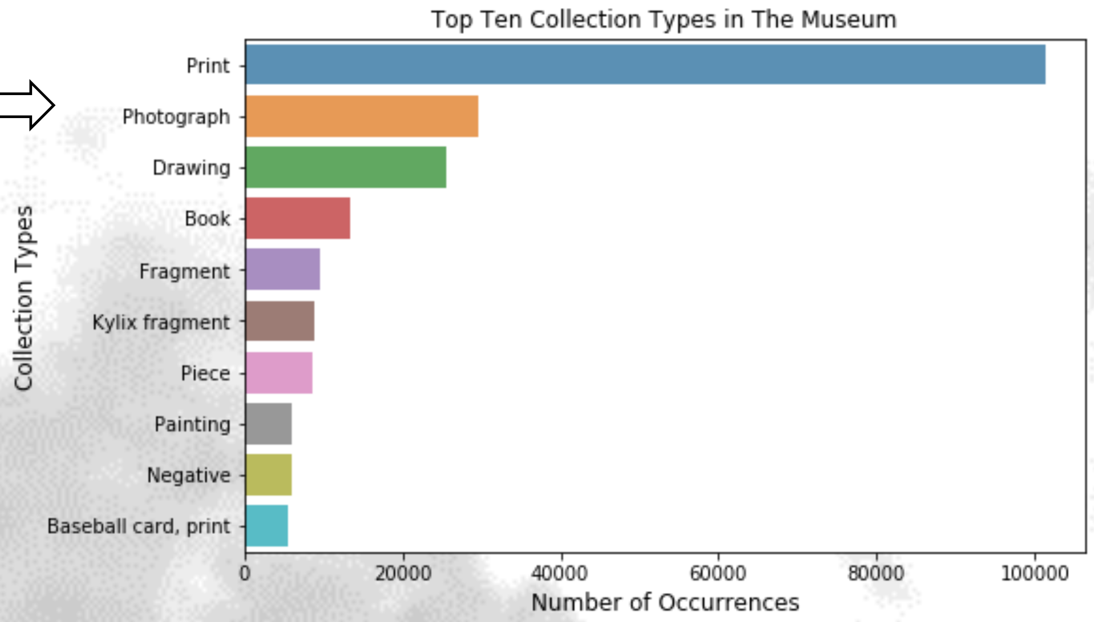
Distribution of artworks’ types

Cleaning: The names of types are quite random defined. So we tokenize the string content, detect and remove the words which are not noun by NLP.

Analyzing: The value counts of types of all the artworks.

Results: Print is the top one. No surprise.

Visualizing:



Distribution of Credit lines by timeline

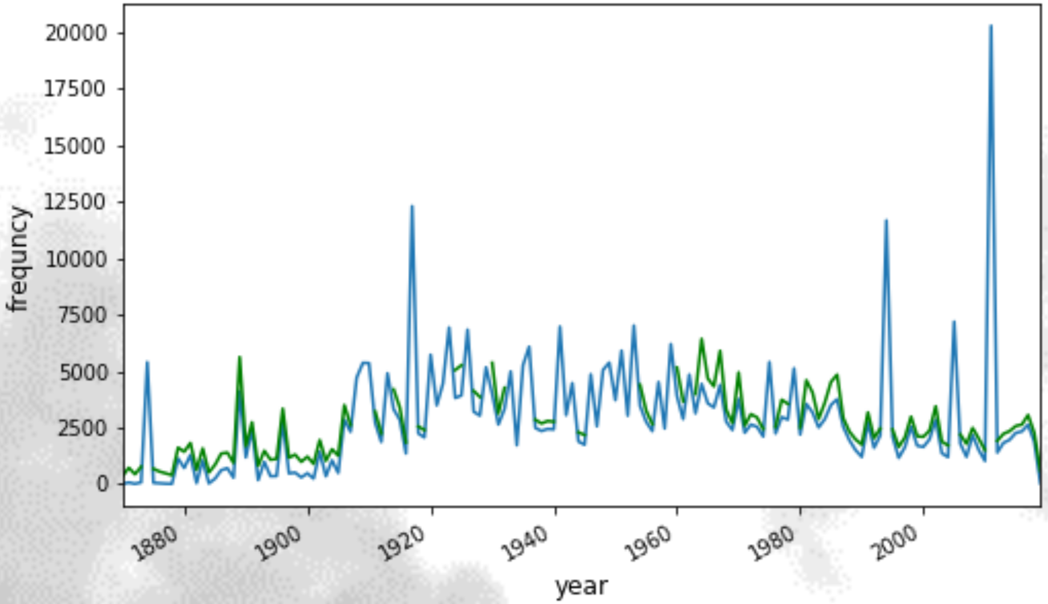
Cleaning: To build the timeline, the year information in content should be extracted.

1. Use re module to get the year.
2. Remove those abnormal data which exceed the history of the museum.
3. Change the data type to Datatime

Analyzing: The value counts of the all the new coming artworks per year. Both actual amount (blue) and the mean of 30 years (green)

Visualizing:

Credit_Line
Gift of Weston J. Naef, 1974
The Jefferson R. Burdick Collection, Gift of J...
Brooklyn Museum Costume Collection at The Metr...
Gift of Joseph W. Drexel, 1889
Bashford Dean Memorial Collection, Funds from ...



Results: We could pair each abnormal peak with the events in the history to explore the truth and track the resources of some artworks which lack information extremely.

Distribution of Artist roles

Cleaning: Some artworks could be contributed by several artist separated by “|”. Some contents are not noun.

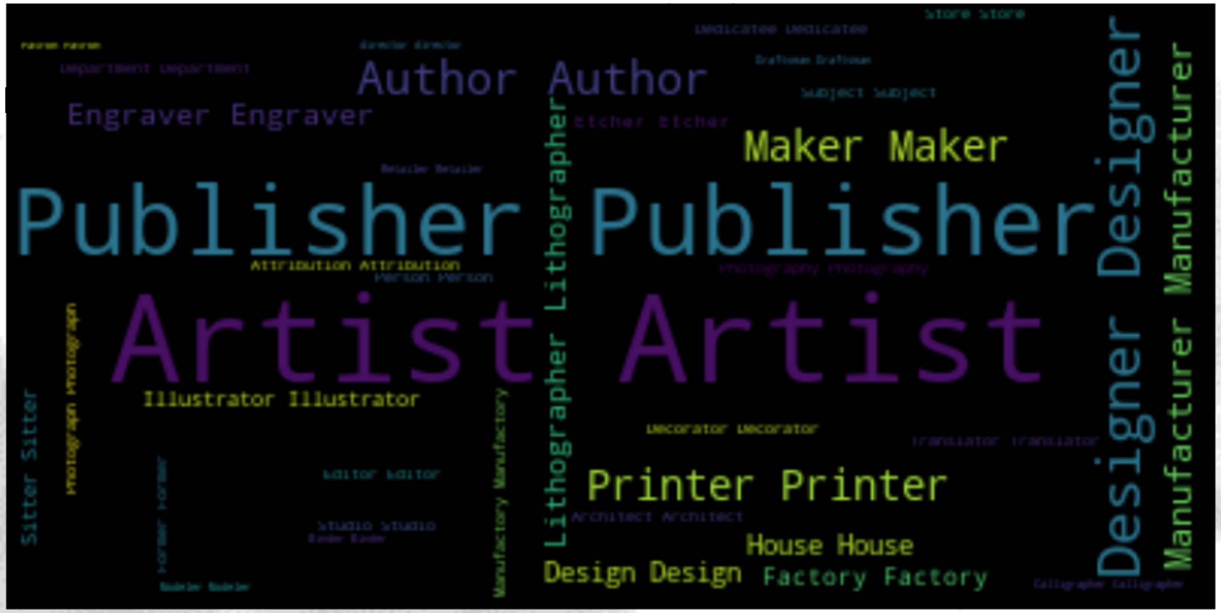
1. Split the content with “|”
2. Tokenize the name of roles and convert them to noun by NLP

Analyzing: We visualize the top ten artists roles

Results: If you are interested in becoming an artist, these should be your right choice.

Visualizing:

	Artist_Role
283688	Publisher
295836	Artist
318102	NaN
208688	Artist
291485	Publisher|Artist



Horrible Data Cleaning Experience

Compare artists' names: The Artist_Display_Name and Artist_Alpha_Sort attributes should have the same content. So we want to find a way to check if there is any spelling mistake of the names.

Artist_Display_Name	Artist_Alpha_Sort
Thomas Ryder I|Johann Heinrich Ramberg|John & ...	Ryder, Thomas|Ramberg, Johann Heinrich|Boydell...
Goodwin & Company	Goodwin & Company
Aristide Maillol	Maillol, Aristide
Salomon de Caus|Jan Norton	Caus, Salomon de|Norton, Jan
Anonymous, Italian, 18th century	Anonymous, Italian, 18th century

Tools:

- from nameparser.parser import HumanName
- from nltk.tag.stanford import NERTagger

"Memorable" Experience:

- Separate name which is mixed with some titles.
- Confirm whether is the alpha sort.
- Compare the full name with nick names or abbreviations
- Foreign letters.

Horrible Data Cleaning Experience

Optimal solution: Set a quantile as confidence level. Then find the proportion of how many common words (not letters) does the two columns have. If it exceed the CL, we considered there is no difference, which indicates no typo.

	Artist_Name_Check	Artist_Display_Name	Artist_Alpha_Sort
337079	False	Troels Wörsel	Worsel, Troels
341661	False	Eugéne Zak	Zak, Eugene
112040	False	Charles Girard	Girard Chales
337912	False	Dulce María Nuñez	Nunez, Dulce Maria
341319	False	Axel Brüel	Bruel, Axel
341654	False	Jacques Thénevet	Thenevet, Jacques
240792	False	Christoph Unterberger	Unterperger, Cristoforo
336630	False	Léna Bergner	Bergner, Lena
246043	False	Bernece Berkman-Hunter	Berkman, Bernese
72109	False	Jean Dessès	Desses, Jean

Notepad

Research Questions:
Mainly focused on the different distributions of data

Data cleaning method:
Mainly implemented with NLP.

Potential of Results:
Museum visiting routine suggestion. Artist roles suggestion. Museum history analysis. Artworks source tracking and filling missing document. Check typo for the museum dataset keeper.

Further to explore:
A better way to identify the names of same person. Solution of how to compare the full name with nick names or abbreviations with less cost of computing power.

Thanks for Watching