IS428 Assignment 2: To Be A Visual Detective

**Student Name:** *Erica Tan* **Student Email:** *erica.tan.2020@smu.edu.sg*

**Section #:** *G2*

# 1 Visualization Tools Used:

*(Write down the tools you used. For example, Tableau, D3.)*

Tableau, Rstudio

# 2 Tasks and Questions

*(When writing the following responses, you may consider* ***highlighting the key observations*** *and* ***try to make your observation as concise and precise as possible****.)*

## 2.1 Using just the credit and loyalty card data, identify the most popular locations, and when they are popular. What anomalies do you see? What corrections would you recommend to correct these anomalies? **Please limit your answer to <= 8 images and <= 500 words**. **(4 marks)**

Chart

Description automatically generated

After analysing the heatmap graph generated from the transaction data, I was able to gain valuable insights into the most popular locations based on transaction count. The heatmap allowed me to visualize transaction density at different locations, with darker-coloured boxes indicating higher numbers of transactions.

I identified three of the most popular locations based on transaction count: Brew've Been Served, Guy's Gyros, and Katerina's Café. These locations were consistently marked by darker-colored boxes, indicating a high volume of transactions.

|  |  |
| --- | --- |
| Brew’ve Been Served |  |
| Upon analysing the transaction data, 7am was the most popular timing for transactions on most days, except for days 10 and 17 where 8am was the more popular timing.  Furthermore, on day 14, the shop saw its highest number of transactions, with a total of 14 transactions taking place at 7am. This suggests that there may have been a special event where large orders were placed in bulk.  There were no transactions on weekends, indicating that the shop may not be open on those days or that customers tend to visit less frequently during weekend days. | |
| Guy’s Gyros |  |
| Upon analysing the transaction data for Guy's Gyros, 1pm and 8pm are the most popular timings for transactions on weekdays.  Additionally on days 9 and 19, there were higher than usual transactions at 8pm. This could be due to a variety of factors, such as a special event where orders were placed in bulk.  Furthermore, on weekends, there were relatively lower transactions, except for day 19. This could be due to lesser foot traffic as there are lesser office workers around. | |
| Katerina’s Cafe |  |
| Upon analysing the transaction data for this café, 1pm and 8pm are the most popular times for transactions, which corresponds to lunch and dinner time. It suggests that the busiest periods for the café are during meal times.  Additionally, Wednesdays have the lowest transactions during the week. This could be due to the office workers having a planned lunch or dinner elsewhere on Wednesdays and hence the workers would not purchase from the café.  Furthermore, the highest transactions throughout the week occurred on day 7 at 8pm and day 14 at 1pm. This could suggest a special event where bulk orders were placed. | |

Anomalies

|  |  |
| --- | --- |
|  | Upon analysing the credit card and loyalty number data by price and transaction time, one of the anomalies that was noticed is that some loyalty numbers are connected to multiple credit card numbers. Ideally, each loyalty number should be attached to only one credit card number. This discrepancy could be due to several factors, such as employees sharing or swapping loyalty cards with each other.  One potential solution could be to combine the count of loyalty transactions with credit card transactions and group them by locations. By doing so, we can compare the actual count of loyalty transactions with the count of transactions associated with credit card numbers, to see if there are any discrepancies or inconsistencies. This analysis could help us identify whether or not there are any issues with loyalty card sharing or misuse. |

## 2.2 Add the vehicle data to your analysis of the credit and loyalty card data. How does your assessment of the anomalies in Question 1 change based on this new data? What discrepancies between vehicle, credit, and loyalty card data do you find? **Please limit your answer to <= 8 images and <= 500 words**. **(4 marks)**

How does anomalies in question 1 change?

|  |  |
| --- | --- |
| Credit card vs loyalty anormaly |  |
| The analysis shows that the transaction counts between credit card and loyalty data are very similar, indicating that there may not be any significant misuse of either the credit or loyalty card data. However, it is important to note that this analysis only provides a high-level overview and further investigation using the GPS data may be needed to ensure the accuracy of the transaction data. | |

Discrepancies between vehicle, credit and loyalty data

|  |  |
| --- | --- |
| 3am transactions at Kronos Mart | *Diagram  Description automatically generated* |
| Upon analysing the data, I observed that there is no GPS data near Kronos Mart at 3am, despite transactions being recorded around that time. This discrepancy could be due to a time lag in transaction recording, or it could suggest that someone made a transaction at a nearby location and then walked to Kronos Mart. It is also possible that Kronos Mart was closed at the time of the transaction, which could explain the absence of GPS data in the vicinity. | |

## 2.3 Can you infer the owners of each credit card and loyalty card? What is your evidence? Where are the uncertainties in your method? Where are the uncertainties in the data? **Please limit your answer to <= 8 images and <= 500 words**. **(4 marks)**

*(Provide your answer and corresponding images here.)*

To analyze the relationship between loyalty number and credit card transactions, I first combined the relevant variables - loyalty number, last 4 credit card numbers, location, and time (day and hour) - into a single sheet. This allowed me to easily compare, and match transactions made with credit cards and loyalty numbers.

To ensure accurate matching, I used the price and location variables to match credit card transactions with loyalty number transactions. By comparing these two sets of transactions, I could identify any discrepancies or irregularities in the data, such as multiple credit cards being associated with a single loyalty number, or vice versa.Table

Description automatically generated

|  |  |  |
| --- | --- | --- |
| Loyalty Number | Credit Card | Full name |
| L1682 | 7253 | Linnea Bergen |
| L2459 | 5010 | Sten Sanjore Jr |
| L2769 | 8642 | Henk Mies |

Evidence/analysis:

To associate Car ID and name with respective loyalty and credit cards, I generated a time retention map to identify cars that stayed in a specific location for over 5 minutes. Then, I compared the location, day, and time to the retention map. For instance, for loyalty number L1682 and credit card 7253, I observed their presence at Coffee Cameleon on days 8, 10, 13, 14, 15, and 17 at 7 am, 7 am, 8 am, 7 am, 7 am, and 7 am, respectively. By utilizing the filter function, I checked for a constant ID during these times, and I found that ID 6 belongs to Linnea Bergen. I repeated this step for multiple observations. This allowed me to link Car ID and name with loyalty and credit cards. The process involved comparing the day, hour, and location with consistent IDs for different observations.

Uncertainties in data and method:

There are a few potential accuracy issues with the combined data, as some employees may prefer to use cash instead of their credit card, which would result in null credit card records in conjunction with some loyalty card records. Similarly, loyalty cards may not be utilized or may be swapped between employees, which could cause issues in accurately identifying the owner of each loyalty card and credit card combination.

Additionally, manually verifying the location based on day and hour for every observation can be a time-consuming task. As a result, to confirm the identity of each loyalty number/credit card combination, I chose to examine a small number of observations from various locations, rather than examining every single record.

|  |  |
| --- | --- |
| Linnea Bergen (Car ID 6) |  |
| Sten Sanjorge Jr (Car ID 31) |  |

Figure. Cross checking day, hour and location to constant car IDs

## 2.4 Given the data sources provided, identify potential informal or unofficial relationships among GASTech personnel. Provide evidence for these relationships. **Please limit your response to <= 8 images and <= 500 words**. **(4 marks)**

*(Provide your answer and corresponding images here.)*

For this question, I decided to focus on a location and analysed for any recurring employees. The two locations I chose was the house on Carnero Street and Chostus Hotel.

|  |  |
| --- | --- |
| Relationship 1 |  |

|  |  |
| --- | --- |
| Based on the observation of the data, it appears that two individuals, Axel Calzes and Lars Azada, were frequently staying together at a house located on Carneo Street. The data suggests that they stayed together for several days, indicating the possibility of an unofficial relationship between them. This assumption is made because the two individuals do not have the same last name, which usually implies a marital relationship.  To support this claim, a summary sheet was created to record the specific hours and days that they spent together in the house. The data may also indicate other possible relationships, such as roommates or close friends. However, without further evidence or clarification, it is difficult to confirm their exact relationship. | |
| Relationship 2 |  |
| Based on the data collected, it appears that there is a possibility of an unofficial relationship or affair between ID 33 (Brand Tempestad) and ID 7 (Elsa Orilla) at the Chostus Hotel. This conclusion was reached after observing that the two IDs spent time together at the hotel multiple times between 12-2pm on different days. To be more precise, I have created a summary sheet that details the specific hours and days the IDs were observed to be together at the hotel. However, it is important to note that this is only a possibility based on the data, and it is not a definitive conclusion. Other explanations or factors may come into play, and further investigation may be needed to confirm the relationship status between the two individuals. | |

## 2.5 Do you see evidence of suspicious activity? Identify **1 - 8 locations** where you believe the suspicious activity is occurring, and why. **Please limit your response to <= 10 images and <= 500 words**. **(4 marks)**

*(Provide your answer and corresponding images here.)*

|  |  |
| --- | --- |
| Jack’s Magic Beans, Kronos Mart |  |
| The data analysis conducted on the GPS records of the drivers revealed that there was an excess of GPS data in several locations, including Jack's Magic Beans, Kronos Mart, and other areas. Further investigation revealed that the driver responsible for this unusual activity was drill technician Isande Borrasca, who was using Car ID 28. The excessive GPS data observed for this driver could be considered suspicious because it suggested that Isande was circling or stalling around these locations instead of travelling to and from different places.  It is recommended that further investigation be conducted to determine the reason behind Isande's actions, or to check if there was a glitch in GPS data collection for Isande’s vehicle. | |
| Kronos Mart | *Diagram  Description automatically generated* |
| The analysis of transaction and GPS data revealed a discrepancy related to the transactions made at 3am in Kronos Mart. Although the transactions were recorded, there was no GPS data during that time, indicating that the purchaser may not have used a vehicle to travel to the location. Assuming that Kronos Mart is open at 3am, it is advisable to visit the mart and confirm the identity of the purchaser to investigate the matter further. This can help to determine if there was a lag in the GPS data recording or if someone made the transaction on foot. Further investigation can also help to ensure the accuracy of the recorded data and identify any potential discrepancies in the transaction records. | |
| Frydo’s autosupply n more |  |
| The behavior of ID 15 (Loreto Bodrogi) was observed to be suspicious as he was seen at Frydo's autosupply n more at 3am, a time that is not within working hours. Following this, he was also observed traveling through Chostus hotel and Ahaggo museum. This could suggest that he was engaging in illegal or suspicious activity. It is important to investigate further and gather more evidence to confirm or deny these suspicions. Additionally, it would be useful to check if any transactions were made during his visits to Frydo's autosupply n more, Chostus hotel, or Ahaggo museum, and if so, to verify their legitimacy. Further surveillance or investigation may also be necessary to identify any potential criminal activity or misconduct. | |

# 3 The Demo of Your VA System **(1 marks)**

*(Provide the Tableau dashboard URL or a brief explanation on how to run your D3-based code)*

[Link](https://public.tableau.com/views/Assignment2_16798258975820/adnormalies?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link)

# 4 Design and Implementation **(4 marks)**

Describe the major data and design challenges faced in accomplishing the tasks, and how you have overcome these challenges with your proposed visualization designs.

## Write a step-by-step description on 1) the necessary data processing and 2) visualization implementation by using d3 or Tableau.

**Biggest challenge**

When working with GPS data, the biggest challenge I faced was extracting the retention time data from the data set. While the GPS data provided information on the movement of employees, it was not clear enough whether the employees actually stayed at a certain location or simply drove past it. This made it difficult to accurately track the retention time at each location, which was a key metric we needed to understand in order to make informed decisions.

To overcome this challenge, I used Rstudio to edit the GPS data and create a retention time data set. First, I grouped the data by employee ID, which allowed me to focus on each employee's movements separately. Next, I created a rule that if an employee's coordinates remained the same for more than 300 seconds (or 5 minutes), it was marked as "retention." This threshold was chosen based on the assumption that an employee would need to stay at a location for at least 5 minutes to be considered as having spent any significant amount of time there.

Using this rule, I was able to identify the IDs of employees who had stayed at a location for more than 5 minutes, and I kept their ID and location data, which allowed me to narrow down the GPS data and gain more insights into employee behaviour. With this data set, I was able to create a map that showed the retention time of employees at different locations, which provided valuable information for decision-making and improving overall performance.

|  |  |
| --- | --- |
| GPS map | Retention Map |
|  |  |

*Figure: The difference between GPS and retention map*

**Data processing**

Date time formatting

When working with the GPS data, one of the challenges that rose was that the date format in the excel sheet did not plot accurately in Tableau. The date format was in “month-date-year”, but Tableau interpreted it as “date-month-year”. This caused issues when trying to filter the data based on specific days, as the dates were not being read correctly. To fix this issue, I used Rstudio to wrangle the date data. I used the lubricate package in Rstudio which manipulated the data into a format that Tableau could interpret.

Map and geometry

We were provided with two maps: the Abila geometry map and the tourist map. The Abila dataset contained the longitudes and latitudes, while the tourist map was an image that showed the major landmarks. To get a better understanding of the geography of the area and landmarks, I had to superimpose the Abila map on the tourist map image. To accomplish this, I first had to upload the tourist map into background images, and generate the longitude and latitude coordinates. This was done by observing the extreme ends of the Abila geometry map.

Once the coordinates were obtained, I plotted the Abila geometry on an empty sheet and used the GPS/retention data as a marks layer.

The final step was to switch the background maps from light to none, which resulted in the tourist map image being revealed with the GPS/retention data. This step gave us clarity on where the landmarks were and how the GPS and retention data interacted with it.

In addition to the above, I had to create a time retention data set from the GPS data mentioned above, which is included in the data processing.

**Visual implementation**

|  |  |
| --- | --- |
| Popular locations and timings |  |
| In order to effectively communicate information about popular cafes and their corresponding busy periods, I utilized two visual aids - a heat map and bar graphs. The heat map was particularly helpful in providing a quick and clear overview of the locations that experienced the highest transaction volumes.  In contrast, the bar graphs I used were helpful in conveying information about popular timings for these cafes.  Together, these two visual aids provided a comprehensive understanding of the most popular cafes and the busiest times to visit them. | |

|  |  |
| --- | --- |
| Unofficial relationships and suspicious activity |  |
| I utilized maps to provide a visual representation of the GPS routes and retention data in the context of monitoring transactions around Abila. Maps were the most suitable medium for this task as they could accurately display the geographical locations of the transactions. However, due to the high number of employees involved, there were many ID colours that overlapped, making it necessary to use filter functions to differentiate between them. By using maps with filter functions, I could clearly observe and analyze the transactional patterns in the area. | |