2021 Vast mini-challenge 2

Hassan Mustafa

August 30, 2023

1 Introduction

The vast challenges is an annual competition designed to advance the field of visual analytics through a competitive platform. This report aims to solve the mini-challenge of 2021 by analyzing and visualizing a given dataset detailing the movement, transaction and information of Gastech employees on the Island of Kronos after several employees has gone missing. By using different analyzing and visualization techniques the goal of this report is to find strange anomolies, relationships between employees and infer ownership of credit-cards given the raw data and give value to the data to help the law enforcement to conduct a thorough investigation by answering the following questions:

1. What are the most popular places and when? is there any anomalies found?
2. What discrepancies are there between the vehicle and transaction data that can be found?
3. Can you infer ownership of each credit and loyalty card?
4. what formal or informal relationships between the GASTech personel are they?
5. do you see any suspicious activity? identify the locations that those activities might occur

2 Data characterisation and preprocessing

To address the identified problem, the following datasets and mapping resources have been provided:

2.1 Data

1. `car-assignment.csv`:

   This dataset enumerates the vehicle assignments allocated by Gastech to its employees. It encompasses details such as: Employee Last Name, Employee First Name, Car ID, Current Employment Type, Current Employment Title

2. `gps.csv`:

   This dataset captures the GPS movements associated with each vehicle assigned to employees. Specifics include: Timestamp, Car ID, Latitude, Longitude

3. `loyalty_data.csv`:

   This dataset reveals the transactional activities recorded on the loyalty cards of each employee. Featured data entails: Timestamp(date), Location, Price, Loyalty Number

4. `cc_data.csv`:

   This dataset chronicles individual credit card transactions and contains: Timestamp(date and time), Location, Price, Last 4 digits of Credit Card Number

2.2 Mapping Resources

For geographical context:

- Map of Abila (JPEG Format): A comprehensive map of Abila has been provided, showcasing various significant locations.
- Shapefile of Abila and Kronos: A shapefile has been included, outlining the geographical contours and significant features of the islands of Abila and Kronos.

2.3 Preprocessing

In this section an overview of the choices made in the preprocessing phase of the project is described. The goal of the preprocessing phase is to extract and simplify the data such that the process is made easier once the analyzing and visualizing stages are initiated.

2.3.1 comparing similarity between card data

An effort is made to compare the two datasets of credit cards and loyalty card to see how they are similar. As an effort to support the local business in Kronos each employee is offered a loyalty card to use for accessing discounts and benefits. In exchange for the loyalty card the credit card transactions information is stored as recorded on the loyalty card. This leads to the hypothesis that there seems to be some linking relationship between the creditcards and loyalty cards. To view the relationship a scatterplott of the two cards are plotted based on price see figure 1. although there are some outliers it is safe to conclude that the

Fig. 1. scatterplot of prices for credit cards and loyalty cards

loyalty card and the credit card seem to describe the same transactions. it is thus plausible to infer that the loyalty cards are connected to the credit cards. The two datasets can thus be merged into one dataset of transactions and we can infer the loyalty card connected to each creditcard by viewing the duplicates in the merging set. All the outliers are also dropped in this phase to ensure a clean data to work with. After this process there is now a single dataset describing the transactions that we are sure to be made by a creditcard or loyalty card that is registered in both data sets.

2.3.2 adding missing locations to the map

The transaction data includes locations but it does not include any geospatial descriptions of these locations. The provided map is thus

used with the shapefile to georeference each location and extract its latitude and longitude data points. These points are stored in a new locations data set so that the locations can be visualized correctly in the visualization stage. These points are also added to the transaction data sets to describe the geospatial position of each transaction. this is needed to later be able to infer ownership of each credit card by comparing to gps tracking data.

2.3.3 concatenating gps and car-assignment data

Many of the employees have access to company cars witch they can use for both business and personal use. Some employees have the option to rent a company truck, which can only be used for business and does not contain a carID. This means that we only track the movement of the company cars. With this the two datasets are merged into one dataset that stores the gps movement and adds the employee information from the car-assignment.csv dataset. this leads to the truck drivers to be discarded in this case and we only examine the drivers that use the car for personal use.

2.3.4 infering creditcard owners

when the datasets have been concatinated into two main datasets(transaction data and gps tracking data) it is now possible to infer ownership of the credit cards finalizing the connection of all the data. This is done by a majority vote algorithm that examines each transaction and searches the closest person to the transaction within a time frame. The result is stored in a list and later a majority vote is performed, assigning a person to the card that they where closest to the most times.

2.3.5 final cleaning

To finalize the preprocessing the choice to remove all NaN values for the merged transaction data is removed for the instances is where there is no latidude and longitude registered with the transaction. This is done to ensure that we remove unnecessary outliers and to only analyze the transactions made in locations we have identified.

3 Analysis and Methodology

In this section, the outline and strategy employed to adress the problem is presented. We'll delve into the tools and methodologies used for data analysis and visualization, as well as the rationale behind the visualization choices made throughout the project. Our approach primarily hinges on two components. Firstly, an interactive dashboard for engaging with the data and in-depth analysis. Secondly, Jupyter Notebook is utilized for swift analyses and graph generation.

3.0.1 plotly dashboard

The dashboard is comprised of two primary sections. The initial component is dedicated to analyzing card transactions. It offers users the flexibility to toggle between various visualization layouts by adjusting the axis values and exploring different dimensions. This segment aims to facilitate a thorough examination of the card transactions and the corresponding transaction locations, as shown in Figure 2.

Fig. 2. Image of the first part of the dashboard viewing data of transactions and locations they occur in

Users have the flexibility to filter by location, individual, day, and time, as well as to modify the plot dimensions based on specific inputs. These interactive features ensure a comprehensive exploration of the transaction data, revealing details about who is making a transaction, the nature of the transaction, and its timing. Additionally, a side panel displays the counts of transactions for the selected locations on a given day. This segment of the dashboard aids in identifying anomalies and finding relational patterns among transactions.

The dashboard's second section features a spatiotemporal visualization of car GPS movements. In this view, users can filter by employee and apply the same day and time filters as in the dashboard's first section to adjust the time dimension displayed. Additionally, users have the option to project the movements onto a 2D plane, simplifying the visualization of an employee's movement, as showcased in Figure 3.

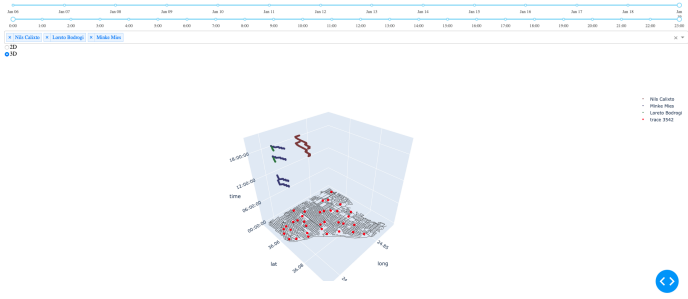

Fig. 3. Image of the second part of the dashboard viewing space time cube of gps tracking

3.0.2 Jupyter Notebook

While the dashboard provides extensive insights into the analysis, it primarily operates on pre-processed data. For the initial question posed by the project, a holistic view of all transaction is necessery. This general perspective, encompassing both credit card and loyalty card transactions, offer a clearer understanding of location popularity compared to the processed data presented in the dashboard. To achieve this in-depth analysis and generate the requisite visuals, A heatmap using Jupyter notebook was generated.

The two parts of the dashboard connect the two datasets and allows for a comprehensive analysis to be made.

4 Results

In this section the results of the analyze will be presented and the process followed to answer each question

4.0.1 what are the most popular places and when? is there any anomalies found?

In Figure 4, daily transactions throughout the city are depicted. The most active locations are immediately clear. Notably, there is a significant decline in transactions on the 11th-12th and 18th-19th of January, which align with weekends. This explains the surge in activity at locations like Katerina's Café and the decrease at Guy's Gyros. Further analysis of top locations reveals that Brew've Been Served and Hallowed Grounds experience traffic earlier in the day. In contrast, Guy's Gyros, Hippokampos, Katerina's Café, and Ouzeri Ellian observe peaks around 14:00 and between 20:00-22:00, as illustrated in Figure 5.

4.0.2 What discrepancies are there between the vehicle and transaction data that can be found?

As depicted in Figure 6 and Figure 7, numerous transactions are recorded at 12:00, yet all these employees are located at Gastech during this period. This anomaly is recurrent at some of the more frequented locations indicating some anomaly in the machinery.

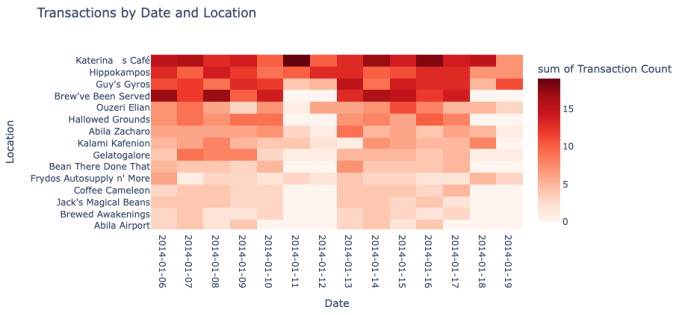

Fig. 4. heatmap showing the most popular places in Abila

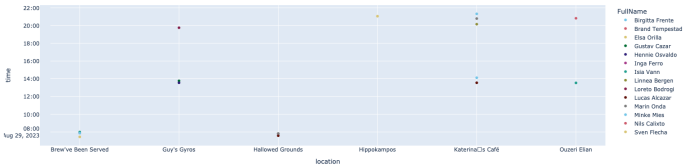

Fig. 5. Scatterplot showing when transactions occur in the most popular locations

4.0.3 can you infer ownership of each credit and loyalty card?

During the preprocessing phase, the assignment of credit card ownership was inferred by examining a transaction and then identifying the closest car to that transaction within a 30-minute window. After considering all transactions, a majority vote determined the credit card owner, based on which car was most frequently in close proximity.

In cross-referencing the transactions on the dashboard with the GPS tracking, they often align. However, discrepancies occasionally emerge, suggesting potential errors in inferred ownership. For instance, as illustrated in Figures 8 and 9, Minke Mies appears to make a transaction at Brew've Been Served at 07:52, which is confirmed by the GPS movement. Yet, another transaction is recorded at Katerina's Café at 14:07. However, the GPS data indicates Minke Mies left work precisely at 14:07, pointing to a potential misattribution of the credit card transaction.

4.0.4 what formal or informal relationships between the GASTech per- sonel are they?

Employees can be categorized into four distinct employment types: Engineering, Executive, Facilities, and Security.

When we filter the GPS movements based on these employment categories, certain patterns and relationships emerge. For instance, GPS data indicates that executives meet at the golf course on the 12th of January around 13:00, as illustrated in Figure 10.

The engineers showcase a synchronized pattern in their daily routine. They consistently arrive and depart from work around the same times in the morning, at lunch, and in the evening. Outside of work hours, they seem to socialize in two distinct groups at separate locations within the city, as showcased in Figures 11 and 12.

4.0.5 do you see any suspicious activity? identify the locations that those activities might occur

The GPS data reveals that Nils often arrives at work quite late in the evening and departs well after midnight on multiple occasions.

Specifically, Nils Calixto appears to have visited Gastech on January 6th, arriving around 22:00 and leaving after midnight. A similar pattern is observed on January 15th.

Furthermore, there's a notable rendezvous between Henke Mies and Hennie Osvaldo at an undisclosed location on January 14th, approximately at 03:30.

Isande Borrasca's GPS tracking seems to be malfunctioning, as the data appears to be slightly off, as depicted in the referenced figure 13. Additionally, there's an instance of suspicious activity associated with Isande on January 19th, just past midnight, at an unidentified location.

5 Design & Implementation of VA Solution

The design and implementation of this project required several decisions to optimize the analyzing and visualizing processes. The most central of these decisions was the adoption of Python as the primary programming language. This choice was rooted in my familiarity with Python, ensuring that I could focus directly on the project at hand without the overhead of learning a new language's intricacies.

To complement Python's capabilities, I used Jupyter Notebook during the preprocessing phase and for generating certain analytical graphs. Jupyter notebook has an iterative nature to its analyzing approach, which allows for an iterative approach to data analysis and processing. This allowed for real-time adjustments and engagement with the data.

For the dashboard creation, Dash by Plotly was chosen because of its compatibility with Python and the robust visualization capabilities of Plotly. The inherent advantages of Plotly, particularly its extensive library of predefined graph templates, simplified the visualization creation process. This combination ensured a coherent platform where I could construct detailed visualizations and delve deep into data analysis.

6 Discussion

While the approach and decisions made in this project largely succeeded in addressing the challenge's questions, there are room for improvement. One significant area of concern is the inference of credit card ownership. The methodology employed isn't foolproof. A more extensive effort to correlate transactions with GPS movements could solidify the link between individuals and their respective credit cards. Unfortunately, time constraints hindered this additional analysis.

Moreover, while the dashboard was developed with the intent of offering a comprehensive exploration of the data, it's evident that the visualization choices could be better tailored. A more pointed approach would align the visualizations directly with the specific questions posed. Future iterations of the project could benefit from crafting visualizations that directly cater to each problem question, rather than emphasizing a broad, interactive exploration tool. This shift in focus might yield more precise insights.

7 Conclusions

In summary, the project addresses the initial questions, although with areas that could benefit from further improvements. Given additional time, a deeper dive into the data might uncover even more insights and enhance the project's value. Nevertheless, considering the constraints of the project's timeframe, it has effectively met the objectives set out in the introduction.

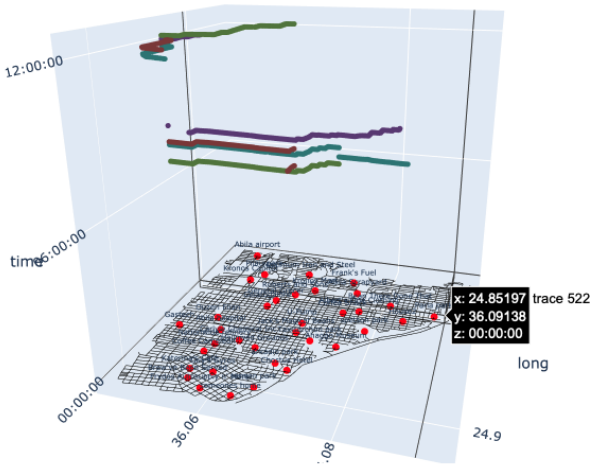

Fig. 6. timespace visualization showcasing where employees are at 12

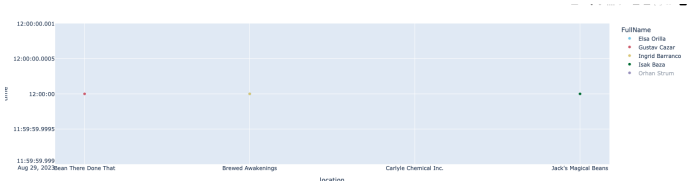

Fig. 7. scatterplot of strange transactions at exactly 12:00

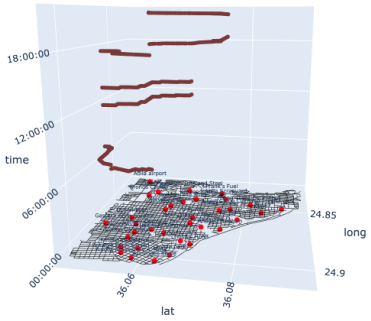

Fig. 8. Gps visualizaiton showing inconsistency in credit card with gps tracking

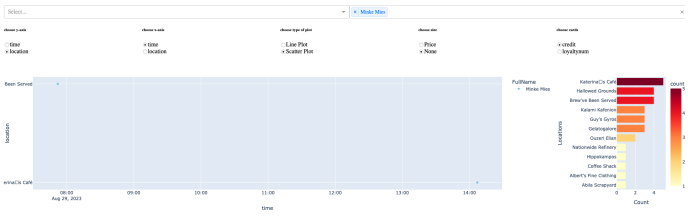

Fig. 9. Gps visualizaiton showing inconsistency in credit card with gps tracking

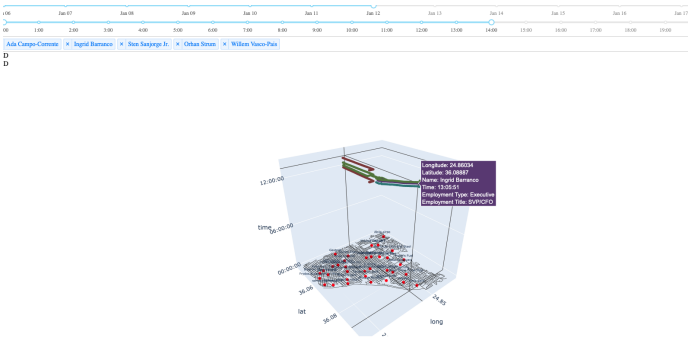

Fig. 10. visualization of executives meeting at golf court

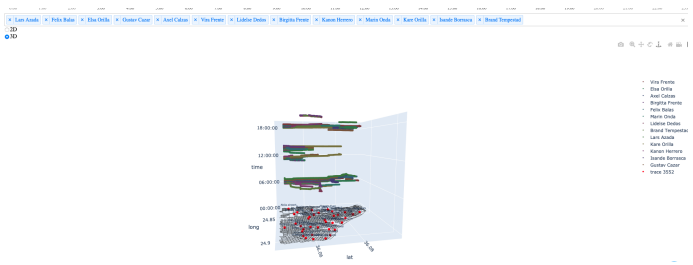

Fig. 11. the overall routine of engineers

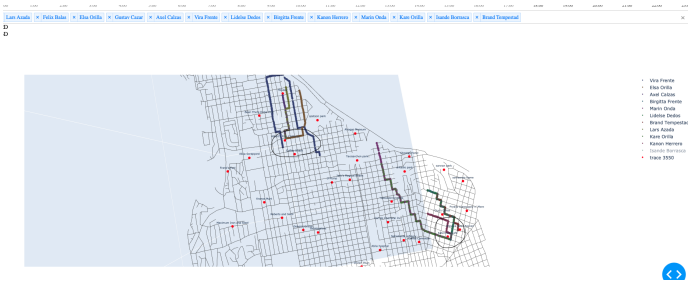

Fig. 12. late night routines of engineers

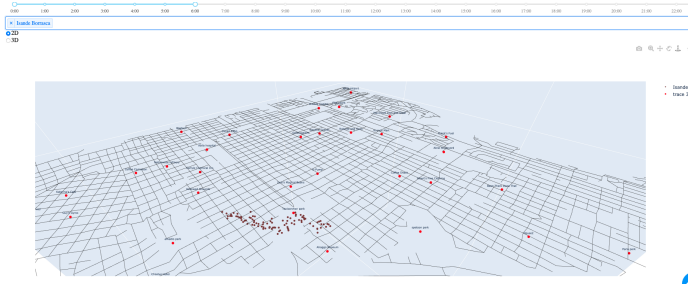

Fig. 13. visualizing gps malfunction and suspicious activity of isande borrasca