VAST Challenge 2014

Mini-Challenge 2 Guide for Reviewers

Mini-Challenge 2 uses a collection of spreadsheets and geospatial files to detail the activities of GASTech employees, including staff that are members of the Protectors of Kronos (POK) as they prepare for their daring kidnapping of GASTech employees. Hidden among the daily routines of regular employees are activities such as practicing the kidnapping route and monitoring the targets’ homes to gather information. This dataset identifies the potential members of POK and their intended targets.

# The Task

The contestants are asked to provide answers to three questions, using visual analytic techniques to support their investigation:

**MC2.1** – Describe common daily routines for GASTech employees. What does a day in the life of a typical GASTech employee look like? Please limit your response to no more than five images and 300 words.  
  
**MC2.2** – Identify up to twelve unusual events or patterns that you see in the data. If you identify more than twelve patterns during your analysis, focus your answer on the patterns you consider to be most important for further investigation to help find the missing staff members. For each pattern or event you identify, describe

1. What is the pattern or event you observe?
2. Who is involved?
3. What locations are involved?
4. When does the pattern or event take place?
5. Why is this pattern or event significant?
6. What is your level of confidence about this pattern or event? Why?

# Please limit your answer to no more than twelve images and 1500 words. ****MC2.3**** – Like most datasets, the data you were provided is imperfect, with possible issues such as missing data, conflicting data, data of varying resolutions, outliers, or other kinds of confusing data. Considering MC2 data is primarily spatiotemporal, describe how you identified and addressed the uncertainties and conflicts inherent in this data to reach your conclusions in questions MC2.1 and MC2.2. Please limit your response to no more than five images and 300 words.

# The Data

The following datasets are provided to contestants:

* A comma separated values file (CSV) containing the GPS tracking data
* A CSV file of credit card data for GASTech employees
* A CSV file of loyalty card data for GASTech employees
* A spreadsheet listing employees, their car, and their position at GASTech
* Geospatial files for the city and country which provide context for the GPS data

## Notes to Reviewers

There are numerous statistical analyses which could be done to this dataset, but they are not compelling in and of themselves. There must be an explanation of what the statistics mean. There must be a “so what?”

Also, though the majority of this data was generated randomly, it is likely that contestants will read meaning into events where it was not intended. If they follow a certain train of logic, then that could still be interesting. If they identify an event but don’t attempt to make sense of it, then that is not interesting.

# Answers to Questions

## Question MC2.1

*Describe common daily routines for GASTech employees. What does a day in the life of a typical GASTech employee look like? Please limit your response to no more than five images and 300 words.*

The people represented in this dataset are very much creatures of habit and, in general, have a fairly predictable routine. On weekdays, they get up in the morning to drive to work while stopping off for coffee along the way. At lunch they will go to lunch then return to work. In the evening they may go out for dinner then back home. On weekends, they may go to lunch, go shopping, and go out to dinner in the evening.

GPS data for the trucks show them running through a set itinerary which varies depending on the driver. They start and end at the GASTech office, frequently go to the airport, and may stop for lunch.

The credit card and loyalty card data will also reflect the daily routines of the employees as they make purchases on each of their stops. Each type of location (food, store, coffee shop, etc.) has a certain cost range that purchases will fall within.

Employee residences are not explicitly provided but can be gleaned from the data provided (cars begin and end each day at the same location). This information is essential to identify interesting patterns for other questions.

## Question MC 2.2:

*Identify up to twelve unusual events or patterns that you see in the data. If you identify more than twelve patterns during your analysis, focus your answer on the patterns you consider to be most important for further investigation to help find the missing staff members. For each pattern or event you identify, describe*

1. *What is the pattern or event you observe?*
2. *Who is involved?*
3. *What locations are involved?*
4. *When does the pattern or event take place?*
5. *Why is this pattern or event significant?*
6. *What is your level of confidence about this pattern or event? Why?*

## *Please limit your answer to no more than twelve images and 1500 words.*

## Several anomalous patterns have been inserted into the routine data. Not all these patterns support the threat in the scenario. The following lists the employee, their car id #, and the dates the event occurs.

* Anomalous but benign patterns:
  + Employees Isande Borrasca (#7) and Brand Tempesta (# 33) having an affair. Multiple times they go to a hotel over lunch, usually leaving and returning at times offset from each other. Happens on days: January 8, 10, 14, and 17
  + Employee Bertrand Ovan (#29) cruising around town on Jan 11
  + Kanon Herrero (#22) (badging office) and Elsa Orilla (drill tech) (#28) are dating. They go to lunch together every day. On Jan 18, they go to several locations together. Kanon usually pays so Elsa will not have corresponding card data for those occasions.
    - Elsa drives on January 9,14, and 17
    - Kanon drives on January 6, 7,8,10. 13, 15, and 16
  + Executives have a practice golf session on the January 12 then golf again with Sanjorge on the January 19
  + Employee Lucas Alcazar (# 1) works after hours on January 6,7,8,15,16, and 17
  + A large party for all the engineers and IT is held on on January 10. They all go to one house for the party around the same time.
  + IT group manager - Linnea Bergen - takes the IT group out for lunch on January 17 so she has a high cost lunch, high loyalty points. The participants carpooled with non-GPS car so there are no GPS records.
  + Axel Cazas (#9) has a spotty GPS with frequent gaps in his data.
* **Anomalous and part of plot:**
  + Inga Ferro (#13), Loreto Bodrogi (#15), Hennie Osvaldo (#21), Minke Mies (#24) do several practice drives from the office to various safe houses
  + Loreto Bodrogi (#15), Isia Vann (#16), Hennie Osvaldo (#21), Minke Mies (#24) carry out surveillance on executive houses in 3 hour shifts on January 6,7,9,10,11,13, and 14 in the middle of the night when nobody else is driving
  + Minke Mies steals the credit card from Lucas Alcazar (a help desk worker). Lucas still has his transactions as planned up until January 13 COB. Then he makes no card purchases till January 16, when his replacement credit card arrives. He still has loyalty data but no credit card data. On Jan 13. Minke makes a couple of small test purchases - couple of dollars online. Then he buys gas. These take place while Lucas is demonstrably elsewhere. Then he makes the large outsized purchase ($1000) at Frydo’s, the store that is the front for the APA. Then he makes no other purchases on the Lucas Alcazar card.

## Question MC 2.3:

### *Like most datasets, the data you were provided is imperfect, with possible issues such as missing data, conflicting data, data of varying resolutions, outliers, or other kinds of confusing data. Considering MC2 data is primarily spatiotemporal, describe how you identified and addressed the uncertainties and conflicts inherent in this data to reach your conclusions in questions MC2.1 and MC2.2. Please limit your response to no more than five images and 300 words.*

This dataset contains data imperfections both intentional and unintentional. The important point is not that they found a data issue, but what they did with it. For instance:

* How did they handle missing or bad data? Did they drop a car completely? Did it cause false positives?
* Did they require a human in the loop at certain stages to coordinate the data sets and/or deal with issues in the data?
* Did they read patterns into data quality issues or otherwise assign meaning to truly random events?
* How did they resolve ambiguity to further their analysis?

# Supporting Information

The following information is provided for additional background.

## A Tour through the Data

*A spreadsheet of GPS tracking data for cars used by employees*

This comma separated file (CSV) contains tracking data for the company provided cars.

*Car assignment spreadsheet*

This file matches up the car ids with the employee assigned to it. It also provides the employee’s position in the company.

*A spreadsheet of credit card data for employees*

Purchases on the company card are captured in this file. It is coordinated with the GPS file so that purchases are made when someone has gone to a location.

*A spreadsheet of loyalty card data for employees*

This spreadsheet matches up with the credit card data and indicates purchases made with a loyalty or “points” type card.

*Geospatial files providing context for the GPS data*

These files are usable in applications such as Google Earth to place the GPS data on the fictitious island of Kronos.

*Tourist map of the city (Abila)*

This map provides overall context for the city and locations in the GPS file but is not meant to match up to the GPS data.

## Data Relationships

In general, the geospatial files provide context to the GPS data. The credit card and loyalty data are based off the GPS data. People generally drive to a location first and then make a purchase. Credit card and loyalty data are generally the same except in cases where someone used only of the two or got cashback leading dollar amounts to not match up.