R Shiny Investigative Tool into GASTech Personnel Disappearance

Aryah Umralkar Chopra
School of Computing and Information Systems,
Singapore Management University
aryahc.2020@mitb.smu.edu.sg

Rhoda Tong Min Ting
School of Computing and Information Systems,
Singapore Management University
rhoda.tong.2020@mitb.smu.edu.sg

Davmes Tan Chee Seng
School of Computing and Information Systems,
Singapore Management University
davmes.tan.2020@mitb.smu.edu.sg

ABSTRACT

A fictitious scenario was created as part of VAST Challenge 2021. A group of staff members from GASTech, an oil and gas company situated on an island known as Abila on Kronos, had gone missing mysteriously. A group known as Protectors of Kronos (POK) was the prime suspect into the disappearance. A series of unprocessed data were made available to the law enforcement agencies to investigate on. The data were split across three mini challenges, which our team had undertaken Mini Challenge 1 and 2.

Mini Challenge 1 consists of email correspondences, employee records and resumes, historical documents and news articles. The ojective of this mini challenge is to identify the complex relationships among the people and organisations, and possibly infer the disappearance of GASTech to any individuals/group who might be involved. Mini Challenge 2 consists of gps tracking, transaction and loyalty card records, together with car assignment records - linked to the gps tracking data. The objective of this mini challenge is to discover anomalies and suspicious activites that may require additional investigating.

The objective of this research paper is to share on the methods and models used to develop an online investigative tool where a law enforcement at Kronos and Tethys could use, to piece the raw data into useable information and evidences.

1. INTRODUCTION

A fictitious scenario was created as part of VAST Challenge 2021. A group of staff members from GASTech, an oil and gas company situated on an island known as Abila on Kronos, had gone missing mysteriously. A group known as Protectors of Kronos (POK) was the prime suspect into the disappearance. While Mini Challenge 1 and 2 provided a set of raw data that allow investigators to establish and identify complex relationships among the people and organisations, discover anomalies and suspicious activities, such investigative work may require humongous man hours and effort, without data analytics and visualisation.

We would be using Shiny R to develop an online investigative tool to aid in the analysis into the disappearance of GASTech Personnel, allowing investigators to explore information and inferential statistics derived from the unprocessed data available.

2. MOTIVATION

The motivation of this project would be two-fold. First, the data presented to the investigators were raw and unprocessed, and to link and derive insights from these data would require tremendous man hours and effort. Second, while insights could be derived and useful information could be formed, there would be a need to present the information in a visually appealing format to facilitate information dissemination and to allow quick collective appreciation of events among the investigators.

To this end, we would be looking to develop a R Shiny app based on three principles: (a) informative; (b) intuitive and; (c) interactive. Our 3Is principles would have the data undergoing baseline cleaning, making them into suitable formats for subsequent processing for information delivery. The user-interface would be made intuitive so that the investigator would be able to use the application without much references to our user guide. The online investigative tool would also be interactive, such that the investigator would be able to provide varied inputs into the formation towards the final visualisation report.

The R Shiny app would comprise of two main modules: (a) Exploratory Data Analysis allowing investigators to draw information such as transaction records, employee records, email correspondences and such; (b) Inferential Statistics allowing investigators to infer relationship linkages among user-selected employees, possible coded words within email correspondences within an identified group of personnel, their movements towards identified locations and possible anomalies at the locations and transaction analysis using both credit card and loyalty card data.

3. REVIEW AND CRITIC OF PAST WORKS

Taking reference on a submission by students from International Institute of Information Technology Hyderabad, they had derived an interface that uses geotools to show the paths of the cars, moving at a particular time or date. While it was a useful tool to visualise the movement of the vehicles, it was unable to plot the locations where the vehicles may had visited. This would be addressed in one of our module.

While credit card data and loyalty card data were provided for, to the investigators, there are no direct linkages between both set of data, except for the locations and price of item. The timestamp for both data are recorded differently with the credit card indicating the time of transaction while loyalty data indicating date of transactions only. Reviewing past work from members of University of Calgary, they attempted Parallel Coordinates plot by linking the locations, timestamp, price and employees, to establish possible linkages among the four variables. Inspired by their work and the sharing from Prof Kam on using Parallel Coordinates Plot on R, we would attempt to create parallel plots and allow the investigators to decide the variables that they are keen to establish linkages with.

This application and charts, focused on the GPS movements of the vehicle. This provides the investigator with the time and location dimension, however, there will be the question unanswered,'What were they doing there?' This question was addressed in the first mini-challenge data of The Vast Challenge.

Submission from the University of Buenos Aires documented suspicious emails by combining the use of a visual network plot of the participants in an email chain and corresponding a table of emails. Like with most submissions, they were static and only highlighted emails that they deemed as suspicious with the search by employee. The merging of a visual and table representation of the data would be enhanced in our application so that the investigator can not only search by person, but also by key words (or both) that his analysis has deemed suspicious.

In addition to that, in the submission by Tianjin University, their application showed the movement of the email with time between recipients. This was narrowed down between 2 recipients only when certain emails are sent to many people at once or replied to many other people. This inspired us to build a network that shows the entire network of the particular subject.

While the email subjects raise suspicion and the analysis was subsequently based their search, the role that each of the employee plays in the network was not studied in the submissions. This will be built on in our application.

4. DESIGN FRAME

Our R Shiny Investigative Tool's design would be based on our 3Is principle.

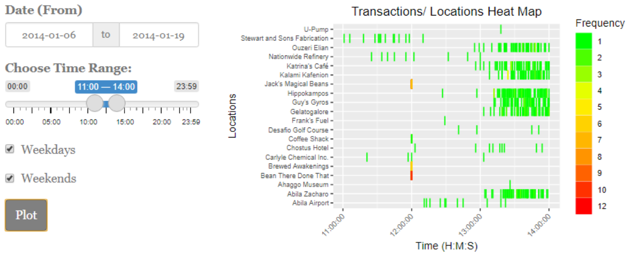
Informative - A series of informative modules would be made available under the Exploratory Data Analysis (EDA) tab, comprising of: (a) Locations Exploration; (b) Transactions Exploration; (c) Cards Exploration; (d) GASTech Employees Information; and (e) Email Correspondences. Details of each modules would be covered under the sub-sections later.

Intuitive - The entire UI design would be simple and intuitive, allowing investigators to use it without frequent references to our user guide. This would be done by ensuring that appropriate input methods were used and that the UI would be kept as clean and simple as possible.

Interactive - The respective modules would require specific inputs from the investigators before a representative visualisation could be produced. It would be made interactive such that the investigators would have the flexibility to select various variables or inputs to create the desired visualisation. Similarly, the investigators would have the option to save the produced visualisation separately.

5. EXPLORATORY DATA ANALYSIS

5.1 Locations - For Exploration!



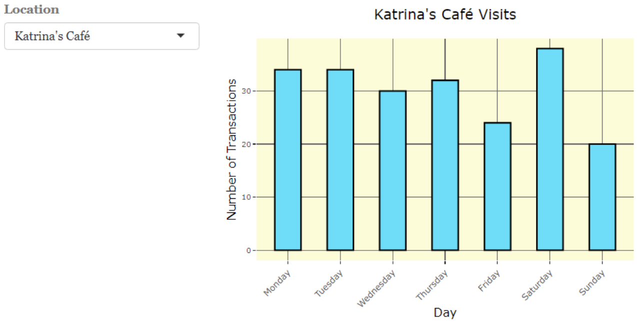
Date (From): 2014-01-06 to 2014-01-19

Choose Time Range: 11:00 — 14:00

☑ Weekdays

☑ Weekends

The intention of this module would be to allow the investigators to explore the popularity of the locations through the analysis of transaction records. It would be conducted in a manner by aggregating the transaction records based on the time of transactions. The transaction counts would be plotted against the aggregated time of transactions and locations using a heatmap to infer the peak periods for the locations. The selected records would be filtered based on the range of date inputs, time range of transactions, with the ability to toggle between weekdays, weekends or both.

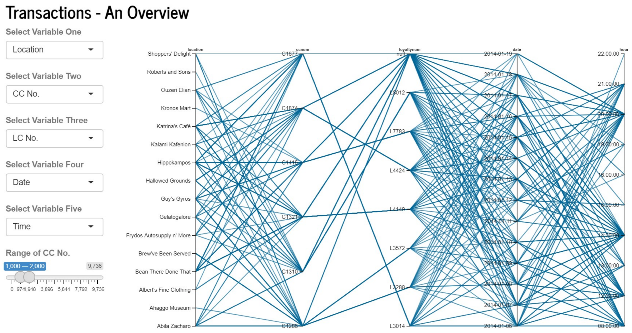


Location: Katrina's Café

The function of this module would be to identify the operating days of the locations through the analysis of credit card transactions, through the aggregation of records based on the day of transactions. The output of this visualisation would suggest the days where the locations would be having transactions, thus inferring their operating days of the week.

5.2 Transactions - For Exploration!

The intention of this module would be to allow investigators to explore all credit card and loyalty card transactional records at each location. The type of visualization chosen enables the investigators to see the linkages between all variables in one view.



Transactions - An Overview

Select Variable One: Location

Select Variable Two: CC No.

Select Variable Three: LC No.

Select Variable Four: Date

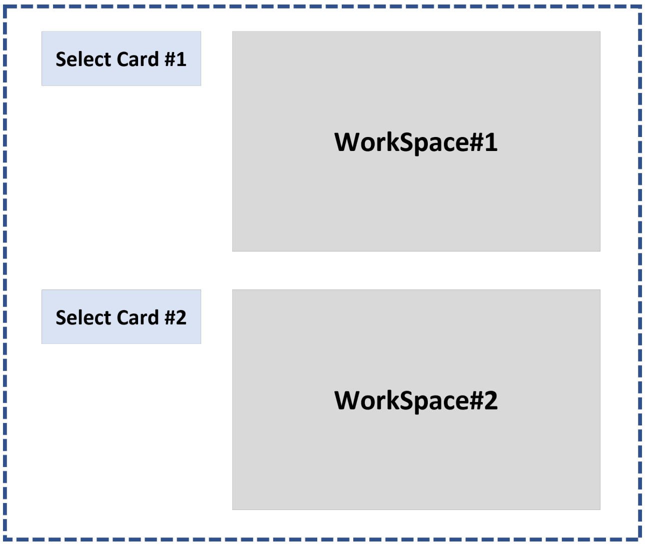
Select Variable Five: Time

Range of CC No.: 1,000 — 2,000

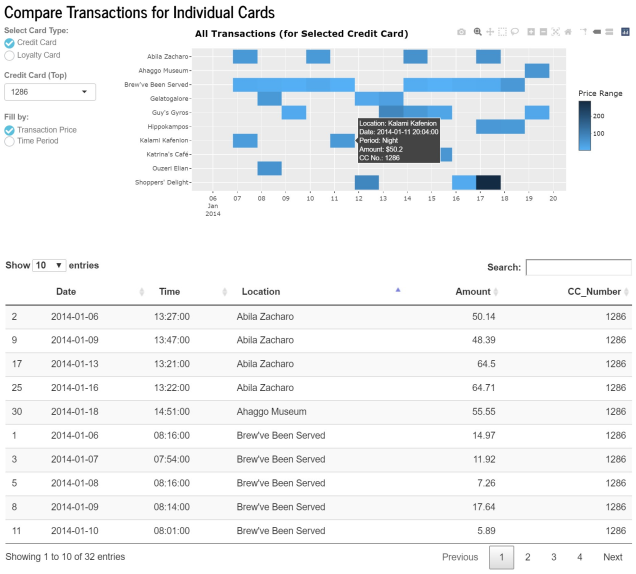
The plot has multiple functions which would allow the investigator to customize his/her view and get more insights, such as (1) Variables to be included can be selected, (2) Variables can be reordered, (3) Range of Credit Card Numbers to display records for can be selected, (4) Specific ranges of each variable can be selected so that only associated links would be shown, (5) Selected ranges can be glided along the axis to see how the connections change dynamically. These customizations and flexibilities are important so that the user can zoom in and out on selected areas of focus, especially when exploring raw data sets. If a certain value, say for a specific credit card number, appears in multiple observations, its associations would be clearly displayed in the form of the series of lines.

5.3 Individual Card Transactions

This module allows the investigator to visualize the transactions made by individual cards for a more in-depth exploration. The page allows the investigator to select two different cards to be displayed, akin to a workspace for comparison between cards. This is essentially a deep-dive compared to the parallel coordinate plot, which gives an overview.



The investigator could select the card type, credit card or loyalty card. If credit card is selected, an additional field would appear to allow selection of type of fill for the tiles.



Compare Transactions for Individual Cards

Select Card Type:
● Credit Card
○ Loyalty Card

Credit Card (Top): 1286

Fill by:
● Transaction Price
○ Time Period

Location: Kalami Kafenion
Date: 2014-01-11 20:04:00
Period: Night
Amount: $50.2
CC No.: 1286

Show 10 entries

Search:

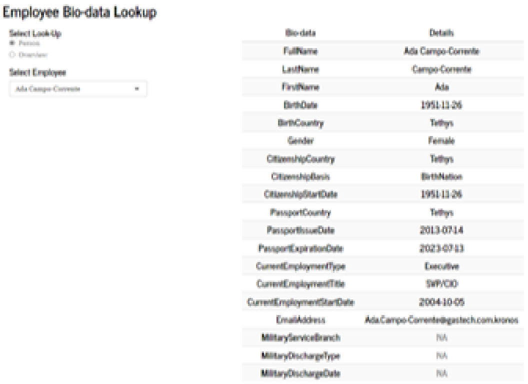
	Date	Time	Location	Amount	CC_Number
2	2014-01-06	13:27:00	Abila Zacharo	50.14	1286
9	2014-01-09	13:47:00	Abila Zacharo	48.39	1286
17	2014-01-13	13:21:00	Abila Zacharo	64.5	1286
25	2014-01-16	13:22:00	Abila Zacharo	64.71	1286
30	2014-01-18	14:51:00	Ahaggo Museum	55.55	1286
1	2014-01-06	08:16:00	Brew've Been Served	14.97	1286
3	2014-01-07	07:54:00	Brew've Been Served	11.92	1286
5	2014-01-08	08:16:00	Brew've Been Served	7.26	1286
8	2014-01-09	08:14:00	Brew've Been Served	17.64	1286
11	2014-01-10	08:01:00	Brew've Been Served	5.89	1286

Showing 1 to 10 of 32 entries

Location is plotted against each day. If there was a transaction at the location for that day, a tile would appear. For this module, the tile plot is chosen as its uniform tiles allow a straightforward view of whether transactions were made consecutively for each day, or on separate days. The color intensity of the tile would give an indication of the magnitude of the transaction amount - the darker the intensity, the higher the amount. Hovering over each tile would display a tooltip indicating the details of that particular transaction. A data table for the selected card would also appear below the tile plot to list down all observations, it can be sorted and filtered, to allow for the investigator's further exploration.

5.4 Employees of GasTech

This module explores the general data of the employees. Like with most investigative portals, this tab will allows the investigator to understand basic bio-data of the employee in question. As shown below, the tab will react to the drop down under *Select Employee* and update to display the respective information.



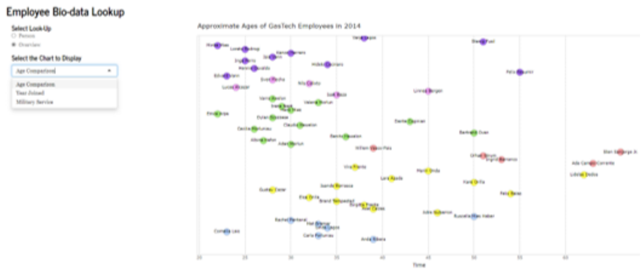
Employee Bio-data Lookup

Select Look-Up
● Person
○ Overview

Select Employee: Ada Campo-Corrente

Bio-data	Details
FullName	Ada Campo-Corrente
LastName	Campo-Corrente
FirstName	Ada
BirthDate	1951-11-26
BirthCountry	Tethys
Gender	Female
CitizenshipCountry	Tethys
CitizenshipBasis	BirthNation
CitizenshipStartDate	1951-11-26
PassportCountry	Tethys
PassportIssueDate	2013-07-14
PassportExpirationDate	2023-07-13
CurrentEmploymentType	Executive
CurrentEmploymentTitle	SVP/CIO
CurrentEmploymentStartDate	2004-10-05
EmailAddress	Ada.Campo-Corrente@gastech.com.kronos
MilitaryServiceBranch	NA
MilitaryDischargeType	NA
MilitaryDischargeDate	NA

The tab was extended with three charts. On selection of *Overview*, this allowed the comparison of biodata that is comparable accross, for example, the year joined in the company, ages and military service. This are useful as they allow the investigator to understand possible relationship or acquintances outside of GasTech.



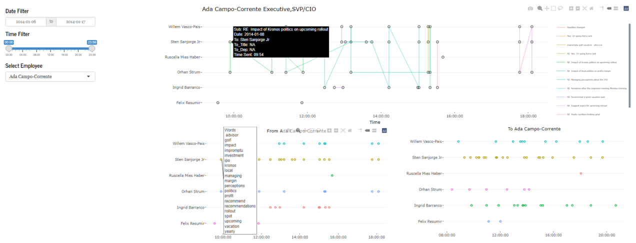
Employee Bio-data Lookup

Select Look-Up
○ Person
● Overview

Select the Chart to Display
Age Comparison
- Age Comparison
- Year Joined
- Military Service

5.5 Email Correspondence

This module is a exploratory part, setting the ground for the Inferential Statistics in the network analysis. Investigators will be able to filter by data, time and employee to understand their email conversation. The main plot will show all emails to and from the employee across the day, while the 2 subplots highlight the unique words shared between filtered employee and their recipients.



Date Filter
2014-01-06 to 2014-01-17

Time Filter
00:00 – 23:59

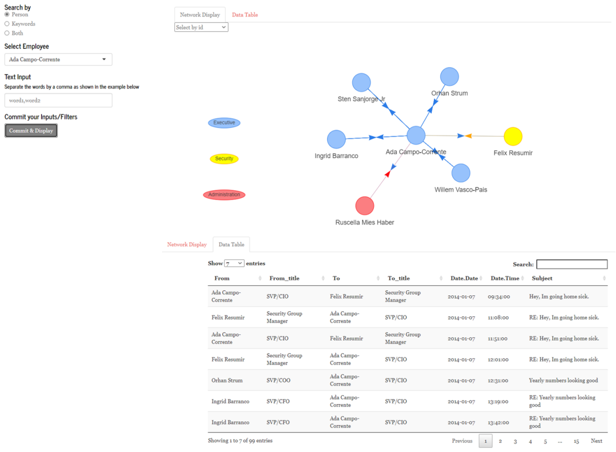
Select Employee
Ada Campo-Corrente

The subplots also highlight oddly timed emails and frequency of conversation between the filtered employee and others.

6. INFERENTIAL STATISTICS

6.1 Email Network Analysis

This module will allow the investigator to make searches by person, keyword or both. The search will bring up both the visual representation of the conversation and also the corresponding data table.



Search by
● Person
○ Keywords
○ Both

Select Employee
Ada Campo-Corrente

Text Input
Separate the words by a comma as shown in the example below
word1,word2

Commit your Inputs/Filters
Commit & Display

Network Display | Data Table

Show 7 entries

Search:

From	From_title	To	To_title	Date.Date	Date.Time	Subject
Ada Campo-Corrente	SVP/CIO	Felix Resumir	Security Group Manager	2014-01-07	09:34:00	Hey, Im going home sick.
Felix Resumir	Security Group Manager	Ada Campo-Corrente	SVP/CIO	2014-01-07	11:08:00	RE: Hey, Im going home sick.
Ada Campo-Corrente	SVP/CIO	Felix Resumir	Security Group Manager	2014-01-07	11:51:00	RE: Hey, Im going home sick.
Felix Resumir	Security Group Manager	Ada Campo-Corrente	SVP/CIO	2014-01-07	12:01:00	RE: Hey, Im going home sick.
Orhan Strum	SVP/COO	Ada Campo-Corrente	SVP/CIO	2014-01-07	12:31:00	Yearly numbers looking good
Ingrid Barranco	SVP/CFO	Ada Campo-Corrente	SVP/CIO	2014-01-07	13:19:00	RE: Yearly numbers looking good
Ingrid Barranco	SVP/CFO	Ada Campo-Corrente	SVP/CIO	2014-01-07	13:42:00	RE: Yearly numbers looking good

Showing 1 to 7 of 99 entries

Previous 1 2 3 4 5 ... 15 Next

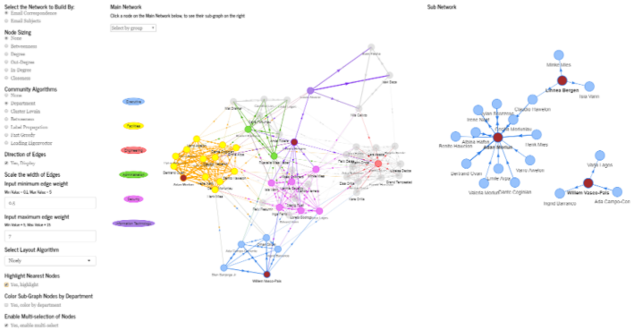
In the submission by University of Abila, the network was colored by suspected POK members too. As this application is meant for analytic deduction by the user, the nodes are colored by department to see inter-department activities. The network is directed. This will highlight nodes of different nature: Authoritative(tend to receive more emails), hubs (tend to receive more emails) and other critical intermediary nodes that facilitate the transfer of information. In addition to that, the network is interactive, the nodes highlight their nearest neighbours on click, showing flow of information. The nodes can be moved around to handle overlap issues.

This tab works in cohesion with the *Networks* tab.

6.2 Networks

This tab will allow the investigator to see the overall email network at GasTech, and run layout, community detection algorithms and identify key nodes through centrality measures. There are one of 2 options the investigator can see, network by email converstaion or network by content of the email subjects.

The side panel contains all the measures that can be changed, this includes node sizing, community algorithm and layout algorithms.



Select the Network to Build By:
● Email Correspondence
○ Email Subjects

Node Sizing
● None
○ Betweenness
○ Degree
○ Out-Degree
○ In Degree
○ Closeness

Community Algorithms
○ None
● Department
○ Cluster Lovain
○ Betweenness
○ Label Propagation
○ Fast Greedy
○ Leading Eigenvector

Direction of Edges
☑ Yes, Display

Scale the width of Edges
Input minimum edge weight
Min Value = 0.1, Max Value = 5
0.5

Input maximum edge weight
Min Value = 5, Max Value = 15
7

Select Layout Algorithm
Nicely

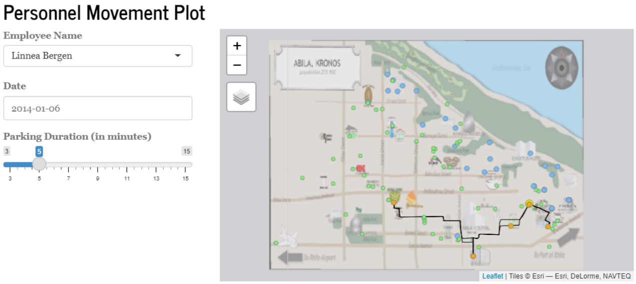
Highlight Nearest Nodes
☑ Yes, highlight

Color Sub-Graph Nodes by Department
☐ Yes, color by department

Enable Multi-selection of Nodes
☑ Yes, enable multi-select

This tab allows the use of network analysis algorithm to identify key nodes within the network. Similar to previous, the user has freedom to enable multi-select, highlight nearest nodes and also direct interaction with the network.

6.3 Employee Movement Plot



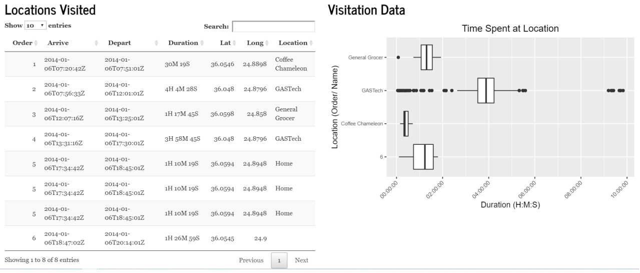
Personnel Movement Plot

Employee Name: Linnea Bergen

Date: 2014-01-06

Parking Duration (in minutes): 5

The data preparation for this module would be to identify Points-of-Interests (POI), using the gps records. Since the recording of gps would only occur while the vehicle is in motion, whenever there is a time interval (3g: 3 minutes), would be a good assumption that the vehicle is in parking mode, thus indicating it as a POI. The time interval may be adjusted based on the investigators' tolerance between 3 and 15 minutes. The POIs would be differentiated between homes of employees and other locations, thus allowing tmap to plot the route taken, and the stops taken.



Locations Visited

Show 10 entries Search:

Order	Arrive	Depart	Duration	Lat	Long	Location
1	2014-01-06T07:20:42Z	2014-01-06T07:51:01Z	30M 19S	36.0546	24.8898	Coffee Chameleon
2	2014-01-06T07:56:33Z	2014-01-06T12:01:01Z	4H 4M 28S	36.048	24.8796	GASTech
3	2014-01-06T12:07:16Z	2014-01-06T13:25:01Z	1H 17M 45S	36.0598	24.858	General Grocer
4	2014-01-06T13:31:16Z	2014-01-06T17:30:01Z	3H 58M 45S	36.048	24.8796	GASTech
5	2014-01-06T17:34:42Z	2014-01-06T18:45:01Z	1H 10M 19S	36.0594	24.8948	Home
5	2014-01-06T17:34:42Z	2014-01-06T18:45:01Z	1H 10M 19S	36.0594	24.8948	Home
5	2014-01-06T17:34:42Z	2014-01-06T18:45:01Z	1H 10M 19S	36.0594	24.8948	Home
6	2014-01-06T18:47:02Z	2014-01-06T20:14:01Z	1H 26M 59S	36.0545	24.9	

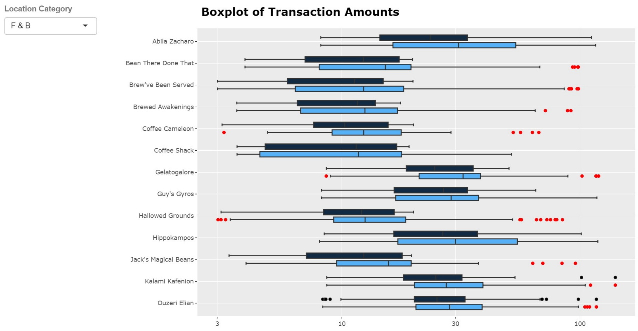
Showing 1 to 8 of 8 entries Previous 1 Next

Visitation Data

The sub-module would be an extension where the selected gps plot would be sequenced based on the time of activity, and further identified based on marked locations such as their homes, GASTech or prominent locations that were named prior. The subsequent box plot would be an analysis of the duration of stop, by visitors, to the locations.

6.4 Transaction Amount Analysis

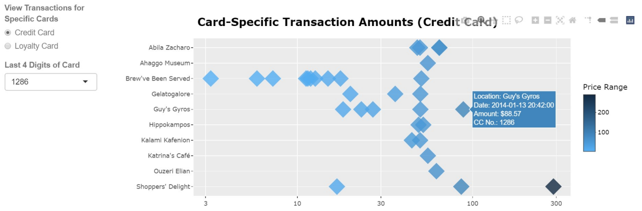
This module provides statistical analysis on the transactions made at each location, by each card type. The distribution of the spread of the transaction amounts is shown by boxplots. Outliers are highlighted in red to gain the investigator's attention.



Location Category: F & B

The plots are categorized by the location types, as we believe transaction amounts should be compared with those made at similar locations (For e.g. F&B with F&B, Industrial with Industrial). It would not be logical to compare a transaction made at a supermarket, with one made at a chemical production factory. This would be useful if the investigator wishes to understand what is a normal range of transaction price for each location, and which transactions deviated from the normal ranges. An additional value-added feature that can be derived from this plot, is the investigator can identify credit card and loyalty card pairs as they would have the same transaction price, on the same day.

A further deep dive can be performed with the specific card transactions scatter plot positioned below the overall boxplot.



View Transactions for Specific Cards

● Credit Card

○ Loyalty Card

Last 4 Digits of Card: 1286

Location: Guy's Gyros
Date: 2014-01-13 20:42:00
Amount: $88.57
CC No.: 1286

This is a scatter plot of all transactions made by the card selected by the investigator. Each point represents one transaction, and positioned according to the transaction amount. How the amounts made at the same location are relative to each other can be seen by their colors as well - the darker the intensity, the higher the amount. As it is possible to have multiple points clustered close together, some transparency is set so that any overlaps are clearly visible.

7. FUTURE WORKS

For the email network analysis, the data did not provide us with the actual content of the email. While subjects can be useful for the investigator to make reasonable judgments, there could be confirmation biases and logical fallacies with that for e.g. suspected members of the POK using similar subjects for different email content.

This can be addressed with the provision of the email content data. In addition to that, the text network can become more accurately with better data.

The parallel coordinate plot in the Transactions Analysis Overview page is currently only based on pure transactions data. Further insights could be drawn, if the plot could be joined with data on vehicle IDs which made stops at that particular location, in that particular period. This would allow the investigator to narrow down which are the possible vehicle IDs that could have made the transaction, or owned the credit card or loyalty card. Linking to vehicle ID would enable the investigator to link to the personnel based on who the vehicle is assigned to.

An improvement towards the Employee Movement Plot could include further analysis on the movement route, in terms of the distance/ speed, which may suggest the vehicle were kept in idle mode, which is currently unable to detected under the current module. In addition, the streets of Abila could be identified based on the movement routes, and thus, and analysis could be conducted whether the vehicle was travelling in a suspicious manner.

8. ACKNOWLEDGEMENTS

The authors wish to thank Professor Kam Tin Seong of Singapore Management University for his extensive guidance and support during this project.

References

[1] Avil, Pilar; Burgos, Valeria; Chikka, Guaymás Amalia 2014. VAST Challenge 2014 Mini Challenge 1. *https://www.cs.umd.edu/hcil/varepository/VAST%20Challenge%202014/challenges/MC1%20-%20Disappearance%20at%20GASTech/entries/University%20of%20Buenos%20Aires%20-%20Avila/*.

[2] Kam, Tin Seong 2020. Hands-On Exercise 8: Creating Parallel Coordinate Plot. *https://rpubs.com/tskam/PCP*.

[3] Pochampally, Yashaswi; Yarrabelly, Navya; Chikka, Veera Raghavendra and Karlapalem, Karlapalem 2014. VAST Challenge 2014 Mini Challenge 2. *http://visualdata.wustl.edu/varepository/VAST%20Challenge%202014/challenges/MC2%20-%20Patterns%20of%20Life%20Analysis/entries/International%20Institute%20of%20Information%20Technology%20Hyderabad/*.

[4] Sahaf, Zahra et al. 2014. VAST Challenge 2014 Mini Challenge 2. *http://visualdata.wustl.edu/varepository/VAST%20Challenge%202014/challenges/MC2%20-%20Patterns%20of%20Life%20Analysis/entries/University%20of%20Calgary/*.

[5] Zhuang, Cai; Mengyao Chen;, Hanqing Zhao; Ying Zhao; Fangfang Zhou; Kang Zhang; 2014. VAST Challenge 2014 Mini Challenge 1. *https://www.cs.umd.edu/hcil/varepository/VAST%20Challenge%202014/challenges/MC1%20-%20Disappearance%20at%20GASTech/entries/Tianjin%20University%20-%20Cai/*.