Group 15: Information Visualization UE 2024W

Code: `https://github.com/sueszli/artvis/`

Godun Alina, 01569197 Lin Tingyu, 12334199 Jabary Yahya, 11912007

Contents

In this project we visualize the ArtVis dataset[1] derived from the Database of Modern Exhibitions (DoME) from the University of Vienna. This dataset contains information about approximately 14,000 modern painters, their exhibitions and the paintings they exhibited between between the years 1905 and 1915. It is provided in a structured CSV format and accessible in this project's public GitHub page in addition to documentation, code and a reproducible environment.

Exercise 1

In the first exercise we conceptualize an interactive information visualization for the ArtVis dataset.

Data Characterization

The ArtVis dataset contains rich information about modern art exhibitions and artists in early 20th century Europe. The data is structured as a relational table connecting artists to their exhibitions through shared identifiers, with each row representing a unique artist-exhibition pairing.

The dataset is of spatiotemporal and multivariate nature and includes biographical information about artists such as their full names, gender, birth and death dates, birthplace, deathplace, and nationality, all stored as a mix of text and standardized date fields.

- `a.id`: Discrete numerical (unique identifier)
- `a.firstname`, `a.lastname`: Nominal (text)
- `a.gender`: Binary categorical (M/F)
- `a.birthdate`, `a.deathdate`: Temporal (YYYY-MM-DD format)
- `a.birthplace`, `a.deathplace`: Nominal (city names)
- `a.nationality`: Nominal (country codes)

Exhibition details are captured through multiple parameters including the exhibition title, venue name, start date, type (group or solo), number of paintings displayed, and precise geographic location using both city names and coordinates. The temporal coverage spans from 1905 to 1915, with dates stored in a standardized YYYY-MM-DD format for artist lifespans and YYYY format for exhibition dates. The geographic scope primarily encompasses European cities, with exhibition titles appearing in multiple languages including English, German, Russian, and French, reflecting the international nature of the modern art scene. Some venue locations are marked as "exact location unknown," indicating gaps in the historical record.

- `e.id`: Discrete numerical (unique identifier)
- `e.title`: Nominal (text)
- `e.venue`: Nominal (text)
- `e.startdate`: Temporal (YYYY format)
- `e.type`: Nominal categorical (e.g., "group", "solo")
- `e.paintings`: Discrete numerical (count)
- `e.country`: Nominal (country codes)
- `e.city`: Nominal (text)
- `e.latitude`, `e.longitude`: Continuous numerical (geographic coordinates)

An interesting characteristic of the dataset is the presence of duplicate exhibition entries with different venues but the same exhibition ID, suggesting traveling exhibitions or shows that took place simultaneously in multiple locations.

Preprocessing

During the data preprocessing phase, a significant number of records had to be omitted due to formatting issues in the original CSV file. The preprocessing script identified 10,472 invalid lines out of a total of 72,078 records, representing approximately 14.5% of the dataset. The main issue stemmed from unquoted delimiters within text fields, particularly in location entries such as "US, Pittsburgh" which caused incorrect splitting of the data rows. The cleaning process enforced strict data validation rules for various fields and also handled quotation mark standardization and delimiter issues where possible, but complex cases involving embedded commas in unquoted fields could not be automatically resolved. These problematic records, which would have required manual inspection and correction, were excluded from the final cleaned dataset to maintain data integrity. The resulting cleaned dataset contains 61,606 valid exhibition records, providing a reliable foundation for subsequent analysis while acknowledging the trade-off between data completeness and quality.

...

```
invalid in line 72072: expected 19 but got 20
 0: a.id -> 13997
 1: a.firstname -> Louis David
 2: a.lastname -> Vaillant
 3: a.gender -> M
 4: a.birthdate -> 1875-01-01
 5: a.deathdate -> 1944-01-01
 6: a.birthplace -> Cleveland
```

[1] Bartosch C., Mulloli N., Burckhardt D., Döhring M., Ahmad W., Rosenberg R.: The database of modern exhibitions (DoME): European paintings and drawings 1905-1915. Routledge, 2020, ch. 30, pp. 423–434.

```
 7: a.deathplace -> Ohio
 8: a.nationality -> null
 9: e.id -> US
10: e.title -> 296
11: e.venue -> Fourteenth Annual Exhibition
12: e.startdate -> Carnegie Institute
13: e.type -> 1910
14: e.paintings -> group
15: e.country -> 1
16: e.city -> US
17: e.latitude -> Pittsburgh
18: e.longitude -> 40.4333
19: null -> -79.9833


total: 61606, invalid: 10472
```

User & Domain Characterization

The data structure allows for multiple types of analysis, including network analysis through artist co-exhibition relationships, temporal patterns in exhibition frequency and spatial distribution of modern art activities. The dataset exhibits both hierarchical aspects in the organization of exhibitions and venues, and network characteristics in the connections between artists through shared exhibitions.

Tasks and Goals

Visual Queries

Exercise 2