iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: IV Month of publication: April 2025

DOI: https://doi.org/10.22214/ijraset.2025.69734

Vast Challenge 2021: Mini-Challenge 2 "Developing a Visual Analytics Tool for Detecting Unconventional Behavioral Patterns in GasTech Employees”

Abraha tesfay Beyene[1], Mr M.Navalan[2], Dr K.Nandha Kumar[3]
[1]PG Researcher, [2]Assistant Professor, [3]Associate Professor, Sri Venkateswara College Of Engineering & Technology (SVCET), Chittoor, Andhra Prades, India

Abstract: In the current era of data-driven decision-making, organizations are increasingly reliant on visual analytics tools to uncover hidden patterns and anomalies in their data. This project focuses on addressing the VAST Challenge 2021: Mini-Challenge 2, which centers on the analysis of employee behavioral data from GASTech, a fictional natural gas company. The objective of this research is to develop a comprehensive visual analytics tool capable of identifying and highlighting unusual patterns in employee movements, access logs, and interactions, which could indicate insider threats, security breaches, or other anomalies.The challenge provides a rich dataset that includes GPS movement data, keycard access records, and travel logs over a simulated period of time. By combining advanced machine learning techniques with visual analytics, this project seeks to build an interactive tool that allows users to explore employee behavior in real time, detect deviations from normal behavior, and flag potential security concerns.These subtle indicators, when analyzed in isolation, may not seem significant. However, using visual analytics, this tool will overlay GPS heat maps, network graphs of employee interactions, and access frequency histograms to highlight John Smith's anomalous behavior. The tool will allow users to explore these visualizations and understand how changes in movement patterns and access logs correlate with potential security risks.Through this approach, the research will demonstrate the power of interactive visual analytics in handling complex, multivariate datasets and delivering actionable insights for organizational security. The results of this study will not only provide a solution to the Mini-Challengebut also contribute to the broader field of insider threat detection and employee monitoring using visual analytics.

I. INTRODUCTION

In recent years, the analysis of human behavioral patterns within organizational environments has become critical for both security and operational efficiency. The VAST Challenge 2021, Mini-Challenge 2, provides an ideal platform to explore this issue by focusing on the employees of GAStech, a fictional company, whose behavior data requires in-depth analysis to uncover unconventional patterns. This challenge reflects real-world scenarios where companies aim to maintain the integrity of their operations while identifying any anomalies that could pose potential risks.

The objective of this thesis is to develop a practical visual analytics tool tailored to detect and interpret unusual behaviors among GAStech employees. By leveraging advanced data visualization techniques and behavioral analytics, the tool is designed to enhance the identification of patterns that might otherwise go unnoticed. Whether these patterns relate to deviations in routine, communication anomalies, or suspicious collaborations, the system will assist analysts in recognizing and investigating critical behavioral shifts.

A key aspect of this project involves merging multiple data sources to create a comprehensive, user-friendly interface capable of delivering actionable insights. The thesis will also focus on the practical application of machine learning and data analysis techniques to improve detection efficiency, providing a robust solution that could be adapted for real- world use in corporate environments. This work will not only address the technical challenges posed by the VAST 2021 Challenge but also contribute to the broader field of visual analytics by showcasing a method for scalable, practical behavioral anomaly detection. The project aimed to develop a practical visual analytics tool capable of identifying unconventional patterns in the behaviors of GASTech employees.

By leveraging transaction data from credit and loyalty cards, along with geospatial positioning satellite (GPS) tracking records for their vehicles, our primary goal was to extract valuable insights, pinpoint anomalies, and enhance decision-making processes. The focus centered on utilizing structured datasets and geospatial maps to create an application empowering users to discern and analyze data patterns efficiently. This approach ensured a scalable solution for addressing missing employee cases and similar challenges across various industries.

II. LITERATURE REVIEW

The VAST Challenge is a prestigious annual competition aimed at advancing the field of visual analytics, promoting the development of new tools, techniques, and methodologies to analyze complex data sets. The 2021 Mini-Challenge 2 (MC2) focused on analyzing employee behaviors at a fictional company, GASTech, to identify anomalies and suspicious patterns. This literature review explores the theoretical foundations, existing methodologies, and related work in the fields of visual analytics, anomaly detection, and behavioral analysis.

Visual analytics combines automated data analysis with interactive visualizations, enabling users to explore large datasets effectively. According to Keim et al. (2008), visual analytics enhances human cognitive abilities by integrating data mining, machine learning, and visualization tools to uncover hidden patterns and gain actionable insights. This approach is particularly relevant to VAST Challenge 2021 MC2, where understanding and interpreting multi-dimensional behavioral data is crucial.In the context of behavioral analysis, visual analytics has been employed in various domains:

Insider Threat Detection: Xu et al. (2015) demonstrated how visualizations can identify anomalies in user activities, such as unusual access patterns or file transfers, which align with the objectives of identifying suspicious GASTech employee behaviors.Movement Data Visualization: Andrienko et al. (2010) developed techniques for visualizing spatiotemporal data, such as GPS logs, to detect movement patterns and anomalies—a key requirement in the GASTech scenario.Anomaly detection, a critical component of the VAST Challenge, involves identifying patterns in data that deviate significantly from the norm. Various techniques have been applied in visual analytics challenges:

Unsupervised Learning Approaches: Techniques like clustering (e.g., K-means, DBSCAN) and density-based methods are widely used for identifying unusual patterns without prior labels (Chandola et al., 2009). These methods can help detect anomalies in GPS movements and badge access data.Rule-Based Anomaly Detection: Systems that rely on predefined rules or thresholds to flag unusual activities. For example, a rule might flag an employee who accesses restricted areas outside working hours.Time-Series Analysis: Techniques for identifying deviations in temporal patterns, such as unusual frequency of visits to specific locations over time (Cook et al., 2007).Insider Threat Prevention: Many studies (e.g., Eberle& Holder, 2009) highlight how behavioral analysis can detect insider threats by identifying deviations from normal patterns.Workforce Behavior Monitoring: Understanding employee behavior can inform organizational policies and operational improvements.

III. METHODOLOGY

The VAST Challenge 2021: Mini-Challenge 2 (MC2) is a data analysis competition that encourages participants to identify unusual patterns in employee behaviors at GASTech, a fictional company. This challenge provides various data sets and requires participants to leverage analytical and visualization techniques to answer specific questions. Below is a detailed. For this project, we utilized four structured CSV-based datasets alongside geospatial maps of Abila and Kronos Island. These datasets and maps provide critical insights into the behaviors and movements of GAStech employees, allowing us to detect unconventional patterns that may indicate security breaches or other anomalies. Below is a detailed description of each dataset used in the project:We used four csv based structured datasets and geospatial maps of Abila and Kronos Island. The four csv-based datasets consist of car assignments, gps, credit and debit card, and loyalty card.

This dataset is essential for understanding the mobility patterns of employees in relation to their work responsibilities and personal activities. By analyzing car assignment data in conjunction with GPS data, it is possible to track which employees are driving company vehicles and identify any deviations from expected routes or locations.The GPS dataset records the real-time movement of GAStech vehicles, providing location data at regular intervals. This dataset is crucial for mapping the routes employees take during business or personal trips and detecting any deviations from typical patterns. Key attributes in this dataset include, By combining the GPS dataset with geospatial maps, it is possible to visualize the movement patterns of each vehicle on Abila and Kronos Island. This aids in identifying abnormal or suspicious activities, such as vehicles entering restricted areas or frequent trips to unplanned destinations.Analyzing this dataset allows us to detect unusual spending habits or transactions at suspicious locations, particularly when combined with other datasets such as the GPS data.

For instance, a purchase at a remote location might coincide with unusual vehicle movements, signaling potential anomalies.By analyzing loyalty card usage, we can gain insights into the regular spending habits of employees and identify any unusual activity. This dataset, when cross-referenced with the credit/debit card dataset and GPS data, can provide a more comprehensive view of an employee’s movements and behaviors.

The car assignments dataset contains the information about the employees who own GasTech cars for personal and business purposes. It contains 44 data points with 4 columns each. Adetailed dimension, dimension description, dimension category for each dataset including the geospatial maps is tabulated in table 1.

Table 1: Datasets and their dimensions along descriptions

Dataset	Dimension	Dimension Description	Dimension Type
tourist map	.jpg image	Image of notable locations inthe city.	Image
abila	abila	Street coordinates andlabels for the city.	ESRI project files
kronos	kronos	Geospatial data with islandshape.	ESRI project files
car- assignments	LastName	Last name of employee(text),38 unique labels.	Categorical44 non-unique labels
	FirstName	First name of employee(text), 43 unique labels.	Categorical44 unique labels
	CarID	Numeric label.	Categorical(0-35) or blank (if employee title is “truck driver”)
	CurrentEmployeeTyp e	Text label of employeeclassification.	Categorical44 non-unique labels
	CurrentEmployeeTitle	Text label of title.	Categorical44 non-unique labels
cc_data	timestamp	Time (date, hour andminute).	Interval1490 non-unique values.
	location	Text label of a store, restaurant or establishment.	Categorical1490 non-unique values.
	price	Numeric value for the costcharged to a specific card.	Interval1490 non-unique values
	last4ccnum	Numeric label.	Categorical4 digit label, 1490 non-unique labels.

The association of the datasets is illustrated in figure 1 below with some important attributes and metadata.

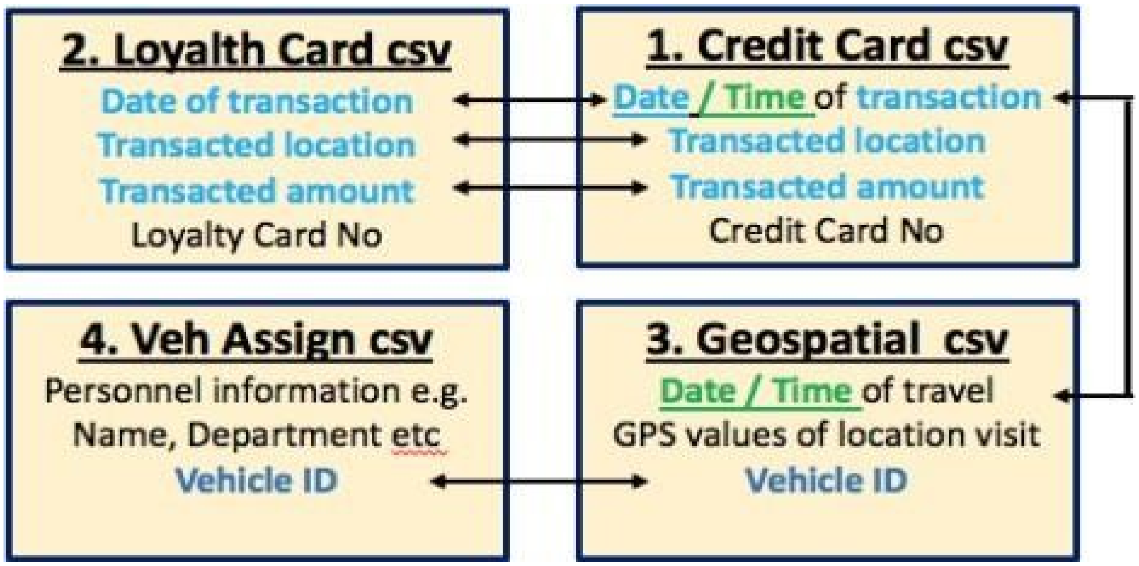

Figure1: Datasetrelationship

The loyalty_data.csv dataset contains 1,392 rows and 4 columns.

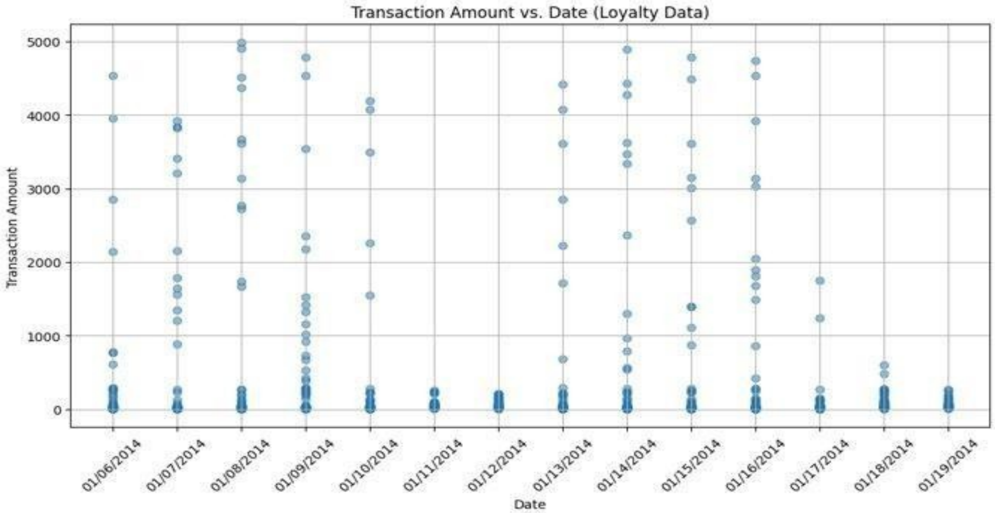

Figure3:Scatterplotofloyaltytransactionsand date

The cc_data.csv dataset contains 1,490 rows and 4 columns.

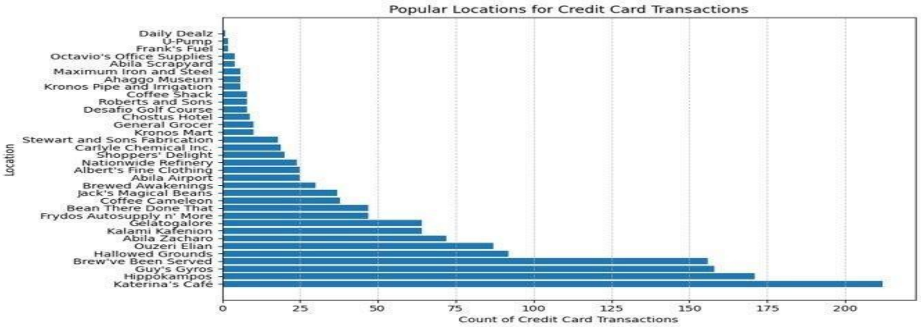

Figure4:Distributionofcredit-basedtransactionsondifferentlocations

We also checked the clean of the datasets as most of the datasets are categorical (non- nominal) values which were exposed to contain some special characters like apostrophes and other characters that were not recognized by our processing program. We have carefully handled this without losing the content of the data and removing important information that can affect our data. However, overall, the data quality at hand is good and does not contain a major issue that can impact our project work.

IV. RESEARCH QUESTIONS AND GOALS

The primary goal of this project was to develop an engaging visual analytics application specifically designed for GAStech employees. This application enables users to identify patterns and gain insights from the structured datasets detailed in Section 3 of the Data Report. Our key focus was to enhance user accessibility, allowing for intuitive exploration and analysis of the data. This was accomplished by implementing visual methods that present a clear and easily comprehensible summary of crucial data characteristics.

Each of these goals serves a critical function in building a robust visual analytics tool tailored for analyzing GAStech employee behavior. The key challenge across all goals is maintaining a balance between functionality and usability while ensuring impartial analysis and practical applications for security and risk management. Ultimately, the project should yield a tool that is both powerful for data analysis and easy enough to use, providing clear, unbiased, andactionable insights to improve organizational security.

V. SOLUTION DESIGN

The design of our GasTech Employee Behavior Analysis tool prioritizes user- friendliness and visual appeal. Our primary aim is to assist unusual pattern identification and understanding patterns for GasTech employees, focusing on transactions and locations. The layout is structured into three main components: a header, side menu, and main content, promoting easy navigation. Our commitment to unbiased analysis remains a key consideration throughout the design.

The header serves as the central hub for global navigation and high-level information. It includes quick access to essential tool features, such as filtering options, global search functionality, and visual cues to guide the user. The design is kept clean and simple, avoiding clutter, while ensuring important information is always visible.

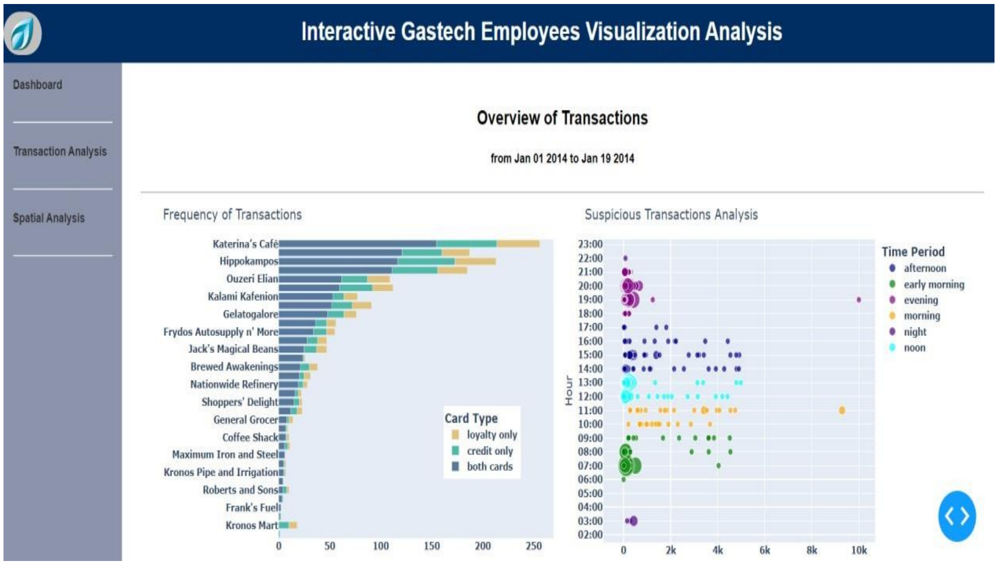

Figure5:Partialviewoflandingpageoftheapplication(dashboard)

VI. IMPLEMENTATION DETAILS WITH FINDINGS

As we discussed in the solution design, the project is divided into three main functionalities. These are dashboard, transaction analysis, and spatial analysis. Within the dashboard four overview visualizations are presented to give general insight regarding popular locations by transaction, suspicious analysis based on time period (hour) and price amount, hourly transactions by time period, and time-based transaction trends for different card types.

Using Dash Plotly's tools, the function dynamically constructs multiple bars on the horizontal chart. Each card type receives a different color, enhancing clarity for GasTech users.

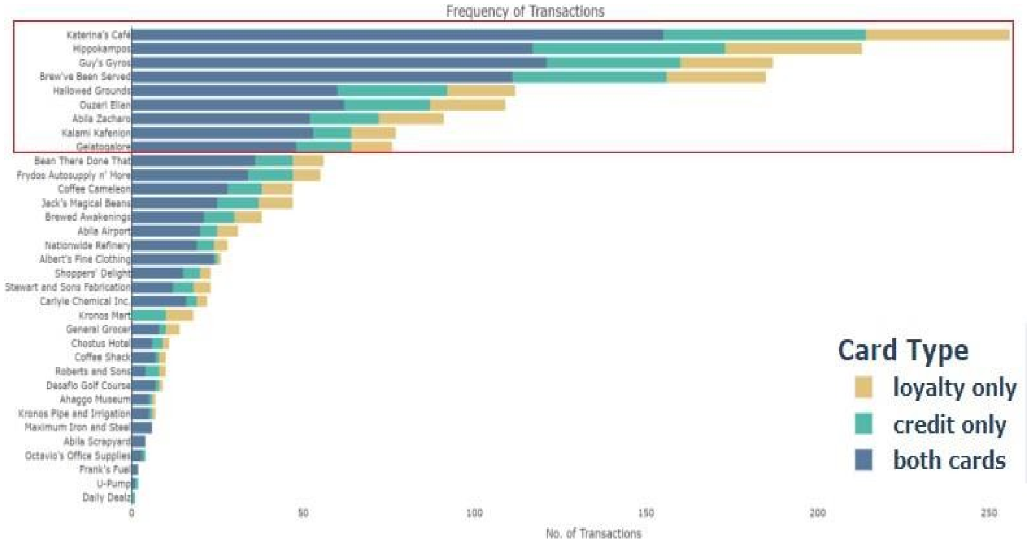

Figure8: Popularlocations

GasTech users can interact with the chart to check transaction frequencies at different locations. The chart's layout and stacked bars make it visually engaging for understanding transaction data. This empowers users to explore transaction data, gaining valuable insights into popular transaction locations and card types. The core goal is to make this exploration process accessible and meaningful for GasTech users, regardless of their level of technical expertise. Furthermore, the engaging interactive functionality allows users to view details upon hover. Figure 8 visually showcases popular locations across all transactions, including credit card transactions, loyalty card transactions, and both card transactions. This visualization aids in discerning patterns and trends related to transaction frequencies, providing a comprehensive understanding of card usage at various locations within the GasTech ecosystem.

The second core part of the GASTech interactive visualization analysis system is the transaction analysis. The transaction analysis mainly focuses on finding detailed transaction analysis at individual card level. In addition, this section also contains a second tab to visualize the mapping of credit card and loyalty card pairing both in a one to one and multiple pairing wise.

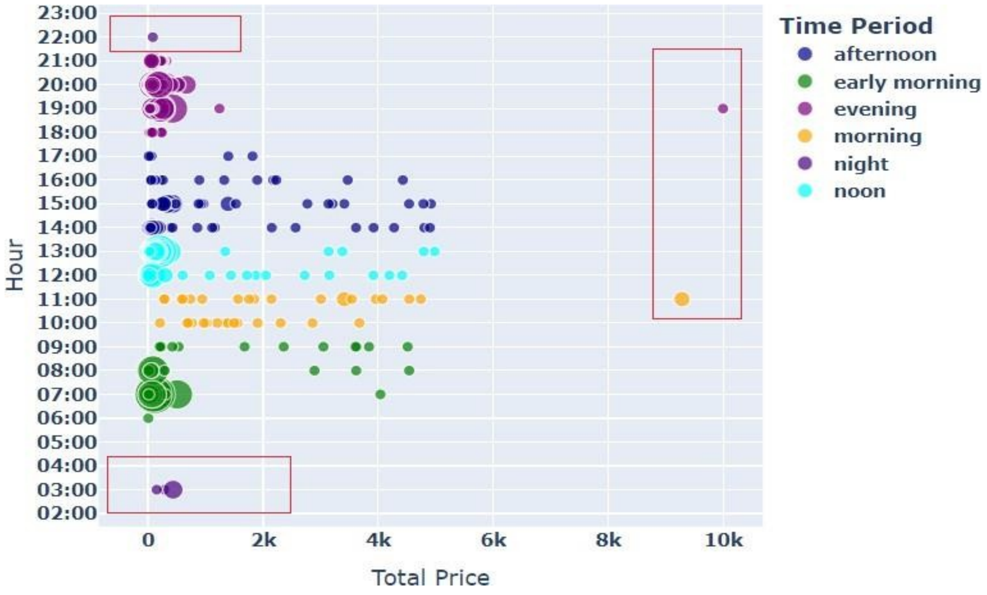

Figure9:Suspicioustransactionsbytimeandprice

As depicted in figure 9 above we found two types of suspicious transactions. Which unusual time period and amount transactions. GasTech users can interact with the scatter plot, gaining detailed information on specific data points by hovering over them. Essential details, including card type, total price, transaction count, hour, date, day, week status, location, and location type, are conveniently accessible. We can see the details by mouse hover and selection of a specific scatter plot. For example, we can see in figure 10 below for a scatter plot selection detail.

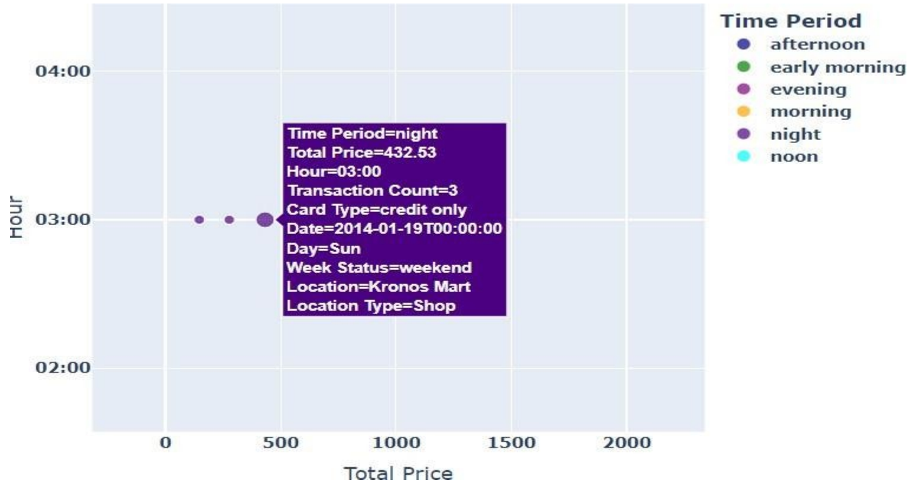

Figure10:Timebasedsuspicioustransactionwithdetails

In general, as we can observe the suspicious transactions encircled by reddish rectangle in figure 9 there are more 6 time-based anomalies which made in the night 1 transaction at 22:00 p.m. and 5 numbers of transactions at 3:00 a.m. The second type of transaction anomaly is based on the consumed amount that is two unusual transaction amounts 9k and 10k respectively.

Additionally, the interactive nature of the Plotly chart allows users to hover over any bar to reveal detailed information, such as the exact number of transactions for that hour and the corresponding time period. This interactivity enhances the user's ability to explore the data in greater depth, making it easier to identify patterns and anomalies.

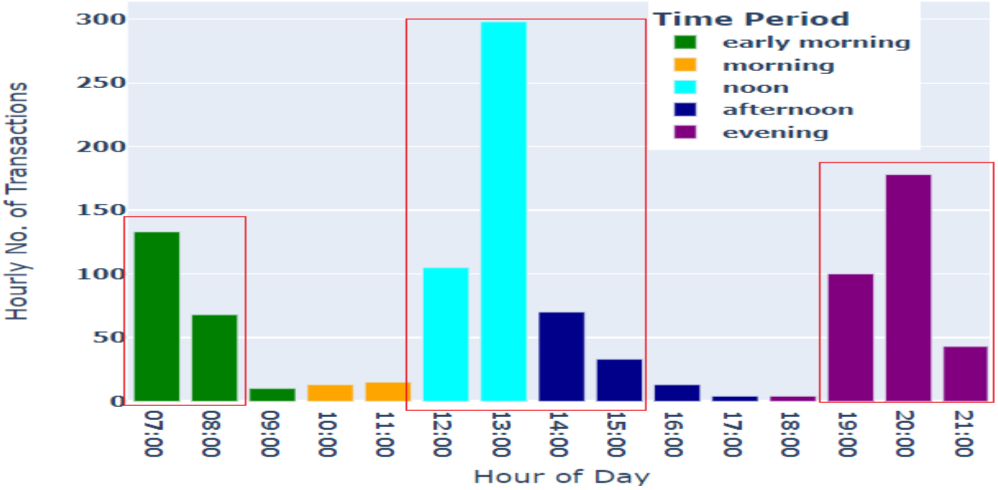

Figure11:Hourlytransactionbytimeperiod

GasTech users can interact with the bar chart, gaining detailed information on transaction counts for each hour. Essential details, including time period, card type, and hourly transaction counts, are conveniently accessible upon hovering over specific bars. This empowers users to explore the temporal aspects of transaction activity, facilitating a deeper understanding of when most transactions occur throughout the day for both credit and loyalty cards. Findings in figure 11 shows most transactions happen in morning, noon, and evening which is expected. Especially, number of transactions at 13:00 p.m. is the highest. This is the period most people it their lunch and it is reasonable to see high transaction at this time period.

Line chart dash plotly visualization is used to get insights on the trends of transactions overtime for each card type. The x-axis represents dates, the y-axis depicts the sum of transaction prices, and distinct colors differentiate between card types. Users can interact with the line chart, gaining detailed insights into the cumulative transaction amounts for each card type. The interactive features allow users to hover over points, view specific data points, and focus on individual card types by toggling the legend.

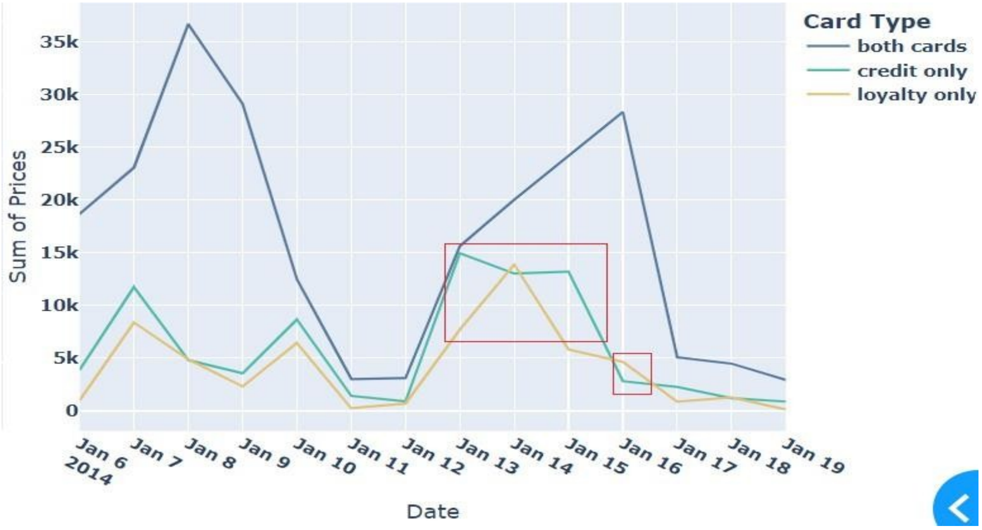

Figure12:Transactiontrendsbycard type

The scatter plot visualization tool plays a crucial role in aiding GasTech users to detect anomalies within individual credit card and loyalty card transactions. This interactive tool allows GasTech users to explore suspicious transactions by focusing on essential parameters such as time and transaction amount. Users can seamlessly interact with the scatter plot, selecting specific dates, locations, and credit card numbers for analysis. The tool provides a user-friendly interface, enabling users to hover over data points and gain detailed insights into transaction behavior. By scrutinizing deviations in time and amount parameters, GasTech users can effectively identify anomalies in transactions. This user-friendly tool aims to empower GasTech users with actionable insights and enhance their ability to analyze suspicious transactions effectively within GasTech's operations. An example highlights transactions associated with credit card numbers 4948, 8129, and 9551, which raised suspicion due to unusually high amounts in the thousands on a specific day.

VII. EVALUATION RESULTS

We conducted evaluation sessions for the GasTech interactive visualization tool with 2 participants via zoom meeting.ProcedureFirst a 10-minute comprehensive demonstration of the GasTech visualization tool was conducted, providing an overview of its features and functionalities.Participants were given 15 minutes each to explore the tool independently. Following the exploration, participants were assigned specific tasks to evaluate their understanding and usability of the tool.Task 1: Identify Anomalies: Both participants successfully identified anomalies using the suspicious transaction analysis based on time period and transaction amount.Task 2: Overall Transaction Trend Observation: None of the participants accurately observed the overall transaction trend by card type.Task 3: Open-Ended Observation: Participants provided insights based on their observation of individual card transaction analyses. Participant 1 answered approximately 80% correctly, while Participant 2 achieved roughly 65% accuracy.Feedback emphasized the need for improved theme and coloring, addressing visibility issues caused by background color.Participants recommended avoiding redundancy in visualizations across different pages, especially those directly mirrored in the dashboard.A suggestion was made to use static color mapping for categorical variables, as Plotly's random color generation caused inconsistency across visualizations.

VIII. DISCUSSION

The following discussion explores the effectiveness of anomaly detection based on transaction data, focusing on two key dimensions: time-based and amount-based anomalies. Additionally, it examines how integrating transaction data with GPS information enhances the ability to detect irregularities.

The integration of transaction and GPS data proved to be a valuable asset in anomaly detection. Aligning financial transactions with the physical movements of employees allows for a holistic view, enhancing the detection of irregularities. During the project, challenges related to inconsistent transaction records were addressed through the integration of GPS data. While initial improvements were made, refinement was necessary to address timestamp granularity disparities. Refinement efforts focused on disregarding seconds in timestamps, leading to a more accurate alignment between stop time, stop location, and transaction details. This meticulous approach enhanced the integrity of analytical insights. Despite improvements, doubts persist, particularlyconcerning the transaction data in the table. Rigorous comparisons with stop time and stop location were conducted, emphasizing the need for continuous improvement in data alignment and anomaly detection.The project's findings demonstrate the effectiveness of both time-based and amount-based anomaly detection. The integration of GPS data provided crucial contextual information that enhanced the system's overall ability to detect anomalies. While substantial improvements were made to overcome data inconsistencies, further refinement is necessary to achieve even greater precision in future anomaly detection efforts.

IX. CONCLUSION AND FUTURE WORK

GasTech's anomaly detection capabilities, particularly in time and amount-based anomalies, demonstrate effectiveness in enhancing transaction security within the gas industry. The integration of transaction and GPS data has proven to be a valuable asset, providing a holistic view of employee activities and transactions. The discussion highlighted successful identification of anomalies based on time, uncovering irregular transactions during non- standard hours, and amount, detecting unusual transaction amounts.Enhanced Data Alignment: Future efforts will focus on further refining data alignment, addressing any remaining discrepancies in transaction records. Continuous improvement in timestamp synchronization will contribute to the overall accuracy of anomaly detection.Advanced Anomaly Detection Models: Exploring advanced anomaly detection models and techniques will be essential for improving the platform's sensitivity to subtle irregularities. This includes leveraging machine learning algorithms to enhance anomaly identification.

REFERENCES

[1] P. Moreno-Sanchez, M. Zafar, and A. Kate, 'Listening to Whispers of Ripple: Linking Wallets and Deanonymizing Transactions in the RippleNetwork', Proc. Priv. Enhancing Technol., vol. 2016, Feb. 2016
[2] H.Wickham,Ggplot2:Elegantgraphicsfordataanalysis,2nded.Cham,Switzerland: Springer International Publishing, 2016.
[3] 'Re-Identification of "Anonymized" Data', Georgetown Law Technology Review, Apr. 12, 2017.
[4] R. Sen and S. Borle, 'Estimating the Contextual Risk of Data Breach: An Empirical Approach', J. Manag. Inf. Syst., vol. 32, pp. 314–341, Apr. 2015
[5] E. Tarameshloo, M. H. Loorak, P. W. L. Fong, and S. Carpendale, 'Using Visualization to Explore Original and Anonymized LBSN Data', Comput. Graph. Forum
[6] L. Rocher, J. M. Hendrickx, and Y.-A. de Montjoye, 'Estimating the success of re- identifications in incomplete datasets using generative models', Nat. Commun., vol. 10, no. 1, Art. no. 1, Jul. 2019
[7] L.Liu, M. Han, Y. Wang, and Y. Zhou, 'Understanding Data Breach: AVisualization Aspect', in Wireless Algorithms, Systems, and Applications, Cham, 2018
[8] A. Schouten, 'AILiteracy 101 — what is it and why do you need it?' Medium, Aug. 25,2020.
[9] C. Nwosu, 'Visualizing The 50 Biggest Data Breaches From 2004–2021', Visual Capitalist,Jun.01,2022.
[10] S. Schmeelk, 'Where is the Risk? Analysis of Government Reported Patient Medical Data Breaches', in IEEE/WIC/ACM International Conference on Web Intelligence - Companion Volume, New York, NY, USA, Oct. 2019
[11] 'DatesandTimesMadeEasywithlubridate|JournalofStatisticalSoftware'.
[12] H.Wickham,R.François,L.Henry,andK.Müller,'dplyr:AGrammarofData Manipulation'
[13] 'Buy GIS Software| ArcGISProduct Pricing - Esri Canada Store'.