Taller Ciencia de Datos

Letra tarea Laboratorio 2019

1. Información Administrativa

La tarea comienza el **lunes 19 de agosto** y finaliza el **Jueves 29 de agosto,** con una entrega parcial el **Jueves 22**. La tarea consiste en la carga y análisis de un conjunto de datos abiertos utilizando las herramientas vistas en el taller. La entrega final debe realizarse antes del **Jueves 29 de agosto** a las **23:59hs**. Consideraciones sobre la tarea:

- Se debe entregar un único archivo Jupyter Notebook (.ipynb) con todo el análisis
- Incluir dentro del archivo el nombre de los integrantes del grupo
- La tarea debe realizarse en grupos de 3 a 5 integrantes máximo
- Presentar el trabajo en clase el día **viernes 30 de agosto**, en no máximo de **15** minutos.

2. Objetivos

Esta tarea pretende la puesta en práctica y profundización por parte el estudiante de los siguientes puntos vistos en el taller:

- Entender cómo se implementa una ingesta de datos
- Poner en práctica técnicas para el análisis de la calidad de un conjunto de datos
- Practicar con herramientas de visualización de datos
- Aplicar la metodología de trabajo vista en el taller

3. Descripción

De cara a un año electoral muy reñido, de acuerdo a todos los pronósticos, el prestigioso diario "*El Durazno Post*" se propuso acercar a sus lectores un informe especial sobre las compras realizadas por el estado en el último período de gobierno. Como el diario no cuenta con científicos de datos, le ha encomendado a usted y su equipo la tarea de realizar un estudio inicial sobre dichos datos para validar la viabilidad del informe periodístico.

Los datos a analizar, provienen del sitio de Compras Estatales[1], el cual pone a disposición varios conjuntos de datos sobre las compras realizadas por el estado, a través del Catálogo de Datos Abiertos del Estado [2]. En particular, de todos los datos que publica esta agencia, en esta etapa solamente interesa analizar el conjunto de datos de Adjudicaciones 2018 [2], el cual contiene todas las adjudicaciones de compras realizadas por el estado en dicho año. El conjunto de datos es un archivo de tipo *.csv* y utiliza codigueras para los valores de algunas de los campos. Para interpretar correctamente la información de estos códigos, se cuenta con la información de todas las codigeras utilizadas, en archivos de tipo *.xml* [4].

4. Requerimientos y visión del negocio

En reunión con el director del diario, este nos transmitió la siguiente visión de negocio:

"Queremos brindarle a la ciudadanía información de primera mano sobre en qué se gasta el dinero de los contribuyentes, y en un lenguaje que puedan entender. Siempre cuidando la objetividad que nos caracteriza. Además, queremos buscar irregularidades en el proceso de compra, analizando los casos en que se incumplen las condiciones impuestas por el T.O.C.A.F, en el proceso de compra".

A su vez se relevaron los siguientes requerimientos de negocio sobre el análisis a realizar.

1. **Calidad de los Datos**
   Se quiere conocer el estado de la calidad de los datos en el set de datos de Adjudicaciones [3]. En particular interesa:
   - Contar con un data profiling convencional, discriminando el análisis en variables continuas y categóricas.
   - Cantidad de adjudicaciones con un precio inválido, discriminando por caso inválido (ej: " ", NULL, $ 0).
2. **Análisis**
   Se quiere obtener la siguiente información a partir de los datos
   - TOP 20 de las organizaciones que gastaron más dinero (ordenados de forma descendente).
   - TOP 50 de las compras más caras.
   - Agregación de monto total por tipo de producto
   - TOP 10 de las unidades ejecutoras que más compraron en el periodo
   - Compras directas mayores $436.000 [5].

5. Se pide

Siguiendo las etapas de la metodología vista en el taller, realizar:

- La ingesta de los datos desde los archivos .csv y .xml a un data frame de R.
- Validación y limpieza de datos con mala calidad, documentando los casos y justificando en cada caso.
- Análisis del conjunto de datos, de acuerdo a los requerimientos de negocio, especificados en el punto 4 de este documento. Puede ayudarse con herramientas de visualización como las vistas en el taller (tablas, gráficos, etc).

Todas las etapas del proceso, deben incluirse en un único documento Jupyter Notebook, con el código R utilizado, resultados del proceso y justificaciones. Para esto se sugiere fuertemente utilizar secciones para dividir las etapas del trabajo y utilizar celdas de tipo *"Markdown"* para comentarios y justificaciones (puede utilizar los notebooks del taller como ejemplo).

6. Calendario de Entregas

Para organizar el trabajo, se planteó el siguiente calendario de entregas:

Nombre	Descripción	Fecha
Ingesta de los datos	Entregar en un Jupyter Notebook y utilizando el lenguaje R, la carga de los datos de adjudicaciones desde los archivos .csv, junto con los datos de las códigeras que crea necesarios para resolver los requerimientos de la tarea. Incluir en una sección a parte, conclusiones preliminares que pueda deducir analizando los datos, tales como: (i) cuántos datos cuenta para el análisis, (ii) cuáles columnas cree que serán de utilidad para resolver los requerimientos planteados y (iii) que problema identifica con los datos cargados si es que los hay.	Jueves 22/08/19
Análisis de los Datos	Agregar al Jupyter Notebook utilizado en la primera entrega: I. Análisis de los objetivos del Negocio y como se mapean a los datos disponibles II. Data Profiling III. Transformaciones de los datos IV. Análisis y visualización para los requerimientos de negocio planteados.	Jueves 29/08/2019
Presentación	Presentación del trabajo en equipo	Jueves 29/08/2019

7. Referencias

1. Agencia de Compras del Estado
   https://www.gub.uy/agencia-compras-contrataciones-estado/
2. Catálogo de Datos Abiertos del Estado
   http://datos.gub.uy/
3. Adjudicaciones del estado
   https://catalogodatos.gub.uy/dataset/datos-de-compras-en-linea
4. Datos y recursos de datos publicados en Compras Estatales
   https://catalogodatos.gub.uy/dataset/compras-estatales

5. Topes de Montos para Compras Directas T.O.C.A.F
http://www.ine.gub.uy/c/document_library/get_file?uuid=649a078c-b658-4c71-9f6f-96931efcaee8&groupId=10181
6. T.O.C.A.F
https://www.tcr.gub.uy/archivos/nor_122_TEXTO%20ORDENADO%20actualizado%20junio%202019.pdf