Group 15: Information Visualization UE

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.

The data structure allows for multiple types of analysis, including (1) **network analysis** through artist co-exhibition relationships, (2) **temporal patterns** in exhibition frequency and (3) **spatial distribution** of modern art activities. The dataset additionally exhibits both hierarchical aspects in the organization of exhibitions and venues and network characteristics in the connections between artists through shared exhibitions.

User Characterization

The ArtVis dataset primarily serves art historians, museum curators and academic researchers studying early 20th-century modern art movements. These users typically hold advanced degrees in art history, museum studies, or related fields and are familiar with academic research methodologies:

- *Academic Researchers*: These users possess extensive knowledge of art historical methods and are comfortable with complex data visualizations. They regularly publish in peer-reviewed journals and need to support their arguments with empirical evidence. They are familiar with network analysis tools and geographic information systems.
- *Museum Professionals*: Curators and exhibition designers use this data to inform their understanding of historical exhibition practices. They have practical experience in exhibition organization and are particularly interested in understanding historical presentation contexts. They are comfortable with timeline-based visualizations and collection management systems.

[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.

- *Digital Humanities Scholars*: These users combine computational methods with traditional humanities research. They are experienced in working with structured data and visualization tools, particularly those focusing on temporal and spatial analysis. They regularly use data visualization software and are familiar with common visualization techniques.

In short, the users are domain experts with a need for a dense and detailed visualization that allows them to explore the dataset in depth and aren't easily cognitively overloaded by complex visualizations.

Domain Characterization

The application domain is cultural heritage analytics, specifically focusing on exhibition history and artist networks in European modernism between 1905-1915. This period marks a crucial transition in art history, characterized by the emergence of avant-garde movements and changing exhibition practices. Domain specifics include the multilingual nature of exhibition titles (appearing in English, German, Russian and other European languages), the geographic distribution across multiple European cultural centers and the complex network of relationships between artists, venues and exhibition organizers.

A nice touch for our domain could be the use of serif fonts and a color palette inspired by the modernist art movements of the early 20th century, such as the Bauhaus school or the Russian avant-garde. This would help to create a visual connection between the data and the historical context it represents. The visualization could also be designed to accommodate the multilingual nature of the dataset, allowing users to switch between different languages for labels and annotations.

User Tasks/Goals

The primary tasks and goals of users interacting with the ArtVis dataset include:

- *Historical Network Analysis*:

  Users need to analyze and visualize the connections between artists through shared exhibitions. This includes identifying key figures in artistic movements, understanding collaboration patterns and mapping the spread of artistic influences across Europe. For example, tracking how artists like Kandinsky participated in multiple exhibition groups across different cities.

- *Geographic Movement Patterns*:

  Researchers aim to understand the spatial distribution of modern art exhibitions across Europe. This involves mapping exhibition locations, analyzing artists' travel patterns and identifying major cultural centers. The data shows significant activity in cities like Berlin, Paris and Moscow, with varying degrees of international exchange.

- *Exhibition Practice Evolution*:

  Users seek to understand how exhibition formats and venues changed during this period. This includes analyzing the frequency of group versus solo shows, the emergence of new exhibition spaces and the relationship between traditional and avant-garde venues. The data reveals a predominance of group exhibitions and the importance of certain galleries like Der Sturm in Berlin.

- *Temporal Development Analysis*:

  Researchers need to track changes in exhibition practices over the decade 1905-1915. This includes identifying peak exhibition years, understanding seasonal patterns and analyzing how political events affected exhibition schedules. The data shows varying levels of activity across different years and seasons.

- *Institutional Network Mapping*:

  Users want to understand the relationships between different exhibition venues and organizations. This includes analyzing how certain galleries specialized in particular artistic movements and how institutional networks facilitated the spread of modern art. The data reveals important roles played by venues like the Berliner Secession and various artist-run spaces.

- *Cultural Exchange Documentation*:

  Researchers aim to document and analyze international cultural exchange patterns. This includes tracking how artists exhibited across national boundaries and how different artistic movements spread throughout Europe. The data shows significant cross-border activity, particularly among avant-garde artists.

Concept Design 1: Artist Tracker

The "Artist Tracker" visualization concept presents an innovative approach to exploring the geographic and temporal patterns.

Core Visualization Components The central element is a geographic time-series visualization combining a map of Europe with temporal tracking of artists' movements. The design emphasizes the spatial relationships between exhibition venues while maintaining temporal context through animated paths and a timeline interface.

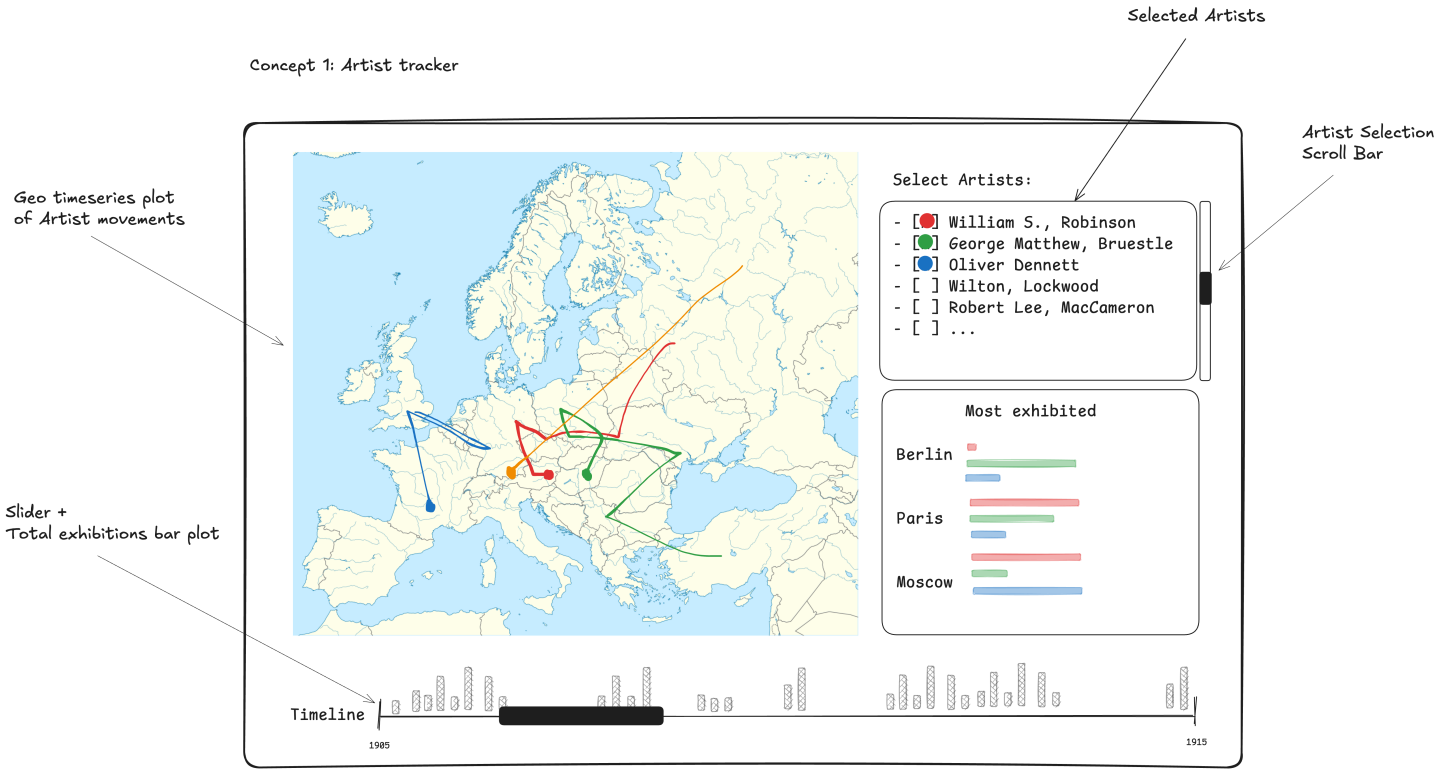


Figure 1: Artist Tracker

The main map employs a light, minimalist base design that clearly shows country boundaries and major cities without overwhelming the movement data. Artist trajectories are represented as colored paths connecting exhibition locations chronologically, with each artist assigned a distinct color to maintain visual distinction.

Visual Encodings and Mappings The temporal dimension is encoded through multiple complementary methods. The primary encoding appears in the path animations showing artists' movements across Europe, while a secondary encoding exists in the timeline bar chart at the bottom showing exhibition frequency over time. This dual representation allows users to understand both the spatial and temporal patterns simultaneously.

Exhibition locations are encoded as points on the map, with their size potentially varying based on the number of exhibitions held at each venue. The color-coding system assigns unique colors to each selected artist, maintaining consistency across all visualization components including the paths, timeline highlights and the selection interface.

Interactive Features and User Support The interface supports several key interaction methods: (1) The artist selection panel allows users to choose which artists to track, with a scrollable list showing all available artists. Selected artists are indicated with colored checkboxes matching their trajectory colors. (2) A timeline slider at the bottom serves dual purposes - it shows the total exhibition activity through bar heights while allowing users to filter the time period of interest. Users can drag the slider to focus on specific years or months. (3) The "Most exhibited" section provides a dynamic bar chart showing the relative exhibition frequencies in major cities, updating based on the selected artists and time period. This helps users identify important cultural centers for different artists or movements.

Design Specifics and Rationale The visualization prioritizes clarity in showing movement patterns while maintaining access to detailed information. The paths use curved lines rather than straight connections to better represent travel routes and reduce visual clutter when multiple paths cross.

The timeline component uses a consistent height scale to show exhibition frequency, with subtle gridlines helping users align temporal patterns. The bars are segmented by artist colors when multiple selected artists exhibited in the same period.

The city comparison bars use a horizontal layout to accommodate city names clearly while showing relative exhibition frequencies through length. The bars maintain the artist color-coding scheme to show which artists exhibited most frequently in each location.

Potential Improvements Several enhancements could strengthen the visualization:

- Adding filters for exhibition types (solo vs. group shows) could reveal different patterns in artists' career trajectories.
- Implementing a brushing and linking feature between the map and timeline would allow users to select specific time periods or locations and see related information highlighted across all components.

- Including additional metadata about exhibitions (such as number of works shown or exhibition duration) could be encoded through variations in path thickness or point size. This could also be provided on hover in addition to color encoding to avoid visual clutter.
- The interface could benefit from a search function in the artist selection panel to help users quickly find specific artists of interest.
- Adding the ability to compare different time periods side by side could help users identify changes in exhibition patterns over time.

This design successfully balances the complexity of the underlying data with accessibility and usability, though there remains room for additional features that could provide even deeper insights into the patterns of early modern art exhibitions.

Concept Design 2: Artist-Exhibition Network

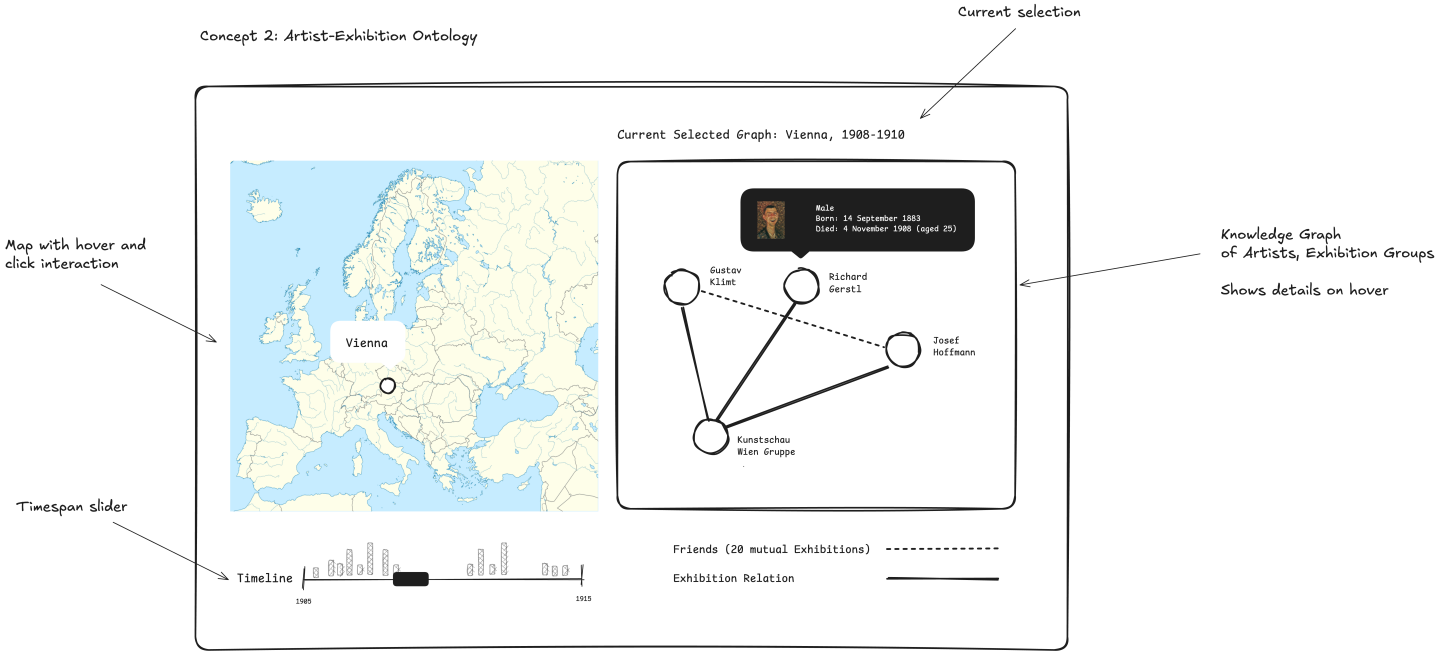


Figure 2: Artist-Exhibition Network

The "Artist-Exhibition Knowledge Graph / Ontology" concept sketch presents a visualization system designed to explore and analyze the complex network of artists and exhibitions.

Core Visualization Components The visualization integrates three main components working in concert: a geographic map of Europe, a network graph visualization and a temporal timeline. This combination allows users to explore the spatial, relational and temporal aspects of the exhibition data simultaneously.

The geographic map serves as the primary spatial reference, displaying exhibition locations across Europe. The map employs a minimalist design with subtle country boundaries and focuses on the relevant geographic region. This design choice helps users maintain spatial context while minimizing visual clutter.

The network visualization represents the relationships between artists and exhibition groups. The visualization uses a force-directed layout where nodes represent either artists or exhibition groups and edges represent participation relationships. The decision to use different edge styles (solid lines for exhibition relationships and dashed lines for frequent collaborations) helps users distinguish between different types of artistic connections.

The timeline component at the bottom provides temporal context and filtering capabilities. It displays exhibition frequency through bar heights, with a draggable selection window that allows users to focus on specific time periods. This design enables users to observe temporal patterns while maintaining the broader historical context.

Visual Encodings and Mappings The system employs several careful visual encoding decisions: The network visualization uses circles for nodes, with consistent sizing to maintain visual clarity. The choice to keep node sizes uniform rather than mapping them to variables like exhibition count helps maintain readability in the force-directed layout.

Edge thickness in the network could represent the number of shared exhibitions, though this isn't explicitly shown in the mockup. The distinction between solid and dashed lines for different relationship types provides clear categorical separation without requiring color differentiation.

The hover functionality reveals detailed information about artists, including birth dates, death dates and age information. This information is displayed in a dark overlay card with white text, ensuring good contrast and readability while maintaining visual hierarchy.

Interaction Design The system implements multiple levels of user interaction:

- *Geographic Navigation*: Users can hover and click on locations in the map to focus on specific cities or regions. This interaction likely filters or highlights relevant connections in the network visualization.
- *Temporal Selection*: The timeline includes a draggable selection window that allows users to focus on specific time periods. This selection presumably updates both the map and network visualizations to show only exhibitions within the selected timeframe.
- *Network Exploration*: Users can interact with the network visualization through hover interactions that reveal detailed information about artists and their relationships. The current selection shows details about an artist named "Male" with specific birth and death dates.
- *Coordinated Views*: All three visualization components are linked, meaning interactions in one component affect the others. For example, selecting a time period in the timeline likely updates both the map and network displays.

Design Specifics and Rationale The layout places the map and network visualization side by side, with the timeline spanning the full width below. This arrangement allows users to easily compare geographic and network patterns while maintaining temporal context.

The choice of a light background with dark features ensures good contrast and readability. The minimal use of color helps avoid visual overload, allowing the structure of the relationships to take precedence.

The current selection indicator in the title ("Current Selected Graph: Vienna, 1908-1910") provides clear context about the visualization's current state. This helps users maintain awareness of their current focus within the broader dataset.

Potential Improvements Several enhancements could further strengthen the visualization:

- The system could incorporate additional filtering capabilities based on artist nationality, exhibition type, or artistic movement. This would help users explore specific aspects of the artistic network.
- The network visualization could benefit from community detection algorithms to highlight clusters of closely connected artists and exhibition groups.
- The map component could use sized markers to indicate the number of exhibitions in each location, providing immediate visual feedback about important artistic centers.
- The timeline could incorporate additional visual encodings, such as color-coding for different types of exhibitions or overlays showing political events that might have influenced artistic activities.

These improvements would enhance the system's analytical capabilities while maintaining its current clarity and usability.

While the proposed visualization concepts differ in their focus and design, both aim to provide users with a powerful tool for exploring the rich ArtVis dataset. By combining spatial, relational and temporal perspectives, the visualizations offer a comprehensive view of early 20th-century European art history. The designs prioritize clarity, interactivity and user control, ensuring that domain experts can extract valuable insights from the data while maintaining a high level of engagement.

Exercise 2

In this exercise, we will focus on implementing a specific visualization technique to not only gain hands-on experience with visualization tools but also to critically assess how effective these techniques are within our particular data domain.

Preprocessing

The first step in any information visualization project is data preprocessing. This phase involves cleaning, transforming and structuring the raw data to make it suitable for visualization.

During the data preprocessing phase, a significant number of records had to be omitted due to syntactical 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 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
```

```
 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
```

The preprocessing begins by loading the raw data and defining paths for input and output files. Non-ASCII characters and newline symbols are removed from the header and missing values represented by `\\N` are replaced with "null". For each line, commas within fields are enclosed in quotation marks to maintain data consistency. Lines that don't match the expected column count are dropped. Basic validation ensures that numeric fields, such as `a.id` and `e.paintings`, contain only positive integers; similarly, `a.gender` is constrained to "M" or "F" and `e.type` to "group," "solo," or "auction." Geolocation data for `e.latitude` and `e.longitude` is validated to fall within acceptable ranges and invalid values are set to "null". For dates, `a.birthdate`, `a.deathdate` and `e.startdate` are formatted and checked against specific historical ranges and lifespan validity is verified. If latitude and longitude are valid, city and country data could be inferred (though this part was omitted due to time constraints; retrieving 70k items from the geo API would take over 6 hours). Finally, the cleaned data is written to the output file. Then further processing with pandas refines data types and handles missing values. Null values in integer columns, such as `a.id`, `e.id` and `e.startdate`, are removed or filled with default values. String columns like `a.firstname`, `a.lastname` and `a.gender` are filled with appropriate placeholder values and categorical data in `a.gender` and `e.type` is standardized with additional "Unknown" categories for missing entries. Date fields are converted to datetime format and invalid dates are coerced to "null." Additionally, geographic fields, `e.latitude` and `e.longitude`, are converted to floats and set to 0 if missing. Finally, a summary is printed, detailing each column's data type, unique values and missing value count.

Data Exploration

The next step in the visualization process is data exploration. This phase involves understanding the dataset's structure, identifying patterns and relationships and formulating hypotheses for further analysis.

To select the right visualization library, we evaluated several options based on their functionality, ease of use, popularity and flexibility[2] [3]. After examining resources like Northwestern University's research computing guide and a comparative notebook on interactive Python plotting packages, we narrowed down our choices. Some tools, like `Streamlit` and `Pygal`, were ruled out due to limitations such as dependency on cloud services, poor HTML export capabilities, or restricted usability outside of notebooks. Ultimately, `Plotly` and `Altair` emerged as top contenders because of their strong interactivity, ease of coding and customization potential. `Plotly`, paired with `Dash`, offers high-quality 3D graphics and a vast library, while `Altair`'s concise syntax allows for rapid and intuitive visualizations, making both highly effective for this assignment's goals. For rapid prototyping and exploration, we also used `Matplotlib` and `Seaborn` as well as the R-`ggplot2` library based library `Plotnine` for Python.

Several visualizations were created to uncover patterns and insights. Five of our many exploratory visualizations are shown below.

- *Geographic Distribution of Exhibitions*: The scatter plot of geographic distribution shows that most exhibitions were concentrated in Europe, particularly around central longitudes and latitudes. This suggests a focus on European art scenes during the early 20th century, with fewer exhibitions occurring in North America and Asia.
- *Distribution of Artist Lifespans*: The histogram of artist lifespans reveals that most artists lived between 60 and 80 years. This distribution indicates a relatively normal lifespan for artists during this period, with a few outliers living beyond 100 years. This needs to be further investigated to understand the reasons behind these outliers.
- *Number of Exhibitions Over Time*: The line chart depicting the number of exhibitions over time highlights a significant increase around 1904, followed by fluctuations. This suggests a growing interest in modern art exhibitions during this decade, possibly influenced by cultural movements or societal changes.
- *Exhibition Activity Across Major Cities*: The bubble chart illustrates exhibition activity across major cities from 1913 to 1926. Paris and Munich appear as major hubs for exhibitions, indicating their importance in the modern art world. Other cities like Vienna, London and New York also show significant activity, reflecting their roles in the international art scene.

Overall, these visualizations provide insights into the geographic focus, artist demographics, temporal trends and city-specific activities, allowing for a deeper understanding of the ArtVis dataset. The next step involves selecting the most appropriate visualization technique for further analysis and storytelling.

[2]Northwestern University Research Computing - https://sites.northwestern.edu/researchcomputing/2022/02/03/what-is-the-best-interactive-plotting-package-in-python/

[3]Comparative Notebook on Interactive Python Plotting Packages - https://github.com/ageller/comparePythonInteractives/blob/main/compareInteractive

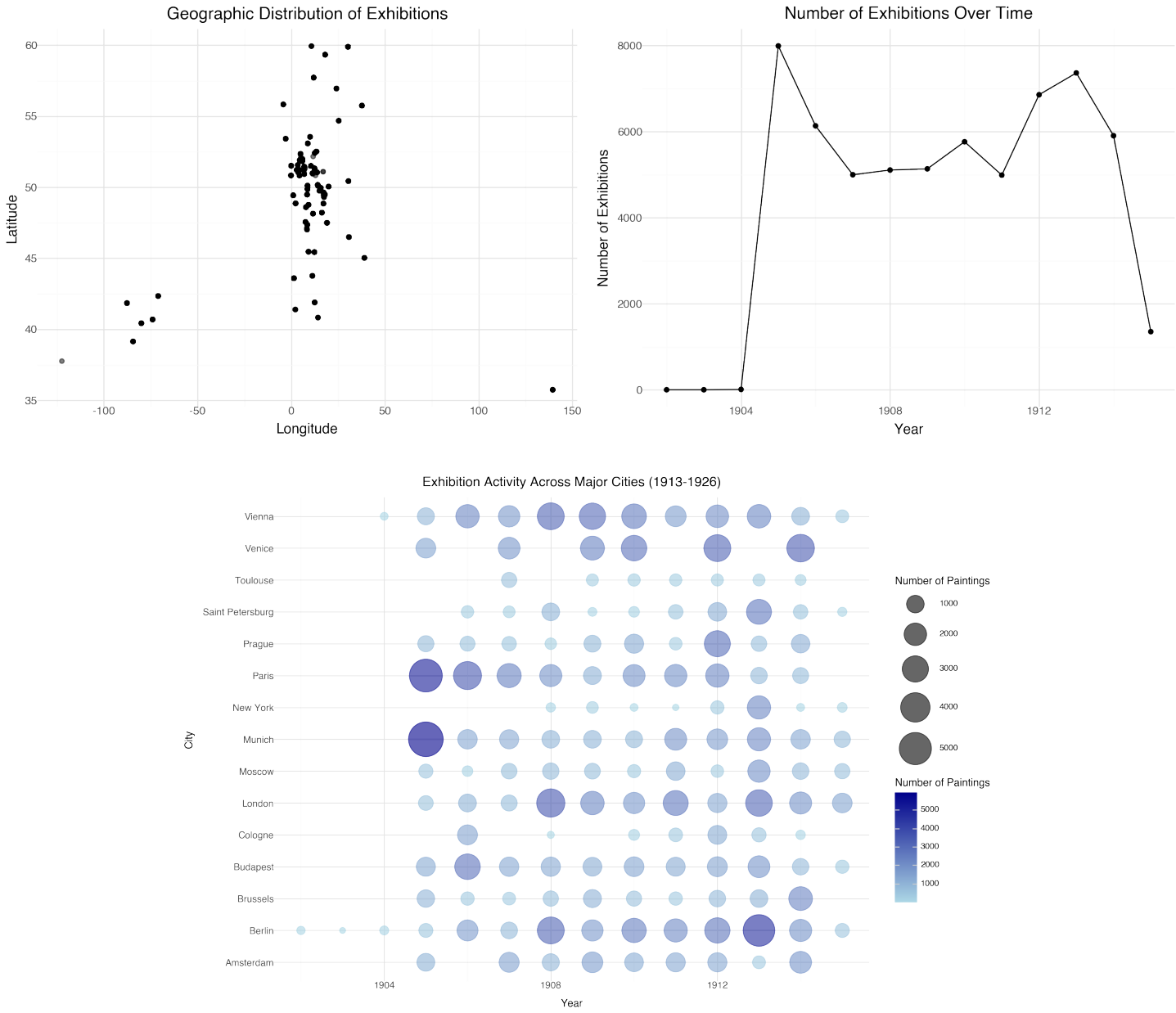
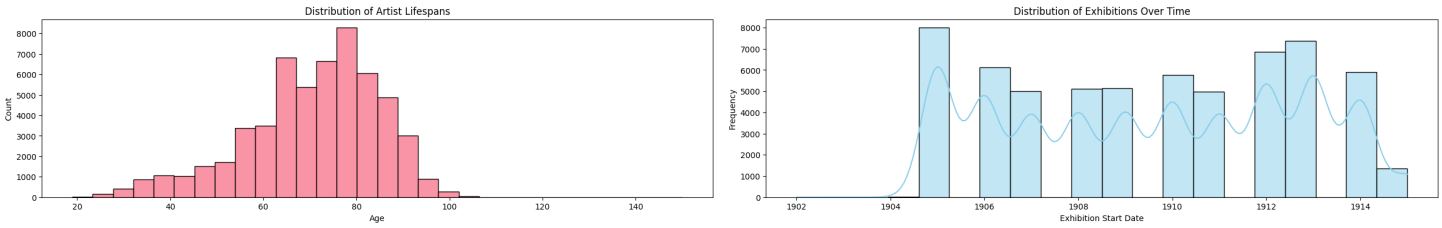


Figure 3: Exhibition Activity Across Major Cities



Visualization 1: Geospacial Querying

The culmination of our efforts are 2 interactive interactive visualizations, each designed to highlight different aspects of the dataset while ensuring a cohesive user experience.

Insights At the heart of the first visualization lies a world map that serves as a spatial reference point for exhibition locations. This map is complemented by a series of interactive charts that provide insights into various dimensions of the data, including temporal trends and comparative analyses across countries and venues. By employing brushing and linking techniques, users can select specific regions on the map to filter data in real-time across all visual components. This interactivity not only enhances user engagement but also facilitates deeper exploration of the dataset.

The map itself is designed with clarity in mind, utilizing light gray fill colors for countries and distinct circles for exhibition locations. The size of these circles corresponds to the number of exhibitions held at each venue, providing an immediate visual cue regarding the density of artistic activity in different regions. Tooltips are integrated into the map, offering detailed

information about each exhibition location when users hover over the circles. This feature enriches the user experience by providing context without cluttering the visual space.

In addition to geographic distribution, our visualization includes several charts that delve into temporal trends and comparative metrics. The line chart depicting exhibition counts over time reveals significant fluctuations in exhibition activity throughout the decade. This timeline not only allows users to observe peaks in artistic activity but also invites them to hypothesize about external factors that may have influenced these trends, such as political events or cultural movements.

Moreover, we implemented bar charts that compare exhibition frequencies across countries and venues. These charts serve as an effective means to identify which nations were most active in hosting exhibitions and which venues became pivotal in shaping modern art discourse during this period. The ability to brush selections on these charts further enhances interactivity, allowing users to focus their analysis on specific countries or venues while observing corresponding changes in other visual components.

User Engagement and Usability Throughout the development process, we prioritized user engagement and usability. Our target audience—art historians, museum curators, and academic researchers—requires a visualization that balances complexity with accessibility. By utilizing Altair's concise syntax and robust features, we were able to create an intuitive interface that caters to both novice users and experienced researchers alike.

Feedback from initial testing sessions ($n = 5$ colleagues) indicated that users appreciated the straightforward navigation and interactive elements of the visualization. They found value in being able to dynamically filter data based on their interests, whether they were examining specific artists' contributions or exploring geographic trends in exhibition practices. This level of interactivity empowers users to derive personalized insights from the dataset, fostering a deeper understanding of early 20th-century modern art.

Critical Assessment of Visualization Techniques Reflecting on our choice of visualization techniques, it is evident that employing linked views significantly enhances data exploration capabilities. The brushing technique allows for coordinated views across multiple visualizations, enabling users to draw connections between disparate data points seamlessly. This approach aligns well with our goal of facilitating historical network analysis and geographic movement patterns among artists. However, it is important to acknowledge certain limitations inherent in our design. While we aimed for clarity and interactivity, some users expressed a desire for additional contextual information or metadata about specific exhibitions or artists directly within the visualizations. Future iterations could benefit from incorporating more detailed annotations or expandable panels that provide richer narratives without overwhelming users with information.

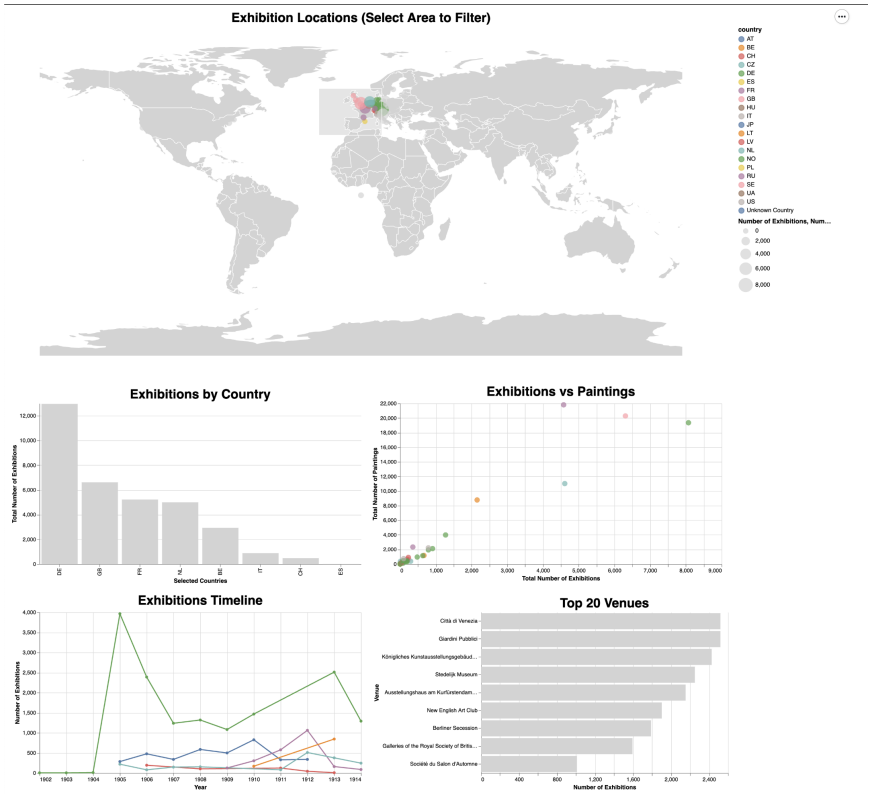


Figure 4: Visualization 1: Geospacial Querying

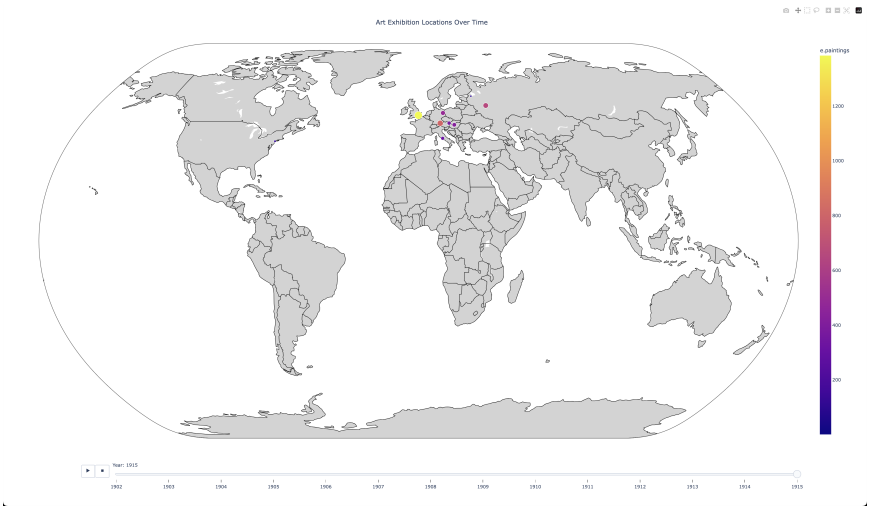


Figure 5: Visualization 2: Temporal Querying

Visualization 2: Temporal Querying

Conclusion This project was a step towards enhanced cultural heritage analytics. It has successfully demonstrated how effective visualization techniques can transform complex datasets into engaging narratives that resonate with diverse audiences. Our interactive linked and brushed visualizations serve not only as a tool for analysis but also as a gateway for exploring cultural heritage through the lens of modern art exhibitions. As we deployed our final product on GitHub Pages, we opened up opportunities for further exploration by other researchers and enthusiasts interested in early 20th-century art history.