University of Pisa

Visual Analytics

VAST Mini-Challenge 2 (2021)

Author:
Francesco Di Cursi

Supervisor:
Salvatore Rinzivillo

June 19, 2022

1 Description of data and presentation of the pattern or model to communicate

1.1 Data description

The 2021 VAST's second Mini-Challenge data folder consists in 4 csv (giving informations about credit and loyalty cards, car assignment and GPS tracking), one brief description of these ones as word document, a JPG map of Abilia city and geospatial data such as Krono's and Abila's shape files. Reasoning about the 4 csv files, exploiting their similarities, two macro groups can be observed:

1. **cc_data + loyalty_data**;
2. **car-assignment + GPS**.

More precisely:

1. the two cards's datasets are equal on location and price columns but they diverge in timestamp column and in the last digits one:
   (a) the timestamp in cc_data contains also the hour in which the purchase happens, while loy_data contains just the date;
   (b) the last digits in cc_data are totally numerical while in loy_data they are alpha-numerical;
   (c) Some outliers are present in the price column, as there are few employees which spend much more than many others.
2. Car-assignment and GPS datasets are equal only on the ID column, the rest changes:
   (a) in the first one, last and first name as well as the type and the title of employment are given;
   (b) in the second one, latitude and longitude as well as the timestamp (m/d/y h:m:s) are reported. Coming to car ID, the trucks driven by anonymous employees go from 100 to 107, while the known ones, as said before, go from 1 to 35.

1.2 Model to communicate

Given these data and the aim of the second VAST Mini-Challenge (i.e. searching information about employees kidnapping by a terrorist organization), an efficient way of proceeding consists in, by order:

1. Use intensively both cards data to understand where and when a purchase happen, which location is the most popular and when (by frequency, price and ratio between these two last ones) and, on the basis of any dimension, exploring dynamics at a deeper level (with a set of filters, one for each dimension) in order to have an overview of the purchase history of each card. In this way, this interface leads to the precise timestamp to search in GPS data;

2. Using car assignment and GPS data to display the state in time of the employees, in order to identify car owners on the basis of the purchases happened in a given place at a given time.

In order to explore these data, as previously said, similarities have been exploited: the preprocessing has been made using python. More precisely:

1. **Merging cc_data and loyalty_data:** these 2 datasets have been merged by standardizing the differences in timestamp and last numbers (timestamp has been split into Date, Hour and Minute columns, giving "Not specified" as value to loyalty_data's rows both for Hour and Minute as this information was only given for credit card. A column "num" has been created by extending credit card column ('last4ccnum') with the loyalty one. Finally, a column "price_count_ratio" reports the ratio between the frequency of a place with the money spent in it;

2. **Merging car_assignment and GPS data:** in this case data have been joined on the car ID column (*nans* have not been dropped as those rows contains valuable information, thus they are handled directly in Javascript using a conditional ternary operator). This objective has been

reached by a dictionary with car_ID as key and full name (last+first name), CurrentEmploymentType and CurrentEmploymentTitle. In this way, car owners can be clearly identified. Again, the timestamp column has been divided into *atomic columns* (Date, Hour, Minute, Second) and, for each timestamp, a progressive number is given as id (n_timestamp) in order to easily use them on an integer slider;

3. **Sunburst bonus:** even if sunbursts have been omitted for clarity sake (i.e. too much dimensions), the function used to format data in order to plot them in Plotly.js has been retained into the python file, as it seems not to be present by any mean on the internet. Again, for clarity sake, only the function for cc_data has been displayed, as it is easily adaptable.

1. In order to visualize cards data, three plots have been used:
   - a barchart to visualize frequencies for each place;
   - an heatmap to evaluate the correlation between time and space dimensions according to frequency, price or their respective ratio;
   - a categorical parallel plot, exploiting the hierarchical structure of the data in an investigative point of view: who, when, where and how much (not necessarily in this order, as it will be said further in this report).
2. In the case of car assignment and GPS data, instead, an SVG map made with d3.js has been realized. In particular:
   - - the Abila JPG has been used as background of the map;
   - - the shapefiles have been merged using an apposite web tool (i.e. *mapshaper.com*) and the resulting topojson file has been used to plot the paths of the streets. I decided to avoid a cluttered representation by using just essential data in the topojson (i.e. coordinates of the streets) as the names are useless (on the map just few of them are actually displayed, and the map itself converges the essential odonomastics). In the case of locations, by the way, only few shops are on the map: this difficulty can be overcome by using heatmap and parallel plot (i.e. by associating some given coordinates to the place in which, at that time, a purchase has happened, and the same goes for identifying the owner of a vehicle).
   - A slider to control time (by n_timestamp);
   - Some d3 rectangles as a layer in the map, in order to see the area around each car to clearly detect overlapping (i.e. using the absolute result of the subtraction between point and fixing a threshold of 0.002 to classify proximity, using an opacity of 0.6 to make the SVG rectangles more 'tangible' if they overlap);
   - A list of clusters's components, if any, dictated by the threshold said in the previous point (half of the rectangle is sufficient at 0.002 to detect a reasonable proximity, as half of a rectangle corresponds to a range of approximately one block on the map) in which id, full name, type and title are provided, together with coordinates;
   - A set of buttons to move easier in time, as 2 weeks of timestamps are not feasible with a simple slider. This multi speed forward and backward motions have been realized with a series of async functions.
   - The GPS data at each timestamp as points on the map with their relative informations at their side (SVG text doubles its size on hover; in case of full overlapping the list at the side of the map can be used to disambiguate);
   - An heatmap layer with a relative Date filter, to look at dynamics from a overall or a daily point of view.

2 Design choices: colors, interactions, shapes, transformations

In order to facilitate the understanding and the experience of the analysis, two simple horizontal lines have been added to create a tripartite vertical page.

- The first section contains the heatmap with the marginal distribution of the x axis (i.e. locations). In order to do this, a bar chart of locations has been added above the heatmap (shrinking its heigth, adapting the width and the left margin and hiding the x axis. The “Hot” colorscale given by plotly.js has been used as it is perfect for continuous values and it has been used in both plots to enhance the perception of integrity of the two plots. In this way, altogether with the marginal distribution, outliers can be easily detected. This dynimic is not possible without normalizing the price or the ratio. In this case the square of the values are taken. In order to visualize better the top outliers (e.g. Abila Airport) the colorscale of the marginal distribution, just in the cases of ratio and price filtering, are augmented by 100, making what would be literally white just a bit yellowish (this doesn’t alterate the color of the lower elements in distribution because the added number is relatively small if compared with the length of the colorscale. By filtering the first section with the heatmap filter, the marginal distribution shape always reflects the frequency of a given place but its color change according to the filtering choise, just as the heatmap does. This, again, helps in the perception of integrity of these two plots. Finally x and y are ordered lexicographically to facilitate the search for dates and locations. By clicking on a specific square on the heatmap, an automatic filtering of the next section happens.

- The second section contains the categorical parallel plot, a bit too complex at first glance, but easily manageable through two types of interaction:

  1. on its top there is the multiple multiselect filter, giving to the user absolute freedom of research (its usability will be treated in the next work section, fully dedicated to the interaction), shrinking the results represented in the plot;
  2. selecting a rectangle or a link, the relative complete path is highlighted across the plot, selecting again the relative part of the selaction can be returned to default, giving the possibility to research both from a bottom-up and a top-down approach.

  The height of the rectangles and the links in the parallel y axis stands for the representativity of that element in the whole according to price (the more the price, the more the height). The eventual discrasy element heights can be easily overcome by hovering on the element to display information: by hovering on the element it displays the absolute count of the price and some other informations. Even if this type of representation slows a bit down the site if fully displayed, it has been used because thanks to the double filtering it can be easily reduced to a smaller plot, as looking the whole ensamble it may be useful to spot some interesting patterns. Also the height of the div containing the plot has been dynamically implemented, according to the card numbers present on the y first axis. A grey color has been chosen as default one to create a background on which better see the firebrick coloured selections.

- The third section is composed by the map and the list of vehicle owners. At the bottom there is the time handler (in descending order: the timestamp, the set of buttons to alter the speed and the slider, to move fast between days).The background of the map is a bit transparent to make the grey paths of the streets more visible. The cars are displayed as blue dots with info at their side (the size, as stated before, can be doubled by hovering on the text, while the color of the points become yellow by hovering on the points, in order to highlight points at a given timestamp. Around each point, a red square is displayed in order to better catch their positioning on the map. Half of the diameter of the square corresponds roughly at one/two city blocks. The choice of making the squares a bit transparents helps in case of visualization of overlapping (also thanks to the automatic update in the list if the halves of the diameters overlaps). The time handler is composed by some buttons placed above the slider (this one has a light blue background in order to facilitate its use, given the light color of the bar). In the buttons, the icon of the arrow specify visually the side of the motion, while a simple “x N” represent the acutal speed of the motion. Also here these two filters are positioned in a way to enhance the perception of a whole, rather than distinct handlers. Coming to a better visualization of patterns on the map, an heatmap layer has been added, manually setting the color bar: it is univariate, mapping white to 0 in order to have a clean background, using a descending order to make the most frequents coordinates easily spottable (from blue to red).

3 State-of-art: similar tools or interfaces for the same problem

In this case, as already said, the task can be broken into 2 steps:

1. explore the cards data;

2. given cards informations, retrieve information from GPS data.

The distribution of dimensions in cards data could be displayed as boxplots, violin plots, histogram+KDE plots or density plots. Another way to investigate correlation, similar to the one chosen in this project, would be a 3D scatterplot where the x and y axis remain unchanged but the z axis becomes the deepness rather than a color (frequency, price or ratio). 3D because, in order to maintain each point as a distinct purchase, the plot must avoid overlapping. This kind of visualization has been tried and abandoned beacuse of the poor representation of ticks of all axis (just few of them are actually displayed and, even forcing the displaying of the missing ticks, the 3D visualization is not feasible with too much data (as in this case). A 2D heatmap has been chosen for this reason, leaving to the categorical parallel plot further inspections. Also sunbursts or treemap could be used to explore the conceptual hierarchy of the investigation but, again, the "curse of dimensionality" makes this representation too cluttered (two sunbursts, one for each card given the differences in time, are useful for a finer inspection but could lead to confusion given the 4-level representation). In this case, categorical parallel plot has been chosen for its capacity to display all the information on plain sight, making the curse of dimensionality more handable through the double filtering system (multiselects to filter data and selections on plot to highlight interesting elements). Coming to the second point, and given the nature of the files present in the challenge folder, using a map seems mandatory. But in this case a different approach would be the use of a clustering algorithm to automatically detect the groups of people at a given timestamp (this has not been implemented, in order to maintain a pure visual approach to the task, without using machine learning to resolve the problem). The same use of machine learning can be the mean through which detect car owners: the car is active at a given moment? Is there a purchase in the given location close to that timestamp? These questions, which guides the manual investigation, could be automated through a decision tree.

4 Detailed description of the visualization with a description of the interaction

In the first section of the web app it's possible to have a preview of the overall correlation between dates and locations according to three dimensions: price, frequency and their relative ratio. Starting from frequency dimension, it's possible to observe that the most representative locations are pubs and bars, namely in descending order of frequency: Katerina's Cafè, Hippokampos, Guy's Gyros and Brew've Been Served. According to price, looking at location's marginal distribution, it's clear that the previously stated locations are highly frequented yet with few purchases: this is consistent with the fact that customers spends only to eat and drink the necessary, without any outlier. On the contrary, there are some poorly frequented places in which there are expensive purchases, moreover done by very few people (e.g. the airport is frequented by the same 3-4 people on the whole span of the 2 weeks). These places are: Abila Airport, Carlyle Chemical Inc., Nationwide Refinery and Stewart and Sons Fabrication. Finally, looking at the ratio dimension, a more concise representation is given: the colors on the marginal distribution doesn't change that much but those places and times in which there are few people and big purchases are highlighted, thus the color for the previously stated locations (i.e. Abila Airport et all) is enhanced. Some interesting patterns also arise for Kronos Pipe and Irrigation and Maximum Iron and Steel, as well as for some isolated cases such as the day 13 at Frydos Autosupply and More or the day 17 at Albert's Fine Clothing. After this first phase of exploration, a more detailed view of these data is provided in the second section. In particular the user is free to explore without filtering or by manually filtering each dimension with the set of multiselect widgets: by the simple click it's possible to highlight a single name, while others can be added by ctrl+right click (also by dragging onto the list) and also deselected always by the same event (i.e. ctrl+right click). A third and last interaction has been developed by linking the first and the second section: by clicking on a square of the above heatmap, the given location and the given time are automatically highlighted in the relative multiselect widgets (to filter only by this two dimensions, the user has to manually deselect the unwanted filtering. Moreover, in the ‘Card type’ filter, if anything different than “All” is selected, then the “ID” filter gets automatically deselected. If an empty selection is created, then the plot will simply not appear. The user is free to change the order of the parallel y axis in the parallel plot, and also the order of elements on each y axis. Finally a manual selections of the paths between y axis is possible by clicking on the relative element that want to be displayed. The user can then proceed in two ways: a top-down approach (on a cluttered representation, clicking a macro area and then deselecting the unwanted data) or a bottom-up one (by initially filtering from the heatmap and then spotting some patterns to analyze, such as some particular cards). In this way the history of each purchase can be tracked and, if there are few people, also the association between cards can be inferred.It seems that if the customer uses both cards for a purchase, then splits the purchase in perfect halves. This dynamic is another track which helps in the investigation. The price is visible by hovering on the wanted element (on paths for the single purchase and on the rectangles for the overall price, for any desired dimension). The “density” of each element is given by plotly.js though the probability. Finally, in the third section, an heatmap can be displayed as a layer in the map. The user is free to hide it through the relative button and to filter it according to the date dimension. Beside the heat layer, there is another button to display the Abila image (and to hide the SVG map) to easily disambiguate locations given the opacity and the scaling of the map. Thanks to the heatmap layer is easy to notice that there are some roads with a noticeable activty: the most visited place is Gastech, with a total cap of 20.000, in the street right in front of Cup o’Joe. Other interesting streets are Egeou Avenue (leading to the airport at the left and to the port at the right) and Ipsilantou Avenue (at the end of Egeou Avenue, toward the port, where the most frequented places are located, i.e. Katerina’s Cafѐ, Guy’s Gyros, Brew’ve Been Served and also Frydo’s Supply, this last one not so frequented but with an anomalous purchase). Other interesting streets are those starting from the Gastech hotspot and going up and left through Ouzeri Elian, Carly’s Coffee, General Grocer, Roberts and Sons and Kronos Mart. The other one starts always from the hotspot going up through Taxichon Avenue, Arkadiou street (UPump and Jack’s Magic Beans) up to Albert’s Fine Clothing and Been There Done That. This trend keeps almost the same in time but for the weekends: in these days, there’s no hotspot at Gastech (apparently no activity at all) with the more frequented locations changing day by day but almost in the same range. Finally, on the basis of all these informations, a final inspection can be done by looking at a given timestamp around the time of a purchase (or by any other intention) to deduce who made the purchase according to the where and when obtained in the previous sections. The optimal use of the slider consists in hitting first the x1 speed and then moving to other speed if wanted (if any anomaly happens, clicking the pause button is sufficient to fix the problem). It occurs to stop reasonably sooner than the timestamp needed, to slowly move with x1 to the objective. Finally, by hovering on points is possible to double the SVG text size and is also possible to highlight the points by hovering on them (without changing timestamp). The list at the right of the map displays those points whose distances are approximately one city block, giving all the informations about the car driver (car ID as badge, full name, type and title of the employee and its relative coordinates.

5 Case example for an analytical task

In order to be concise, the most clear examples are taken. As first thing to find suspicious activities, a place with few purchases must be identified: the fewer the purchases, the simplest the investigation becomes. This selection should have also high price, thus an high ratio. Two different case examples are described in the following subsections.

5.1 Finding card owners

- the day 17 at Albert’s Fine Clothing there are only 2 purchases. On the price filter there’s no anomaly but it is highlighted if filtered by ratio. By clicking on the orange square, the selection is made on the second section. As displayed from the categorical parallel plot, only two purchases happen in this day, one with credit card (1321) and one with loyalty one (L419), having both the same amount of price. As stated before, this dynamics could probably imply that the purchase has been made at the same moment by the same person, splitting the amount in perfect halves or to save money, or maybe to try to hide expansive purchases. Looking at the credit card path, it’s possible to see that the purchase happens at 19:44.By looking at the map in the third section at that given timestamp, it's possible to infer the card owner. Before proceeding, the heuristic that a purchase is followed by one or two minute pause after the car starts on moving again it is chosen by an inductive analysis. In this case Campo-Corrente Ada arrives at Albert's Finest Clothes at 19:44:40, making the purchase almost impossible in only 20 seconds. At 19:46:00, by the way, Calzas Axel exit from the shop. He may be the author of purchase. A way to be sure about this is to go back to the second section of the web app and deselect all filtering, refiltering only by cards (in this case L4149 and 1321). By doing this, we see that most expansive purchase made by these two cards is the one previously observed (the same goes if the second section is filtered only by Albert's Fine Clothing, making this purchase the most expensive for that location in the whole 2 weeks span). Having the plot filtered only by card number, the pattern of the amount of price split in perfect halves keeps on occurring. As an example, considering (thus manually selecting) Abila Zacharo or Guy's Gyros for a simpler inspection as they have both only two purchases, one with the credit card and one with the loyalty one). The amount is equally split on cards (two purchases of 27.3 Abila Zacharo the day 13 at 13:34 and two of 31.05 the day 10 at 13:52 at Guy's Gyros). Considering the case of Abila Zacharo, looking at the map, it's impossible to locate the shop as it is not displayed on the map, so the location of the shop must be inferred on the basis of car movement (if any). The day 13 at 13:35 Calzas Axel is at Roberts and Sons (maybe Abila Zacharo is nearby [24.85097804, 36.06349268]). At 13:59:00 Calzas Axel starts moving from Ipsilantou Avenue, amid Katerina's Cafè, Brew've Been Served and Guy's Gyros. Probably he has parked nearby, making the purchase at 13:52 and taking 7 minute to hop again in the car (way more than the 1 or 2 minutes chosen as heuristic but still reasonable , for example as the car is a bit far from the shop). Based on this, the following association can be inferred: Calzas Axel, Engineer (Hydraulic Technician), riding the car with 11 as ID, is the owner of the credit card 1321 and the loyalty card L4149.

5.2 Detecting informal bonds

Finally, a way to identify informal relationships between personnel may be searching for group activities during the weekend. The following example looks at Desafio Golf Course.
By looking at the frequency in Desafio Golf Course, it's possible to say that the course is attended by the Gastech personnel only on Sunday. Moreover, the 19 the purchases almost double (from 6 to 10). Other filters on the first section are irrelevant for this analysis. By inspecting both days (12 and 19), on the basis of the perfect halves observation, it's possible to see that effective people in these days are, respectively, 3 and 5 people. The day 19 two groups can be observed: 8156 (L5224) is probably at the course with 7688 (L4164) as they pay at the same hour (12) but at a slightly different minute (the first at 12:41, the second at 12:57). In order to do this is necessary to click on the needed hour rectangle on the parallel plot (thus selecting the active cards at that time). To deselect all, just click on the selected place. The second group is made up by 5010 (L2459) and 2463 (L6886) which, has the group before, do the payment both at the same hour (15) but at a slightly different minute (15:46 for the first, while 15:15 for the second). Then 8332 (L2070) attend the course alone at 13:51.

Going to the map. on the day 19, at 12:31 **Strum Orhan and Barranco Ingrid**, both Executive, go together to Desafio Golf Course **(this is a clear sign of an informal bond)**, then another Campo Corrente Ada (another Executive) follows them. entering at 12:39:36. In few minutes also Vasco-Pais Willem (again, an Executive) joins (at 12:41:12). Meanwhile, even if it is off topic, Borrasca Isande (Engineering) does strange circular movement near the point in which Strum Orhan and Barranco Ingrid starts their trip to Desafio Golf Course, then stops after few minutes at Spetson Park (12:34:32). At 12:51:01 Strum Orhan moves just for a second, then it disappears. At 13:19:45 , Sanjorge Jr. Sten (Exectuive and President) arrives at the site from the Chostous Hotel. At 15:15 Frente Vira (Engineering as Borrasca) exits from General Grocer and goes nearby Desafio, then stops at Parla Park. At 15:26, Borrasca Isande starts again to do strange circles on the same place as before. It is as Borrasca disappears from one street to appear on the closest one for each second while doing these strange movements.She stops at 15:28:51. Sanjorge exits from Desafio Golf at 15:50 and goes to the Chostus Hotel.

Giving another suitable example of card owner detection, *by doing a further inspection on the hotel, looking at the first section of the web app and filtering by ratio, no activity appears the day 19l, but the day before there is a huge transaction. Going to the parallel after the previous selection on the heatmap, just two cards with perfect halves are used, thus a single person does the check-in the day 18 at 12:03 using 5010 (L2459) (thus he payed at 15:46 at Desafio Golf). It is the only day in which there* *is a single check-in and it i also the same day in which there is the highest transaction. To double check, going to the map the day 18, at 12:03 there's nothing at the hotel. But at 12:39 Sanjorge is the first to exit from the Hotel. Reasonably Sanjorge owns cards 5010 and L2459. The same can be done for any card used in Desafio but it is not part of this very example.*

Going back to the day 19, at 16:01 Barranco Ingrid exits from the Desafio, followed at 16:02 by both **Campo Corrente Ada and Vasco-Pais Willem (another informal bond, they are overlapped while they move before separating)** then they all go to the start position (structure between Septson Park and the Museum). The same can be done on the previous Sunday (the day 12) to keep on investigating relationships on Desafio Golf. By the way, besides the two apparent informal bonds, a greatest view emerges: **all and only the Executive team spend their Sunday morning together at Desafio Golf Course.**