Intellidata: Visual Analytics for Bank Data

- App ID: inteli-data
- URL: http://www.weblab.deusto.es/intellidata/
- Contacto: Aitor Almeida (aitor.almeida@deusto.es) y Pablo Orduña (pablo.orduna@deusto.es)

Intellidata es una aplicación para la exploración de los datos de consumo mediante diferentes visualizaciones. Intellidata permite identificar diferentes patrones de consumo, ya sean geográficamente o temporalmente.

Procesado inicial de datos

Para poder procesar los datos de manera sencilla el primer paso ha sido realizar un volcado de los datos que se encontraban en la API. Esto permite agregarlos y procesarlos de manera más potente. Para ello se ha creado un programa que scrapea el API permutando todas las posibles combinaciones de consultas.

Una vez se han recuperado todos los datos se ha procedido a agregarlos para poder realizar consultas sobre ellos de manera más sencilla y eficiente. Se ha seleccionado MongoDB como base de datos sobre la que volcar toda esta información. El objetivo es poder beneficiarse de las agregaciones y el soporte de map-reduce para poder generar fácilmente la información disponible en diferentes vistas, optimizadas para cada búsqueda. De esta manera, la información está almacenada de manera redundante, pero permite ejecutar las búsquedas más rápidamente. Por ejemplo, en una sóla búsqueda optimizada se puede obtener toda la información de cubos de un código postal concreto sin necesidad de ejecutar nada más ni en el propio servidor de MongoDB ni en la capa de aplicación.

Los procesos que más tardan en ejecutarse en toda la aplicación son la generación de mapas. Se consideró el uso de Google Maps y de CartoDB. A pesar de que este último era especialmente interesante por mostrar gráficos avanzados de calidad y facilidad de uso, se optó finalmente por una solución offline que no dependiese de un contrato con un proveedor externo. El principal problema es que hay que generar los mapas, y un mapa con muchos datos puede llegar a tardar del orden de 7 segundos (y luego en el navegador puede llegar a tardar dos segundos en mostrarlo). En estos momentos se utiliza una caché para no generar un mismo mapa si ya está generado, lo cual no evita que el navegador siga tardando en renderizarlo. Esto es así porque el mapa descarga un gráfico vectorial por el cual el usuario puede navegar, haciendo zoom y moviéndose por el mapa para ver los datos. Se ha considerado llenar la caché con todas las combinaciones, ya que el resultado ocuparía menos de medio terabyte.

Aplicación web

La aplicación se divide en tres categorías, "*Global*", "*Local*" y "*Search*". Dentro de la categoría *Global* están aquellas visualizaciones relacionadas con los datos en su conjunto, sin dividirlos por zipcodes. Dentro de *Local* se lleva a cabo un análisis pormenorizado por cada zipcode. Dentro de la categoría *Search* se puede realizar búsquedas de zipcodes dependiendo del grupo de edad, sexo, categoría y tipos de gastos para ver cuáles son las localizaciones donde más ha gastado ese grupo.

Global > Relationships

Esta visualización muestra las relaciones entre diferentes códigos postales. Para crear esta visualización se han realizado varios pasos:

1. Se han recuperado las coordenadas geográficas de los códigos postales.
2. Se han calculado las relaciones entre códigos postales. Esto se ha hecho haciendo uso del método *customer_zipcodes* del API.
3. Con estas relaciones se ha creado el grafo que indique las relaciones "sociales" que se originan entre los diferentes zipcodes. El zipcode del cliente es el origen de la relación y el zipcode de la tienda donde se ha efectuado la compra el destino. El peso de la arista será el total del gasto realizado.
4. Se ha creado una visualización del grafo usando como layout las coordinadas geográficas. Se ha decidido utilizar este método en vez de usar otros layouts de visualización de grafos como los basados en repulsión de fuerzas para que sea más sencillo poder apreciar las relaciones que se crean entre las diferentes regiones de España.
5. Se ha aplicado un algoritmo de detección de comunidades para poder identificar aquellos zipcodes más relacionados entre sí. Después de analizar los resultados de varios algoritmos (K-core, K-corona, K-cliques…) se ha decidido usar el método de Louvain basado en alcanzar aquella partición del grafo que maximice la modularidad.

El resultado es el que se puede observar en la Figura 1. El mapa interactivo que se ha creado permite su navegación haciendo uso del ratón (rueda para hacer zoom, arrastrar para moverse por él). Este mapa permite observar fácilmente que localizaciones centran sus gastos en las zonas de Madrid y Barcelona. Como se ve en la Figura 1 aquellos zipcodes de color verde son los relacionados con Madrid y los de color amarillento con Barcelona. En resto de zipcodes (azul y morado) son los que se encuentran relacionados con ambos núcleos y su color variará dependiendo si tienen más relación con uno u otro.

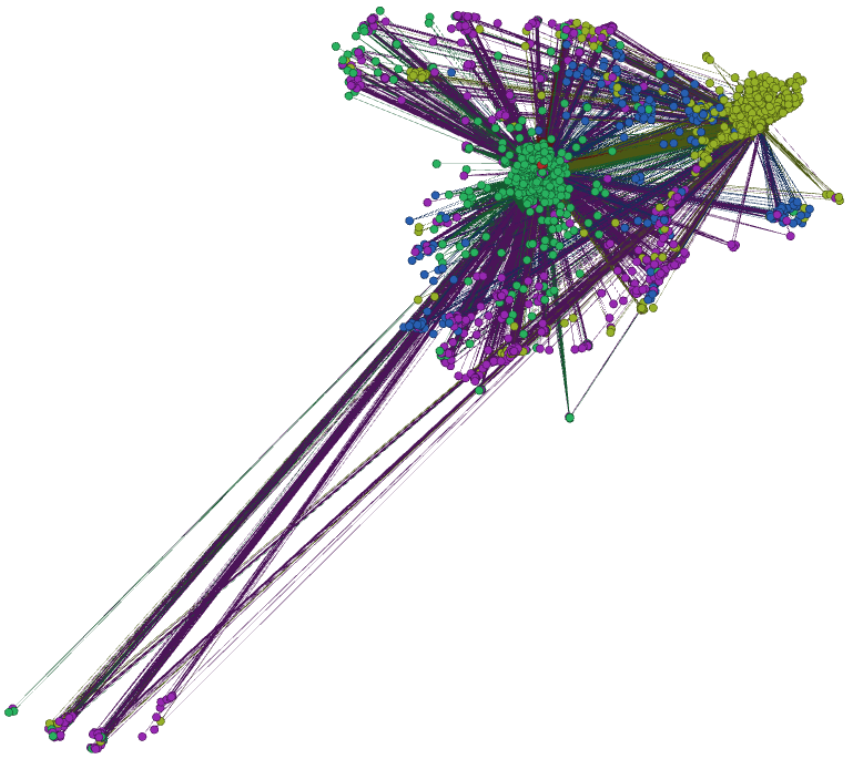


Figura 1 Visualización de las relaciones entre zipcodes

Global > Adyacency Matrix

Esta visualización permite explorar las relaciones entre zipcodes por categorías y filtrándola por distancia geográfica. La aplicación permite seleccionar una categoría y una de las distancias mínimas predefinidas (50, 100, 200, 300 y 400 kilómetros). Se ha calculado de nuevo el grafo de relaciones, filtrándolo por la categoría seleccionada y usando las coordenadas geográficas para filtrar aquellas relaciones que superan la distancia mínima. En este caso se han vuelto a identificar las comunidades haciendo uso de nuevo del método de Louvain.

Con ello se ha creado una matriz de adyacencia que permite visualizar estas relaciones (ver Figura 2). Los colores de la matriz identifican la comunidad a la que pertenece el nodo de origen (en este caso los nodos de la columna derecha). La transparencia del color indica el peso de la relación (estando indicado este por el número de pagos efectuados). La matriz puede ser ordenada por nombre de los nodos (Name), por el peso de las relaciones (Frecuency) o por la comunidad a la que pertenecen (Clustering).

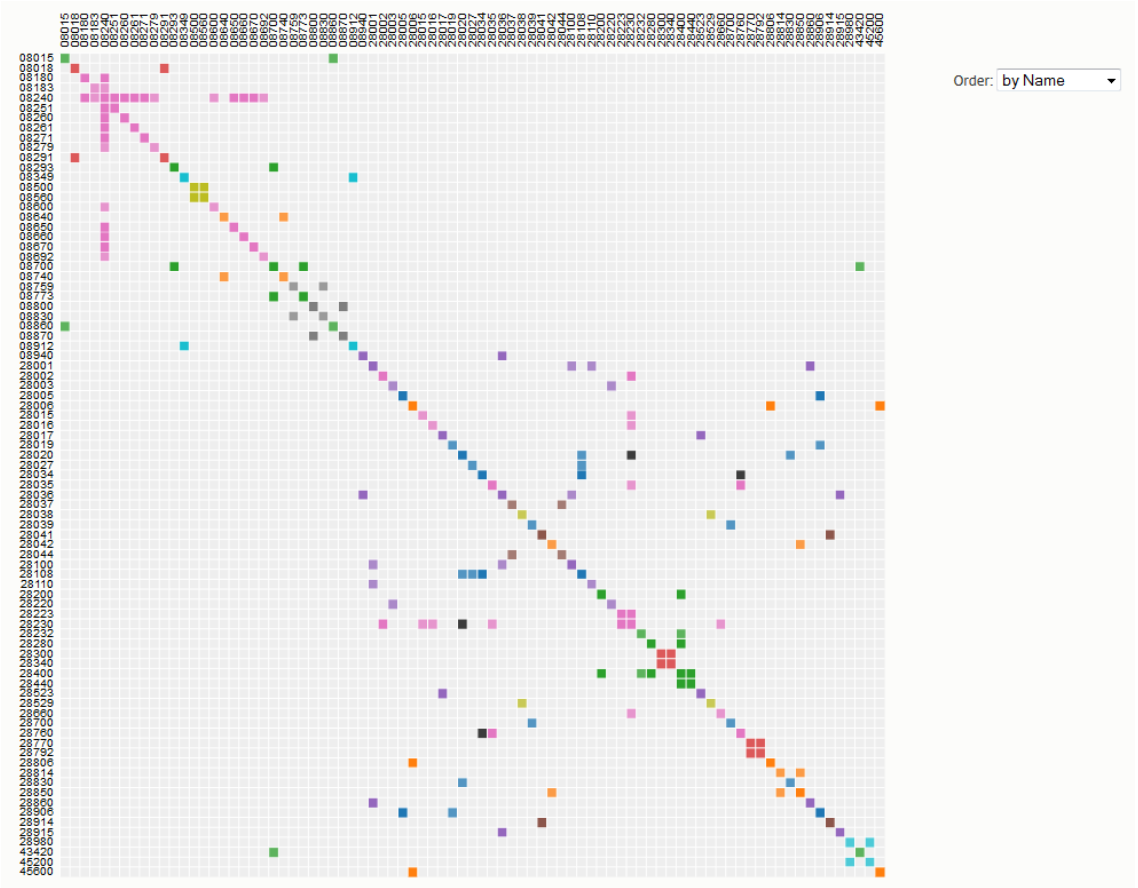


Figura 2 Matriz de adyacencia para la categoría es_auto para más de 50 kilómetros

Esta visualización permite explorar como cambian las relaciones de los zipcodes dependiendo de la distancia que existen entre ellos y de la categoría a la que pertenecen esas compras. Es posible observar cuales son las categorías en las que la distancia geográfica no es un problema y extraer conclusiones Por ejemplo se producen muchos gastos de *bar & restaurants* en los desplazamientos, pero no así en la categoría *home*.

Global > Summary

La pestaña de summary contiene un resumen general de los datos analizados. Para ello se crea un grafo de las relaciones entre zipcodes como se ha explicado anteriormente para cada categoría. Por cada categoría se han identificado los siguientes zipcodes:

- Zipcode más importante en la categoría. Para identificar el zipcode más importante se ha procedido a efectuar un análisis de las centralidades del grafo. Después de realizar varias pruebas se ha decidido utilizar la eigenvector centrality[1] frente a otras centralidades (degree, betweenness, closeness…) y a otros métodos para el cálculo de la importancia de nodos (HITS, PageRank…).
- Zipcode que más desplazamientos ha realizado en esa categoría. El zipcode con mayor outdegree[2] en el grafo.

[1] http://en.wikipedia.org/wiki/Eigenvector_centrality#Eigenvector_centrality
[2] http://en.wikipedia.org/wiki/Outdegree#Indegree_and_outdegree

- Zipcode que más desplazamientos ha recibido en esa categoría. El zipcode con mayor indegree en el graf.
- Zipcode que más dinero ha gastado en la categoría.
- Zipcode que más dinero ha recibido en la categoría.

Esta visualización permite identificar fácilmente los zipcodes más importantes dentro de cada categoría, pudiendo ser utilizada para identificar aquellos zipcodes que pueden resultar más interesantes para diferentes campañas.

Además se muestran tres timelines que muestran la evolución del total de gastos, el total de pagos y el gasto medio en el tiempo, pudiendo identificar aquellos momentos del año en el que los gastos suben o bajan. También es posible visualizar estos timelines por categoría, pudiendo de esta manera ver cómo afecta la fecha a cada categoría. Un ejemplo de esto se ve claramente en la Figura 3, mientras que los gastos en tecnología tienen un pico importante en navidades, los gastos en hoteles se mantienen relativamente constantes durante el año para viajar después de navidades.

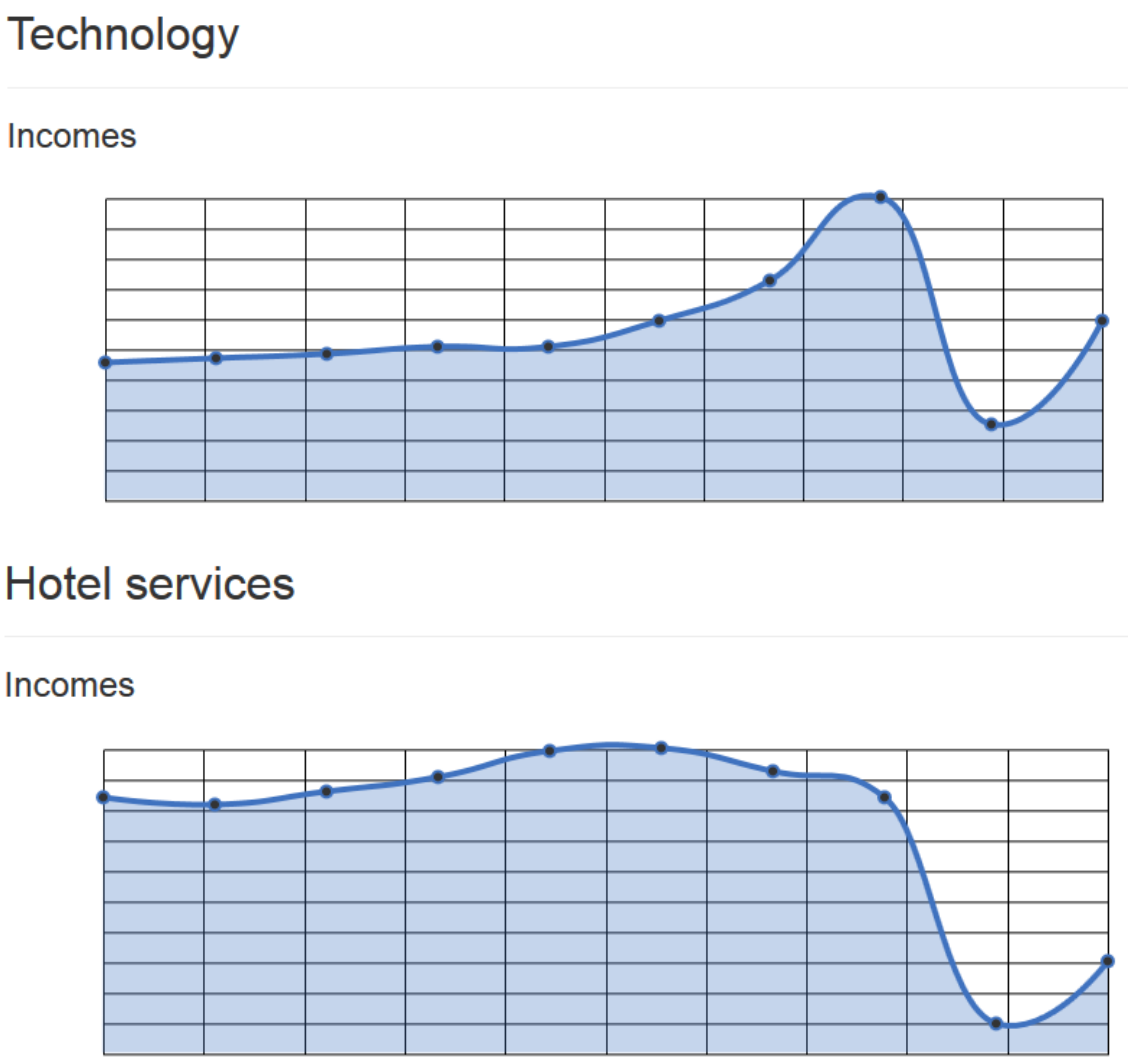


Figura 3 Ejemplo de variación entre categorías

Local > Timelines

La sección de timelines permite ver la evolución de diferentes métricas para cada zipcode a lo largo del tiempo. Actualmente se contemplan tres métricas. El pago medio, el gasto total y el número total de pagos.

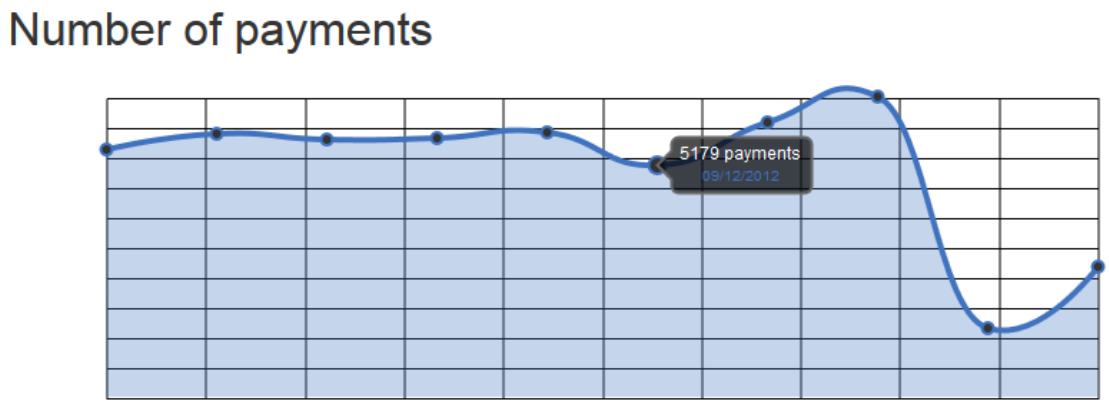


Figura 4 Evolución de la cantidad de pagos en un zipcode

Esta visualización permite ver como evoluciona el gasto en un zipcode en los diferentes momentos del año, para identificar los picos de gastos y aquellas épocas en las que los gastos disminuyen (por ejemplo, después de navidades).

Local > Timetables

La sección de timetables permite visualizar los patrones en cuanto a horas y días de la semana que siguen los gastos de un zipcode. De esta manera se pueden identificar aquellos momentos de la semana más importantes de una categoría en una localización. Por ejemplo en la Figura 5 se puede ver el patrón temporal de gastos para la categoría *Bars & Restaurants* para el zipcode que incluye Sol y Gran Vía en Madrid. Se puede ver cuáles son las franjas horarias y los días con mayor gasto. En este caso siendo la hora de la comida y las noches y sobre todo centrándose los gastos en fin de semana.

Total

	M	T	W	T	F	S	S
00:00	10747.18	11906.47	20908.40	27504.45	37679.04	62815.48	63871.98
01:00	6908.09	3295.38	7448.36	9637.27	12630.03	27525.27	28178.58
02:00	4846.25	1286.66	6073.10	6002.93	7265.17	14335.31	16501.01
03:00	4828.18	3820.55	1626.66	5829.55	6476.84	11248.64	16005.29
04:00	2460.77	1691.99	1708.54	3460.40	3585.15	10168.81	14758.84
05:00	2462.14	1694.97	2245.97	2562.25	3687.28	6958.02	10340.09
06:00	117.30	291.02	247.00	203.32	160.00	2135.94	2478.08
07:00	108.17	47.94	113.64	112.42	120.46	594.18	214.58
08:00	270.12	292.71	318.07	288.55	301.29	513.12	247.06
09:00	719.39	460.41	553.79	826.89	1074.26	554.13	330.66
10:00	1047.95	830.86	875.46	881.81	929.90	735.37	596.33
11:00	1276.98	1356.75	1878.11	1671.70	1691.32	1408.25	1203.54
12:00	2064.75	2040.95	3083.76	2706.91	2905.32	2399.52	1885.62
13:00	8285.28	9214.70	10526.69	10378.76	8093.14	6911.43	6292.22
14:00	24255.22	27865.01	26922.60	31531.36	25916.07	31006.14	28159.54
15:00	47961.49	52412.47	54927.39	63886.82	54663.49	92690.39	80594.96
16:00	41351.01	43940.26	50769.12	67274.01	66869.43	117980.05	99167.39
17:00	14707.95	17221.91	13770.97	24024.36	27088.65	59937.23	37896.93
18:00	6556.44	6146.13	4806.48	8284.17	11775.43	15109.35	8804.17
19:00	6882.94	6070.71	6562.71	6421.78	7642.59	10276.77	7718.40
20:00	6643.55	6801.64	8058.95	10339.99	9873.40	13075.62	9226.38
21:00	10879.38	14487.33	14412.96	19434.34	27286.91	31946.74	16711.49
22:00	19390.64	25953.96	28308.99	34455.20	49884.15	58899.43	25826.12
23:00	23505.33	34386.96	36738.88	53638.61	66437.37	75579.01	22192.09

Figura 5 Patron temporal de gastos totales para el zipcode 20013

Se han creado mapas de calor temporales para el gasto máximo, el gasto medio el gasto total y el número de pagos.

Local > Map

Esta última sección proporciona información detallada de todos los pagos realizados en un zipcode. Por un lado proporciona datos sobre el gasto total, el total de pagos y el número de tarjetas utilizados en el zipcode. También muestra el top de zipcodes que más han gastado, más pagos han hecho y mayor pago medio han tenido en el zipcode analizado (ver Figura 6).

#	Value
Incomes:	9569334.32
Payments:	225566
Cards:	175307

#	Top zipcodes	Incomes
1	28005	717567.66 €
2	28045	440512.09 €
3	28004	431006.22 €
4	28013	430995.38 €
5	28012	410451.57 €

More

#	Top zipcodes	Payments
1	28002	25953 payments
2	28005	14738 payments
3	28013	12047 payments
4	28004	10253 payments
5	28012	9205 payments

More

#	Top zipcodes	Average
1	45001	343.61 €
2	37900	313.40 €
3	35013	308.66 €
4	46007	301.01 €
5	35200	280.93 €

More

Figura 6 Resumen del zipcode

Por otro lado muestra un mapa de calor (ver Figura 7) de todos los zipcodes de origen de todas las personas que han realizado algún gasto en el zipcode. Este mapa es navegable, permitiendo el desplazamiento y zoom sobre el mismo. El mapa permite mostrar datos sobre gastos, pagos y número de tarjetas. Los colores de los zipcodes variarán de acuerdo con la métrica. Es posible cambiar el algoritmo de cálculo de estos zipcodes (ranked, linear y logarithmic), lo que permitirá obtener un coloreado más gradual o resaltar los outliers.

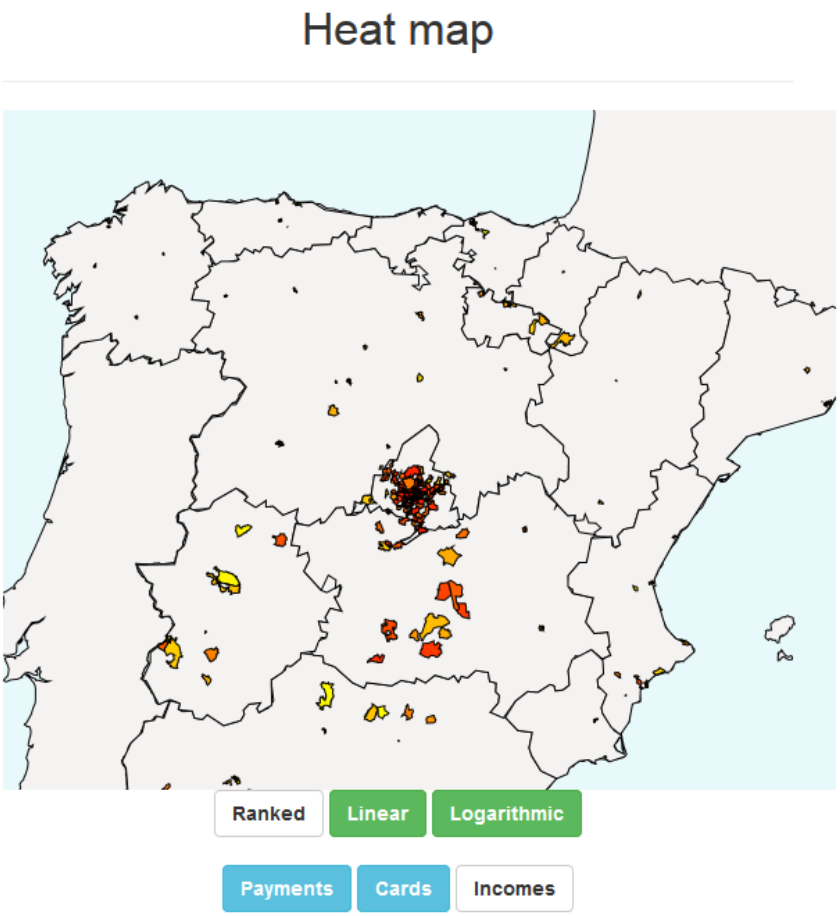


Figura 7 Mapa de calor

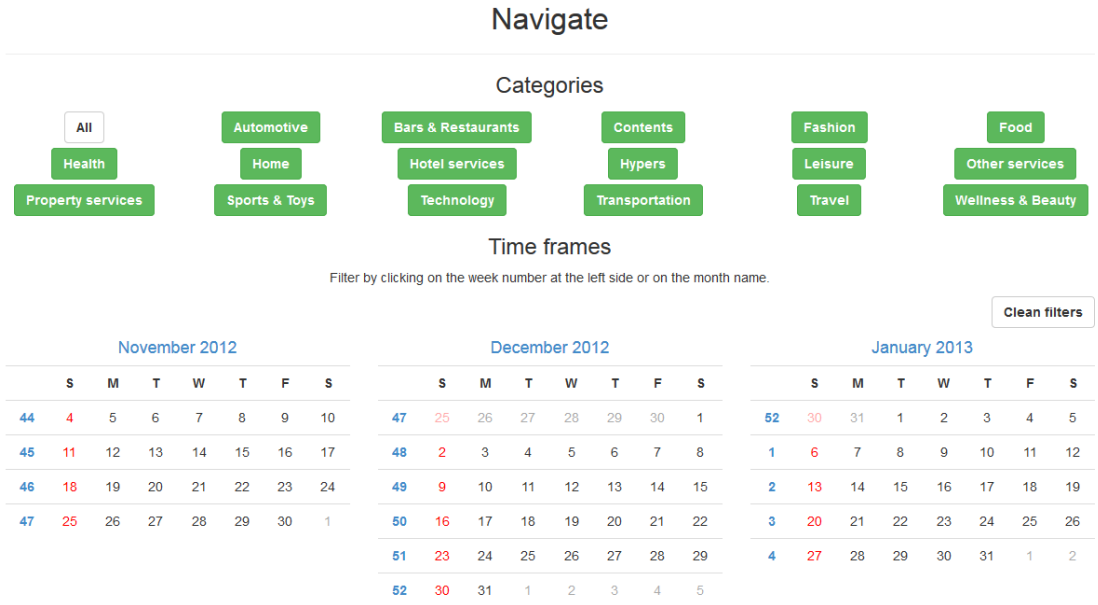
También se ha calculado una tabla de frecuencias para los datos demográficos de los usuarios que han realizado algún gasto en el zipcode. Esta tabla de frecuencias permite ver las diferencias entre sexos y entre franjas de edad en las compras realizadas.

Demography

Age	Male	Female	Company	Total
0-18	0.09 %	0.09 %	0.00 %	0.18 %
19-25	3.17 %	3.45 %	0.00 %	6.63 %
26-35	13.23 %	12.60 %	0.00 %	25.83 %
36-45	12.99 %	11.00 %	0.00 %	23.99 %
46-55	10.70 %	9.70 %	0.00 %	20.40 %
56-65	7.57 %	6.20 %	0.00 %	13.77 %
66+	4.76 %	3.79 %	0.00 %	8.54 %
Total	52.51 %	46.82 %	0.66 %	100.00 %

Figura 8 Tabla de frecuencia para los datos demográficos

Todos estos datos pueden ser filtrados con la herramienta de navigate al final de la página. Esta herramienta permite filtrar los datos visualizados por categoría y fecha (meses o semanas), pudiendo de esta manera refinar la búsqueda.



Esta sección permite explorar de manera muy sencilla los gastos efectuados en una localización, pudiendo conseguir todo tipo de información sobre ella: datos de los pagos, zipcodes de origen, distribución geográfica, información demográfica…

Search

Esta sección permite realizar búsquedas sobre los datos agregados. Los criterios de búsqueda son:

- Franja de edad: Franja de edad de los usuarios.
- Sexo: Sexo de los usuarios
- Categoría de gastos: Categoría en la que se quiere realizar la búsqueda
- Criterio de ordenación:: Gastos totales, números de pagos y gasto medio

Como resultado se obtendrá los 20 zipcodes más relevantes para el criterio de búsqueda. En la Figura 9 se pueden ver los 20 zipcodes más relevantes para la categoría *Bar & Restaurants* para mujeres en la franja de edad de 19 a 25 años teniendo en cuenta el criterio de número de pagos.

Zipcode	Expenditure	Number of payments	Average payment
28013	48591.11 €	3446.0 payments	14.10 €
28004	42857.15 €	2394.0 payments	17.90 €
28008	25956.44 €	2366.0 payments	10.97 €
28006	30951.88 €	1898.0 payments	16.31 €
28015	24389.80 €	1800.0 payments	13.55 €
28012	22560.92 €	1706.0 payments	13.22 €
28001	33792.62 €	1609.0 payments	21.00 €
28020	22009.81 €	1557.0 payments	14.14 €
28029	14591.20 €	1265.0 payments	11.53 €
28010	16196.48 €	1063.0 payments	15.24 €
28003	16900.18 €	942.0 payments	17.94 €
28046	17160.05 €	911.0 payments	18.84 €
28042	8663.30 €	822.0 payments	10.54 €
08005	8260.52 €	821.0 payments	10.06 €
28043	13225.85 €	806.0 payments	16.41 €
28036	15023.08 €	800.0 payments	18.78 €
08007	11757.54 €	777.0 payments	15.13 €
28916	13243.60 €	719.0 payments	18.42 €
28014	11445.84 €	665.0 payments	17.21 €
08030	9455.28 €	652.0 payments	14.50 €

Figura 9 Ejemplo de búsqueda