UNIVERSITÀ DI PISA

LAUREA MAGISTRALE IN INFORMATICA UMANISTICA

VISUAL ANALYTICS

Vast Challenge 2021: Mini Challenge 2

Gastech International employees' data exploration and visualization

Jacopo Gasparro
Matricola: 620038
Anno accademico 2021/2022

Introduction

The following report presents an Interactive Dashboard implemented to solve the Mini Challenge 2 of the VAST Challenge 2021, a Visual Analytics competition designed to help researchers develop software for a variety of analytics tasks. Such challenge aims at identifying the identities and behaviours of the employees of the GAStech International, a Kronos-based company settled in the fictional city of Abila. In this project it is possible to recognize the most frequent locations visited by the employees, as well as typical patterns and issues of concern, analysing their transactions and car movements on Abila's territory during a two-week period in January 2014.

Data

The data available for the competition include a .jpg map of Abila with locations and roads, geospatial data provided as shapefiles and database files useful to build a bidimensional projection of the city map along with the streets and their names, and a series of .csv files with employees' information:

- the first file is **car-assignment.csv**, a dataset with 45 records, corresponding to GAStech employees with their *FirstName, LastName, CarID* (a unique id for each person), *CurrentEmploymentType* (working area) and *CurrentEmploymentTitle* (actual work): this file has been lately modified since there were missing IDs for GAStech truck drivers. The data were introduced manually to match the IDs present in the gps dataset, which conversely provided this kind of information;
- the second is **gps.csv** which includes 685710 signal records from a gps device installed on employees' vehicles: each record is denoted by a *Timestamp*, an *id* of the moving employee and a pair of coordinates as *latitude* and *longitude*;
- next we have the **cc_data.csv** and **loyalty_data.csv** with all cards transactions made by employees, as each record is a transaction linked to a particular credit card or loyalty card: in the first case the database provides a *timestamp* with day and time of the transaction, a *location*, a *price* (money spent in transaction) and a *ccnum* (credit card number); in the second the *timestamp* only consists of information on the day in which the transaction was made, while we have *loyaltynum* (loyalty card number) instead of the cc number.

The jpg of Abila has been edited in order to hide those elements which could disturb the visualization process, like the white box behind the name or the compass on the top right corner of the image; concerning the geospatial files, these have been merged by a web tool into a single Topojson file which is useful to manipulate geographic data for projection and representation (**Abila.json**). Finally, the cards datasets have been merged on mounting, after a little preprocessing, so to have all the data available in a single object.

Dashboard

Il visualization project has been carried out exploiting *Vue*, a component-based framework for building user interfaces built on top of HTML, CSS and Javascript, in its 3.2 version. For this reason, it was also used *Bootstrap Vue 3*, a Vue integrated version of the front-end toolkit which utilizes prebuilt components to help building the web page structure. The additional libraries employed are:

- **Topojson**, an extension of GeoJSON that encodes topology, that was needed to read the map's json and for the representation of all the geographical data at disposal, like streets or employee's coordinates in space;
- **D3**, a javascript library designed for an efficient manipulation of the DOM, as it allows to apply data-driven transformations to the document with flexibility;
- **Crossfilter**, a library for exploring multivariate and high dimensional datasets and performing data filtering or grouping;
- **Plotly.js**, an open source graphing library with a detailes API reference which offers the possibility to create charts of all kinds.

Structure

The structure of the page has been projected to be very symmetrical and to be particularly focused on the centre, where the map is located. In fact, the dashboard implements a centred view on the map containing all the information on employees' movements throughout the two-weeks period spanning from the 6th to the 19th of January 2014. Right below the map there is a dropdown menu which enables the choice for a particular day to filter the GPS features to be represented and a time bar for a 24h exploration of cars' displacement. On its sides we can find the most popular locations represented by a horizontal bar chart and the list of employees of the GAStech with their names, a unique ID and their occupation in the company. On top of the locations' chart we have the Card selector, one of the main components inside the page since it offers the possibility to filter all charts data to match the selected card type, granting high reactivity to the whole page.

In the second row we can find a parallel coordinates graph centred in the middle and two couples of vertical histograms on its sides.

Each section is introduced by a title, while the chart for the most popular locations also presents a subtitle containing some information on the interactivity of the chart itself. In the very centre of the page there is also an indication of the card type chosen and the location inspected, so to let the user know the filters applied to data even on page scrolling.

Map

The map represents the main item in terms of interactivity and visualization power, and it is structured on different levels: we can find the Abila modified map as background image, which works as a base reference for the roads structure. This is implemented by defining a topology through Abila's topojson. Once defined the projection, it is possible to represent all the streets above the jpg, inside a SVG element. The opacity of this kind of geographical features has been reduced a little so that the road system is naturally integrated with the background.

The second features level includes the representation of the GPS tracking system installed on employees' cars moving on Abila's streets, which is recorded in the **gps.csv**: each employee is represented as a coloured point indicating the actual position of the car at a particular timestamp, where the colour is determined by the working area of the employee among IT, engineering, executive, security, and facilities. In the default setting, their trajectories are displayed too, so that only the moving employees have a tail indicating the direction of movement. These are coloured as black with a low opacity, in order to avoid that the overlapping could hinder the visualization of lower trajectories; the opacity trick can also give an idea of what are the most popular roads, since the same overlapping of points leads to a more intense and visible colour: this happens when the top right selector is turned on, enabling the visualization of the whole vehicles' trajectories. Thus, there are few differences between the head of every trajectory and its body. First of all, the body is reduced progressively when an employee stops for longer periods; secondly, the radius of head points is wider than in the tails, which makes these latter less conspicuous, especially towards the middle and end of the day. The problem of overlapping is also tackled by a mouseover event applied to all features: in fact, hovering on a specific trajectory turns it to yellow, while the head becomes light blue and increases in radius, thus highlighting the whole path for the chosen employee. The D3 library also allowed to associate a click event to each employee, so to let the user identify such person, bringing it on top of the list of employees on the right side of the map.

Along with these main features, it has been implemented a way to solve a specific question proposed in the mini challenge, that is the one referring to the unofficial relationships between employees. In fact, during the exploration of the movement of the employees, it is possible to notice circles emerging in particular positions all over the map, signalling the locations where two or more people meet either randomly or intentionally, which effectively helps identifying many recurrences in each person's behaviour. This is achieved in different consecutive steps that take as argument the heads of the trajectories: first, the distance between each couple of heads is calculated and compared to a *minimum difference* determined arbitrarily: if the absolute value of the difference is minor w.r.t. such a value, then the mean coordinates between the two points are computed and added to a dictionary with unique points, where each key is a centroid coordinates pair and each value is a list of ID's corresponding to the employees that are nearly in the same place. In this way, it is possible to spot those populated areas moment by moment and to check who belongs to that grouping.

In this case, the circles have been filled with a colour between red and pink with a low opacity, not to hide the background completely. Moreover, the circles are given a transition property which entails them in a black line whenever the mouse pointer enters the spot area, so to give a visual response to the user.

List of Employees

On the right side of the map there is a list of employees with their First and Last name and their working area, ordered by a unique ID, corresponding to the ID in the GPS data; the working area determines a colour assigned to employees, so to include a first visual grouping both in the map and in the list (*IT*: Blue, *Engineering*: Orange, *Executive*: Green, *Security*: Red, *Facilities*: Purple).

The ID matching mentioned early allows to select specific employees on the map to show them on the top of the list, thanks to some event handlers linked to the visual features inside the map. The first interactive event is realized by an onclick handler which is triggered when the cursor clicks

on an employee's trajectory: this emits the respective ID and brings the employee on top of the ordered list, colouring its box with a warm yellow. The same thing happens when the click is performed on the circle deriving from the encounter between two or more employees: in this case we have a group of people moving all up, keeping an ID-based order.

Timeline and day selector

The interactivity of the map depends on two elements: the day selector and the timeline, both under the map. The day selector is a dropdown menu that let the user choose the day to inspect for the GPS data to display on the map; the timeline, instead, allows to move along the time dimension on a 24h base, 5 movements at the time, starting from the first recorded in the day to the last one, since the bar adjusts its maximum value adaptively.

Location Chart and Card Selector

The card selector and the *LocationsChart* represent the component that drive all the interactivity of the statistical charts belonging to the page. This kind of charts are focused on the data transactions contained in the **cc_card.csv** and **loyalty_card.csv**, which are merged and mapped in order to keep a separation between *date* and *time* dimension. The *timestamp* feature, in fact, is altered to split into two other dimensions in the same way it was performed in the case of the GPS data.

The card selector let the user choose which card to use as filter for all the transactions in both credit card and loyalty card datasets, while the second allows to look for the transactions made in a specific place and to check the overall popularity of the chosen location. In particular, the location chart is a horizontal bar chart which counts the total times each location has been visited by all employees, ordered from the most to the least popular. It is itself subject to the choice of the card type, but it works also as a filter, since the user is given the possibility to click on a location bar to filter line charts, histograms, and the parallel coordinates plot in the bottom part of the dashboard. The clicked bar acquires the same red colour as those following graphs and the name of the location chosen is displayed in the middle of the screen, together with the selected card type.

The location chart also provides a way to clear all the filters by locations. In fact, double clicking on the background of the plot we can see that all charts are restored, and the only filter applied remains the one with respect to the card type.

Info Charts

The *InfoChart* component represents either a line plot or a histogram, based on a specific id. The four charts in the bottom, included into the two columns on the sides of the parallel coordinates plot, in fact, are implemented with the same component: the top plot is a line plot inspecting the daily trend of either the visits or the money spent during the two-week period, while the second plot is a histogram representing the distribution of the same variables on a hourly basis, giving insights about those parts of the day where the transactions take place the most. Concerning the histogram, the bin size has been modified to match an hourly interval, which seems to be a fair range of comparison.

Parallel coordinates plot

The *Parcoords* component implements a parallel coordinates plot, which represents all the transactions given their card identifier, the date and time in which they are recorded, the location, the payments. In an early version of the dashboard the data plotted by this chart were picked on the basis of the day selection linked to the map representation; later, an overall analysis of the transactions looked more appropriate, since the date dimension could always be inspected thanks to the plot dimension filtering: in fact, it was included among the dimensions in the plot building, making the plot reactive in the same way the others are.

In this case, since the records relative to the loyalty cards are missing with respect to the *time* dimension, the loyalty cards lines hit the 0 value on the *time* y-axis, but the card type selector allows either to filter them out or to inspect them alone.

Concerning the line colour, it is determined by a colour scale based on the *price* dimension, that is the amount of money spent in a single transaction: bigger values correspond to a colour span that ranges from yellow to green and blue for the biggest, while little amounts are assigned a dark red colour. This feature helps having a quick look at either the locations with higher overall prices or the days of the week with more movements linked to money spent, specific patterns regarding the card types.

Analytical task: MC2

This section is dedicated to the discussion of the questions proposed in the Mini Challenge 2 of the VAST Challenge 2021.

- **Most popular locations**

From the locations' chart on the left side of the page it is possible to assert that the most popular location overall is Katerina's Café, followed by Hippokampos and Guy's Giros, which all seem to be places where to have a meal. This is confirmed first from the line plot, since it always indicates a high number of visits throughout all the weeks; secondly, the histograms show a very high affluence at lunchtime or dinner, from 13 to 14 and between 19 and 22 in the evening. This may also be motivated by the fact that such places have far lower prices if compared to the locations lower in the ranking.

- **Inferring owners of cards and detecting informal relationships**

The system of circles for detecting the main groups of people around Abila is already a measure to find out relationships between employees. Anyway, crossing the GPS coordinates with the transactions' data it is often possible to infer the owner of a credit card or a loyalty card and to confirm relationships spotted on the map. An explicative example of this feature is given by two particular transactions that occur the 18$^{th\ of}$ January at 19:06: they have the location (*Robert and Sons*) and the money spent in common, but they belong to the credit card **9617** and to the loyalty card **L5947**. If we look at the map in that moment, we can see that there pops up a circle on *Robert and Sons*, in the center-left side of the map; inspecting further it is possible to see that only two people are there in the same place inside the circle, namely Ingrid Barranco and Brigitta Frente.

This may mean that they are somehow related and that they may have bought the same product from that shop.

At this point, if we filter the transactions on the parallel coordinates by the credit card **9617** and we choose a timespan around 21:00, we can see that only two transactions match the search, both at Katerina's Café. The interesting one is the one occurring the 6th of January: in fact, if we check on the map at that day and that particular timestamp, we can see that only two people are there in the Cafè, Brigitta Frente and Kare Orilla. This might suggest that the credit card **9617** belongs to Brigitta Frente. Obviously, the process may be repeated for all the transactions and GPS data at disposal to retrieve all the correspondences between employees and cards.

These data can be double checked through a brief analysis of the transactions and GPS made in python in a Jupyter notebook inside the *Analysis* folder of the project repository. The criterion used to associate the cards to employees consists in checking the movements of each employee in a brief span of time around each transaction. In particular, the person should be standing in the same place for around 10 minutes before the transaction and moving instead before 5 minutes after the transaction. The method is inductive, and it is based on ranges in time that have been determined arbitrarily, so it works only as an analytical comparison with the evidence in the visualization tool, as it shows the same case of the credit card **9167**. The python analysis, in fact, links that card exactly to Brigitta Frente.

In the same way, we can determine, for example, that the credit card **9220** belongs to Albina Hafon, since she is the only employee at the Abila Airport around the 8:15 of the 6th of January. Filtering the parallel coordinates plot we can see that the only transaction is made by the credit card **9220**. If we look for other transaction regarding that identifier, it is possible to spot another transaction at Abila's airport on the 13th, indicating that she may have left Abila for a week and come back on that same day.

A hint on the unofficial relationship between employees can also be their first and last positions of the day: in fact, inspecting the GPS data it is possible to see that each point starts almost from the same location in the morning and ends in the same place, which might coincide with residential locations or the place they are being hosted. An overall look at the last timestamps of every day, in fact, highlights that there are certain patterns among employees, concerning the place they live in: most of the engineering employees live in a wide zone in the north of Abila, above Carnera Street; the executives, instead, all live between Spetson Park and Taxiarchon Park, while the IT employees live near Arkadiou Street between Jack's Magic Beans and Arkadiou Park; finally, the security employees are concentrated in the east of Abila, near Sannan Park. Actually, there is a highly inhabited area around the most popular locations, always in the east of Abila, where employees from different working area seem to be finally meeting at night: this could suggest that they are roommates or that they simply live in the same building

- **Suspicious activity**

Sometimes GPS data gives us the possibility to explore some strange paths on Abila's map, which may be due either to wrong signal detections or a to manumission, as well as to a strange behaviour of the employee in question. In this regard, it is interesting to check how the 19th day of January begins, with the first GPS records that are just past the midnight: we can see the employee Isande Borrasca moving haphazardly from one side of the Taxiarchon Park and crossing it to the

other towards Barwyn Street in a span of time that goes exactly from the midnight to 2 in the morning.

The same behaviour can be registered for Isande Borrasca on almost every day of the two weeks: the car seems to have always the same distorted path, going from the already mentioned park to an unknown location between Ouzeri Elian and Abila Hospital and coming back to the start, which allegedly is its home location.

Another suspicious event is a large gathering involving most of all engineering and IT employees, which takes place the 10th of January at Lars Azada's home. This may suggest an intricate relationship system between a wide group of employees, as well as an activity to monitor: anyway, since that day is a Friday it is possible that Lars Azada has just organized an informal meeting or a party indoor.

The last suspicious event may be linked to the transactions happening in the night: in particular, there are four transactions at Kronos Mart between 3 and 4 AM, especially on the 19th of January. These transactions are not confirmed by any movement, since there not seem to be any employee near that position in the early morning: in fact, the nearest recorded car in the end of the day on the 18th is four or five blocks to the south, near Kronos Capitol. It might be possible that the employees have moved on foot after a night out, which may be justified from the fact that the 18th is Saturday, also given that there is no residential area in the surroundings, but it is impossible to say for sure which is the origin of such transactions.