

Universidad de los Andes



Visual Analytics

Recaudo Bogotá

Entrega 2

Docentes

Tiberio Hernández

John Alexis Guerra

Integrantes

Meili Vanegas

Anamaria Mojica

Laura Cortés

Bogotá- 2016

Caracterización 2

What - Original 2

What - Procesado 3

WHY 3

HOW 3

Estado del arte 4

La olla a presión del transporte público en Bogotá 4

LIVE Singapore! 6

Caracterización

What - Original

Dataset info

Number of variables	7
Number of observations	3749798
Total Missing (%)	4.6%
Total size in memory	200.3 MiB
Average record size in memory	56.0 B

Variables types

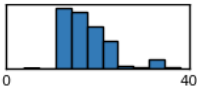
Numeric	3
Categorical	4
Date	0
Text (Unique)	0
Rejected	0

Warnings

- 2012-07-31 15:05:41 has a high cardinality: 3367899 distinct values Warning
- Unnamed: 5 has 605444 / 16.1% missing values Missing
- Unnamed: 5 has a high cardinality: 2802689 distinct values Warning
- Unnamed: 6 has 806394 / 21.5% zeros
- Unnamed: 6 has 605444 / 16.1% missing values Missing

id_estado
Numeric

Distinct count	28	Mean	17.032
Unique (%)	0.0%	Minimum	4
Missing (%)	0.0%	Maximum	38
Missing (n)	0	Zeros (%)	0.0%
Infinite (%)	0.0%		
Infinite (n)	0		



estado
Categorical

Distinct count	28	NUEVO	566749
Unique (%)	0.0%	EN PROGRESO	534393
Missing (%)	0.0%	CERRADO	523317
Missing (n)	0	Other values (25)	2125340

fecha_abierta
Categorical

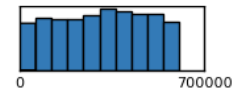
Distinct count	3367900	2016-05-30 10:45:07	40
Unique (%)	89.8%	2016-04-24 04:45:03	40
Missing (%)	0.0%	2016-08-07 14:45:02	40
Missing (n)	0	Other values (3367897)	3749679

fecha_cerrada
Categorical

Distinct count	2802689	2016-04-24 04:45:03	40
Unique (%)	89.1%	2016-08-07 14:45:02	40
Missing (%)	16.1%	2016-05-30 10:45:07	40
Missing (n)	605445	Other values (2802685)	3144234
		(Missing)	605445

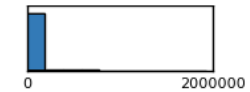
id
Numeric

Distinct count	605445	Mean	308430
Unique (%)	16.1%	Minimum	2
Missing (%)	0.0%	Maximum	605481
Missing (n)	0	Zeros (%)	0.0%
Infinite (%)	0.0%		
Infinite (n)	0		



tiempo
Numeric

Distinct count	191799	Mean	18664
Unique (%)	6.1%	Minimum	0
Missing (%)	16.1%	Maximum	1931400
Missing (n)	605445	Zeros (%)	21.5%
Infinite (%)	0.0%		
Infinite (n)	0		



What - Procesado

Se van a utilizar únicamente los registros en los que el estado ya ha sido cerrado.

- Se crea un DataSet en el que se almacenan únicamente el identificador, el estado y su duración.
- Se crea un segundo DataSet con la suma del tiempo de cada ticket, su identificador y su estado.

WHY

We identified 6 different tasks:

T1: **Present trends** of the total response time of all tickets (by type of incident).

T2: **Compare similarity** of the average response time, for each state in a defined range of time (week, day or hour), with the expected, which is the exact same range in the past month.

T3: **Identify outliers** from the response time in a defined range of time.

T4. **Browse outliers** from the response time filtering by operator, rush hour VS valley hour.

T5. **Compare distribution** of time for different tickets choosing the combination of states, based in the time that is selected in the filter.

T6 **Lookup** a single ticket with the corresponding delay between each of the different states.

HOW

For all tasks it is necessary to have in mind that in every moment the user can filter by any combination of states of the incident, with an starting point of all the states.

For T1 we are going to use a stacked Bar Chart, using the type of incident as a categorical attribute for **align, separate and order** in the y axis. On the other hand, we are using the states of the ticket as a second categorical attribute for **separating** the line using the **color hue** as the channel. Finally, the quantitative attribute of time is going to be encoded with the **length** of the **line**.

For T2 the user will have a filter to choose the period of time (month, day or hour) which depends of the availability of the data, using the same graph but adding a new stacked bar for each type of incident, representing the expected value¹. This new stacked bars are going to have their name in the 'y axis' to difference the expected from the actual period.

For T3, T4 we will use a scatterplot in order to display all the tickets generated in an specific range of time. The user will be able to filter the data according to each state. Using this idiom, we will be representing the amount of time, ordinal value, with the **position in the vertical scale**, the **position in the horizontal scale** will represent the timestamp. The **color hue** will categorize each point in an specific incident. The range filters will be given as a result of the **interaction** with the first visualization. In addition, the user can click in an specific ticket in order to **present the details**.

On the other hand, for T5 we are using a line chart, the user will select a maximum of 5 tickets from the scatterplot. We will use the categorical attribute state of the ticket, on the x axis it will be **aligned, separated and ordered**. Also, we will use the **vertical position** of the line to represent the time that took a ticket on that state. Finally, there is going to be a line for each ticket and are going to be differentiated by **color hue**.

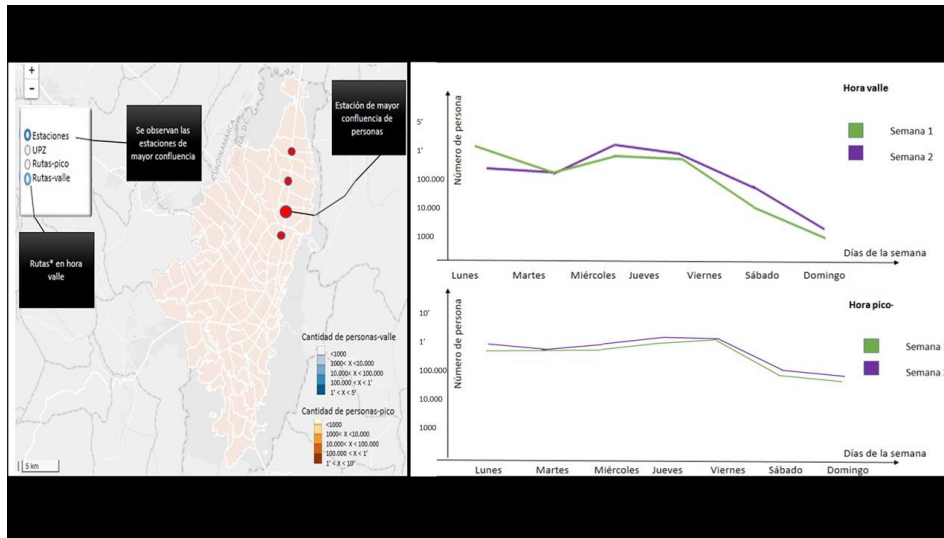
Finally, for T6 we will use a table and single horizontal bar. The table will present the principal information of the selected ticket like: the operator, the type of incident, dates (initial and final). And the single bar chart will be used to represent the total delay. The **length** will be the principal channel, representing the total delay and the different states of the incident will be represented by the **color hue**. For this task all the states will be represented, but the states that are no chosen by the user will be colored grey.

Mockups

A continuación, se muestra el proceso de mockups que se ha venido trabajando bajo las correcciones, observaciones de nuestros pares (otros estudiantes), y docentes del curso.

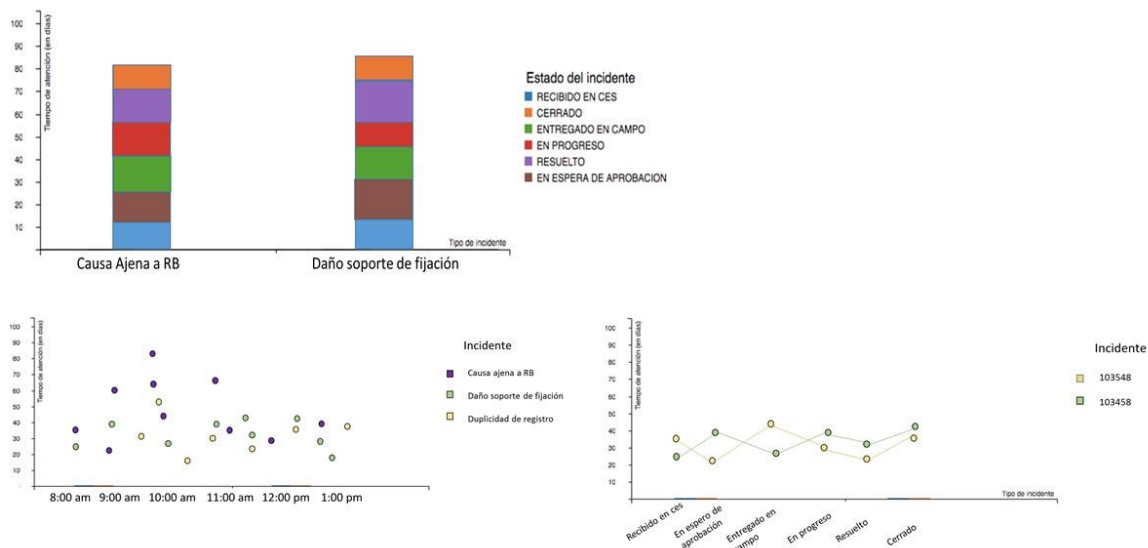
Mockup 1:

¹ Expected value: Is the data from a similar period of time that happened before. For example: the same week selected of the month before.



Se muestran las estaciones más influyentes, se comparan en los diagramas de línea las horas pico y las horas valle de las personas (entrantes). Luego de la retroalimentación recibida tanto en la clase (role-table) y de hablar con nuestro cliente (Recaudo Bogotá), se cambia el what, why, how a un enfoque de un problema operacional (np completo), en el que se manejan los estados de los diferentes tiquetes. Dado a esto, se genera un segundo mockup para tener la idea de lo que se quiere generar.

Mockup 2:



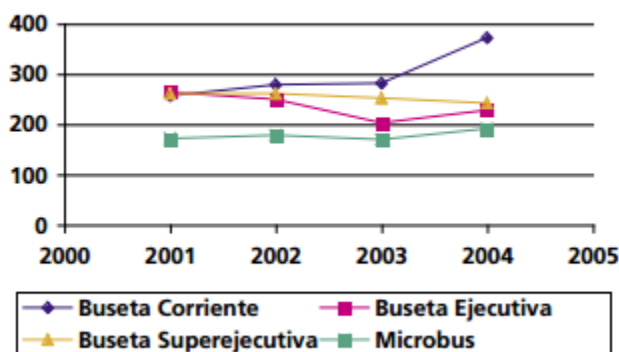
En este mockup 2 se habla del problema operacional de estados. Donde lo más importante es el manejo de estados, para que se pueda evidenciar retardo en los tiquetes y donde es generado.

Estado del arte

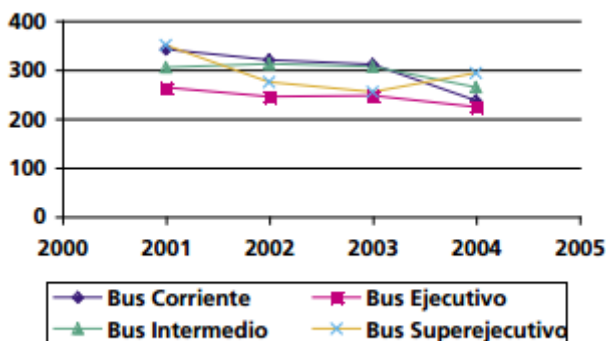
La olla a presión del transporte público en Bogotá

Este paper, realizado por Árdila Gomez en el 2005, habla del problema de Bogotá con la sobreoferta de vehículos en el transporte público. Lo que tiene lleva y es consecuencia de varios aspectos que afectan a los Bogotanos en el diario vivir. Los cuatro aspectos principales de lo que se habla son: primero una tarifa inflada, costosa, la cual mantienen la sobreoferta; segundo un manejo por parte del gobierno de manera débil; tercero un arreglo institucional para la prestación del transporte público plagada de incentivos perversos, lo cual incentiva la sobreoferta; cuarto trae actores porque la economía no ofrece mejores alternativas de inversión y trabajo. Con base a esto plantean una solución que se fundamenta en fortalecer el gobierno, un recaudo centralizado y cambiar el sistema de pagos de rodamiento por parte de los transportadores.

Para la solución, justificación de la misma se realizan visualizaciones acerca del número de pasajeros promedio que se censan en un periodo de días (ver ilustración: Pasajeros por día por buseta y microbús, Pasajeros por día por bus)



Pasajeros por buseta y microbús. Fuente: Datos del DANE y cálculos del autor (Tomado de: Árdila, La olla a presión del transporte público en Bogotá)



Con estas ilustraciones se quiere mostrar la baja demanda de transporte público en la ciudad y como esta ha disminuido a lo largo de los años.

Observaciones a la visualización:

La visualización carece de precisión, y dado que los datos en la ilustración de busetas y microbuses oscilan en un rango de 150 a 400, se podría manejar un rango de Y adecuado, ya que al empezar en 0, y terminar en 400 no permite que el lector distinga con claridad los datos. Lo mismo sucede con la segunda visualización, en la cual los datos oscilan entre 200 y 400. Por otra parte, el canal escogido, el *color hue*, no es el más adecuado, ya que se usan colores que no impactan o permiten la identificación de cada tipo de bus. La escala escogida tampoco permite ver el impacto de la disminución a lo largo de los años, lo cual es la tarea principal de esta.

Ardila Gómez, A. (2005). La olla a presión del transporte público en Bogotá. *Revista De Ingeniería*, 0(21), 54-65. doi:10.16924/riua.v0i21.404

LIVE Singapore!

El proyecto 'LIVE Singapore!' tenía como objetivo utilizar datos en tiempo real para convertirlo en herramientas dirigidas a los ciudadanos en las que pudieran basar sus decisiones de las actividades en la ciudad. Una de las características más interesantes, es que no hay casos de uso restringidos. En este proyecto, le dan valor al uso de la analítica visual dado que permite manejar grandes cantidades de datos e interpretarlos no solo de la manera tradicional, por el medio de la representación numérica y verbal, sino por medio de representación gráfica.

Por ello, en el proyecto se hace uso de diagramas espaciales en los que se representan datos como los trayectos de los taxis en Singapur, con el fin de mostrar el tiempo que se toma de ir de un punto a otro. Para esto, el mapa cambia de tamaño dependiendo del tráfico y el tiempo que se toma de llegar de un punto a otro, es así como el mapa reduce su tamaño si el tiempo que se toma de ir de un lugar a otro aumenta, en caso contrario, el tráfico disminuye, el mapa se agranda. (Cada área del mapa se reduce de manera independiente)



Figure 6

Observaciones de la visualización:

Teniendo en cuenta la visualización y el fin con la que se realiza los canales escogidos para la misma no son los mejores (movimiento, área). El movimiento (reducción del área), no permite identificar de manera correcta que tanto disminuye o aumenta el tráfico. Por otra parte, el área de cada parte del mapa al reducirse de manera independiente, no es de manera visual la más resaltante. Esto se da ya que sí se reduce el mapa en su totalidad, así un área individual no lo haga, se verá la impresión de que esto sí ha sucedido.

Fuente: Journal of Urban Technology Fecha: April 1, 2012

Referencias

- Fuente: Journal of Urban Technology Fecha: April 1, 2012
- Ardila Gómez, A. (2005). La olla a presión del transporte público en Bogotá. *Revista De Ingeniería*, 0(21), 54-65. doi:10.16924/riua.v0i21.404