Visual Analytics project proposal

Alessio Palma
Sapienza University of Rome
1837493

Antonio Andrea Gargiulo
Sapienza University of Rome
1769185

In this document, we illustrate our proposal for the final project of the Visual Analytics course at Sapienza University of Rome, a.y. 2023/24.

1 Dataset

Our dataset comes from the VAST Challenge 2022, which focuses on performing an urban planning analysis for the fictitious city of Engagement in Ohio, USA. The analysis has its main focus on addressing the following problems for the city: finding patterns in the demographics of the city, characterizing the different areas of the city by the different attributes or problems they have, understanding the more prominent and less prominent businesses and the cost of living in the city. This involves a comprehensive assessment of the city's current status and potential areas for future development.

All the data in this experiment is artificially generated, 1011 residents are chosen to represent the city's demographics and are contributing to the dataset, which is made of 18 GBs of data divided into:

- **Attributes** folder: contains 9 csv files, each containing static informations about the main entities present in the city (general context about participants, buildings, apartments, jobs, restaurants, pubs);
- **Activity Logs** folder: contains 72 csv files, logging every 5 minutes the financial, hunger, sleep status and location of each of the participants, for the whole duration of the 15-month data collection period;
- **Journals** folder: contains 4 csv files, which are summaries of the activity logs divided into financial, social, traveling and check-in information.

Since using all 18 GB of data is not possible due to computational reasons, we will mainly use the Attributes folder and some aggregated data from Journals and Activity Logs, generating a dataset with an AS index ≥ 31000.

2 General idea

Our visual analytics application offers a comprehensive exploration of urban dynamics, with a primary emphasis on two main parts: the demographics and economics of participants (Figure 1) and the recreational activities in the city (Figure 2). We want to clarify that the only recreational activities present in the city are restaurants and pubs. The user can easily switch interactively back and forth from the participant view to the activities view using a toggle. The interface allows the user to delve into geographical locations, analyze various trends through histogram visualizations and find new insights from the dimensionality reduction and clustering plot. Moreover, the parallel coordinate plots have the advantage of both range brushing and supporting high-dimensional data. The intended user of the system is a member of the town council who must decide where to perform economic interventions for the renewal of city areas.

3 The Visual Analytics cycle

In this section we go more in-depth discussing the main parts of the Visual Analytics cycle of our system:

- **Analytics**: as a preprocessing step we do some aggregation of data from the logs, computing e.g. the annual turnover of the activities, the total expenses of each participant, the total number of visitors for each recreational activity, the distribution of distance traveled by clients of each activity. The dimensionality reductions are used to ease the subsequent use of clustering algorithms, in order to obtain helpful insights into the groups of participants and also for the recreational activities. We believe that using t-SNE could be better for the *participants* view, but it may not be suitable for the *activities* one; in that case, we will probably use PCA. After the dimensionality reduction, we plan to use K-Means to gain insights into the different clusters that can be obtained. We also want to calculate some real-time statistics (based on current selection in the system) such as the average Engel's coefficient of people, where each person's coefficient is obtained as: Engel's coeff. $= \frac{\text{Total food expenditure}}{\text{Total personal consumption expenses}}$ to express some inner component of the provided data, since it is a good indicator of the standard of living, and explains better patterns and correlations between different data attributes.

- **Visualizations**: in the *participants* view (Figure 1) we have on the left a map showing the residence of each person and the statistics interactively computed for selected people; in the center a column containing some histograms (or other types of plot, we still have to decide if they are a good fit) showing mainly aggregated statistics about spending and jobs during the data collection period; on the right a scatter plot showing the result of dimensionality reduction and clustering; in the bottom part we have a parallel coordinates plot showing mainly static attributes (age, education level, joviality, etc.). The *activities* view (Figure 2) is conceptually similar to the previous one, but now all the plots and the dimensionality reduction refer to restaurants and pubs and their attributes, with the central part now focusing on aggregated statistics about the number of visitors and earnings.

- **Interactions**: the user can brush directly from the map the area or the single entity of interest for which to perform the analytics and the visualizations, restricting the view and having more granular insights via spatial filtering. It is also allowed to perform brushing on the parallel plot to interact with some groups that are not easily highlightable on the map view. From the dimensionality reduction view, the user can click on the legend to select only entities that belong to the specific cluster. The interaction between visualizations is tightly coupled: on the user's selection of a subset of entities in one plot, all the plots change to display only the information for the chosen subset.

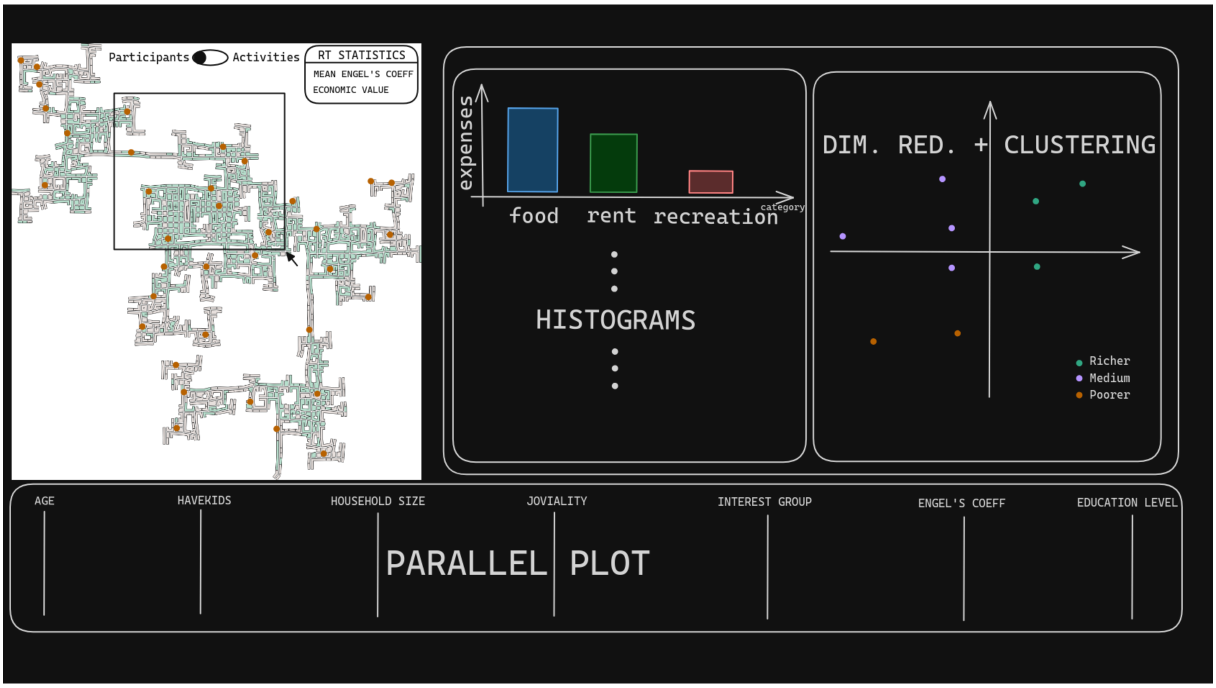

Figure 1: Draft mock-up of the user interface for the pattern analysis of the inhabitants participating in the study.

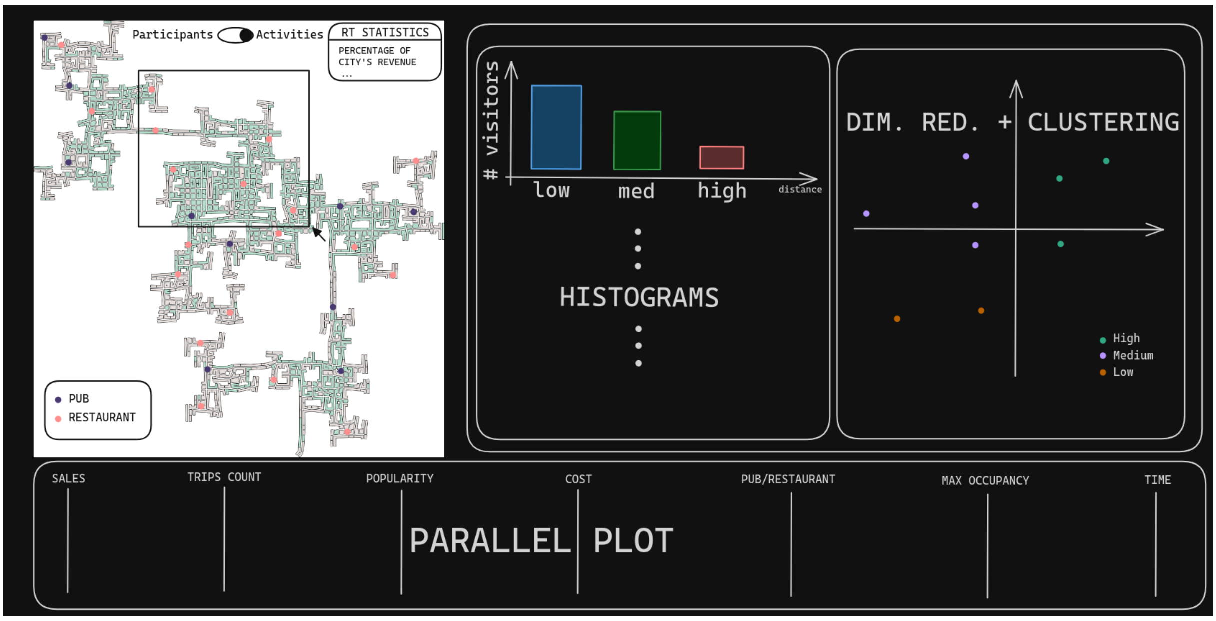

Figure 2: Draft mock-up of the user interface for the pattern analysis of recreational activities present in the city.