Aproximación a la Calidad

Alvaro Machicado, Diego Santos, Maximiliano Cabezón Alvarez

Calidad de Datos - Departamento de Computación - Universidad de Buenos Aires

1. Análisis e interpretación del dataset

De la información provista por los metadatos, la nomenclatura utilizada para nombrar las columnas y los datos mismos podemos realizar las siguientes observaciones:

- Si hay un campo denominado X seguido por un X-desc estos campos estan relacionados. El campo X representa un código y el campo X-desc representa la descripción asociada. Desde el punto de vista de base de datos, estos dos campos deben representar a una tabla categórica en la base de datos, con al menos este diseño para la tabla X:

Codigo	Descripcion
1	Descripcion para codigo 1
2	Descripcion para codigo 2
3	Descripcion para codigo 3

Siendo en el csv el campo X el codigo y X-des la descripción

Por ejemplo para los campos Jur y Jur-desc la tabla sería:

Jur	Jur_desc
1	Legislatura De La Ciudad De Buenos Aires
2	Auditoria General De La Ciudad De Buenos Aires
3	Defensoria Del Pueblo
5	Ministerio Publico
6	Tribunal Superior De Justicia

Se tomó solo un extracto de 5 registros de los datos del csv

Y esto es análogo para todos los pares de campos del archivo.

- El csv es el resultado de la combinación de varias tablas donde una registra la transacción con los codigos, respetando las formas normales y se agrega el campo descripción para facilitar la lectura del usuario.

- Hay jerarquías entre las columnas de las tablas entre las columnas del tipo YY con SYY, el comienzo con S indica que es una subcategoría dentro de la categoría principal.
Los columnas con este tipo de relación son: Jur y SJur , Prog y SProg , Par y SPar.

Par_desc	Spar_desc
Retribución Del Cargo	Retribución Del Cargo
Sueldo Anual Complementario	Sueldo Anual Complementario
Contribuciones Patronales	Contribuciones Patronales
Retribución Del Cargo	Retribución Del Cargo
Sueldo Anual Complementario	Sueldo Anual Complementario
Contribuciones Patronales	Contribuciones Patronales

Ejemplo de la relación entre Par y Spar

,

- Hay jerarquías y relaciones entre los atributos:

1- Car - Jur - Sjur - Ent - Og - UE:

Representan la estructura jerarquía del Gobierno de la Ciudad de derecha a izquierda. Donde a la derecha se encuentran los de mayor jerarquía.

Car_desc	Jur_desc	Sjur_desc	Ent_desc	Og_desc	UE_desc
Administracion Central	Legislatura De La Ciudad De Buenos Aires	Legislatura De La Ciudad De Buenos Aires	Legislatura De La Ciudad De Buenos Aires	Legislatura De La Ciudad De Buenos Aires	Secretaria Administrativa
Administracion Central	Jefatura De Gobierno	Jefatura De Gobierno	Jefatura De Gobierno	Jefatura De Gobierno	Jefe De Gobierno
Administracion Central	Jefatura De Gobierno	Jefatura De Gobierno	Jefatura De Gobierno	Jefatura De Gobierno	Subsecretaria De Contenidos
Administracion Central	Jefatura De Gobierno	Jefatura De Gobierno	Jefatura De Gobierno	Jefatura De Gobierno	Dir.Gral.Contenidos Y Marcas
Administracion Central	Jefatura De Gobierno	Jefatura De Gobierno	Jefatura De Gobierno	Jefatura De Gobierno	Direccion General De Opinion Publica
Administracion	Jefatura De Gobierno	Jefatura De Gobierno	Jefatura De Gobierno	Jefatura De Gobierno	Direccion General

Ejemplo del organigrama

2- Prog - Sprog - Proy - Act - Ob:

Representan la estructura jerárquica de un programa, es decir la jerarquía de la cual depende una obra.

Prog_desc	Sprog_desc	Proy_desc	Act_desc	ob_desc
Formacion Y Sancion De Leyes	Formacion Y Sancion De Leyes	Formacion Y Sancion De Leyes	Conducción	Conducción
Formacion Y Sancion De Leyes	Formacion Y Sancion De Leyes	Reparacion Y Puesta En Valor Edificio De La Legislatura De La Ciudad De Buenos Aires	Reparacion Y Puesta En Valor Edificio De La Legislatura De La Ciudad De Buenos Aires	Restauracion De Fachadas Palacio De La Legislatura
Formacion Y Sancion De Leyes	Formacion Y Sancion De Leyes	Reparacion Y Puesta En Valor Edificio De La Legislatura De La Ciudad De Buenos Aires	Reparacion Y Puesta En Valor Edificio De La Legislatura De La Ciudad De Buenos Aires	Reordenamiento Y Diseño Para Nueva Distribucion De Despachos
Conduccion Superior	Conduccion Superior	Conduccion Superior	Conduccion	Conduccion
Conduccion Superior	Conduccion Superior	Conduccion Superior	Administracion Y Servicios Generales	Administracion Y Servicios Generales
Actividades Comunes A	Actividades Comunes A	Actividades Comunes A Los	Conduccion	Conduccion

Ejemplo del detalle de Programa

3 - Inc - Ppal - Par - Spar:

Estos campos detallan el objetivo de la obra de un programa.

Inc_desc	Ppal_desc	Par_desc	Spar_desc
Gastos En Personal	Personal Permanente	Retribución Del Cargo	Retribución Del Cargo
Gastos En Personal	Personal Permanente	Sueldo Anual Complementario	Sueldo Anual Complementario
Gastos En Personal	Personal Permanente	Contribuciones Patronales	Contribuciones Patronales
Gastos En Personal	Personal Transitorio	Retribución Del Cargo	Retribución Del Cargo
Gastos En Personal	Personal Transitorio	Sueldo Anual Complementario	Sueldo Anual Complementario
Gastos En Personal	Personal Transitorio	Contribuciones Patronales	Contribuciones Patronales
Gastos En Personal	Asignaciones Familiares	Personal Permanente	Personal Permanente
Gastos En Personal	Beneficios Y Compensaciones	Beneficios Y Compensaciones Sin Discriminar	Beneficios Y Compensaciones Sin Discriminar
Bienes De Consumo	Productos Alimenticios, Agropecuarios Y	Alimentos Para Personas	Alimentos Para Personas

Ejemplo del destino de la obra

4 - Sanción - Vigente - Definitivo - Devengado

Permiten establecer el costo estimado ejecutado y rendido de la obra.

2. Calidad del Dataset

El dataset cumple con las categorías de calidad de forma aceptable.

1. **Lógica:** los atributos y entidades cumplen con las reglas lógicas.
2. **Semántica:** las entidades representan de manera fidedigna las situaciones en el mundo real.
3. **Metadatos:** todos los atributos estan nomenclados y tienen una defición. Sin embargo la definición de muchos casos es ambigua y dificil de comprender sin previa explicación.

2.1. Medidas intrínsecas de la calidad de datos

Analizamos cada una de las medidas que hacen a la calidad de los datos pertenecientes al dataset. Encontramos problemas en todas las categorías, hay casos donde los problemas abarcan más de una, como por ejemplo cuando un error de acentuación genera registros duplicados para una columna.

Precisión

Si bien la mayoria de los campos corresponden a datos tabulados encontramos diferencias en las abreviaturas y faltas de ortografía que van desde falta de acentos a errores de escritura.

Por ejemplo tomando la columna Jur-desc:

Jur_desc
Ministerio De Desarrollo Urbano Y Transporte
Min.Hàbitat Y Des. Humano
Min.Modern.Innovacion Y Tecnologia

Diferencias de criterio en distintos registros

O que generan una duplicación del registro como en la columna Ent-desc:

Ent	Ent_desc
216	Consejo Economico Y Social De La Caba
217	Consejo Economico Y Social De La C.A.B.A

Diferencias de criterio que duplican registros

Estos problemas se encuentran en todas las columnas de descripción.

Completitud

No podemos afirmar que a nivel de negocio se cumpla con los requisitos de completitud ya que solo están presentes las entidades que hicieron gastos en el período, sin saber si existen otras.

Respecto a nivel lógico encontramos que algunos atributos en las columnas que se relacionan de a pares tipo codigo/descripción, varias descripciones comparten mismo código y misma descripción posee distintos códigos por lo que no queda clara la forma que se utiliza para identificar por estos pares.

A modo de ejemplo tomando las columnas Prog y Prog-desc, el código 13 se repite para varias descripciones. Además ilustra el problema de presición y diferencia de criterios a la hora de usar acentos.

Prog	Prog_desc
13	Comision Preservacion Del Patrimonio De La Ciudad
13	Planeamiento Educativo
13	Administracion De Bienes
13	Administración De Bienes
13	Gestión De La Comuna 13

Problema: Múltiples descripciones asociadas a un código

,

Y tomando las columnas UE y UE-desc se puede ver de la misma descripción asociada a distintos códigos.

UE	UE_desc
2199	Consejo Economico Y Social De La Caba
2191	Consejo Economico Y Social De La Caba
108	Corporacion Del Sur S.E.
120	Corporación Del Sur S.E

Problema: Misma descripción asociadas a distintos código

,

Este problema se repite para datos entre los pares Sjur y Sjur-desc, Ent y Ent-desc, Prog y Prog-desc, Sprog y Sprog-desc, Proy y Proy-desc, Actividad y Act-desc, Ob y Ob-desc, Fun y Fun-desc, Ppal y Ppal-desc, Par y Par-desc, Spar y Spar-desc

De lo observado en el set de datos podemos ver que a nivel físico se cumplen los requisitos de completitud.

Consistencia

Respecto a las columnas Definitivo y Devengado en los metadatos son declaradas como valores enteros, pero los datos son decimales. Además en estas columnas encontramos discrepancias en los valores de esta columna que deberían ser iguales, pero en 5494 no lo son.

	Definitivo	Devengado	Val_Dif
15	546851.52	546851.22	0.30
103	7247836.00	3540148.00	3707688.00
131	82950.00	0.00	82950.00
140	35984260.00	17319512.00	18664748.00
156	32116700.00	19978985.00	12137715.00
...	...	...	...
48719	59244.33	41585.26	17659.07
48725	449397.55	435941.86	13455.69
48770	3427705.00	1738831.00	1688874.00
48774	20343.32	343.32	20000.00
48785	9133770.47	5839914.59	3293855.88

5494 rows × 3 columns

Discrepancias entre Definitivo y Devengado

,

Unicidad

En lo referido a registros completos no hay duplicados. Pero si encontramos que a nivel columna y pares de columna hay datos duplicados como se presentó en las categorias anteriores.

Actualidad

Los registros no tienen una fecha de creación o actualización por lo que solo podemos guiarnos por la información provista por el link de descarga.

3. Análisis Univariado

El análisis de cada una de las variables del dataset se encuentra dentro de la carpeta anexo bajo el nombre de analisis-univariado.html. También se adjunta un notebook de jupyer para poder realizar la ejecución del mismo.

4. Análisis Bivariado

El análisis bivariado se encuentra dentro de la carpeta anexo bajo el nombre analisis-bivariado.html También se adjunta un notebook de jupyer para poder realizar la ejecución del mismo.

5. Nos llamó la atención

1. Los metadatos son dificiles de entender.
2. Consideramos que no haría falta incluir las columnas de código.
3. La cantidad de veces que el mismo código estaba asociada a distintas descripciones.
4. El valor de los códigos en algunas columnas alcanzaba los 6 dígitos, a pesar de que no había códigos mas chicos. Mientras que en otras columnas se utilizaba una númeracion baja.
5. La falta de un criterio de abrevación para las descripciones.

6. La falta de acentos.
7. Las faltas de ortografía.
8. La gran cantidad de dependencias y reparticiones estatales.
9. La mayoría de los gastos son sobre salarios.