Projects
Visual Analytics

Summer semester 2025
Project 2

This project will need to be uploaded in ILIAS before **27.06.2025** at **18:00 o'clock**.
The presentation will be done during your tutorial in the week of **02.07.2025**.

- Exercises will be done in groups of 3-5 students. **Every** group member needs to be present at the presentation.
- All students should know the reasoning behind design decisions and are able to articulate this during the presentation.
- Additional libraries/dependencies/frameworks are not permitted.
- 25 out of 40 points must be obtained over the 2 exercises in order to be allowed to take the exam.
- At least one person per group needs to upload the presentation and the code used to generate the visualizations in Ilias before 18:00 27.06.2025. Failure to upload will result in 0 points for this exercise.

Task 1: Using dirty data

In the previous assignment, you worked with a relatively small and cleaned dataset. In this assignment, you will learn how to work with larger and more realistic data, and apply the techniques learned in the lecture to work with it.

There are two data files to be used for this. First is the *recommendations-2021-12-31.csv* file which contains the top 1000 boardgames as rated by users from boardgamegeek.com from 2021 (Data set from https://gitlab.com/recommend.games/bgg-ranking-historicals by Markus Sheperd). Added to the original dataset are the fans also likeboardgames as provided on www.boardgamegeek.com.

Second is the *bgg_Gameitems.csv* file (Data set recombined from https://www.kaggle.com/datasets/mshepherd/board-games by Markus Shepherd). For all games from 2021 and earlier, this contains similar metadata as the previous project, but now for the full dataset.

Note that these files are large and contain more data than your system will need. Document how you pre-process the data in these files on the server to support your systems. Ensure that your system can handle different top-X (i.e. top 100, top 200, top 1000) selections from this dataset.

For the remainder of the tasks, you can start using the initial dataset from the previous project, before switching towards the cleaned and compressed datasets as a result of this task.

Task 2: Tightly-integrated visualization via clustering via k-means

In this task, you will analyze the cleaned data from Task 1 using k-means to cluster the data in your visualization. A file is provided `kmeans.js` which can be placed in the server folder that implements that k-means algorithm for multidimensional data.

Explore the full dataset, and come up with a 5-tuple task using the Trend target that could be interesting for an analyst and can be supported by k-means in particular. Then design and implement a visualization that supports this analysis task using k-means. For this dataset in particular, attention should be paid to the distance measure used (Including representing categorical/ordered strings as numbers) and the choice of k. Ensure that in the implemented visualization, both k and the distance measure can be interacted with to allow for human control of the clustering algorithm.

Task 3: Subsequent significant object visualization for graph analysis

For this task, you will use the cleaned data from Task 1. An analyst wants to analyse why boardgames are recommended form other board games. In order to do this, they suggest using automatic significant object detection based on the recommendations to find key boardgames in the recommendation network. Using these key-boardgames, they would then be able to compare similarities and differences between the recommenders. Design and implement a visual analytic system that allows the analyst to do this.

For this task assignment, you can use the pagerank algorithm from https://github.com/alixaxel/pagerank.js/. This automatically assigned significance scores to nodes in a graph network. Your design should use both the significance scores and the recommendations as core components. Using your design, the analysts should be able to describe which features have a strong impact on a boardgame being recommended from one of the key-board games (Describe, Locate, Correlate, Significance score + recommendations + input features, All).

Task 4: Interacting with linked visualization.

In this task, you will link all previous visualizations (including those from Project 1) together to view the same data from different perspectives.

Implement highlighting *from* both the cluster visualization and the graph visualization towards all other visualizations. Ensure that the cost of interaction is kept low as possible to perform this highlighting. In your presentation, show an example how linked the visualizations through highlighting can be used to support one of the 5-tuples from earlier analysis tasks.

Task 1 Deliverables

There are two deliverables for this task again, a working version of the code and a presentation. Both have to be uploaded to Ilias **before 18:00 27.06.2025**. Do not upload the `node_modules`-folder and its content.

Code

A working version of the code should be uploaded to Ilias. This includes all files used to generate the visualizations, as well as any files that have been used to preprocess the data and the preprocessed data itself.

Presentation

During the tutorial in the week of 02.07.2025, a presentation session will be held. Each group will present their work in at most 15 minutes, with an additional 3 minutes for questions. A single individual from each group should present their work, while the remainder of the group should be ready to address the questions. All individuals should be present during the presentation.

The presentation should consist of a PowerPoint (or similar) presentation going explaining the design and why design decisions were made, and a small live demo of your implemented visual design. At the end of the presentation, questions will be asked to individual members of the group for further clarification and to test understanding of the proposed design, see below for a list of example questions.

Presentation first slide constraints

The first slide of your presentation should have at least the following details:

- The names of the people in the group.
- The tasks that you came up with.
- The distribution of the work over the members of the group.
- The rough time spent on each part of the work.
- A screenshot of the final visualization.

Example questions

Below, we present examples of the type of questions that you may be asked during the presentation.

1. What did you initialize notice in the data when exploring it?
2. Why did you preprocess the data in this way?
3. What data-cleaning did you do?
4. Can you explain why you have chosen this task to explore?
5. How is this task supported?
6. Why have you chosen for this type of visualization for this task?
7. Why is this basic element (not) included.
8. Did you find any interesting insights using your visualization?
9. What changes did you make to the original visual design?
10. What was the most difficult part to design?
11. Given more time, what would you add?
12. If you did the exercise again, what would you have done differently?
13. What would you have liked to incorporate into your design from the presentations seen so far?
14. Are there questions, that can not be asked with your visualizations, but can be answered with the data itself?
15. What are the limitations of your technique?
16. Ideas how to deal with these limitations?
17. Why did you choose this particular strategy to handle the data issues?
18. How did you handle categorical and ordered data for the k-means algorithm?
19. How does this 5-tuple correspond to the task?

20. How did you preprocess the data for the purpose of the pagerank algorithm?
21. How did you filter the data?
22. How does the highlighting support the analysis?
23. Were there any issues with clustering the data using k-means?

Grading scheme

Below we present the list of criteria that the group will be graded on. Each criteria awards 1 point when adequately addressed.

- ☐ The new dataset is incorporated into the visualizations from Project 1 and Project 2
- ☐ It is clear which data quality issues were present and how they were handled.
- ☐ It is clear how the data preprocessing is done and well argued how this supports the tasks from Project 2.
- ☐ The dataset can be dynamically set to filter to the top-X.
- ☐ A 5-tuple analysis task using the Trend target is designed for Task 2, and it is well argued how it could be interesting for an analyst.
- ☐ A visualisation is implemented that shows the results from the clustering.
- ☐ It is well-argued how the visualization is appropriate for the designed analysis task.
- ☐ Interaction with the parameter k from k-means is implemented and intuitive.
- ☐ The distance metric used for k-means can be adapted dynamically and with low cost of interaction.
- ☐ A visualization has been implemented that uses pagerank to show important boardgames in the recommendation network.
- ☐ It is well argued how the visualization supports describing the correlations.
- ☐ The significance scores and the recommendations are core components in the visualization
- ☐ The visualization supports describing the correlations between a key boardgame and its recomendees.
- ☐ Boardgames can be selected from the visualizations of Task 2 and Task 3 of Project 2 for highlighting in the other visualizations.
- ☐ Highlighting is implemented for the implemented visualizations of Project 1 and Project 2.
- ☐ It is shown how the highlighting can be used to support the analysis for one of the tasks of Project 1 or Project 2.
- ☐ Argumentation for why basic elements are included and excluded are clear and sound.
- ☐ The live demo during the presentation worked and matched the designs presented.
- ☐ Most questions during the presentation were answered to satisfaction.
- ☐ All questions during the presentation were answered to satisfaction.

Note that for an individual to get any points at all, they have to upload a working version of the code and the presentation to Ilias before the deadline.