SAPIENZA
UNIVERSITÀ DI ROMA

Visual Analytics for Engagement: an Urban Analysis Task

Alessio Palma
Sapienza University of Rome
1837493

Antonio Andrea Gargiulo
Sapienza University of Rome
1769185

Contents

1 Introduction **2**

2 Related Works **2**

3 Dataset and domain **3**
3.1 Data Preprocessing . . . 3

4 Our solution **4**
4.1 Map Plot . . . 4
4.2 Central Area Charts . . . 5
4.3 Clustering and PCA . . . 6
4.4 Parallel Coordinates Plot . . . 7

5 Insights **7**
5.1 Where is it better to live? . . . 7
5.2 Which recreational activities to invest in? . . . 7
5.3 Does money buy happiness? . . . 8

6 Conclusions **8**

1 Introduction

This report illustrates our final project for the Visual Analytics course at Sapienza University of Rome, a.y. 2023/24. Our project addresses the VAST Challenge 2022, which is about Urban Planning in the fictional town of Engagement, Ohio (USA). The city council is looking for the best visual analytics application that can help to decode the patterns of its community, with the final goal of understanding where to allocate a very large city renewal grant they have recently received. As Engagement prepares for future growth, our mission is to harness the power of data to illuminate pathways toward sustainable development and a better quality of life for its residents. Our solution 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. The **intended user** of the system is a member of the town council who must decide where to perform economic interventions to improve the city. Our solution offers a double view that allows the user to focus the analysis on the demographics and expenses of the participants ("participants view"), or on the earnings and popularity of recreational activities ("activities view") separately on demand.

The code for the application is available here.

2 Related Works

ClusBridges (Fei et al., 2022) focuses on the analysis of traffic data and patterns of life divided into various time intervals (daily, weekly, and monthly) and provides a clustering of individuals showing similar life routines. The system abstracts and summarizes large volumes of travel records, clustering also the city into 15 regions based on the number of visits and average travel time in minutes needed to reach that region. Inspired by ancient Chinese bridges, ClusBridges encodes traffic volume, speed, and direction using bridge-like arcs displayed on the map. Although intuitive, this encoding introduces a lot of cluttering and this solution only provides an analysis of the city's traffic and people's daily routines, ignoring the economics and demographics of the city.

Comprehending City Economics from Heterogeneous Data (Splechtna et al., 2022) focuses only on the economic side of the city, developing an interactive application that allows for the analysis of participants and activities separately. Similarly to our solution, for the participants, they show the wage and the expenses for each category, while for each activity they show the distance the guests have traveled to reach that location. The application lacks the possibility of spatial brushing on the map, which can be very useful for location-centric analysis. Moreover, participants' demographics were not taken into account, and it is not possible to analyze pubs and restaurants together, which could be useful because they show similar patterns.

EconomicVis (Liu et al., 2022) designed a Visual Analytics tool that can display a large amount of data on demand, mainly the financial health of the city and the daily life patterns of people. The system consists of 4 plots, the map plot provides a spatial perception of the city's economic situation, the bipartite view shows participants' employment relationships and turnovers, the calendar heatmap displays time series attributes at different granularities and the parallel coordinates plot allows brushing on both non-time series and time series attributes. The console allows for entity switching, clustering and choosing the time series to visualize. The system effectively allows to analyze a lot of data but lacks visual efficiency (i.e. if we select for example the Pubs as the entity to analyze, we will still have on screen the bipartite view regarding participants' employments, which is not needed at that moment and can be misleading) and brushing on the map.

Visual Analytics for Demographics, Social Networks and Business Base Pattern (Li et al., 2022) focuses on the demographic characteristics of the city, social relationship patterns and main business area patterns. Along with the demographic of each participant, they cluster the activities' locations into 5 groups to identify the main business areas of the city, and for each area, they show the number of workers and customers. Specific temporal intervals can be selected from a calendar and the system also includes a node-link diagram to visualize the social network of the city. This solution does not allow the user to define filters on demographic attributes freely but only allows them to apply predefined filters.

Visual Analytics for Urban Data Analysis

(Burmeister et al., 2022) presents a view with the map of the city and 4 dashboards. The demography dashboard summarizes the population composition with bar and donut charts, coupled with financial status visualization. The social network dashboard shows a node-link diagram, similar to (Li et al., 2022). The business dashboard summarizes the job market of the city, showing histograms about employers, jobs and employed participants. The gastronomy dashboard describes the health of pubs and restaurants with monthly visit histograms, sparklines for venue visit trends and map depiction of venue connections with guest's residences. Although highly interactive, this system offers map brushing only in the business dashboard and omits earnings data in the gastronomy dashboard.

Analyzing demographics and patterns-of-life using SAS Visual Analytics (Falko et al., 2022) involved a detailed analysis using the SAS Viya tool, which is a SAS-developed automated Visual Analytics tool that does not require programming skills. This solution presents many plots about demographics, traffic data, job distributions, participants' expenses, daily life routines, visits and earnings from recreational activities. Even if comprehensive, it lacks multiple coordinated views because interacting with a plot will cause no effect on the other ones, since each plot lives in its own window on the SAS Viya platform.

3 Dataset and domain

As we anticipated, our dataset originally comes from the VAST Challenge 2022, which asks to perform an urban planning analysis for the fictitious city of Engagement in Ohio. All the data for this task has been artificially generated by the authors, 1011 residents are "chosen" to represent the city's demographics and all of their actions and information have been recorded for 15 months to form the dataset, which is made of 18 GBs of data divided into:

- **Attributes** folder: contains 9 csv files, each containing static information 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 GBs of data is not possible due to computational reasons, we performed some preprocessing to extract the data useful for our purposes, mainly using the Attributes and Journals folders.

3.1 Data Preprocessing

Performing some analytics through Python, we aggregated information spread among various files and computed some derived data starting from the dataset available from the challenge. Our final dataset has an **AS index** of 63386 and is divided into 5 csv files:

- **ParticipantsAugmented**: contains static information about the inhabitants involved in the data collection and is used for all the plots in the participants view. Here we also cleared out people that dropped out of the experiment after one month, having a total of 880 active participants remaining. Some attributes here were aggregated from the Journals (e.g. total expenses for each category for each person) or derived, like the Engel's coefficient which is an indicator of the standard of living and is computed for each person as:

$$\text{Engel's} = \frac{\text{Total food expenses}}{\text{Total personal expenses}};$$

- **BuildingsAugmented**: contains the GeoJSON formatted information about the shape of buildings in the city, each building is a polygon, and drawing them in an SVG enables one to obtain the map with all the information about the type of building depicted (Residential, Commercial or School);
- **ActivitiesAugmented**: contains static information about the recreational activities (Pubs and Restaurants) in the city, used to draw the parallel coordinates plot, the dimensionality reduction plot and the points on the map in the activities view. The total turnover and number of visitors for each activity during the 15 months were aggregated from the Journals;

- **MonthlyLog**: contains the monthly earnings for each activity for each month of the study period. This information was extracted from the Journals and is used to draw the line plot in the activities view;
- **VisitsLog**: contains the count of visits of each participant to each activity, along with the Euclidean distance between the participant's residence and the activity location. This information was extracted from the Journals and is used to draw the stacked bar chart in the activities view;

4 Our solution

With its dual views on demographics and the economic side of the city, the application presents a detailed analysis of the urban life of Engagement. These two perspectives offer invaluable insights into the inner dynamics shaping this city. The **participants** view offers a panoramic into the demographic landscape, capturing the diverse composition of citizens, and shedding light on various factors such as age, income, education, happiness and family members. Conversely, the **activities** view offers a description of leisure and sociality trends for citizens, providing an economic picture of what people love to do. By analyzing patterns of recreational activities, eating preferences and entertainment choices a full picture of the city's economic vitality and consumer behavior emerges. On a high level, the two views share common characteristics, yet each offers distinct inner meanings derived from different data, together offering a complete understanding of the life in Engagement.

Our application is developed using Node.js, D3.js and a small Python backend for analytics. It is highly interactive, with the possibility of brushing both on the map plot and the parallel coordinates plot, and the possibility of selecting a cluster to analyze from the dimensionality reduction plot. Each interaction triggers the update of all the other plots in the view, which will show only the information of the entities in the current selection. The updating of plots uses transitions to prevent change blindness.

4.1 Map Plot

On the left side of the screen we have the **map plot** depicted in Figure 1, reporting the geographic information of the city. Here one can easily understand where the educational, residential and commercial

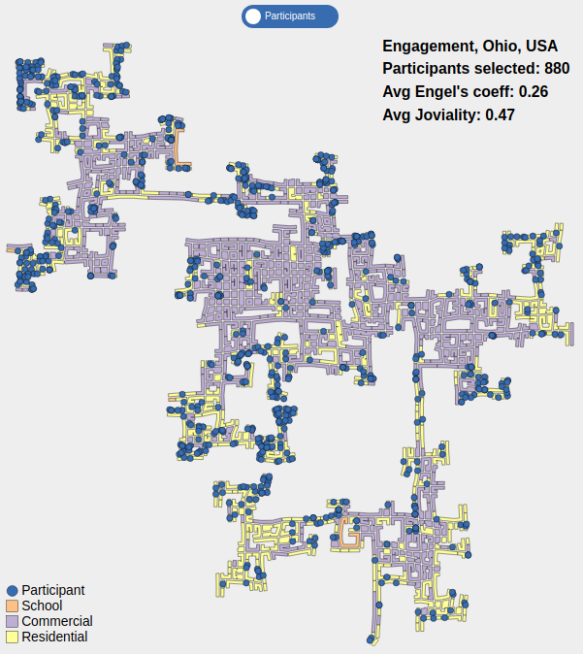

Figure 1: Participants view map.

buildings are located in the city through color encoding of the polygons. Over the map, dots with color encoding are drawn to represent Participants in the participants view and to represent Restaurants and Pubs in the activities view. We used **ColorBrewer2.0** to choose a categorical color scale instead of using the D3.js predefined ones, and the color encoding is consistent through all the charts in the same view. The legend in the lower-left corner of the map shows which color is associated with which category. On the upper right part of the map we display a "*toolbox*" reporting some aggregated statistics about the current selection. In the participants' view the toolbox reports the number of inhabitants in the current selection, the average Engel's coefficient and the average joviality (happiness) of the selection. In the activities view, instead, the toolbox shows the number of businesses in the current selection and the percentage of the city's total revenue to which they contribute.

The map is brushable and, through it, the user can perform various types of analyses that focus only on specific areas of the city, a feature that is missing from almost all the works reported in Section 2. When brushing is performed, the user can see all plots changing in real-time through transitions, displaying only the information regarding the current selection; furthermore, the user can continue the analysis by complementing filters on other plots. When hovering the mouse over the dots, a

tooltip with the entity ID pops up.

4.2 Central Area Charts

In the central part of our web application we have 2 plots for each view, displaying some aggregated data about economic information and giving highly focused insights into specific behaviours.

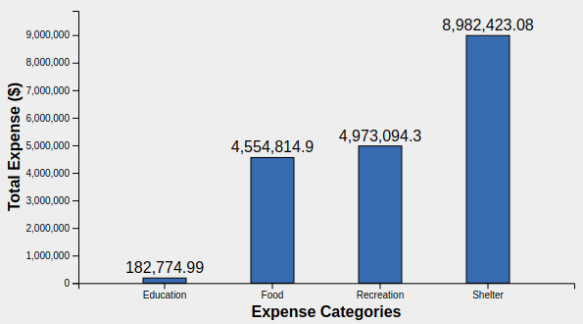

Figure 2: Participants view bar chart, expenses by category.

In the participants view, we have on top the **bar chart** reported in Figure 2, showing the total expenses made by participants in the current selection across the various categories of education, food, recreation and shelter. It gives an overview of the money flow in the city and provides a detailed description of how different groups of people prefer to spend money. This chart gives hints about which city zones privilege some activities instead of others, which neighborhood tends to save more and which tends to spend more, and thus the city council can leverage these information to decide where to invest on what. To ease the understanding, on top of each bar is reported its numerical value. When filtering is performed on the other plots, this chart updates accordingly.

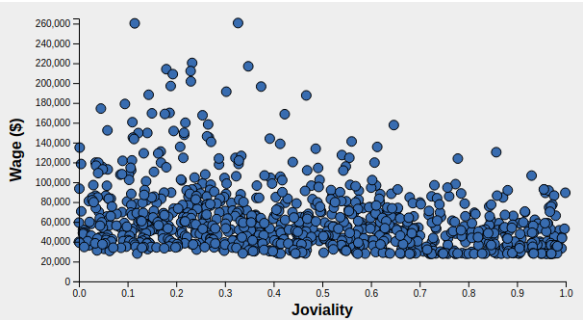

Figure 3: Participants view scatter plot, wage by joviality.

At the bottom, instead, we have the **scatter plot** illustrated in Figure 3, comparing joviality with the total income of each participant during the 15-months period. It provides details on the relationship between joviality and wage that are crucial for understanding the population's well-being. This plot, for the overall city, shows: 1) that Engagement could set a minimum wage of at least 20,000$ per year, and 2) a sort of negative correlation between these two attributes, displaying that the citizens of Engagement rarely seek money and huge salaries do not bring so much happiness here. When filtering is performed on the other plots this chart updates accordingly, and when hovering the mouse over the dots a tooltip with the entity ID pops up.

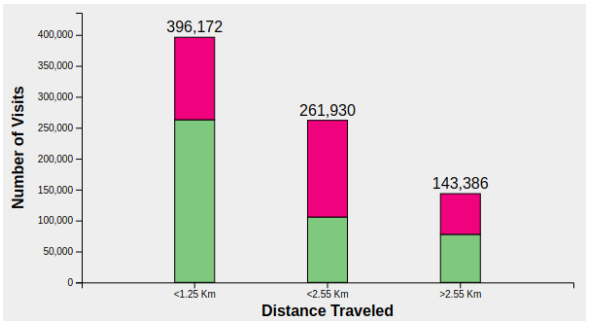

Figure 4: Activities view stacked bar chart, visit frequencies vs distance traveled by customers to reach the business.

Passing now to the central area of the activities view we have on top the **stacked bar chart** shown in Figure 4, depicting the distribution of distances traveled by customers to reach the businesses present in the current selection. The L2 distance is computed for each customer w.r.t. its home location. The fact that it is stacked implies that we can also see the percentage of restaurant and pub visits that contribute to each bar. As one can notice, usually people in Engagement tend to choose recreational activities that are not too far from where they live. So, an activity with more visitors coming from distant locations could be considered a more popular activity, because it is worth traveling a long distance for it. When filtering is performed on the other plots, this chart updates accordingly and so it can be used to determine the most popular areas and the most trending activities. This provides a good indicator of which zones are more appealing for investing in new recreational activities, channeling public resources to where they are most needed.

On the bottom, in the activities view, we have the **multiple lines chart** reported in Figure 5, where

each line represents the monthly income of a recreational activity. From it, the user can easily grasp the time variation in sales in the city and correlate diverse activities based on those trends. Moreover, it shows very clearly the difference between pubs and restaurants in terms of earnings and, from this information, the city council could maybe revise the actual systems of fixed-price meals of restaurants and "pay for each hour of stay" of pubs. Since restaurants make less money than pubs, a first idea may be to just lower pub prices and increase restaurant ones. When filtering is performed on the other plots, this chart updates accordingly.

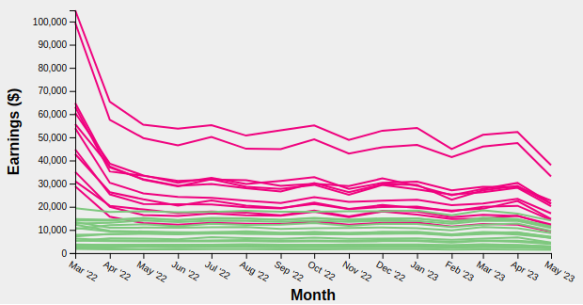

Figure 5: Activities view multiple lines chart, depicting monthly earnings for each business.

4.3 Clustering and PCA

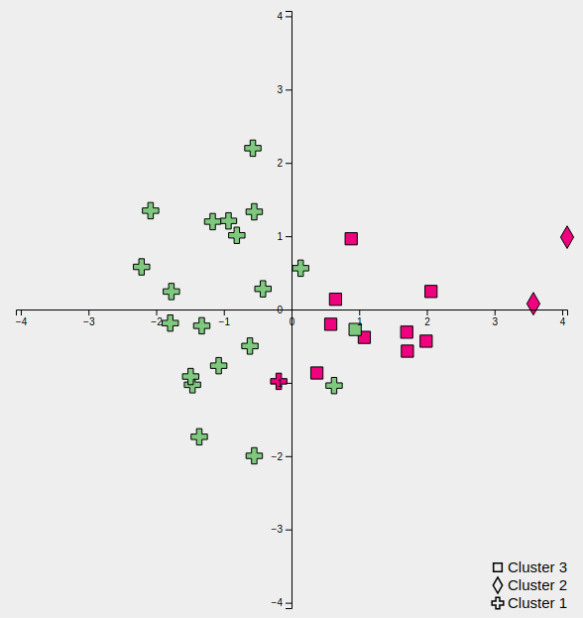

Figure 6: Activities view clustering with K-means and 2D projection using PCA.

On the right side of the web application we have the result of the **analytical** part of our system, which is formed of a real-time clustering computed by a Python back-end, followed by a dimensionality reduction algorithm used to draw a 2D scatter plot of the entities in the current selection. An example can be seen in Figure 6. For the participants view we applied the **Gaussian Mixture Model** (GMM) clustering algorithm, choosing 4 as the number of clusters, on the following attributes: household size, age, education level, joviality and expenses for each of the categories reported in Figure 2. Then **PCA** is applied over the z-score normalized data to reduce the dimensions, and we display the first 2 components of the projected data in our scatter plot. For the activities view we used the same pipeline, replacing the GMM algorithm with a 3-clusters **K-means** over the following attributes: cost, maximum capacity, total number of visits and total earnings. The clustering algorithms were applied in the original space in order to avoid introducing indistinguishable false positives in the visualization, while the number of clusters and the clustering algorithm type were chosen after some empirical attempts. The color encoding of the elements drawn on the scatter plot is consistent with the colors used in all other plots of the same view, whereas the shape of the data point encodes the cluster to which it belongs. A legend is drawn in the lower-right corner of the plot to ease the understanding of the shape encoding. When filtering is performed on the other plots, clusters and PCA will be recomputed and shown in real time. The dimensionality reduction scatter plot is also interactive: the user can click on a point, and the entire cluster to which that point belongs to will be "locked" and highlighted, becoming the current selection of elements in the system. Additional filters can then be added on that cluster through the parallel coordinates plot and the map, but the clusters will not be recomputed. By clicking again on the previously selected cluster, the clusters will be "unlocked". Also, when hovering the mouse over the plotted symbols a tooltip with the entity ID pops up. Having a visualization that clusters data and reduces the complexity of multiple attributes helps to discover similarities that are not well comprehensible at first glance. In this way the user can easily analyze groups of peoples or activities that have common characteristics, making strategic decisions aimed at enhancing not only the city's earnings but also the social and entertainment offerings of Engagement.

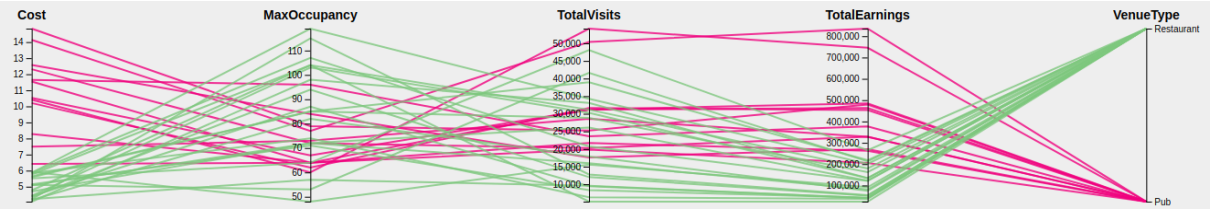

Figure 7: Activities view parallel coordinates plot.

4.4 Parallel Coordinates Plot

On the bottom of the page we have the **parallel coordinates plot** depicted in Figure 7. It allows users to easily visualize and filter many attributes at the same time through parallel axes. This versatile visualization method is well suited for discovering trends, similarities, seeing correlations and unveiling patterns that are crucial for insightful analyses. Although performing all the previous actions is difficult when the parallel plot shows all the data together, the use of brushing is very convenient to set filters on the data we want to see. In this way, users can perform very complicated and detailed filtering with low effort, mitigating the issue of over-plotting that often plagues this visualization. In the participants view, the parallel plot shows for each person: age, joviality, household size, interest group, education level, if one has kids and the Engel's coefficient. In the activities view, instead, it shows for each activity: maximum capacity, venue type, total visits received during the study period, total sales made during the study period and its cost. The cost is the only attribute that carries different meanings for restaurants and pubs: for the former ones it represents the fixed-price menu cost of meals, whereas for the latter ones it denotes the cost that customers has to pay for each hour spent in the pub (due to the so called "blue law").

5 Insights

5.1 Where is it better to live?

By brushing on the map, one can study the characteristics of different city areas. Combining it with the mean joviality value displayed in the real-time statistics toolbox, it is easy to see how this number changes through various areas. The user can notice that areas far from the city center have a lower mean joviality value w.r.t. the central area. The overall joviality of the city is 47%, with higher levels (around 60%) in the city center and lower levels (30% to 40%) in suburban areas. A city councilman can think that Engagement is developing suburban areas that are not very satisfying in terms of life conditions, because those are not bringing happiness to people living there, and plan potential investments in these areas.

5.2 Which recreational activities to invest in?

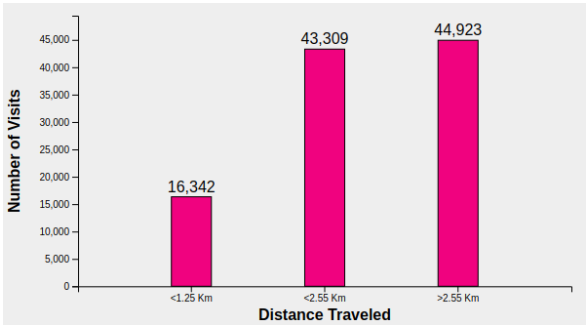

Figure 8: Distribution of visits received by the two most popular pubs.

Thanks to the toolbox present on the map and to the brushable parallel coordinates plot, we can see that we have a total of 32 recreational activities in the city, divided into 12 pubs and 20 restaurants. In general, pubs are the main business area of the city because they bring 69.5% of the total turnover of the activities, while restaurants are less relevant with the remaining 30.5%. By brushing on the map, we can easily discover that the 13 activities in the central area form 52. 4% of the total turnover, which means that the city center is the most profitable area and also the most visited. In particular, pubs with IDs 1342 and 1344 are the two most profitable businesses in the city and also the ones that receive the most visits from customers residing far from them, as we can see in Figure 8. Altough being so popular, pub 1344 is the smaller one and pub 1342 is only the 4th biggest one in terms of maximum customer capacity. Putting ourselves in the shoes of a town administrator, if our goal is to maximize the profits of already popular businesses, we would invest money in expanding pubs 1342 and 1344. If instead we would like to help less profitable businesses it is better to invest in renewing

or promoting the restaurants, maybe also raising their price, because they are in general more visited than pubs but are earning less moneys. Also, investing in suburban activities could bring more people to the peripheral zones and also increase the happiness of people who live there.

5.3 Does money buy happiness?

The Wage vs Joviality scatterplot reported in Figure 9 could reveal a thought-provoking insight: a slight negative correlation between income and happiness. Despite conventional wisdom suggesting that higher wages lead to greater joviality, the plot may challenge this assumption by showcasing that individuals with higher wages exhibit lower levels of happiness and, instead, individuals with lower wages span a higher range of joviality values. Also, by selecting on the map people living in the center of the city (and so with a higher average joviality), we can see that they tend to be not so rich. This observation prompts a reevaluation of the traditional narration linking wealth and happiness, suggesting that money alone does not guarantee happiness, and to find it one must look to other factors like, in Engagement, the position of its residence.

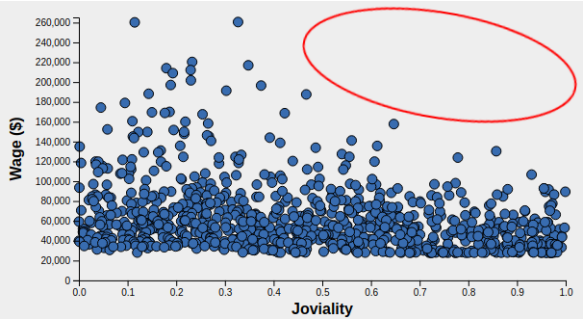

Figure 9: Scatterplot of Wage vs Joviality, we added a red oval in this image to highlight the fact that no people is very happy and very rich at the same time.

6 Conclusions

The application's objective revolved around equipping Engagement's town council members with a useful tool that aids in formulating reasoned hypotheses and well-grounded conclusions. Our system allows to analyze the demographic characteristics of the population and the economic life of both participants and business activities in the city, gaining insights and providing actionable recommendations for the future primary decisions of the city government. Through the application functionalities, the user can effectively verify his theories and gather empirical evidence about which areas or activities need investments to improve the population's quality of life, making the system act as a reliable compass, guiding council members towards impactful interventions for the citizens of Engagement. Overall, this application serves as a cornerstone for evidence-based governance, empowering decision-makers with actionable insights to prioritize investments and initiatives that truly resonate with the needs of Engagement's citizens.

References

Jan Burmeister, Jilin Liao, Jieqing Yang, Qingtian Wei, and Kexin Wang. 2022. Visual Analytics for Urban Data Analysis.

Schulz Falko, Sztukowski Stu, LeSaint Cheryl, Harenberg Steven, Chapman Don, and Mayse Chelsea. 2022. Analyzing demographics and patterns-of-life using SAS Visual Analytics.

William C Fei, Yawen Lu, Hao Wang, Xingyu Jiang, Tianyi Zhang, Zhenyu Qian, and Yingjie Chen. 2022. ClusBridges: Clustering Life Routines and Data.

Yuxiao Li, Xuexi Wang, Yue Wang, Ting Liu, Huiting Wang, Ziyue Lin, and Siming Chen. 2022. Visual Analytics for Demographics, Social Networks and Business Base Pattern.

Ting Liu, Huiting Wang, Ziyue Lin, Yuxiao Li, Siming Chen, Xuexi Wang, and Yue Wang. 2022. EconomicVis: Visual Analytics for Financial Health, Employment and Similar Life Patterns Mining.

Rainer Splechtna, Disha Sardana, Denis Grac ̌anin, Thomas Hulka, Nikitha Donekal Chandrashekar, and Kres ̌imir Matkovic. 2022. Comprehending City Economics from Heterogeneous Data.