DEPARTAMENTO
DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

Aproximación a la Calidad de Datos en el Presupuesto de la Ciudad de Buenos Aires

Caso de estudio: 3th trimestre 2018

Mayo 15 del 2020 — Calidad de Datos

Integrante	LU	Correo electrónico
Venegas Ramirez, David Alejandro	783/18	davidalevng@gmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2610 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (++54 +11) 4576-3300
https://exactas.uba.ar

1. Introducción

Los datos cada vez más son una parte crucial de la dirección estratégica de un negocio o proyecto, siendo actualmente considerados en muchos casos el activo más valioso de algunas corporaciones. Sin embargo, este valor está altamente correlacionado con el uso que se le puede dar a los mismos y por ende su veracidad y estructuración a la hora de manejarlos. El término calidad de datos se refiere al estado cualitativo en el que se encuentra la información almacenada, y su fiabilidad para tomar decisiones sobre ésta o representar el mundo real, pues la interpretación incorrecta de los datos puede repercutir en errores costosos, así como apunta el MIT Sloan que el uso de datos deficientes puede llegar a costar un 15-25 % de los ingresos totales de un negocio.

En el presente trabajo, se busca aproximar el estado de la calidad de los datos de la ejecución presupuestaria de la ciudad de Buenos Aires para el 3er trimestre de 2018. En las próximas secciones se analizará brevemente el dataset y sus variables, y se determinará la posible existencia de una correspondencia o jerarquía entre estas. Seguidamente, se realizará una apreciación sobre la calidad intrínseca para dicho conjunto de datos, para posteriormente efectuar un análisis descriptivo de cada variable y finalmente un análisis bivariado de elementos relevantes.

2. Análisis de Correspondencia

En la documentación provista por el sitio web[1] de la Data de la ciudad de Buenos Aires, en particular para el presupuesto ejecutado del tercer trimestre de 2018, se puede observar toda la data relativa a los gastos ejecutados de los órganos del gobierno central, las inversiones patrimoniales y los recursos empleados, entre otros. En la organización de dicha información se halla una directa correlación entre pares de variables que funcionan como 2 categorías distintas pero fuertemente dependientes. En estos grupos se pueden encontrar variables independientes en una categoría y en la otra su código numérico respectivo, el cual es constante y congruente a lo largo del dataset. Estas columnas que se pueden emparejar mantienen el formato de nombrevariable y nombrevariable_desc que detallan el código numérico respectivo. Por ejemplo: Dado el dato **jur** (jurisdicción) definido como *"Código de jurisdicción 1"* y su contraparte jur_desc establecido como *"Son las organizaciones públicas sin personalidad jurídica que representan a cada uno de los poderes establecidos por la Constitución de Ciudad Autónoma de Buenos Aires"* se puede apreciar en el dataset que cada organismo en **jur_desc** tiene un código asignado en **jur**.

En las definiciones de estos pares variables se puede encontrar que algunas contienen un número de referencia en la documentación según como estas se relacionan. Por ejemplo, en el caso de la variable *obra* se tiene la descripción como "Código de obra 3" donde el 3 define su pertenencia a la Categoría Programática así como en el caso de data referente "Código de partida principal 1", "Código de partida subparcial 1" y "Código de fuente de financiamiento 1" se entiende que conforman el Objeto del Gasto.

Existen variables pertinentes al presupuesto que ayudan a aclarar el ciclo de vida del mismo y generar suposiciones sobre los resultados de dichas asignaciones, e incluso si se quisiera, se prestan para hacer estudios de *business inteligence* sobre las obras efectuadas y el presupuesto en cuestión. Por ejemplo: En el caso **sanción** se tiene el monto aprobado originalmente (lo que estaba previsto que se iba a gastar) y **vigente** que intenta salvaguardar el monto en **sanción** de la desactualización de los datos y toma en cuenta las modificaciones que se hicieron en el presupuesto a partir del original. Por último, **devengado** se refiere al definitivo o mejor dicho al monto que se terminó pagando, es decir, los gastos reales de la actividad.

Para concluir, se evidencia que también existe una jerarquía en la asignación de códigos a las variables categóricas. Por ejemplo: En el caso de la variable *var* existe una prioridad para su correspondiente *var_desc* donde la "Administración Central" tiene el código 1 y los "Organismos Descentralizados" el 2, lo cual le da la cualidad a *var_desc* de ser categórica ordinal. En el resto de los pares de variables algunos casos presentan una jerarquía respecto a las instituciones mencionadas y el código asignado.

[1] Data de la ciudad de Buenos Aires en https://data.buenosaires.gob.ar/dataset/presupuesto-ejecutado/archivo/

3. Apreciación de las Cualidades Intrínsecas de Calidad de la Data

3.1. Precisión

El grupo de datos encargado de representar el mundo real en este caso son los de tipo numérico relativos al presupuesto como *sanción*, *vigente*, *definitivo* y *devengado*. Se puede apreciar un intento por representar mejor el gasto público al actualizar el presupuesto mas allá del sancionado inicialmente, con la presencia de estos otros 3 últimos campos. Y con respecto a la exactitud de los cálculos tanto *definitivo* como *devengado* tienen dos decimales de exactitud.

3.2. Completitud

Analizando la presencia o ausencia de características en la tabla, se evidenció que existe un alto nivel en la data del presupuesto, ya que no se encontraron datos faltantes o *NULL*. En algunos campos numéricos estaba presente el valor cero, el cual no se interpretó como faltante sino como un dato definido, por ejemplo en el caso de *sanción*, este implica que casi 1/3 de las actividades fueron aprobadas por el Poder Legislativo, y más tarde por el Ejecutivo con un presupuesto tentativo de 0 pesos. Para el caso de *vigente*, quiere decir que la mayor parte de las actividades fueron actualizadas, mientras que 1/3 también fueron descontinuadas y/o abortadas, con presupuesto *devengado* de 0 pesos.

3.3. Consistencia

Evaluar la coherencia de los datos representados en múltiples copias en este estudio de caso no es posible ya que se está analizando un trimestre aislado, cuyas variables no se referencian dentro del dataset (no hay múltiples copias de un valor). Se puede especificar que el dataset no es redundante en este caso y con respecto a las reglas del negocio es aritméticamente consistente, ya que el presupuesto *devengado* para cada fila es siempre igual o más chico que el aprobado inicialmente. Sin embargo, para hacer un análisis más profundo de la consistencia, sería necesario comparar con respecto a otros trimestres o copias del dataset.

3.4. Unicidad

En el presente presupuesto no se encontraron duplicados, y se considera única cada sanción.

3.5. Actualidad

No existe un campo de fecha para el momento de la asignación de cada presupuesto, ni para su actualización en *vigente* o *definitivo*. Se presume que los datos están actualizados para el momento de la publicación del 28 de mayo de 2019. Y dado que el último cambio en el sitio web oficial es del 28 de mayo de 2019, se considera que no existen enmiendas oficiales en el dataset hasta el momento del presente informe.

4. Análisis Descriptivo

Como se mencionó anteriormente, el dataset tiene un alto grado de completitud y el porcentaje de datos faltantes para cada variable es 0 %. A continuación se listan las principales variables estudiadas con sus respectivas métricas como media, moda, y desviación en en caso de las numéricas y frecuencia en el caso de las categóricas.

La variable *car* se describe como “Código de carácter” y tiene una media de 1,168405 con una desviación estándar de 0,374229, y una mediana de 1,0. Por otra parte la variable *car_desc* (la correspondiente categórica de *car*) tiene 2 únicos valores y su moda es “Administracion Central” con 43287 apariciones.

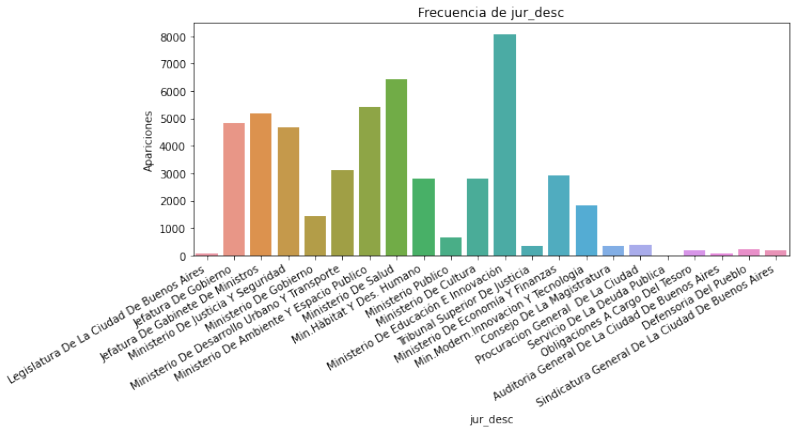

Figura 1: Frecuencia de la variable *jur* en el dataset

La variable *jur* se describe como "Código de jurisdicción 1" y tiene una media de 37,394348, una desviación estándar de 15,752022, y una mediana de 1,0. Su variable categórica correspondiente *jur _desc* tiene 22 valores únicos, cuya moda es "Ministerio De Educación E Innovación" con 8078 apariciones.

La variable *ent* se describe como "Código de entidad 2" y tiene una media de 58,185407, una desviación estándar de 181,812197, y una mediana de 0,000000. Su variable categórica correspondiente *ent _desc* tiene 59 valores únicos, cuya moda es "Ministerio De Educación E Innovación" con 8078 apariciones.

La variable *ogese* se describe como "Código de Oficinas de gestión sectorial" y tiene una media de 147,579525, una desviación estándar de 227,083588, y una mediana de 1,000000. Su variable categórica correspondiente *ogese _desc* tiene 59 valores únicos, cuya moda es "Ministerio De Educación E Innovación" con 8078 apariciones.

La variable *ue* se describe como "Código de unidad ejecutora 3" y tiene una media de 3524,896087, una desviación estándar de 3852,789681, y una mediana de 10,000000. Su variable categórica correspondiente *ue _desc* tiene 345 valores únicos, cuya moda es "Dir. Gral De Educación De Gestión Estatal" con 3943 apariciones.

La variable *prog* se describe como "Código de programa 3" y tiene una media de 41,795535, una desviación estándar de 28,04340, y una mediana de 1,000000. Su variable categórica correspondiente *prog _desc* tiene 513 valores únicos, cuya moda es "Atención Médica General En Hospitales De Agudos" con 2042 apariciones.

La variable *sprog* se describe como "Código de subprograma 3" y tiene una media de 2,463124, una desviación estándar de 8,374948, y una mediana de 0,000000. Su variable categórica correspondiente *sprog _desc* tiene 556 valores únicos, cuya moda es "Educación Primaria" con 1285 apariciones.

La variable *proy* se describe como "Código de proyecto 3" y tiene una media de 1,346954, una desviación estándar de 8,871356, y una mediana de 0,000000. Su variable categórica correspondiente *proy _desc* tiene 808 valores únicos, cuya moda es "Educación Primaria" con 1285 apariciones.

La variable *act* se describe como "Código de actividad 3" y tiene una media de 7965,134344, una desviación estándar de 12621,393726, y una mediana de 0,000000. Su variable categórica correspondiente *act_desc* tiene 1677 valores únicos, cuya moda es "Administración Y Servicios Generales" con 5552 apariciones.

La variable *obra* se describe como "Código de obra 3" y tiene una media de 7965,134344, una desviación estándar de 12621,393726, y una mediana de 0,000000. Su variable categórica correspondiente *obra_desc* tiene 2191 valores únicos, cuya moda es "Administración Y Servicios Generales" con 5552 apariciones.

La variable *fin* se describe como "Código de finalidad 1" y su variable categórica correspon-

diente *fin_desc* tiene 5 valores únicos, cuya moda es "Servicios Sociales" con 25106 apariciones.

La variable *fun* se describe como "Código de función 1" y su variable categórica correspondiente *fun_desc* tiene 20 valores únicos, cuya moda es "Dirección Ejecutiva" con 8045 apariciones.

La variable *inciso* se describe como "Código de inciso 1" y su variable categórica correspondiente *inciso_desc* tiene 8 valores únicos, cuya moda es "Gastos En Personal" con 17258 apariciones.

La variable *ppal* se describe como "Código de partida principal 1" y tiene una media de 3,978330, una desviación estándar de 2,673006, y una mediana de 1,000000. Su variable categórica correspondiente *ppal_desc* tiene 49 valores únicos, cuya moda es "Personal Permanente" con 7026 apariciones.

La variable *parc* se describe como "Código de partida parcial 1" y tiene una media de 4,105719, una desviación estándar de 2,746505, y una mediana de 1,000000. Su variable categórica correspondiente *parc_desc* tiene 173 valores únicos, cuya moda es "Otros No Especificados Precedentemente" con 5287 apariciones.

La variable *sparc* se describe como "Código de partida subparcial 1" y tiene una media de 1,489367, una desviación estándar de 10,855034, y una mediana de 0,000000. Su variable categórica correspondiente *sparc_desc* tiene 367 valores únicos, cuya moda es "Otros No Especificados Precedentemente" con 5287 apariciones.

La variable *eco* se describe como "Código de clasificador económico 1" y tiene una media de $2{,}138180e{+}07$, una desviación estándar de $2{,}893522e{+}05$, y una mediana de $2{,}120000e{+}07$. Su variable categórica correspondiente *eco_desc* tiene 23 valores únicos, cuya moda es "Remuneraciones Al Personal" con 17258 apariciones.

La variable *ff* se describe como "Código de fuente de financiamiento 1" y tiene una media de 11,263866, una desviación estándar de 0,835930, y una mediana de 11,000000. Su variable categórica correspondiente *ff_desc* tiene 7 valores únicos, cuya moda es "Tesoro De La Ciudad" con 45841 apariciones.

La variable *geo* se describe como "Código de ubicación geográfica 1" y tiene una media de 4,61618, una desviación estándar de 5,36082, y una mediana de 1,00000. Su variable categórica correspondiente *geo_desc* tiene 16 valores únicos, cuya moda es "Comuna 1" con 19622 apariciones.

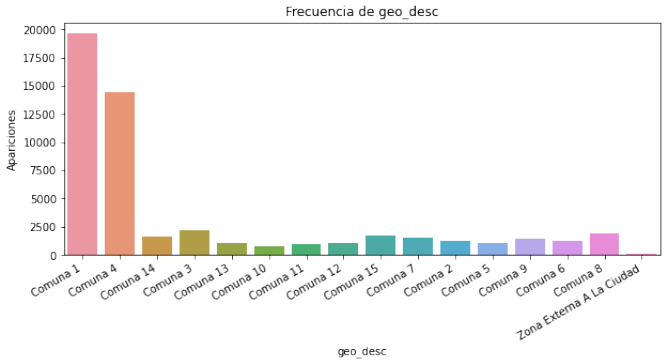

Figura 2: Frecuencia de la variable *geo* en el dataset

La variable *sanción* se describe como "Código de sanción" y tiene una media de 4.280371e+06 con una desviación estándar de 4.165309e+07, y una mediana de 0.000000e+00.

La variable *vigente* se describe como "Código de vigente" y tiene una media de 4.831267e+06 con una desviación estándar de 8.166233e+07, y una mediana de 0.000000e+00.

La variable *definitivo* se describe como "Código de definitivo" y tiene una media de 3.399354e+06 con una desviación estándar de 6.868158e+07, y una mediana de 0.000000e+00.

La variable *devengado* se describe como "Código de devengado" y tiene una media de 3.133693e+06 con una desviación estándar de 6.829328e+07, y una mediana de 0.000000e+00.

5. Análisis Bivariado

Se compararon las variables numéricas utilizando un box plot bivariado que permite observar con facilidad la similitud existente entre las distribuciones, así como percentiles, outliers y la media. Sería interesante comparar gráficamente variables como *sancionado* y $definitivo$ pero por razones de cómputo para el presente caso didáctico en particular se usó *car* y jar.

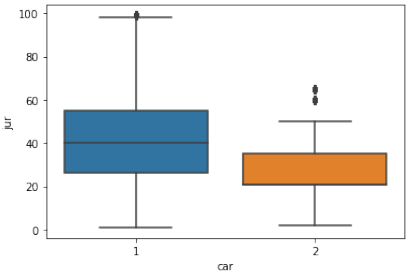

Figura 3: Box Plot comparativo entre *car* vs jur

6. Conclusiones

El dataset del presupuesto de la Ciudad de Buenos Aires se encuentra correctamente estructurado y categorizado en sus 52053 filas y 44 columnas con información consistente y completa; igualmente, es buen ejemplo del gobierno de datos y de cómo mantener variables categóricas definidas de forma estándar. Además, en lo que respecta a la semántica, hace un buen uso de códigos y descripciones para facilitar el manejo del dataset, sin embargo, se podría mejorar la documentación. Asimismo, es interesante resaltar que aunque *definitivo* y *devengado* aparecen en la documentación como números enteros, en el dataset aparecen con 2 decimales de precisión, lo cual es una incongruencia debido a que el tipo de la columna en ese caso debería ser floating-point. Ahora bien, para finalizar se recomienda añadir la fecha de los cambios en vigente a fin de mantener un mejor control sobre el histórico de las transacciones.