

UNIVERSITÀ DI PISA

DATA SCIENCE AND BUSINESS INFORMATICS

PROGETTO DI VISUAL ANALYTICS

L'incidente di Kronos

VAST Challenge 2014: Mini-Challenge 2

Anno Accademico: 2020/2021

Studentessa: Merendi Federica, matr. 584572, f.merendi@studenti.unipi.it

Agosto 2021

Indice

1	Introduzione	1
2	Stato dell'arte	1
3	Dati	2
4	Scelte di design	4
4.1	I componenti grafici	4
4.2	L'interattività	4
4.3	I colori	5
4.4	Il font	5
5	La visualizzazione	5
5.1	Selezione del giorno	6
5.2	Struttura dell'azienda	7
5.3	Timeline degli eventi	8
5.4	Mappa	9
5.5	Confronto spese	11

6	Esempi di utilizzo	12
6.1	La giornata-tipo del dipendente GAStech	12
6.2	Un esempio di analisi tramite la visualizzazione	13
7	Conclusioni	13

1 Introduzione

Come progetto si è deciso di sviluppare la Mini-Challenge 2 di VAST 2014. [11] Il tema principale della challenge riguardava la scomparsa di alcuni dipendenti dell'azienda GASTech International situata ad Abila, nell'isola di Kronos, nel gennaio 2014.

L'azienda stava tracciando segretamente gli spostamenti delle auto messe a disposizione dei propri dipendenti (per uso sia lavorativo che privato) e i dati dei GPS sono stati messi a disposizione per le indagini. Oltre a questi si hanno le transazioni delle carte di credito (Credit Card) e delle carte fedeltà (Loyalty Card) dei dipendenti raccolte nelle due settimane precedenti alla scomparsa, ma non il giorno stesso.

2 Stato dell'arte

Prima dell'ideazione della visualizzazione, se ne sono analizzate tre che hanno partecipato alla Challenge nel 2014.

La prima, dell'Università di Konstanz [10], mostra sulla mappa di Abila le locations, gli incontri e i tragitti percorsi con l'automobile dai diversi impiegati.

La seconda, della Middlesex University & MASS [7], presenta due visualizzazioni principali: una utilizza la mappa su cui vengono indicati i diversi percorsi delle automobili con sottolineati il punto di inizio e fine, mentre la seconda è una tabella in cui per ogni dipendente e giorno della settimana vengono indicati il numero di tragitti compiuti con l'auto e le spese sostenute.

Infine, la terza visualizzazione della Purdue University [1] presenta i dati nella mappa come le precedenti e aggiunge una Dodeca Ring Map per ogni location.

3 Dati

La Challenge fornisce 4 dataset e i dati necessari a costruire la mappa di Abila. Innanzitutto, la mappa è stata convertita in file .geojson tramite il tool online *MapShaper*[8],

I dataset invece sono stati inizialmente analizzati utilizzando Python e la libreria Pandas, poi sono stati modificati per facilitare la rappresentazione dei dati.

Come chiave di unione tra i diversi dataset si è usato il nome completo del dipendente: è stata quindi creata per ogni individuo una variabile *FullName* nella forma "Cognome Nome". Il seguente elenco descrive i file messi a disposizione dall'organizzazione e le modifiche effettuate.

- Il file *car-assignments.csv* presenta l'elenco dei dipendenti con indicato: nome, cognome, tipologia di impiego, titolo attuale di impiego e (nel caso fosse presente) il codice identificativo dell'auto aziendale a loro assegnata.

Analizzando i file relativi alle carte di credito e fedeltà, è emerso che nell'elenco non erano presenti tutti i dipendenti poiché a loro non era assegnata un'auto aziendale. Per ognuno di essi si è cercato il dipartimento e il titolo di lavoro nel file "*GAStechKronos-org-chart.pdf*" messo a disposizione per la Mini-Challenge 1. In aggiunta, per tutti i dipendenti si sono sostituiti gli attributi *FirstName* e *LastName* con la nuova chiave identificativa *FullName*. Infine, si sono aggiunti cinque mezzi di trasporto - probabilmente camion - di cui è stato rilevato il GPS, ma non è stato comunicato il guidatore, questi sono identificati nella forma "Truck 107", dove 107 è il numero identificativo del GPS.

- Il file *cc_data.csv*, relativo agli acquisti pagati con carta di credito, presenta per ogni transazione: il nome e il cognome del proprietario della carta, il prezzo pagato, il locale destinatario del pagamento e il giorno e l'ora a cui è avvenuto il pagamento.

L'unica modifica effettuata al file è stata sostituire gli attributi relativi a nome e cognome con la chiave *FullName*.

- Il file *loyalty_data.csv*, relativo alle carte fedeltà, presenta per ogni transazione: il nome e cognome del proprietario della carta, il prezzo registrato e il giorno in cui è stato effettuato l'acquisto.

Come nei casi precedenti si è aggiunta la nuova chiave *FullName*. Inoltre, dato che non era presente l'ora a cui è avvenuta la transazione si è deciso, per facilitare la rappresentazione, di associare ogni transazione con la carta fedeltà al rispettivo pagamento con carta di credito. Nel caso in cui la transazione non fosse di uguale valore (ad esempio sulla carta fedeltà è stata registrata una cifra di soldi inferiore rispetto alla carta di credito o viceversa) si è deciso di assegnare l'orario "00:00" come default in modo da non incorrere in supposizioni errate.

- Il file *gps.csv*, contiene ai dati raccolti tramite GPS. Gli attributi rappresentati sono: la data e l'ora del rilevamento, l'ID dell'automobile che si stava spostando, e le coordinate geografiche (latitudine e longitudine). Il GPS dell'auto assegnata a Orilla Elsa (ID: 28) risulta difettoso.

Per poter aggiornare il file con la nuova chiave *FullName*, è stato necessario fare un *merge* tra i dati del GPS e quelli degli impiegati.

Al fine di identificare facilmente i rilevamenti che appartengono allo stesso tragitto è stato aggiunto un *PathID*. Per poter calcolare questi percorsi si sono raggruppati i dati per giorno e ID dell'automobile e si è considerato un intervallo massimo di 1 minuto tra un tragitto e il successivo. Infine, data la mole di dati, è stato necessario dividere il file in base al giorno di registrazione dello spostamento: gli spostamenti avvenuti tra il 6 e il 12 gennaio sono stati inseriti nel file *gps-fullname-6-12.csv*, mentre quelli avvenuti tra il 13 e il 19 gennaio nel file *gps-fullname-13-19.csv*.

A questi file sono stati aggiunti due dataset ulteriori:

- Il file *locations.csv*, associa ad ogni locale di Abila le sue coordinate geografiche. Queste coordinate sono state estratte dai dati del GPS, considerando come latitudine e longitudine i punti di arrivo o partenza dei percorsi dei dipendenti che si stavano dirigendo verso (o allontanando da) quel determinato luogo.
- Il file *paths-united.csv* indica per ogni percorso effettuato dalle auto: il numero identificativo del percorso, le coordinate e l'orario di partenza e arrivo, il nome del dipendente e l'ID dell'automobile.

4 Scelte di design

Per il layout generale si è deciso di mantenere una visualizzazione piuttosto compatta, la cui comprensione è facilitata dall'utilizzo di un percorso guidato su come leggere i dati e delle spiegazioni testuali.

4.1 I componenti grafici

Per quanto riguarda il design si è deciso di dividere la visualizzazione in quattro componenti principali in modo da mostrare tutte le caratteristiche dei dati a disposizione:

1. Treemap, utile a visualizzare l'organizzazione aziendale
2. Timeline degli eventi, per monitorare cosa stavano facendo i dipendenti in ogni momento della giornata
3. Mappa, per mostrare gli spostamenti dei dipendenti e la posizione dei luoghi di interesse
4. Bar Chart, per confrontare le spese sostenute dai dipendenti con quelle dei colleghi a diversi livelli di aggregazione. Sono presenti due bar chart uno per i dati relativi alla Carta di Credito e uno relativo ai dati della Carta Fedeltà.

Questi componenti verranno descritti approfonditamente nella Sezione 5.

4.2 L'interattività

Per poter permettere il maggior livello di interattività possibile tutti i componenti grafici sono stati pensati per essere in una qualche misura collegati tra loro.

Più nello specifico la Treemap permette di filtrare i dati visualizzati nella mappa e nei bar chart per il sottoinsieme di dipendenti scelto. La timeline degli eventi permette di selezionare intervalli di tempo più brevi dell'intera giornata e filtrare così anche le altre visualizzazioni. Questa permette anche di selezionare un tragitto e rappresentarlo (in isolamento) nella mappa oppure su una transazione e vedere il confrontato con le spese medie dei colleghi. La mappa invece permette di selezionare un luogo di interesse per filtrare i bar chart.

Le modalità di interazione verranno descritte in dettaglio nella Sezione 5.

4.3 I colori

Per rendere la visualizzazione adatta alle persone affette da Daltonismo, si è deciso di optare per una palette di colori che vanno dal blu al giallo.

Per selezionare i colori presenti nella timeline degli eventi si è utilizzata una palette di colori realizzata appositamente per essere di comprensione ottimale anche per gli individui daltonici [6].

Sfortunatamente nella mappa, data la grande quantità di percorsi da rappresentare (fino a 256 per giorno), non è stato possibile scegliere una palette ad hoc, ma è stato utilizzato un generatore casuale di colori.

4.4 Il font

Per poter rendere la visualizzazione fruibile per le persone affette da Disturbi Specifici di Apprendimento (DSA) si è scelto di utilizzare un font *sans-serif*. In particolare, sono stati individuati i font *Avenir*, *Helvetica* e *Arial*, che verranno selezionati in base al browser utilizzato.

5 La visualizzazione

Per la costruzione della visualizzazione si è utilizzato il framework Vue [13], le librerie D3 [5] e CrossFilter [4] per la gestione dei dati, la libreria BootstrapVue [3] per il layout generale, le librerie ApexCharts [2] e Plotly [9] per la costruzione dei grafici e la libreria Leaflet [12] per la mappa. Il funzionamento è stato testato su browser Chrome.

L'intera visualizzazione è mostrata in Figura 1.

Si procede ora a descrivere per ogni componente, le scelte di design iniziali, l'implementazione e le modalità di utilizzo.

GAStech International employees analysis

GAStech International has been operating a natural gas production site in the island country of Kronos, in particular the city of Abila. In January 2014 several employee go missing after the initial public offering of the company. This visualization has the aim to help identify suspicious behaviors.

1. Select a day

The available data are two weeks before the disappearance of the employees.

Select a day to analyze it:

Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
6	7	8	9	10	11	12	13	14	15	16	17	18	19

3. Analyze the employees daily routine

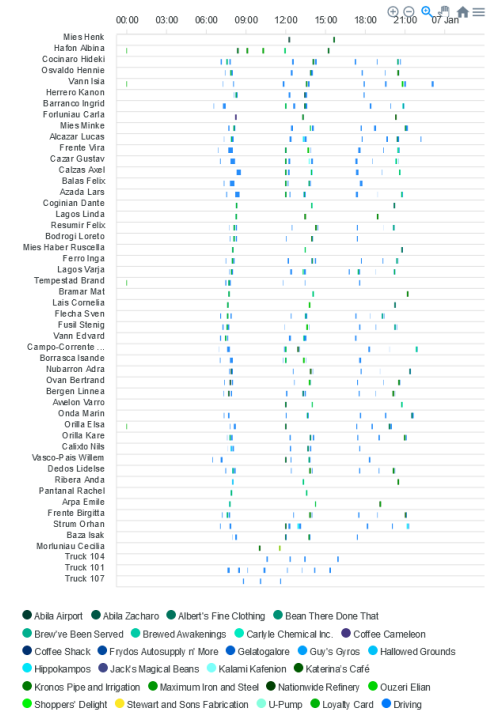
The timeline shows the daily schedule of each employee based on the GPS, the Credit Card, and the Loyalty Card data.

The time zoom permits to highlight a certain time of the day.

It is also possible to click on the data point: if the data refers to the GPS it will show the path on the map, while if it refers to Credit Cards or Loyalty Card transactions it will show an insight of the expense in the respective bar chart.

Finally, the legend of the locations, at the bottom of the chart, lets you filter the bar chart in conjunction to the company's structure TreeMap.

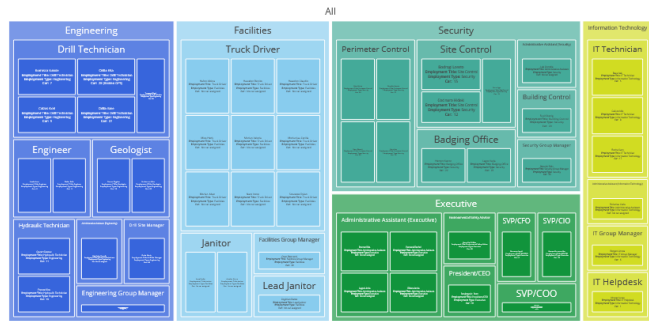
To reset the timeline, click on the house icon at the top-right on the visualization.



2. Company structure

GAStech International has 5 different branches, each employee has an employment type and an employment title. 35 of the employees use a company car.

Select an employment type, title or a employee to show in the map and in the bar charts the relative subset of data. To reset the chart, click on the main category 'All'.

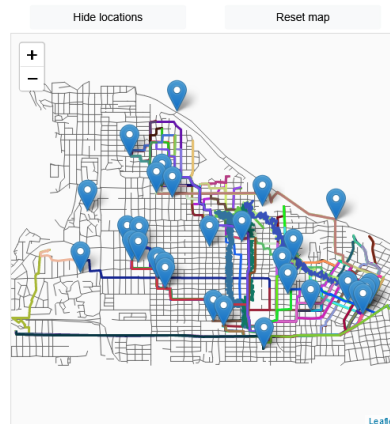


4. Abila map

Abila is the city on the island of Kronos where GAStech International is located.

The colored lines represent the paths of the employees. Each path has a different color. Pass over the lines to see who drove that itinerary.

The locations are also shown: pass over their icons to see their name or click them to filter the bar chart. It is possible to hide the locations or reset the map displaying all the paths of the selected day.



5. Compare the expenses

The expenses charts are useful to compare the average expenses of a certain employee or company branch to the others.

Filter the data using the treemap above, the locations in the map and the timeline. To remove the locations filter click the button below.

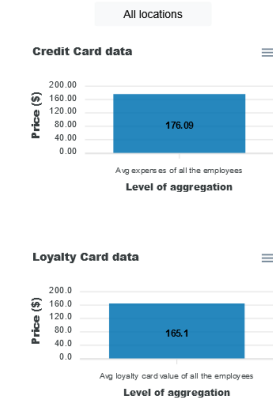


Figura 1: Visualizzazione finale

Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
6	7	8	9	10	11	12	13	14	15	16	17	18	19

Figura 2: Buttons per la selezione del giorno

5.1 Selezione del giorno

Data la grande quantità di dati, per facilitare l'utilizzo della visualizzazione si è scelto di rappresentare di volta in volta solo i dati relativi a un giorno specifico. Si è quindi creato con Bootstrap un *button group* - Figura 2 - in cui è presente un pulsante per ogni giorno che è possibile analizzare.

Selezionando uno dei *button* con le date si aggiorna l'intera visualizzazione e il giorno selezionato verrà evidenziato tramite un colore più scuro.

5.2 Struttura dell'azienda

Per poter mostrare le relazioni tra i dipendenti della GAStech si è utilizzata una TreeMap così da poter mostrare l'organizzazione interna dell'azienda. La *TreeMap* è stata creata utilizzando la libreria Plotly e i dati sono stati forniti in modo statico, poiché, dato che la struttura aziendale rimane stabile nel tempo, non era necessario appesantire ulteriormente l'applicazione e comporla in modo dinamico.

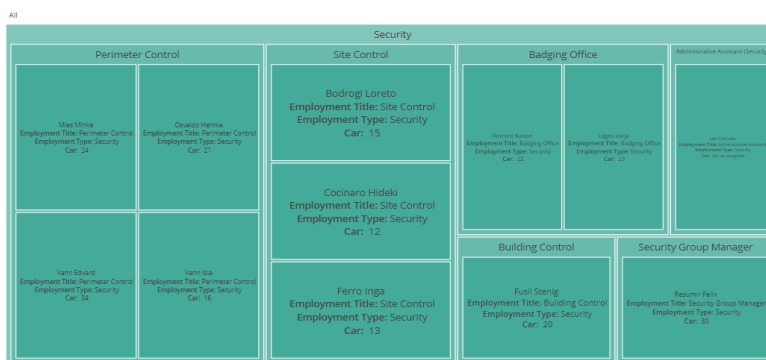


Figura 3: Focus sul reparto "Security"

Per comporre la TreeMap, che è possibile vedere completa in Figura 1, si è considerata innanzitutto l'unità organizzativa di appartenenza dei dipendenti e poi il loro ruolo all'interno di questa unità. Quindi ad esempio, se si considera l'unità organizzativa addetta alla sicurezza in Figura 3: si ha come "parent" Security, questa unità ha sei "child", uno per ogni ruolo, e ognuno di loro è a sua volta "parent" dei dipendenti che svolgono quella mansione. Ad esempio, "Badging Office" è "parent" di Herrero Kanon e Lagos Varja.

Per ogni dipendente è stato poi aggiunto una piccola descrizione riportando ancora una volta i dati relativi all'impiego e specificando se ha o meno un'automobile aziendale e, nel caso, il numero identificativo della stessa.

Cliccando sui riquadri che rappresentano le sotto-categorie si avrà una visualizzazione in dettaglio delle stesse e verrà modificata la visualizzazione della mappa e dei bar chart considerando solo gli individui appartenenti a quella sotto-categoria o il dipendente stesso. È possibile riportare la TreeMap allo stato iniziale, annullando così la selezione per sotto-categorie, selezionando la macro-categoria "All" in alto.

5.3 Timeline degli eventi

La timeline degli eventi è utile per mostrare cosa stiano facendo i dipendenti della GASTech in ogni momento della giornata. In questa timeline, infatti, viene mostrato per ogni ora del giorno quando un individuo sta guidando e quando usa la carta di credito o quella di fedeltà. L'obiettivo principale di questo componente infatti è mostrare quale sia la giornata-tipo degli individui ed evidenziare pattern o comportamenti sospetti. Ad esempio, se tutti i dipendenti effettuano dei pagamenti a un ristorante nello stesso momento possiamo assumere che abbiano pranzato insieme. Inizialmente si era scelto di differenziare le tre categorie di dati (dati del GPS, della carta di credito e della carta di fedeltà) tramite delle forme diverse. Sfortunatamente, a livello implementativo non è stato possibile ottenere questo risultato e si è dovuto optare per una divisione in base al colore aiutata da input testuali (*tooltip* e legenda).

La timeline è stata creata utilizzando la libreria ApexCharts: si è generata una *series* di dati per ogni luogo di interesse e in aggiunta una *series* per identificare tutte le registrazioni della carta fedeltà (identificata nella legenda come "Loyalty Card") e una per i dati del GPS (identificata come "Driving").

Nonostante lo svantaggio dato dalla mancanza di differenziazione tra le forme, questa timeline permette di cambiare il range temporale preso in considerazione, dando la possibilità di focalizzarsi solo su determinati momenti della giornata e rendendo così più visibili anche i diversi colori.

Quando si sceglie un diverso range temporale (sia cliccando le opzioni in alto a destra, sia agendo direttamente sul grafico) si modificano sia la mappa che i bar chart con l'intervallo selezionato. Un esempio è presentato in Figura 4.

Passando il cursore sopra ai nomi dei luoghi nella legenda, questi verranno messi in risalto sul grafico, mentre cliccandoli la location verrà nascosta e filtrerà i bar chart con le sole spese relative ad essa.

Per aiutare ulteriormente nella comprensione, avvicinando il cursore ai diversi simboli del grafo è possibile vedere un *tooltip* con i dettagli dell'utilizzo della carta o la durata del tragitto effettuato con l'automobile. Selezionando le barre relative alle carte di credito o fedeltà, si vedrà nel relativo bar chart un confronto tra la spesa avvenuta in quel momento,

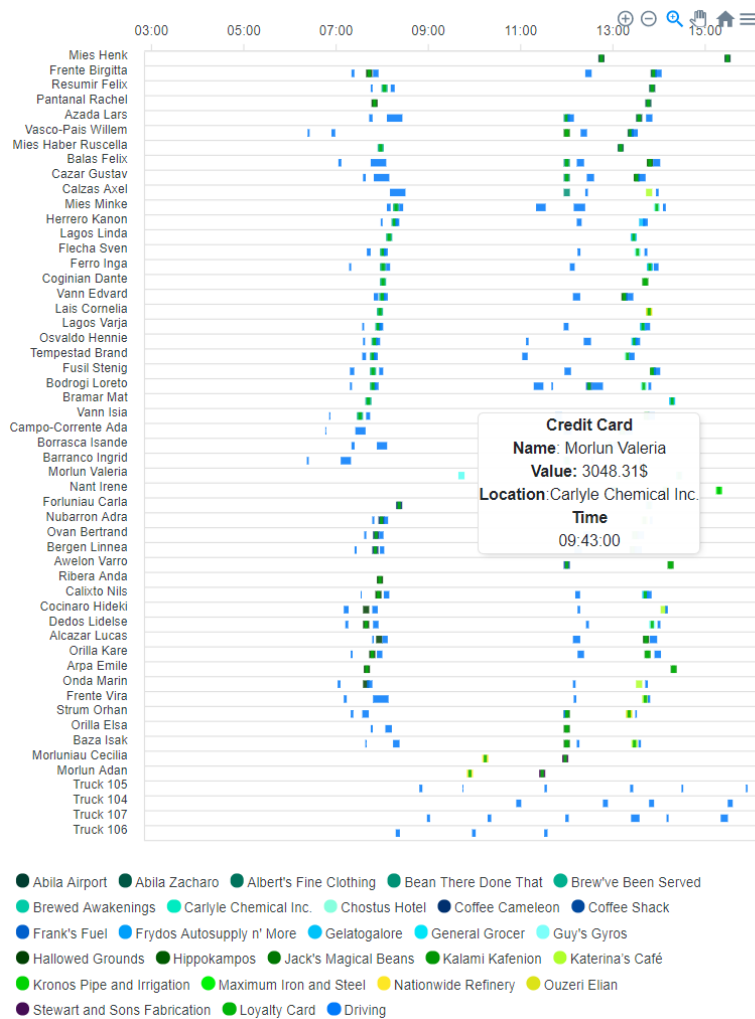


Figura 4: Esempio di zoom sulla timeline e tooltip per una transazione con Carta di Credito

la spesa media del dipendente in quel locale e la spesa media degli altri colleghi sempre ristretta al locale. Se la barra rappresenta un tragitto, cliccando sarà possibile visualizzarlo (in isolamento) sulla mappa.

Per resettare la visualizzazione è necessario premere l'icona della casa in alto a destra.

5.4 Mappa

La mappa è utile a mostrare gli spostamenti dei dipendenti e la posizione dei luoghi di interesse. Per crearla si è utilizzata la libreria Leaflet, in particolare si sono usati: un componente *GeoJson* per la generazione della mappa e delle *Polyline* per i diversi tragitti delle automobili. Per collocare i luoghi sulla mappa sono stati utilizzati dei *Marker*.



Figura 5: Visualizzazione sulla mappa di un sottoinsieme di percorsi e relativo *tooltip*

È possibile vedere il nome delle vie avvicinando il cursore alle stesse. Allo stesso modo è possibile vedere per ogni tragitto: chi era il dipendente che ha percorso quel tratto e l'orario di partenza. Un esempio è mostrato in Figura 5.

Cliccando sui Marker che indicano i luoghi è possibile filtrare i bar chart per il confronto dei costi.

I pulsanti sopra la mappa permettono rispettivamente di nascondere (o mostrare) i luoghi di interesse e di portare la mappa al suo stato iniziale.

5.5 Confronto spese

Infine, si è ritenuto importante offrire un confronto immediato anche in merito alle spese sostenute con carta di credito e l'utilizzo di carte fedeltà. L'obiettivo principale della rappresentazione è quello di evidenziare se sono presenti delle spese "anomale" da parte di uno o più dipendenti.

Si è quindi creato un componente con la libreria ApexCharts per generare il bar chart e si è poi utilizzato due volte: una per le carte di credito e una per quelle di fedeltà.

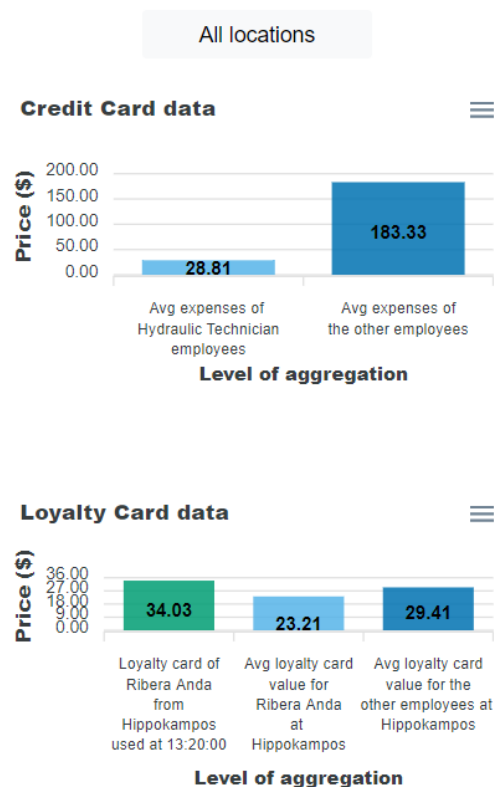


Figura 6: Esempi di barchart

In dettaglio, filtrando i dati come spiegato nelle sezioni precedenti, è possibile visualizzare (sia per le Carte di Credito, che per le Carte Fedeltà):

1. Spesa media di tutti i dipendenti GASTech
2. Spesa media dei dipendenti che fanno parte di una unità organizzativa vs spesa media dei dipendenti delle altre unità organizzative
3. Spesa media dei dipendenti che hanno un determinato ruolo vs spesa media degli altri dipendenti

4. Spesa media di un determinato dipendente vs spesa media degli altri dipendenti
5. Singola spesa di un dipendente vs spesa media del dipendente vs spesa media degli altri dipendenti in uno specifico luogo di interesse

Oltre il punto 5, è possibile filtrare per luoghi di interesse anche i grafici precedenti. Per poter cancellare tutti i filtri per locali è necessario cliccare il pulsante "All locations" posto sopra ai grafici.

6 Esempi di utilizzo

6.1 La giornata-tipo del dipendente GASTech

Semplicemente osservando la timeline degli eventi è possibile notare che nei giorni lavorativi i dipendenti della GASTech International la mattina si recano a fare colazione al bar (i locali più frequentati sono: Bre've Been Served, Coffee Cameleon e Halloweed Ground) e poi al lavoro. Durante la pausa pranzo si recano generalmente a mangiare in ristoranti o bar (Katerina's Café, Brewed Awakenings, Hippokampos, Abila Zacharo, Kalami Kafenion, Bean There Done That, Gelatogalore e Guy's Gyros), successivamente tornano a lavoro. A fine turno tornano a casa, o si fermano a cenare in ristoranti (Katerina's Café, Ouzeri Elian, Hippokampos e Guy's Gyros) prima di recarsi casa.

Invece, nel fine settimana, i dipendenti rimangono in casa fino all'ora di pranzo, quando si recano a mangiare in ristoranti o café (come nei giorni lavorativi) e poi si dedicano ad attività ricreative (ad esempio visitare il museo, comprare vestiti o andare a trovare gli amici). Infine, dopo aver cenato fuori, si recano di nuovo a casa.

6.2 Un esempio di analisi tramite la visualizzazione

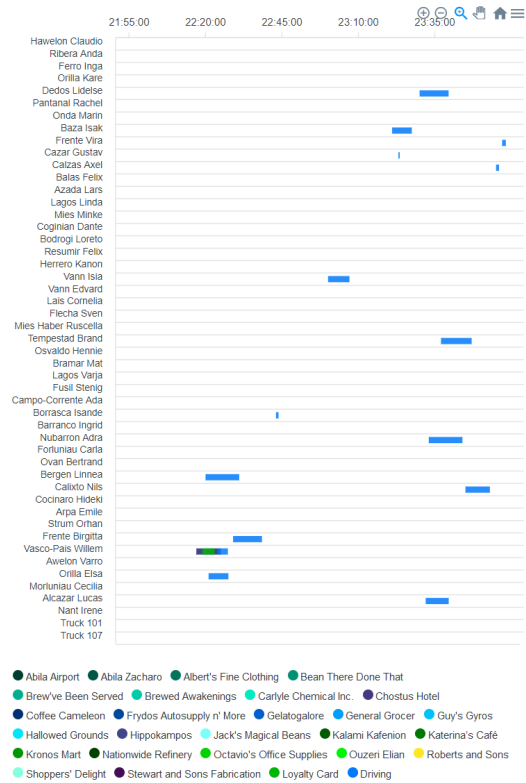
La sera di venerdì 10 possiamo notare che molti dipendenti di GASTech si spostano con l'automobile tra le 22:30 e mezzanotte. Quindi, se si sceglie quell'intervallo di tempo per approfondire l'avvenimento in Figura 7 (a) si vedrà dalla mappa in (b) che i dipendenti si trovavano tutti nello stesso luogo, probabilmente a casa di un collega.

Seguendo lo stesso ragionamento, è possibile individuare anche l'ora di arrivo nel luogo dei diversi dipendenti. Guardando quindi la timeline totale della giornata possiamo identificare quale sia stato l'ultimo spostamento effettuato con l'automobile dai soggetti coinvolti e selezionare quell'intervallo di tempo per eliminare i tragitti superflui dalla mappa - Figura 8 (a). Così facendo anche in questo caso si può notare come il luogo di arrivo sia la casa del collega - Figura 8 (b)- e guardando l'orario (circa tra le 18:30 e le 19:30) possiamo ipotizzare che questo incontro sia avvenuto per cenare insieme.

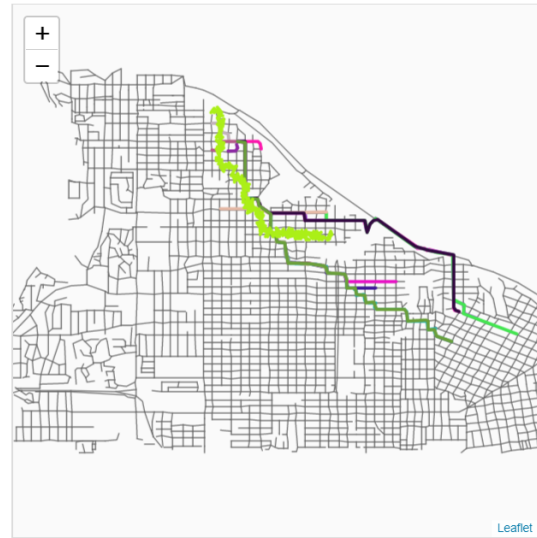
Infine, filtrando la mappa in base alle diverse unità lavorative possiamo notare come i partecipanti a questa cena fossero per lo più delle unità di Ingegneria e Information Technology. I loro tragitti sono infatti diretti verso la casa del collega. Mentre, possiamo anche notare come i dipendenti dell'unità Sicurezza fossero insieme da Frydos Autosupply n' More.

7 Conclusioni

In conclusione, possiamo affermare che la visualizzazione creata permette, seppur con alcuni limiti, di analizzare i dati messi a disposizione considerando diverse sfaccettature.

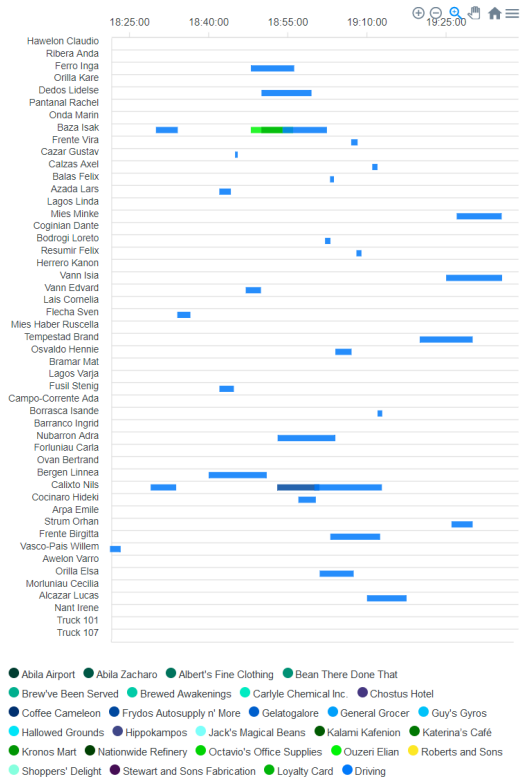


(a) *Timeline*

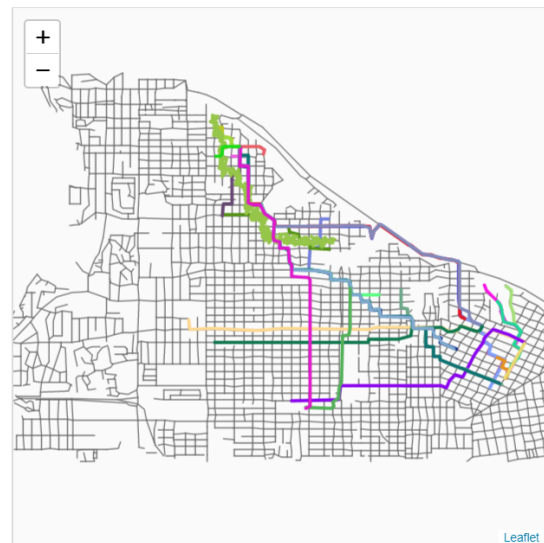


(b) *Tragitti*

Figura 7: Timeline e mappa tra le ore 22:30 e ore 00:00 di venerdì 10



(a) *Timeline*



(b) *Tragitti*

Figura 8: Timeline e mappa tra le ore 18:30 e ore 19:30 di venerdì 10

Riferimenti bibliografici

- [1] Chen Guo et al. *Purdue-Guo-MC2*. 2014. URL: <https://www.cs.umd.edu/hcil/varepository/VAST%5C%20Challenge%5C%202014/challenges/MC2%5C%20-%5C%20Patterns%5C%20of%5C%20Life%5C%20Analysis/entries/Purdue%5C%20University/>.
- [2] *ApexCharts*. URL: <https://apexcharts.com/>.
- [3] *BootstrapVue*. URL: <https://bootstrap-vue.org/>.
- [4] *Crossfilter 2*. URL: <https://square.github.io/crossfilter/>.
- [5] *D3 Data-Driven Document*. URL: <https://d3js.org/>.
- [6] *Designing for Color Blindness*. URL: <http://mkweb.bcgsc.ca/colorblind/palettes.mhtml>.
- [7] IDCMDX. *2014 VAST Challenge - MC2 - Middlesex University & MASS*. Youtube. 2014. URL: <https://www.youtube.com/watch?v=S7BSnZ3SLgY>.
- [8] *Mapshaper*. URL: <https://mapshaper.org/>.
- [9] *Plotly*. URL: <https://plotly.com/>.
- [10] Natascha Siirak. *VAST Challenge 2014 - University of Konstanz, Team mc2*. Youtube. 2014. URL: <https://www.youtube.com/watch?v=THLB7JKYThM>.
- [11] *VAST Challenge 2014*. URL: <https://vacommunity.org/VAST+Challenge+2014>.
- [12] *Vue Leaflet*. URL: <https://vue2-leaflet.netlify.app/>.
- [13] *Vue.js*. URL: <https://vuejs.org/>.