VAST Challenge 2022, MC2

Oliver Lundin, Emma Zettervall

Abstract—In the 2022 VAST Challenge, we explored patterns of life within the fictitious city of Engagement, using data from 1000 residents. This report focuses on the second mini-challenge, which involves characterizing city areas, identifying traffic bottlenecks, examining participants' daily routines, and analyzing changes over time. Our analysis pipeline consisted of data cleaning, preprocessing, visualization, and iterative exploration, utilizing Python, Pandas, d3.js, matplotlib, and plotly. We identified distinct commercial and residential areas, potential traffic bottlenecks, and detailed daily routines of selected participants. Financial and social activity patterns revealed insights, such as a financial boom and a burst of social activities. Despite the effective workflow, areas for improvement include developing standardized data-processing scripts and conducting more exploratory data analysis to mitigate bias.

1 Introduction

The 2022 VAST challenge is a about a fictitious city called "Engagement". The city is conducting a participatory urban planning exercise with approximately 1000 representative residents using an app that tracks their visits, spending, and purchases. This data will assist in community revitalization efforts and the allocation of a large city renewal grant. To aid the city a visual analytics expert must make sense of all the data provided by the participants.

The second mini-challenge, which is the topic of this report, is specifically about the patterns of life within the city. The goal is to describe the daily routines for some representative people, characterize travel patterns to identify potential bottlenecks or hazards, and examine how these patterns change over time and seasons.

To aid the research, four questions/tasks are provided to narrow down to focus:

- Assuming the volunteers are representative of the city's population, characterize the distinct areas of the city that you identify.
- Where are the busiest areas in Engagement? Are there traffic bottlenecks that should be addressed?
- Participants have given permission to have their daily routines captured. Choose two different participants with different routines and describe their daily patterns, with supporting evidence.
- Over the span of the dataset, how do patterns change in the city change?

2 Analysis Strategy Outline

The analysis strategy outline can be described as a pipeline in which data is fed into, processed and cleaned, visualized and analyzed and then if necessary goes back to the beginning. This pipeline is used as basis strategy for all the questions/tasks within this mini challenge.

First step is to clean and preprocess the data that might be relevant for the question at hand. Discussions in the group lead to selecting the most interesting data files to look at and extract the necessary information. It was also necessary for some tasks to purely discard uninteresting data to have smoother workflow without large file sizes. After the preprocessing, data is used to visualize aspects that help us answer the given question through analysis. During this process we might make discoveries that lead us to further process more data or perhaps extract more information that we initially discarded. This was our iterative approach to answering the questions.

• *All authors are with Linköping University, Sweden.*

The goal with this approach is to always be able to go back to the data processing, since the dataset was large and the questions sometimes vague, we needed an approach that allowed us to explore the data and go back and forth between analyzing and working with the data. This approach is also inspired by the infovis and visual analytics reference models [1] [2]

3 Data Characterisation and Preparation

The data given spanned over 13 months and was largely made up of logs for each of the participants that contained detailed information about their location, financial balance and more. The data points are logged on a 5 minute basis and the activity logs were split across 72 files. The data contained along with a png map of the city, a folder with different attributes of the city, such as location of apartments/schools/commercial buildings. The largest part of the dataset was a set of journals, that logged financial, location and social status of each participant throughout the 13 months, also on a 5 minute basis.

Most of the data files contained a timestamp column in the format "2023-05-24T18:05:00Z". Which were read as a string unless turned into a date type. Locations were given as POINT type coordinates which gave us some trouble in interpreting them with our chosen tools for visualization. The approach we took was to convert the coordinates to their pure number form. One of the more important pieces of data to answer the questions, was the map. To be able to have an easier workflow with the map, the coordinates of all buildings was turned into a TopoJSON. TopoJSON is a JSON format for encoding geographic data structures into a shared topology. By doing this we could easily visualize individual buildings and classify them.

4 Results

The results are presented question by question.

4.1 Distinct areas

To find the distinct areas in the city we used a manual cluster based approach with python and matplotlib. By plotting the coordinates for all the different building types onto the city map it becomes quite clear what areas are "city centers" with more commercial buildings and what areas are more residential. Another factor that supports this is that most restaurant lie within the commercial areas. To further investigate the different areas we merged the apartment file that contains rental costs with the buildings file and plotted buildinigs tagged with the residential label into TopoJSON map.

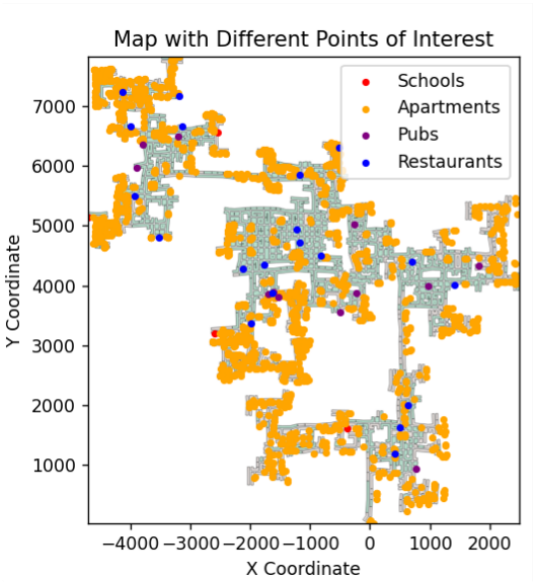


Fig. 1: Using building type to identify distinct areas.

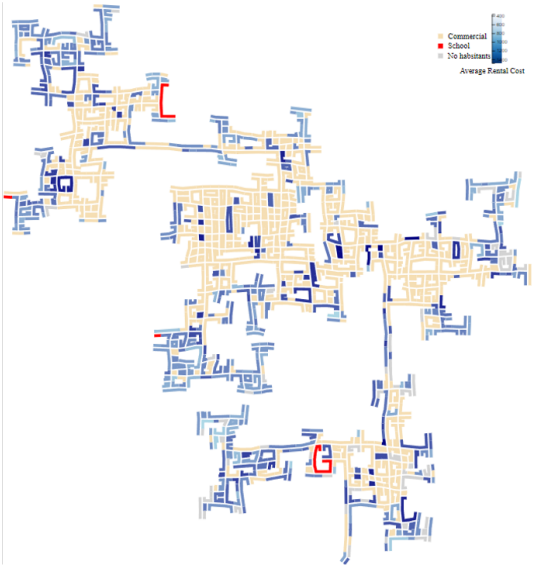


Fig. 2: Map with rental costs and commercial areas.

This map further solidifies the claim that the city centers are commercial areas and also allowed us to identify more expensive areas which seem to be around the commercial areas. From this map we identify 3 distinct areas: Top right, middle and bottom left.

4.2 Traffic

To find the traffic flow in the city we utilized the map from the previous question but this time only using the buildings with tags ”residential” or ”commercial”. We imported the map into a d3.js project and then plotted all datapoints with the label ”transport” from the travel journal and activity logs on a week day that we had previously combined into one data file using python. By then plotting lines between each data point in their respective order, a traffic flow can be seen in figure **??**

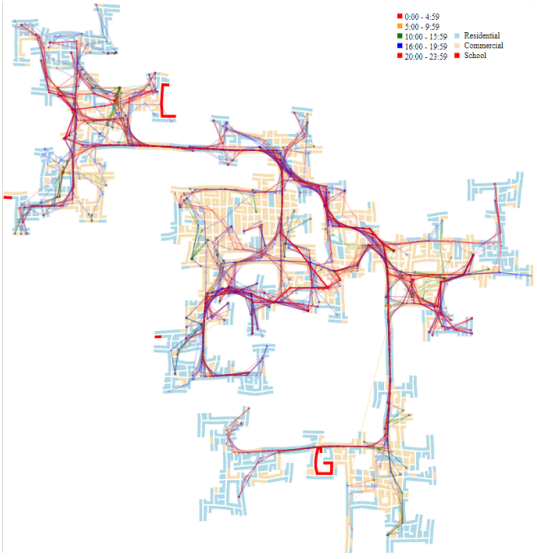


Fig. 3: Traffic through the city on a 2022/03/02.

It is difficult from figure 3 alone the derive the busiest areas and identify potential bottlenecks. To further investigate we divided the day into 4 different time spans: morning, lunch, evening and night. Figure 4 shows the distribution of traffic through out these times.

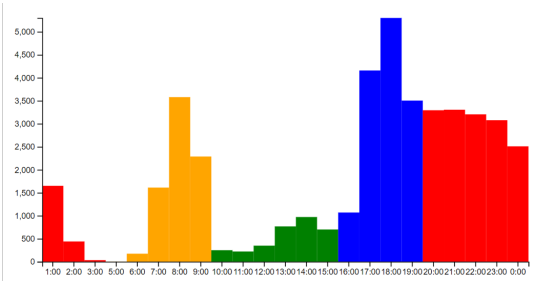
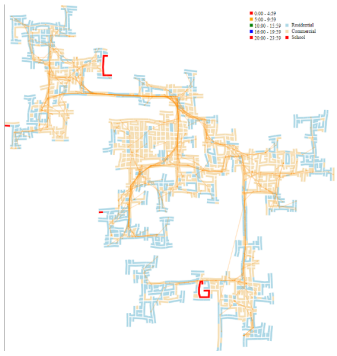
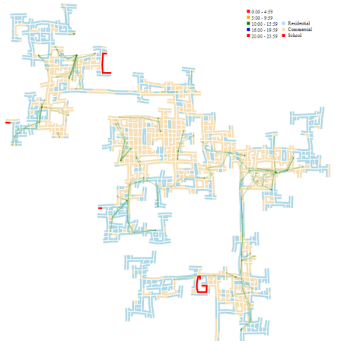


Fig. 4: Traffic distribution over time on 2022/03/02.

From figure 4 graph it is possible to derive that morning, evening and night contain the most traffic. By plotting the same map but extracting only the selected timespans we can see the flow through the city on these busy times in figure 5 and 6

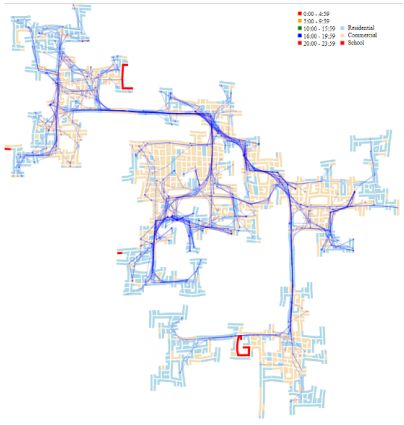


(a) Mornings traffic.

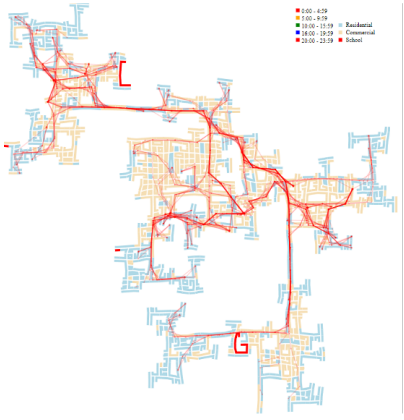


(b) Lunch traffic.

Fig. 5: Traffic during the day



(a) Evening traffic



(b) Night traffic

Fig. 6: Traffic during evening and night

These figures show that there is significant traffic in the morning and evenings/nights which gives us some insight to what roads might be a bottleneck. We also see that during lunch time the participants only travel shorter distances, perhaps to get lunch.

By utilizing our plots we can identify a set of potential bottlenecks shown in figure 7. However this is purely a guessing game based on the figure. Of course some bottlenecks are easy to identify, for example the two roads that are linking the upper and lower parts of the city to the middle is easy to identify as potential bottlenecks.

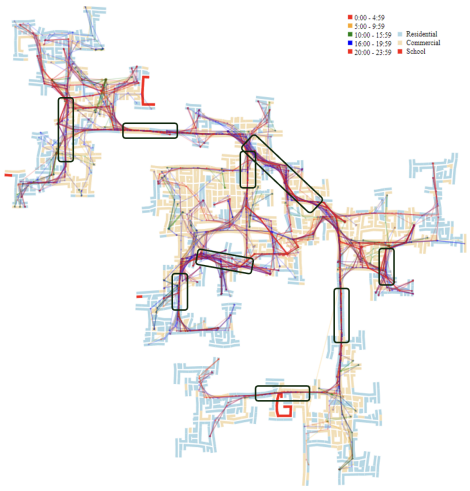


Fig. 7: Potential bottlenecks.

4.3 Daily routines

To capture and analyze the participants daily routines we sampled two random participants and extracted 2 weeks of location and ”mode” data from the activity logs with pandas. The ”mode” column from the activity logs provides 5 categories of locations where the participant is currently at. To see their geographical routines we plotted the location data from the two weeks onto the map. To be able to explore the participants routines on a more detailed level a timeline was also made, the timeline shows the different categories as colors throughout the two weeks which is then plotted with the days as x ticks.

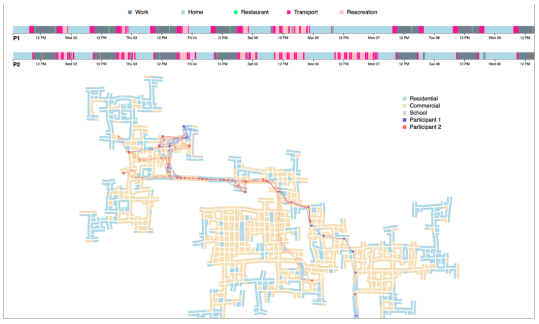


Fig. 8: Dashboard for the daily routines.

The figure 8 shows what seems to be their daily commute from residential area to a commercial area where they might work. We also notice some minor commuting perhaps for recreational activities. The timeline help us understand that the participants are in fact making longer commutes during the morning and afternoon and are making smaller commutes to recreational activities in the evening. We can also from the timeline deduce that one participant seems to be more actively engaging in recreational activities, especially during the second week.

4.4 Patterns

For the patterns we chose to focus on two different aspects, financial and social. For the financial analysis we sampled 5 random participants and looked at their available balance from the activity logs throughout the entire dataset. The reason for choosing only 5 samples is purely based on performance. We then plotted this balance in a a simple line chart seen in figure 9.

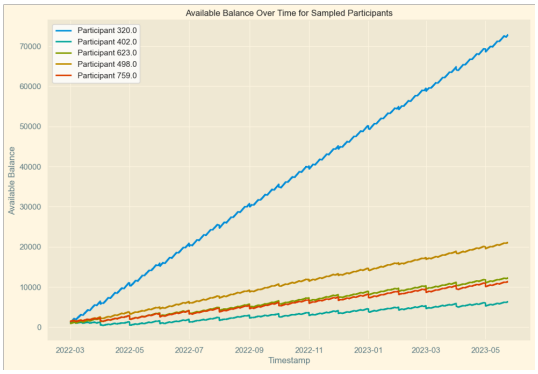


Fig. 9: Balance over time for sampled participants.

Looking at figure 9 the balance seem to be going up for all of the sampled participants which might suggest that the city is in a financial boom, or perhaps inflation is at play. To address the "jaggedness" of the lines we decided to plot the same participants wage income over the entire dataset with the help of the financial journal.

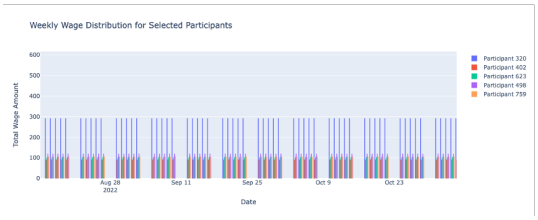


Fig. 10: Wage distribution for the sampled participants.

Figure 10 shows that participants get wage on a weekly basis which might explain the jumps in balance for each month in figure 9. The drop each month is assumed to be related to participants paying rent.

For the social pattern we plotted the number of check-ins per venue for all participant over the entire span of the dataset using the check-in journal. From figure 11 we noticed a burst of check-ins during the first month. To further investigate we made a bar chart using data from the travel journal. All rows were participants were headed to a social gathering were extracted and then plotted as a distribution over time, see figure 12.

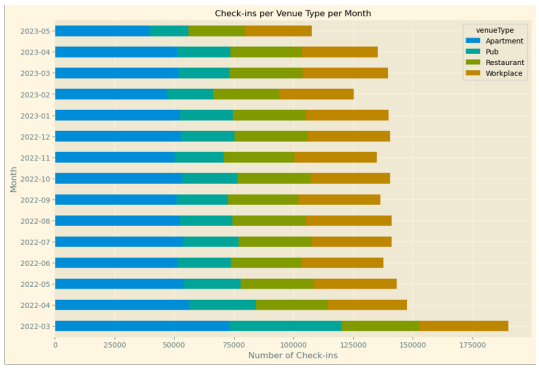


Fig. 11: Check-ins per venue, all participants.

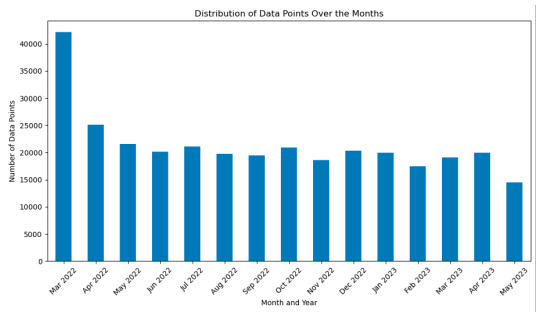


Fig. 12: Number of times participants headed social gatherings, per month.

This chart shows us that there was in fact a burst of activity in the first month.

5 Design & Implementation of VA Solution

The tools used for data processing was python along with the library pandas. These were used together with the jupyter notebook system for python which allows for a more iterative workflow and quicker view of the data without the need of rerunning an entire script each time. The choice of python and pandas was based on previous experience with these tools. For visualization a combination of javascript with d3.js and python with matplotlib and plotly were used. We found it easier to work with map related visualizations in d3 than in python which led us to do all of the map related visualization in d3 and the charts with matplotlib/plotly in python.

6 Discussion

The analysis strategy turned out work well for the questions. Some pros of the method used were, the clarity of the pipeline and the approach of focusing one question at the time. These two factors provided a structured workflow. It also provided a adaptable environment which we made sure to utilize during our analysis. Some cons of the approach were, time consumption and rabbit holes. Since there was no specified goal of when a question was completely answered, it would be possible to go through the pipeline a many times which could be rather time consuming. Another danger with going back to data processing is the potential of bias and assumption during the analysis process. If something was defined as interesting during the pre-processing step and only specific data was extracted this might reflect the overall analysis.

The cons of our strategy could be improved by both having a more rigid and standardized workflow for the data-processing. A lot of questions involved similar steps in terms of extracting data, a script could help automate this process. To address the risk of bias, an exploratory data analysis phase could be used as an initial step. This process would involve examining all of the data without any specific assumptions.

7 Conclusions

Overall the solution used provided some answers to the questions that were given. However there is definitely room for improvement in terms of answering the questions on a more detailed level. The analysis strategy provided a good workflow but had some caveats that could be addressed in the future.

References

[1] E. H.-H. Chi. *A framework for information visualization spreadsheets*. PhD thesis, USA, 1999. AAI9921428.

[2] D. A. Keim, J. Kohlhammer, G. P. Ellis, and F. Mansmann. Mastering the information age - solving problems with visual analytics. 2010.

Contributions from Each Project Member

Emma Zettervall Question 1 and 2, data processing, data merging

Oliver Lundin Question 3 and 4, coordinate projection