Fiscal Clarity: Turning Data into Insights for Engagement, Ohio

2022 Economic VAST Mini-Challenge

Omair Neazi
Arizona State University
Tempe, Arizona, USA
oneazi@asu.edu

Eric Waters
Arizona State University
Tempe, Arizona, USA
eswaters@asu.edu

Collin Wood
Arizona State University
Tempe, Arizona, USA
cjwood8@asu.edu

Adam Miyauchi
Arizona State University
Tempe, Arizona, USA
amiyauch@asu.edu

Bijan Anjavi
Arizona State University
Tempe, Arizona, USA
banjavi@asu.edu

Andrew Murwin
Arizona State University
Tempe, Arizona, USA
amurwin@asu.edu

1 INTRODUCTION

Big data has revolutionized the way businesses, governments, and societies around the globe operate. Data visualization, in particular, has emerged as a critical tool for exploring, analyzing, and communicating complex datasets. Visualizations transform tangled datasets into interactive visual representations capable of uncovering patterns, trends, and relationships contained within them. As a result, end users gain valuable insight into their data, enabling them to make informed data-driven decisions.

In this paper, we present our visualization interface solution for the 2022 VAST Economic Mini Challenge. Utilizing the D3.js Javascript library [2], we design a data visualization to help the city planners of Engagement, Ohio analyze a massive dataset of resident and business financial activities. We have designed our interface specifically to support city planners in identifying patterns and trends in the financial behaviors of its residents. Using our visualization system, city planners can visually contextualize the city's current state. In addition, the planners will be able to identify areas of weakness and opportunities for potential improvement.

We begin by introducing our visualization interface and describe the system's design and interactions. We then outline the 2022 VAST Economic Mini Challenge and review the questions the city planners need answered. Following that, we discuss the dataset of resident and business financial activities, including its size, scope, and preprocessing steps. We then outline several use cases that demonstrate how our system can be used to facilitate data exploration and analysis. These use cases outline how our interface can be used to identify patterns in the financial health of residents and businesses, uncover areas of weakness, and support data-driven decision-making by city planners. Lastly, we discuss some of the lessons we learned along the way and note improvements that can be made in the future.

2 VISUALIZATION DESIGN

The data visualization consists of two pages, with the first page featuring four business-related charts as seen in *Figure 1*, and the second page showcasing two people-related charts as seen in *Figure 2*. Several interactions are available throughout the visualization, including hover tooltips, date and revenue range sliders, demographic attribute dropdowns, on-click events, a lasso tool, and animation.

Chart 1, "Employee Turnover Rate by Region," is a heatmap that displays turnover rates over the selected time frame. Users can utilize the lasso tool to select smaller regions or hover for individual building turnover rates. This lasso filters every other chart across both Business and People pages. This chart takes advantage of query-based actions since the user can narrow down which regions they want to focus on; this allows the user to focus on the distribution of turnover rates since each building is colored according to a sequential color scale. This chart also takes advantage of data reduction since using the lasso allows users to focus on a certain geographic region without having to look at all of the data at once.

Chart 2, "Employee Turnover for Selected Region," is a line chart representing employee turnover rate over time based on the region selected in Chart 1. Users can adjust the date slider to focus on specific periods and hover over the chart for additional daily employee turnover information. The line chart is primarily designed to present trends appearing in the turnover rates for the region selected in the heatmap.

Chart 3, "Total Revenue by Business Type," compares the revenue performance of different business types in a bar chart format. Users can filter data using the revenue and date sliders, as well via the Chart 1 heatmap lasso tool. This chart heavily focuses on allowing the user to discover the distributions present in the revenue data. The user can easily identify which businesses are making the most money for the selected time period and which businesses might be struggling.

Chart 4, "Cumulative Revenue by Business Type," is a normalized stream graph that displays the revenue performance of various business types over time. Users can filter data with the revenue, date, and map options or hover for more details. The primary focus of the stream graph is to allow users to discover both distributions and trends relating to revenue by business type. Since the chart shows the data over time, trends can be observed in what time of month might be most prosperous for certain business types. In terms of Munzner's action and target pairs [3], this chart analyzes the distribution of revenue proportions.

Chart 5, "Average Balance by Demographic," is an interactive bubble chart that illustrates each participant's demographic category and their average account balance over time. Aside from the Chart 1 lasso, the demographic dropdown button, and the date range slider, users can also filter the scatter plot to the right by clicking on a category bubble. The bubble chart is primarily designed to assist

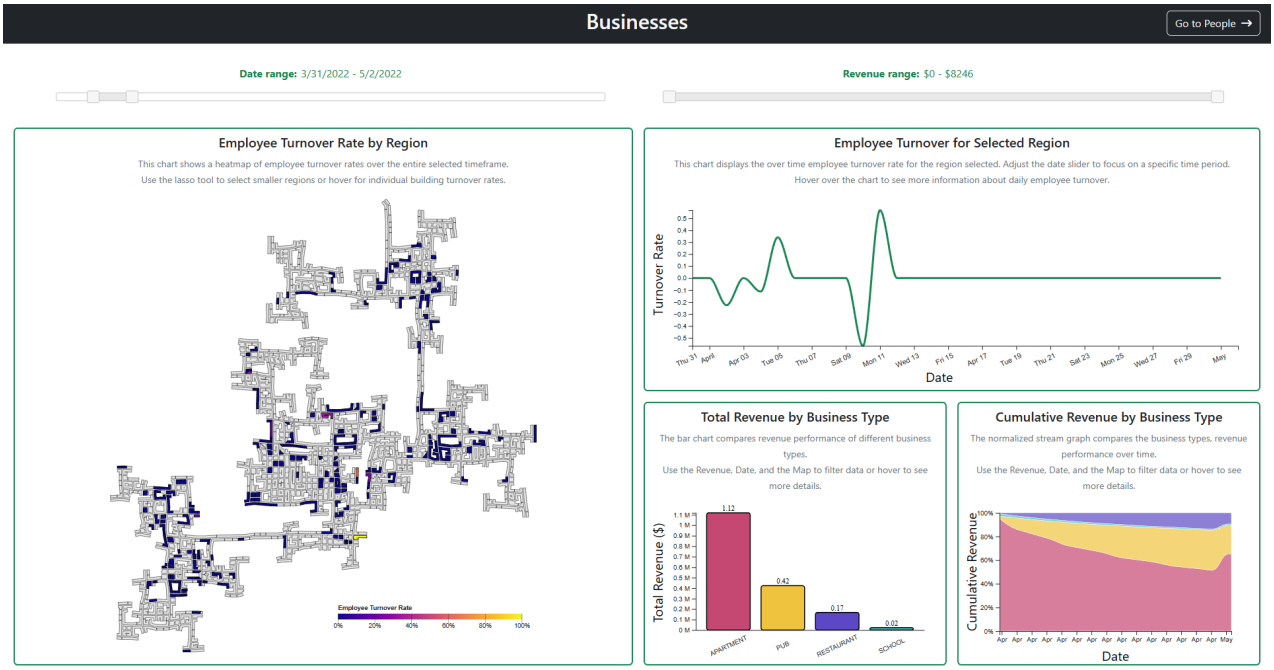

Figure 1: Business Page

Figure 2: People Page

in discovery and presentation tasks. It shows how much money is made by the various selected demographics, so it helps shows the distribution of income levels between groups.

Chart 6, "Wages vs. Cost of Living," is a scatter plot that shows the wages and costs for each resident in the city over time. Users can click on a participant point to highlight its corresponding category in the Bubble chart on the left. The scatter plot is designed to support

browsing the data and finding potential trends and correlations. Since the scatter plot animates the data over time, users can identify trends in how the cost of living and wages change. They can also browse individual data points by hovering over them to gain more insight into how much people make and spend.

3 DESCRIPTION OF VAST MINI CHALLENGE

As the popularity of the small town of Engagement, Ohio continues to rise, city planners must consider how to deal with the surge in population. With new people constantly moving to the city, it is important for city planners to make data-driven decisions regarding what areas and resources to develop. In order to help guide the decision-making process, the city recruited around 1000 residents and tracked their jobs, financial habits, movement around the city, and residences to collect data over the course of more than a year. [1]

When choosing what areas to develop, city planners must carefully consider the economic health of the city. This includes factors such as residential income, business revenue, and employee turnover. By using the financial data included in the dataset, one can begin to understand which businesses are most prosperous and which groups of people are earning the most income based on their demographics. People can be categorized based on their interests, level of education, age, and other factors, so it is important to highlight how belonging to a certain group might affect people's financial prosperity. It is also important to understand how people's income compares to the cost of their rent to ensure people are comfortably able to afford housing. Different areas of the city house various businesses, so analyzing these businesses can help city planners gain insight into which areas might need to be allocated the most resources for development. Visualizing the economic data described above helps city planners make informed decisions about how to invest in the growth of Engagement, Ohio.

4 DATASET DESCRIPTION

For the VAST MC 2022 Economic Challenge, a dataset was provided that contains various types of data located in a multitude of separate CSV sources [1]. The dataset represents a Relational Database (Tabular) [4] that combines temporal data with categorical information, providing in-depth detail about the behaviors, transactions, and attributes associated with the synthetic population of the fictional city of 'Engagement, Ohio'. The bulk of data is found in the seventy-two files (eighteen gigabytes) of *ParticipantStatusLogs*, describing various attributes of the city's residents across a fifteen-month data collection period.

The *ParticipantStatusLogs* contain many foreign keys that reference the other tables in our dataset. Table 1 shows the different data abstraction types that are used in this system: some visuals will be encoded by multiple different abstraction types (categorical, ordered, ordinal) [3]. The *Attributes* and *Journals* data provide additional context and detail about the participants' lives, habits, and interactions.

The *Attributes* set consists of tables that contain information about various static characteristics and properties of the city, its residents, and the available facilities. The *Journals* data, on the other hand, contains tables that provide insight into the dynamic behaviors and transactions of the study participants over time.

Table 1: ParticipantStatusLogs Columns

Name	Data Type	Description
timestamp	Ordinal/Sequential	Time when the status was logged
currentLocation	Quantitative/Diverging	Location of participant within the city
participantId	Categorical	Unique ID assigned to each participant
currentMode	Categorical	Mode the participant is in
hungerStatus	Categorical	Participant's hunger status
sleepStatus	Categorical	Participant's sleep status
apartmentId	Categorical	Apartment ID where participant resides
availableBalance	Quantitative/Diverging	Participant's financial account balance
jobId	Categorical	Job ID held by participant
financialStatus	Categorical	Participant's financial status
dailyFoodBudget	Quantitative/Sequential	Budgeted amount for food
weeklyExtraBudget	Quantitative/Sequential	Budgeted amount for miscellaneous expenses

Table 2: Attributes Reference Table

Table Name	Foreign Key	Referenced By
Apartments	apartmentId	ParticipantStatusLogs
Buildings	buildingId	Apartments, Employers, Pubs, Restaurants, Schools
Employers	employerId	Jobs
Jobs	jobId	ParticipantStatusLogs
Participants	participantId	ParticipantStatusLogs, CheckinJournal, FinancialJournal, SocialNetwork, TravelJournal
Pubs	pubId	CheckinJournal
Restaurants	restaurantId	CheckinJournal
Schools	schoolId	TravelJournal

The *Attributes* data includes the following tables: *Apartments*, *Buildings*, *Employers*, *Jobs*, *Participants*, *Pubs*, *Restaurants*, and *Schools*. These tables provide detailed information about the city's infrastructure, employment opportunities, and resident demographics.

The *Journals* data consists of the following tables: *CheckinJournal*, *FinancialJournal*, *SocialNetwork*, and *TravelJournal*. These tables

Table 3: Journals Reference Table

Table Name	Foreign Key	Referenced By
CheckinJournal	participantId	Participants
CheckinJournal	venueId	Apartments, Pubs, Restaurants, Employers
FinancialJournal	participantId	Participants
SocialNetwork	participantIdFrom	Participants
SocialNetwork	participantIdTo	Participants
TravelJournal	participantId	Participants
TravelJournal	travelStartLocationId	Apartments, Pubs, Restaurants, Employers, Schools
TravelJournal	travelEndLocationId	Apartments, Pubs, Restaurants, Employers, Schools

capture the participants' movement, financial transactions, social interactions, and travel motivations throughout the study period.

By combining the information from both sets of tables, users can gain a comprehensive understanding of the city's economic health and make data-driven decisions to promote its growth and development. They are essential for a comprehensive understanding of the factors influencing the city's growth and development. By analyzing the interconnected data from these tables, city planners can make informed decisions regarding the prosperity of businesses, the economic impact of certain actions, and the demographic makeup of our population.

4.1 Derived Data

The Total Revenue by Business Type and Cumulative Revenue by Business Type charts use business revenue data to analyze the financial health of businesses within the city. A Python script was developed to derive daily revenue in dollars for each school, pub, restaurant, and apartment business. For school revenue, the *FinancialJournal.csv*, which contains financial transactions made by residents was joined with the *Schools.csv*, which details school information, including their location. Doing so allows money spent on tuition each day by school location to be summed. To calculate apartment revenue, the Participant Status Log files, which contain the apartment in which each resident resides, was joined with *FinancialJournal.csv* and *Apartments.csv*, which details the location of each apartment. The rental payments by apartment location were totaled to obtain daily apartment revenue by location. Following a similar approach, *Restaurant.csv* and *Pubs.csv* were used in place of *Apartments.csv* to determine revenue for restaurants and pubs by mapping financial transactions to restaurant and pub locations. The resulting *DailyRevenueByBusinessLocation.csv* contains the following columns: *date*, *unitId*, *buildingId*, *location*, *revenue*, and *businessType*. Each row represents the daily total revenue a given business generates in dollars. Regarding data abstraction, the derived revenue data is static and tabular. The date field is ordered, ordinal, and sequential. The *unitId*, *buildingId*, and *businessType* fields are categorical. The *location* (latitude/longitude) field is ordinal, quantitative, and diverging. The *revenue* field is ordered, quantitative, and sequential. The final preprocessing step for revenue data consists of summing the revenue for the selected date, revenue, and geographic area.

To calculate the average balance by demographic, another Python script combines log files into a single DataFrame, and the average available balance per month for each participant is calculated. Participant metadata is then loaded from the *Participants.csv* file and merged with the average available balance data based on the *participantId*. This process associates the derived average balance attribute with each participant's demographic information from the *Participants.csv* file. The output of the script is saved as *AveragedParticipantBalances.csv*, which is used in the D3.js code for the visualization. In the D3.js code, the demographic information from the merged data is leveraged to group participants by various attributes. The average balance across these participant groups is computed based on the derived *average availableBalance* attribute, which is a quantitative diverging attribute representing the average financial account balance of each participant over time. The visualization dynamically displays the average balance by demographic group over time, providing users with an interactive way to explore data and discern patterns among different demographic groups in Engagement City.

In order to calculate the turnover rate for each of the employers, a Python script was used to calculate how many employees worked for an employer each day. This script took advantage of the Participant Status Log files since these contained information about the job of each participant in the study as well as what date they were employed on. The *Jobs.csv* file was used to map these jobs to individual employers, and the *Employers.csv* file was used to locate which building each employer resided in. The *jobId* and *employerId* in each file are categorical attributes. A resulting CSV file was created containing the following columns: *employerId*(categorical), *buildingId*(categorical), *date*(ordinal/sequential), and *positionsFilled*(ordered/quantitative/sequential). Each row in this file represented how many people were employed by a given employer each day. This count of employees could then be used to calculate the turnover rate using Equation 1. This equation was applied on a daily basis to find the turnover rate for the line chart, and applied to the start and end date to find the turnover for the heat map. The data created by the script was tabular in nature and the *positionsFilled* attribute is ordered, quantitative, and sequential since the minimum amount would be 0. Turnover rate itself is ordered, quantitative, and diverging since the turnover can be positive or negative depending on if employees left the company or joined.

Lastly, in order to compare wages with the cost of living, some simple aggregation preprocessing was used. In order to determine the cost of living, each participant had all of their purchases totaled by month. This included their monthly rent, food, and entertainment, among other expenses. If a participant had no wage or no expenses for any given month, they were omitted from the graph as they would not be able to be displayed properly.

$$\frac{\text{startingEmployees} - \text{endingEmployees}}{\frac{\text{startingEmployees}+\text{endingEmployees}}{2}} \tag{1}$$

5 VAST CHALLENGE EXPLORATION

5.1 Business Prosperity

Over the period covered by the dataset, which businesses appear to be more prosperous? Which appear to be struggling?

In order for City Planners to grasp a more complete understanding of the relative prosperity of different business types in 'Engagement' the development of two visualizations was necessary (see Figure 1). The creation of a bar chart comparing the total revenues of each business type allows the planners to get a breakdown of which businesses have a greater impact on the economy and where participants are making purchases at a higher level. To drill down on this information and get a temporal breakdown of these revenues was key in ensuring that users have all information needed to see how business might be prosperous in certain time ranges. The creation of a normalized stream chart allowed for the ability to conduct an analysis of the cumulative revenue as time throughout the system progresses. The addition of a normalized stream chart (as opposed to a non-normalized stream chart) allows the user to more accurately compare the relative proportions of different business types. The ability to filter our bar and stream charts by revenue, range, and location gives all of the tools necessary for making informed decisions about the prosperity of certain businesses. For example, the date range can be filtered to see only summer months, the revenue range can be set to only show business types making less than a certain amount per day, and the location filter on our map visual's lasso select can allow the user to see a breakdown of the proportions and revenue totals in that region. By filtering the data as described above, the city planner has the ability to explore the relative prosperity of different businesses and can answer questions such as:

- Which businesses are successful in the winter yet struggle in the summer months?
- What is the most successful business that makes over $1000 per day?
- Which businesses were most successful in this neighborhood?
- When do businesses see the most growth in revenue?

These two charts support Munzner's action-target pair abstraction in that they both use the action of *Analysis* on the distribution of *Revenue* to support answering the MC questions.

5.2 Resident Financial Health

How does the financial health of the residents change over the period covered by the dataset? How do wages compare to the overall cost of living in Engagement? Are there groups that appear to exhibit similar patterns?

The *people* dashboard of the application is designed specifically to answer this set of questions. A powerful way to assess the financial health of residents in the city is to compare their wages against their cost of living. If their wages and cost of living are similar, it can be inferred that residents are living paycheck to paycheck. If the wages are much higher than the cost of living, it can be inferred that residents have leftover money to save.

The simplest and most effective way of achieving this was to create a scatter plot of wages against the cost of living. Each point on the scatter plot represents a participant in the city. This scatter plot can be seen in Figure 2. Since wages and rent are on a monthly basis, the data was aggregated by month, i.e, the cost of living for a particular month is the combination of their rent and any other spending on food, entertainment, etc. However, since a scatter plot on its own is unable to capture how the financial health of the residents change over time, two visualization techniques were used: small multiples and animation. The user can access each month's scatter plot by using a dropdown to select which month they would like to view. Selecting a point assists in discovery by making trends in the data more prevalent. The points on the scatter plot will smoothly transition to their updated positions when the month is changed so that the city planner can see precisely how the data changes as time changes. In order to show the whole story of how the data changed over the course of the entire dataset, the user can press the *Play Animation* button to animate the data. This provides the city planners with a timeline of how wages and the cost of living change.

In order to address whether or not groups appear to exhibit similar patterns, a bubble chart was created that shows the average account balance by a variety of demographic characteristics. This bubble chart can also be seen in Figure 2. The user can select between *Household Size*, *Has Kids*, *Age*, *Education Level*, *Interest Group*, and *Joviality*. The user can also use a date range slider to filter the data to a particular time frame. The combination of these tools will allow city planners to find financial patterns by demographic groups.

In order to provide further insight, the bubble chart and scatter plot were linked via two interactions. The first is that the user can click on a bubble in the bubble chart in order to emphasize the participants in the scatter plot that belong to that particular demographic group. This allows the user to see, for example, the wages versus cost of living for only participants that have a household size of three. The second interaction is that the user can click on a participant in the scatter plot and see the corresponding demographic group that the participant belongs to. For example, the user could click on the point *Participant 78* and see that the *Gen Z* bubble is highlighted, indicating that *Participant 78* belongs to the *Gen Z* age group.

5.3 Employer Turnover Rates

Describe the health of the various employers within the city limits. What employment patterns do you observe? Do you notice any areas of particularly high or low turnover?

In order to examine the health of employers, and specifically their turnover rates, two different visualizations were created. Turnover rates can be examined in two different ways: geographically and over time. The Employee Turnover Rate by Region graph displays the turnover rates of different businesses using a heat map scheme placed over the city. By combining the two, the user can observe both the turnover rate of specific businesses and the general trend of turnover rates across the map. Generally, heatmaps have one major issue: outliers can skew the scale. While showing outliers is helpful, and works out well in the case of this map, it makes it hard to differentiate between areas where all buildings are similar colors. To allow for a closer examination of areas, a lasso can be

used to select specific buildings. Lassoing a building queries the list of buildings to select the buildings used by all of the visualizations on the website. These buildings then have their color turnover rates in a particular area that can be observed, rather than just the overarching trends as seen in *Figure 3*.

As important as location is to this question, time is equally important. The Employee Turnover for Selected Region graph shows trends in turnover rate over time. By shifting the time slider and selecting regions on the map, the chart can be adjusted to fit a specific area. Inversely, the chart can provide information on where to examine the map. Focusing the time range on areas of change can reveal new information on the overall map regarding points of interest. Both visualizations are designed to stay synced for exploration.

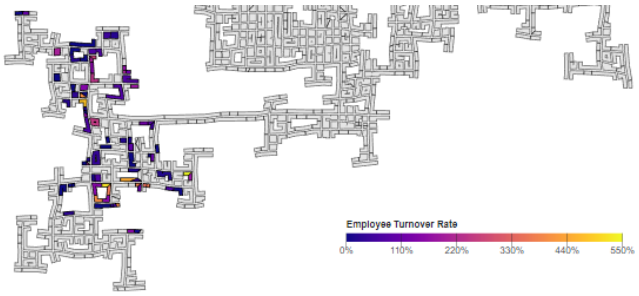

Figure 3: Examining the Turnover Rates of the Southwest Area

6 DISCUSSION

Several important lessons were learned throughout the process of making this system of visualizations. One important takeaway is that data preprocessing is key to making efficient visualizations of that data. There was an initial attempt to use the raw data in making the visualization, deriving needed attributes as needed within the JavaScript code. There was also an attempt to load the data into a postgresql database and use SQL queries to derive important data to visualize. Both techniques were far too slow. Preprocessing the data beforehand and exporting the derived attributes to CSV files was the clear winner. The other significant takeaway was that visualizations, when intended to interact with several other visualizations, should be developed with those interactions in mind. Initially, the graphs were created separately, but this caused difficulties in having them communicate and respond to each other. More time was spent integrating the graphs together than creating the graphs themselves because the graphs were not initially created with the integration in mind.

There are several improvements that could be made to the system in the future. Firstly, it would be ideal if the heatmap zoomed into the area that was selected by the lasso tool. This would provide clarity by omitting data that is not being focused on. Another potential improvement to the system would be to implement animations for all temporal graphs. The scatter plot has an animation to show how the data evolves over time, but it would be ideal if all graphs dealing with temporal data had this functionality. This would help the city planners see how the data changes over time without having to tweak the date range slider excessively. Another issue with the current system is concerned with the employee turnover rates. The data has extremely high rates of turnover in the first months of the data, but then has virtually zero turnover for the remaining time. The system could be improved by finding some sort of innovative way to analyze turnover rates that results in a more useful visualization. Lastly, it would be interesting to use machine learning clustering algorithms to group participants as opposed to grouping them solely by demographic characteristics. These algorithms could derive more insightful groupings of participants which could be useful to the city planners.

REFERENCES

[1] 2022. VAST Challenge 2022. https://vast-challenge.github.io/2022/description.html. Accessed: 2023-04-22.
[2] Mike Bostock. 2023. D3: Data-Driven Documents. https://d3js.org/ Accessed: 2023-04-22.
[3] Chris Bryan. 2023. Data & Task Abstractions. (2023). Lecture slides, Google Drive. Available at: https://docs.google.com/presentation/d/19sqmS7LHXrF5q6nMVROXOExMXQ11Zo6xfsDqnGzdKTA/edit.
[4] Chris Bryan. 2023. Tabular Data. (2023). Lecture slides, Google Drive. Available at: https://docs.google.com/presentation/d/1U89oHlUzE2NQq5En1tXlanNVkEHjL3DZYKZsEza6-ps/edit.