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: 3er 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 esta 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 información relativa a los gastos ejecutados por 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 y 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**.

Los pares que se corresponden son:

- [Car, Car_desc],
- [Jur, Jur_desc],
- [Sjur, Sjur_desc],
- [Ent, Ent_desc],
- [Ent, Ent_desc],
- [Og, Og_desc],
- [UE, UE_desc],
- [Prog, Prog_desc],
- [Sprog, Sprog_desc],
- [Proy, Proy_desc],
- [Actividad, Act_desc],
- [Ob, Ob_desc],
- [Fin, Fin_desc],
- [Fun, Fun_desc],
- [Inc, Inc_desc],
- [Ppal, Ppal_desc],
- [Par, Par_desc],
- [Spar, Spar_desc],
- [Eco, Eco_desc],
- [Fte, Fte_desc],
- [Geo, Geo_desc]

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

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.

Además, 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, tomando en cuenta las modificaciones que se hicieron en el presupuesto a partir del original. Por último, **devengado** se refiere al final 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 como en `Caracter, Jurisdiccion, Subjiccion, Entidad`. En otros ejemplos se tienen también las variables: `Programa` y `Subprograma` que agrupan el propósito de la acción gubernamental a la vez que las variables `Proyecto` y `Obra`. En el caso de la jerarquía `Inciso, Partida principal, Partida parcial, Subpartida` estas determinan el objeto del gasto.

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 más 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. Sin embargo, sin información complementaria sobre el modelo de negocios no se puede afirmar la presencia o ausencia de caracterısticas, atributos o relaciones en el dataset.

3.3. Consistencia

Evaluar la coherencia de los datos representados en múltiples copias, en este caso en estudio, 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*. Aunque 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 %.

Ahora bien, entre la clasificación de las variables se encuentran las numéricas que son: **Sanción, Vigente, Definitivo, Devengado**, y las variables categóricas de a pares que son: **Carácter, Jurisdicción, Subjicción, Entidad, Oficina de gestión, Unidad ejecutora, Programa, Subprograma, Proyecto, Actividad, Obra, Finalidad, Función, Inciso, Partida principal, Partida parcial, Subpartida, Clasificador económico, Fuente de financiamiento, Ubicación geográfica.**

A continuación, se listan las principales variables estudiadas con sus respectivas métricas como media, medianaa y desviación en caso de las numéricas y frecuencia en el caso de las categóricas.

4.1. Carácter

De acuerdo a la documentación oficial, *car_desc* se define como "Cada uno de los agrupamientos en que se divide el Subsector Administración Gubernamental del Sector Público Gubernamental no Financiero". Por otra parte la variable *car* se describe como "Código de carácter'

Respecto al análisis de las variables existe una correspondencia directa entre el código de carácter y *car_desc* la cual tiene 2 únicos valores y su moda es "Administración Central" con 43, 287 apariciones evidenciando el domino de esta en la asignación de presupuestos.

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

4.2. Jurisdicción

Se refiere a 'las organizaciones públicas sin personalidad jurídica que representan a cada uno de los poderes establecidos por la Constitución de Ciudad Autónoma de Buenos Aires."

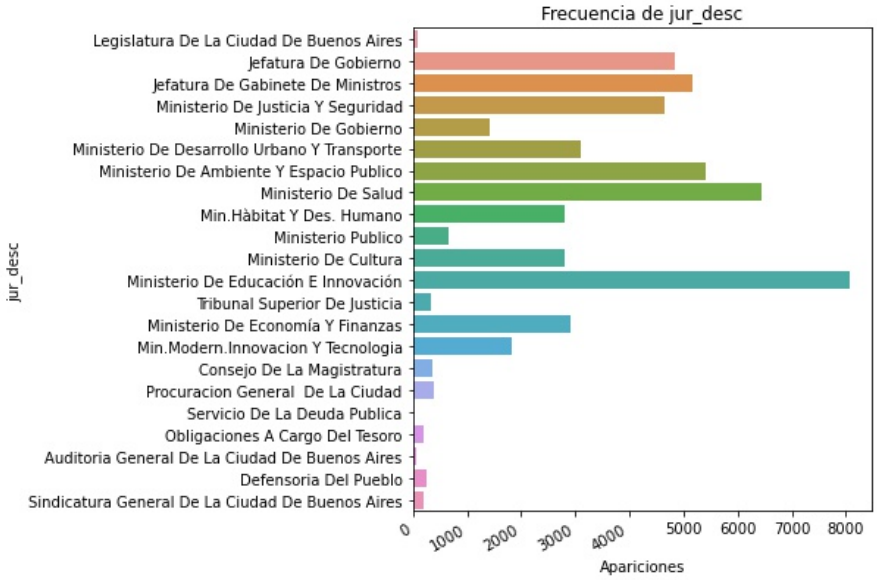


Figura 2: Frecuencia de la variable *jur_desc* en el dataset

La variable *jur* se describe como "Código de jurisdicción 1" y su par *jur_desc* tiene 22 valores únicos, cuya moda es "Ministerio De Educación E Innovación" con $8,078$ apariciones y un $15{,}52\,\%$ de las asignaciones. Luego, sigue el "Ministerio De Salud" con un $12{,}38\,\%$ y en tercer lugar el "Ministerio De Ambiente Y Espacio Público" con un $10{,}42\,\%$. Las organizaciones con menos asignaciones fueron la "Legislatura de CABA", la "Auditoría General de CABA" y el "Servicio De La Deuda Publica".

4.3. Subjiccion

Se refiere a las unidades institucionales que pertenecen a una Jurisdicción que surgen de criterios de ordenamientos en la estructura organizativa del GCBA.

La variable *sjur* se describe como "Código de subjicción" y su variable categórica correspondiente *sjur_desc* tiene 29 valores únicos, cuya moda también es "Ministerio De Educación E Innovación" con $8,078$ apariciones al igual que en *jur_desc*. Esta variable, además, tiene una relación de jerarquía con la anterior y mantuvo valores similares exceptuando la aparición de nuevas categorías como secretarías y "Vías Peatonales"

4.4. Entidad

Con esto se señala: "Toda organización pública con personería jurídica y patrimonio propio, se trate de empresas o sociedades y organismos descentralizados". La variable *ent* se describe como "Código de entidad 2" y su variable correspondiente *ent _desc* tiene 59 valores únicos. Debido a que estos son tantos, se listan a continuación los 10 principales:

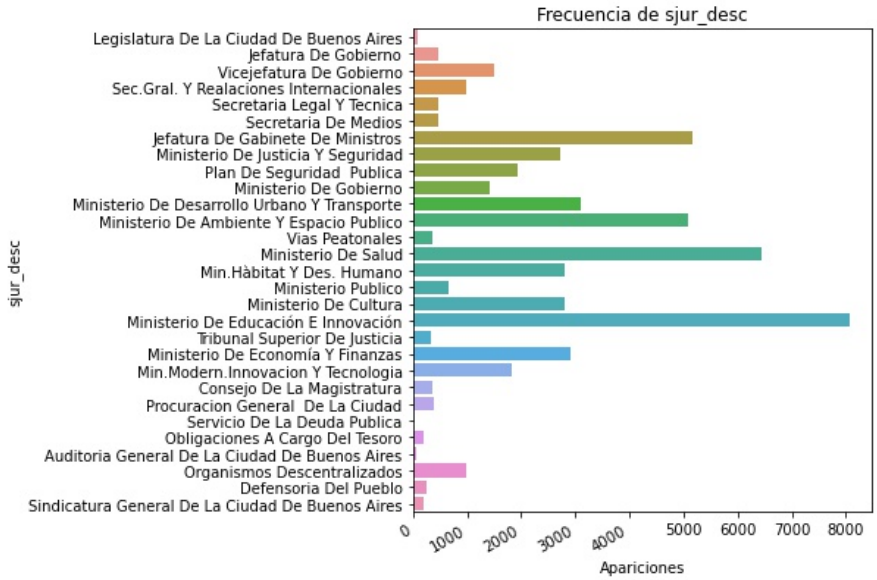


Figura 3: Frecuencia de la variable *sjur_desc* en el dataset

Cuadro 1: Principales valores de la variable end_desc

Entidad	Frequencia	Porcentaje
Ministerio De Educación E Innovación	8,078	15.51 %
Ministerio De Salud	6,447	12.38 %
Ministerio De Ambiente Y Espacio Público	3,478	6.68 %
Ministerio De Desarrollo Urbano Y Transporte	3,114	5.98 %
Min.Hábitat Y Des. Humano	2,797	5.37 %
Ministerio De Cultura	2,515	4.83 %
Jefatura De Gabinete De Ministros	2,244	4.31 %
Ministerio De Justicia Y Seguridad	1,907	3.66 %
Ministerio De Economía Y Finanzas	1,858	3.56 %
Plan De Seguridad Pública	1,843	3.54 %
Otros	17,772	34.18 %

4.5. Oficinas de Gestión

Según la documentación oficial de la página del gobierno de la ciudad[2], la variable *ogese* se describe como "Código de Oficinas de gestión sectorial".

Por otra parte, la variable categórica es *ogese_desc* contiene el nombre de cada oficina de gestión sectorial. Es importante resaltar que en el dataset la variables en realidad aparecen como *og* y *eg_des* respectivamente, lo cual puede ser un error ya que no corresponde a la documentación asociada. Ahora bien, esta última variable tiene 59 valores únicos. Sus principales apariciones se encuentran en la tabla 2. Esta tabla resulta idéntica a la 5 pero esto ocurre porque Oficina de Gestión y Entidad son exactamente iguales para 51, 803 valores, es decir, tan solo difieren en 250 filas, pero curiosamente cuando se revisaron a mano estos valores, son los mismos y la diferencia se debe posiblemente al espaciado. Por ende se concluye que ambas variables son casi idénticas y tienen los mismos valores desde el punto de vista semántico. En la sección del Anexo se puede encontrar un link al spreadsheet con esta comparación de 250 valores.

[2]https://data.buenosaires.gob.ar/dataset/presupuesto-ejecutado/archivo/juqdkmgo-16623-resource

Cuadro 2: Principales valores de la variable eg_desc

Entidad	Frequencia	Porcentaje
Ministerio De Educación E Innovación	8,078	15.51 %
Ministerio De Salud	6,447	12.38 %
Ministerio De Ambiente Y Espacio Público	3,478	6.68 %
Ministerio De Desarrollo Urbano Y Transporte	3,114	5.98 %
Min.Hábitat Y Des. Humano	2,797	5.37 %
Ministerio De Cultura	2,515	4.83 %
Jefatura De Gabinete De Ministros	2,244	4.31 %
Ministerio De Justicia Y Seguridad	1,907	3.66 %
Ministerio De Economía Y Finanzas	1,858	3.56 %
Plan De Seguridad Pública	1,843	3.54 %
Otros	17,772	34.18 %

4.6. Unidad Ejecutora

Son los "centros administrativos responsables de la planificación, programación, asignación formal y utilización de recursos en función de una producción o provisión de bienes y servicios determinada"

La variable *ue* se describe como "Código de unidad ejecutora 3" y tiene como variable categórica correspondiente a *ue_desc* con 345 valores únicos, los cuales la hacen mucho más numerosa que las anteriores y muestra como se van desglosando o especificando las asignaciones de presupuesto. Sus principales valores son listados en la tabla 3.

Cuadro 3: Principales valores de la variable ue_desc

Entidad	Frequencia	Porcentaje
Dir. Gral De Educación De Gestión Estatal	3,943	7.57 %
Agencia Ambiental	777	1.49 %
Ss. Carrera Docente Y Formación Técnico Profesional	670	1.28 %
Subsecretaria De Coordinación Pedagógica Y Equidad Educativa	621	1.19 %
Administ.Gubernamental De Ingresos Públicos	616	1.18 %
Consejo De Los Derechos De Las Niñas, Niños Y Adolescentes	591	1.13 %
Dir.Gral. Fiscalización Del Espacio Público	551	1.05 %
Dir.Gral.Patrimonio Museos Y Casco Histórico	542	1.04 %
Dirección General De Educación Superior	528	1.01 %
Dirección General De Espacios Verdes	518	0.99 %
Otros	42,696	82.02 %

En estos valores se puede apreciar que a pesar de tener una mayor especificidad prevalece la Educación como la principal categoría (con respecto a las variables de mayor jerarquía vistas anteriormente) con las apariciones de la "Dir. Gral De Educación De Gestión Estatal", la "Subsecretaria. Carrera Docente Y Formación Técnico Profesional" y la "Subsecretaría De Coordinación Pedagógica Y Equidad Educativa 621".

4.7. Programa

De acuerdo a la documentación oficial: "Dentro de la clasificación de Categorías Programáticas, el Programa representa una asignación formal de recursos, físicos y financieros, que tiene a su cargo la producción".

La variable *prog* se describe como "Código de programa 3" y tiene de variable categórica correspondiente a *prog_desc* con 513 valores únicos. Se encuentran listadas las que tienen más presupuestos asignados en la tabla 4, donde se puede ver que aun las mayores asignaciones de presupuestos corresponden a los rubros de Salud y Educación, siendo el principal Programa el correspondiente a la Atención Médica en Hospitales de Agudos, y aparte se puede observar como se subdividen los programas de Salud y los presupuestos correspondientes al tener Primaria mayor asignación que Educación Inicial. Por último también se aprecian los presupuestos de la Policía de la Ciudad y como curiosamente tienen más asignaciones la Administración De Infracciones En La Ciudad, sería prudente consultar a un experto en el modelo del negocios si hay alguna relación importante allí o si la Policía debería requerir mayor atención monetaria.

Cuadro 4: Principales valores de la variable prog_desc

Entidad	Frecuencia	Porcentaje
Atención Médica General En Hospitales De Agudos	2,042	3.92 %
Educación Primaria	1,285	2.46 %
Educación Inicial	887	1.70 %
Atención De Salud Mental	747	1.43 %
Educación Del Adulto Y Del Adolescente	595	1.14 %
Atención Médica De Patologías Específicas	521	1.00 %
Cuidado Y Puesta En Valor De Espacios Verdes	518	0.99 %
Administración De Infracciones En La Ciudad	480	0.92 %
Policía De La Ciudad	455	0.87 %
Atención Médica Materno Infantil	449	0.86 %
Otros	44,074	84.67 %

4.8. Subprograma

Estos se definen como: "Un centro formal de asignación de recursos que tiene por finalidad precisar con un mayor grado de desagregación la producción del programa que le da origen."

La variable *sprog* se describe como "Código de subprograma 3" y su correspondiente *sprog _desc* tiene 556 valores únicos. Más adelante en la tabla 4 se listan sus variables más frecuentes, allí se puede observar que algunos items se mantienen con respecto a la tabla anterior como "Educación Primaria" y "Educación Inicial" ya que estas no se subdividen, por otra parte vemos que otras partes del gasto como la Atención Médica ya no aparece entre las primeras, lo cual se debe a que se subdivide en muchas categorías.

Cuadro 5: Principales valores de la variable sprog_desc

Entidad	Frecuencia	Porcentaje
Educación Primaria	1285	2.46 %
Educación Inicial	887	1.70 %
Educación Del Adulto Y Del Adolescente	595	1.14 %
Cuidado Y Puesta En Valor De Espacios Verdes	518	0.99 %
Administración De Infracciones En La Ciudad	480	0.92 %
Policía De La Ciudad	455	0.87 %
Educación Media	448	0.86 %
Sistema Estadístico De La Ciudad.	437	0.83 %
Actividades Centrales	433	0.83 %
Otros	45,674	87.74 %

4.9. Proyecto

Tenemos que "Dentro de la clasificación por Categoría Programática, el proyecto es el proceso de producción de un bien de capital destinado a crear, ampliar o modernizar la capacidad de oferta"

La variable *proy* se describe como "Código de proyecto 3" y su par *proy_desc* sirve para clasificar la asignación del presupuesto. Esta tiene 808 valores únicos, y sus principales valores se listan a continuación en la tabla 6 donde vemos que la Educación se mantiene de primera, al igual que en la variable anterior.

Cuadro 6: Principales valores de la variable proy_desc

Entidad	Frecuencia	Porcentaje
Educación Primaria	1,285	2.46 %
Educación Inicial	887	1.70 %
Educación Del Adulto Y Del Adolescente	595	1.14 %
Administración De Infracciones En La Ciudad	480	0.92 %
Educación Media	448	0.86 %
Sistema Estadístico De La Ciudad.	435	0.83 %
Actividades Centrales	423	0.81 %
Sanidad Y Tenencia Responsable De Mascotas	419	0.80 %
Policía De La Ciudad	389	0.74 %
Otros	45,948	88.27 %

4.10. Actividad

Según la documentación como "Dentro de la clasificación por Categoría Programática, la Actividad refleja los procesos contenidos en las acciones presupuestarias de mínimo nivel cuya producción es intermedia"

La variable *act* se describe como "Código de actividad 3" y su variable correspondiente *act_desc* tiene $1,677$ valores únicos. Los más frecuentes se pueden observar en la siguiente tabla 8.

Cuadro 7: Principales valores de la variable act_desc

Entidad	Frecuencia	Porcentaje
Administración Y Servicios Generales	5,552	10.66
Conducción	5,338	10.25
Conducción	4,816	09.25
Administración Y Servicios Generales	3,297	06.33
Conducción Y Administración	1,518	02.91
Servicios De Diagnóstico Y Tratamiento	1,030	01.97
Servicios Generales De Mantenimiento	557	01.07
Atención Ambulatoria - Consultorio Externo	472	00.90
Servicios De Internación	401	00.77
Mantenimiento Comunal	400	00.76
Otros	28,675	55.08

En la tabla se puede apreciar que la moda es "Administración y Servicios Generales" con $5,552$ apariciones, seguido de "Conducción" en segundo y tercer lugar, pero al examinar el dataset podemos apreciar que estas variables semánticamente son las mismas y deberían unificarse en una (al limpiar el dataset). Este cambio mencionado alteraría la moda de la variable de forma que "Conducción" seria la más común.

4.11. Obra

De acuerdo a la documentación "Dentro de la clasificación por Categoría Programática, la obra representa una de las desagregaciones de los proyectos".

La variable *obra* se describe como "Código de obra 3" y su variable categórica correspondiente *obra_desc* tiene 2,191 valores únicos, es decir, se continúan subdividiendo las asignaciones de presupuestos. Sus principales valores se pueden apreciar en la tabla **??** la cual es idéntica a la tabla 8, esto se debe a que ambas variables solo difieren en 2,150 valores, los cuales se pueden ver en la sección del Anexo.

Cuadro 8: Principales valores de la variable act_desc

Entidad	Frecuencia	Porcentaje
Administración Y Servicios Generales	5,552	10.66
Conducción	5,338	10.25
Conducción	4,816	09.25
Administración Y Servicios Generales	3,297	06.33
Conducción Y Administración	1,518	02.91
Servicios De Diagnóstico Y Tratamiento	1,030	01.97
Servicios Generales De Mantenimiento	557	01.07
Atención Ambulatoria - Consultorio Externo	472	00.90
Servicios De Internación	401	00.77
Mantenimiento Comunal	400	00.76
Otros	28,675	55.08

4.12. Finalidad

De acuerdo a la metadata es "un criterio de clasificación del gasto según la naturaleza de los servicios que prestan las instituciones públicas a la comunidad".

La variable *fin* se describe como "Código de finalidad 1" y su variable categórica correspondiente *fin_desc* tiene 5 valores únicos, cuya moda es "Servicios Sociales" con 25,106 apariciones y casi el 50 % de las apariciones, y luego le sigue la "Administración Gubernamental" con casi un cuarto de las apariciones. Finalmente, también se puede destacar un gasto mínimo menor al 1 % para la Deuda Pública.

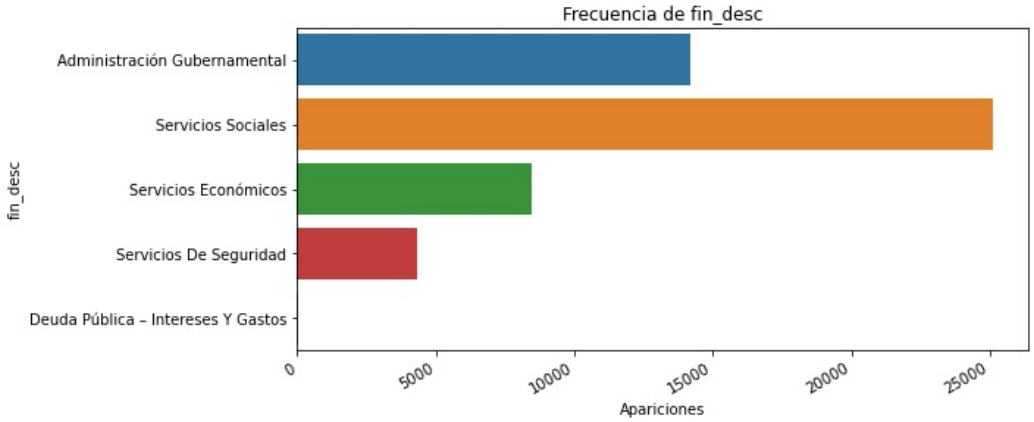


Figura 4: Frecuencia de la variable fin_desc en el dataset

4.13. Función

Se cuenta con la descripción "Es el nivel de cuentas dentro de la clasificación del gasto por Finalidad y Función que permite examinar en el tiempo las tendencias del gasto para las funciones generales del Gobierno."

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 8,045 apariciones, además, en la figura 5 se pueden apreciar mayores asignaciones para el Ejecutivo y la Educación. Igual que en la variable anterior la Deuda Pública se encuentra de última junto con los seguros y finanzas.

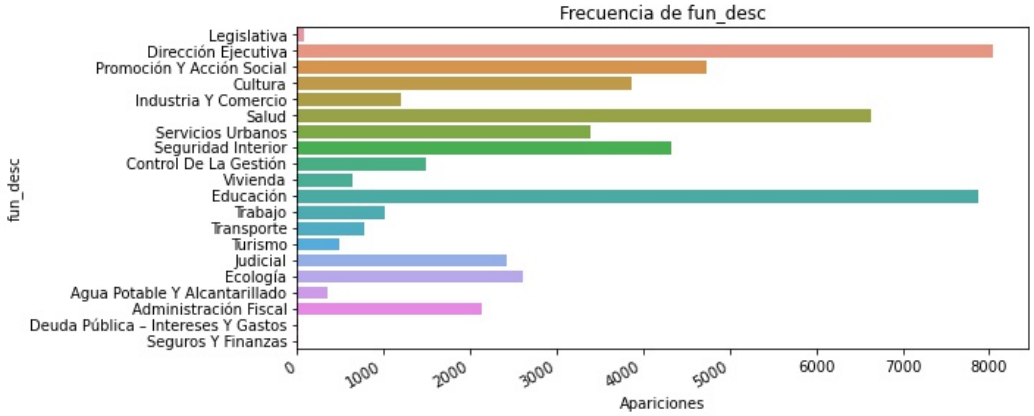


Figura 5: Frecuencia de la variable fun_desc en el dataset

4.14. Inciso

Se define como "Nivel superior de los cuatro niveles de cuentas que conforman el Objeto de Gasto" por lo cual se puede destacar que inicia otra jerarquía.

La variable *inciso* se describe como "Código de inciso 1" y su variable correspondiente *inciso_desc* tiene 8 valores únicos, cuya moda es "Gastos En Personal" con 17,258 apariciones. En la figura 6 se puede apreciar que las principales asignaciones son para gastos en personal, bienes de consumo y servicios no personales en ese orden.

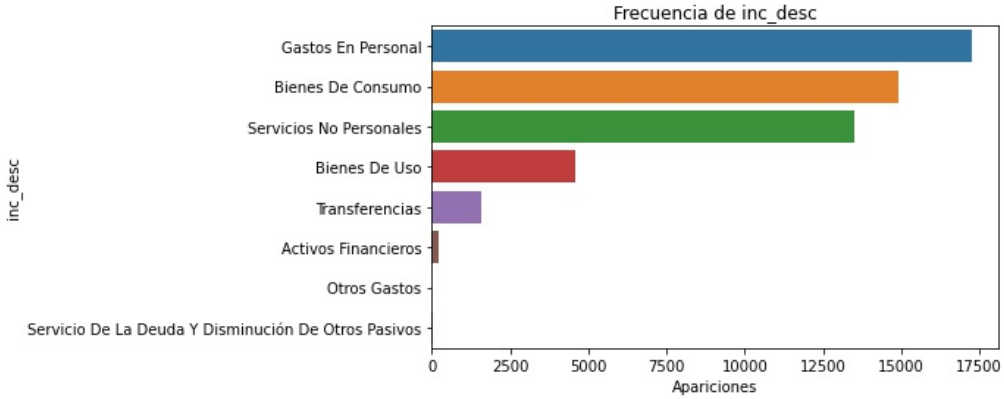


Figura 6: Frecuencia de la variable *inc_desc* en el dataset

4.15. Partida Principal

En la descripción se tiene que "es el nivel de segundo grado de cuentas dentro de las cuatro que conforman el Objeto de Gasto".

La variable *ppal* se describe como "Código de partida principal 1" y su variable correspondiente *ppal_desc* tiene 49 valores únicos, cuya moda es "Personal Permanente" con 7,026 apariciones, en la tabla 9 se puede ver que sobresalen las asignaciones a personal y bienes de consumo.

Cuadro 9: Principales valores de la variable ppal_desc

Entidad	Frecuencia	Porcentaje
Personal Permanente	7,026	13.49 %
Otros Bienes De Consumo	4,866	9.34 %
Personal Transitorio	4,632	8.89 %
Servicios Especializados, Comerciales Y Financieros	3,386	6.50 %
Productos Químicos, Combustibles Y Lubricantes	3,026	5.81 %
Mantenimiento, Reparación Y Limpieza	2,988	5.74 %
Maquinaria Y Equipo	2,969	5.70 %
Pulpa, Papel, Cartón Y Sus Productos	2,792	5.36 %
Otros Servicios	1,974	3.79 %
Asignaciones Familiares	1,926	3.70 %
Otros	16,468	31.63 %

4.16. Partida Parcial

Se define como "Es el nivel de tercer grado de cuentas dentro de las cuatro que conforman el Objeto de Gasto".

La variable *par* se describe como "Código de partida parcial 1" y la variable *par_desc* tiene 173 valores únicos.

En la tabla 10 se puede observar que la moda es "Otros no especificados precedentemente", lo cual resulta curioso y significa que existen muchos desconocidos, por lo cual se debería consultar a un experto del negocio y ver si es de interés o hay alguna explicación para su desconocimiento en el presupuesto oficial de la ciudad. Los valores siguientes tienen que ver con salarios y utilería.

Cuadro 10: Principales valores de la variable par_desc

Entidad	Frecuencia	Porcentaje
Otros No Especificados Precedentemente	5,287	10.15 %
Contribuciones Patronales	3.333	6.40 %
Retribución Del Cargo	3,328	6.39 %
Sueldo Anual Complementario	3,238	6.22 %
Complementos	3,143	6.03 %
Seguros De Riesgo De Trabajo	1,656	3.18 %
Personal Permanente	1,164	2.23 %
Alimentos Para Personas	1,039	1.99 %
Útiles De Escritorio, Oficina Y Enseñanza	1,008	1.93 %
Productos Farmacéuticos Y Medicinales	958	1.84 %
Otros	27,899	53.6 %

4.17. Subpartida

Se conoce que "Es el nivel de mínimo grado de cuentas dentro de las cuatro que conforman el Objeto de Gasto que solo existe para los Incisos de Trasferencias, Activos financieros y Servicios de la deuda".

La variable *sparc* se describe como "Código de partida subparcial 1" y la variable *sparc_desc* tiene 367 valores únicos. La tabla correspondiente es la 11 y es idéntica a la anterior debido a que las variables solo difieren en $1,513$ filas, estas se pueden encontrar en el Anexo.

Cuadro 11: Principales valores de la variable spar_desc

Entidad	Frecuencia	Porcentaje
Otros No Especificados Precedentemente	5,287	10.15 %
Contribuciones Patronales	3,333	6.40 %
Retribución Del Cargo	3,328	6.39 %
Sueldo Anual Complementario	3,238	6.22 %
Complementos	3,143	6.03 %
Seguros De Riesgo De Trabajo	1,656	3.18 %
Personal Permanente	1,164	2.23 %
Alimentos Para Personas	1,039	1.99 %
Útiles De Escritorio, Oficina Y Enseñanza	1,008	1.93 %
Productos Farmacéuticos Y Medicinales	958	1.84 %
Otros	27,899	53.6 %

4.18. Clasificador Económico

Esta variable es de gran utilidad ya que "La clasificación económica del gasto permite identificar la naturaleza económica de las transacciones que realiza el Sector Público, con el propósito de evaluar el impacto".

La variable *eco* se describe como "Código de clasificador económico 1" y par *eco_desc* tiene 23 valores únicos, cuya moda es "Remuneraciones Al Personal" con $17,258$ apariciones.

Figura 7: Frecuencia de la variable *eco_desc* en el dataset

4.19. Fuente de Financiamiento

La variable *fte* se describe como "Código de fuente de financiamiento 1" y par correspondiente *fte_desc* tiene 7 valores únicos, cuya moda es "Tesoro De La Ciudad" con $45,841$ apariciones, de forma que se puede apreciar en la figura 9 que casi todos los presupuestos se financian con el tesoro

de la ciudad. Además es interesante destacar otra inconsistencia entre la metadata y el dataset, ya que en la documentación se refieren a estas variables como *ff* y *ff_desc* respectivamente.

4.20. Ubicación Geográfica

En esta variable es importante comentar que no existe documentación oficial sobre la descripción ya que es un duplicado de otra variable y contiene un error que da a entender que no es el significado original, ya que dice textualmente: "Eco_DescTextoLa clasificación económica del gasto permite identificar la naturaleza económica de las transacciones que realiza el Sector Público, con el propósito de evaluar el impacto"

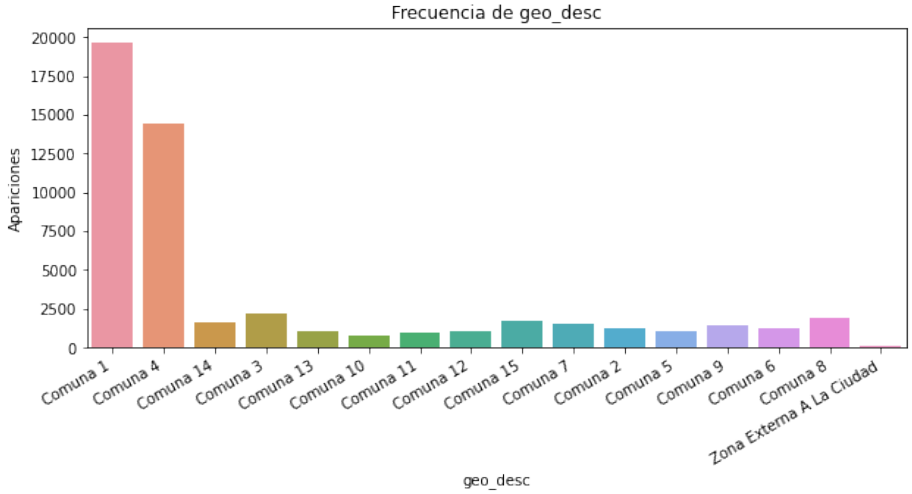


Figura 8: Frecuencia de la variable *geo* en el dataset

La variable *geo* se describe como "Código de ubicación geográfica 1". Su variable categórica correspondiente *geo_desc* tiene 16 valores únicos, cuya moda es "Comuna 1" con 19, 622 apariciones, seguido de la "Comuna 4". Juntas estas dos representan dos tercios de las asignaciones.

Figura 9: Frecuencia de la variable fte_desc en el dataset

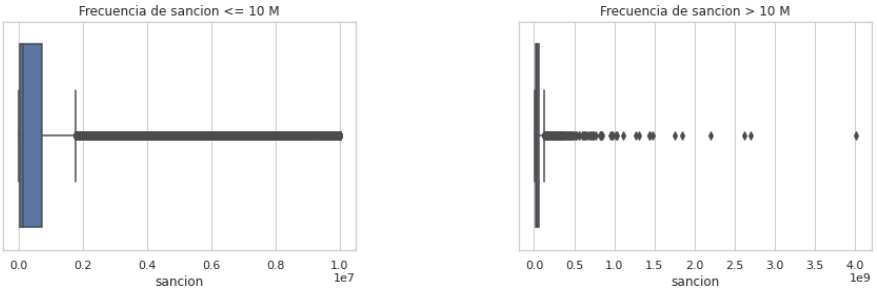
4.21. Sanción

Esta es "el crédito aprobado por Ley de Presupuesto para cada ejercicio por La Legislatura de la Ciudad Autónoma de Buenos Aires y promulgado mediante decreto por el Poder Ejecutivo".

La variable *sanción* tiene una media de $4,280,371e+06$ con una desviación estándar de $4,165,309e+07$, y una mediana de 0 ya que 16, 042 valores tienen 0 como sancionado, es de-

cir, el 30 % de las filas. Por lo cual para analizar mejor la data se separó en dos boxplots en la figura 10 donde se puede apreciar los presupuestos con sanción mayor o menor e igual a 10 Millones y que además sean diferentes de cero (ya que nos interesa visualizar como se distribuye el gasto).

En el caso cuando la sanción es menor o igual a 10 millones la media es $837,750{,}79$, con una mediana de $111,485{,}0$ y una desviación de $1,690,943{,}96$. En el caso contrario está la media de $68,288,004{,}32$, mediana de $27,635,764{,}0$ y la desviación en $165,105,224{,}26$; además se puede apreciar en el gráfico una presencia muy marcada de outliers.



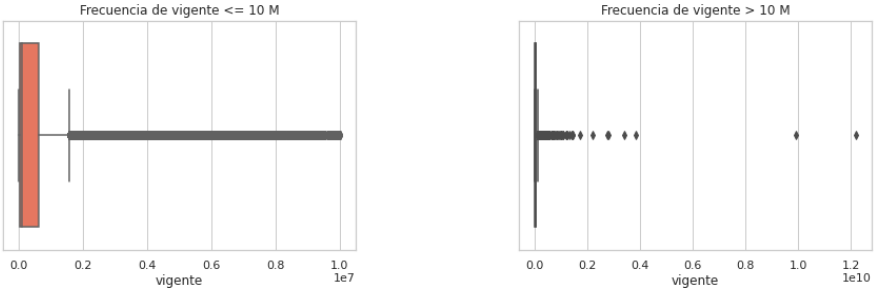
(a) Sanción $\leq$ 10 Millones (b) Sanción $>$ 10 Millones

Figura 10: Comparaciones entre boxplots de la variable sanción.

4.22. Vigente

Este es "el crédito sancionado más/menos las modificaciones presupuestarias realizadas dentro del ejercicio corriente".

La variable *vigente* tiene una media de $4,831,267e+06$ con una desviación estándar de $8,166,233e+07$, y una mediana de $2,700,000e+04$. Al igual que en la variable anterior, esta tiene una cantidad considerable de ceros ($5,585$ exactamente o casi el 11 %), además para una mejor visualización se subdividió en dos boxplots en la figura 11.



(a) Vigente $\leq$ 10 Millones (b) Vigente $>$ 10 Millones

Figura 11: Comparaciones entre boxplots de la variable vigente.

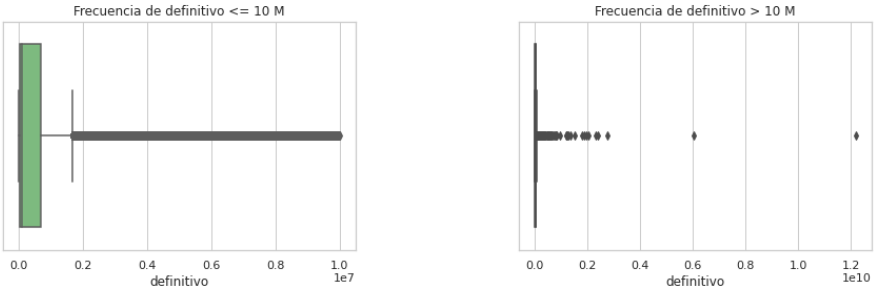
En la sección de menor valor del dataset (vigente acotado por 10 Millones) se tiene la media de $789,175{,}29$, mediana de $89,720{,}0$ y desviación en $1,644,460{,}18$. Su contra parte mayor a 10 millones tiene una media de $69,749,225{,}36$, mediana en $24,007,975{,}0$ y una desviación de $327,029,752{,}86$.

4.23. Definitivo

Se tiene definido como "aquel que dá origen a una relación jurídica con terceros, que originará, en el futuro, una eventual salida de fondos, sea para cancelar una deuda o para su inversión en un objeto determinado".

La variable *definitivo* tiene una media de $3,399,354e+06$ con una desviación estándar de $6,868,158e+07$, y una mediana de $7,209,100e+04$. Esta variable también tiene 16,007 valores en cero, es decir, casi 30 % y de igual forma que la anterior se separó en dos boxplots para su análisis en al figura 12 y se removieron los ceros.

En la primera división cuando la variable es menor a 10 millones se tiene un media de 804,575,82, una mediana de 98,203,68 y una desviación en 1,660,673,36. Es interesante destacar que por último en la segunda parte del dataset cuando definitivo > 10 Millones se tiene una media de 63,386,999,91, una mediana en 23,543,106,75 y la desviación en 316,379,007,24.



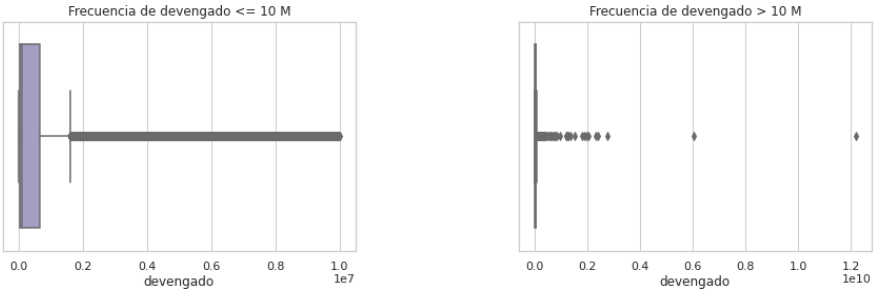
(a) Definitivo $\leq$ 10 Millones (b) Definitivo > 10 Millones

Figura 12: Comparaciones entre boxplots de la variable Definitivo.

4.24. Devengado

Finalmente, devengado es "el surgimiento de una obligación de pago, por la recepción en conformidad de bienes o servicios oportunamente contratados o por haberse cumplido los requisitos administrativos".

La variable *devengado* tiene una media de $3,133,693e+06$ con una desviación estándar de $6,829,328e+07$, y una mediana de $1,455,198e+04$. Esta variable aparece en la documentación como una variable entera y en el dataset aparece como float. Por otra parte, tiene 16,874 valores en cero, es decir, más del 30 % y para visualizarla mejor, igual que en los casos anteriores, se separó en dos boxplots en la figura 13 y se removieron los ceros.



(a) Devengado $\leq$ 10 Millones (b) Definitivo > 10 Millones

Figura 13: Comparaciones entre boxplots de la variable Devengado.

La primera sección donde devengado $\leq$ 10 Millones se tiene una media de 781,999,04, con una mediana de 93,800,69 y una desviación de 1,632,113,62. Y en la segunda (devengado ¿10 Millones) se tiene la media de 63,303,066,80, la mediana de 22,804,711,59 y la desviación en 328,869,474,10

5. Análisis Bivariado

5.1. Categórica vs Categórica

Se compararon las variables categóricas para analizar las frecuencias del carácter del gasto para cada finalidad en la figura 14. De esta forma se confirmó que para cada finalidad el carácter con mayores asignaciones de presupuesto continuó siendo la "Administración Central" (al igual que el dataset en general), es decir, no existe una finalidad con mayores asignaciones por parte de los organismos descentralizados, y en particular se puede apreciar que la gran mayoría de los servicios sociales y de seguridad son dirigidos por la "Administración Central".

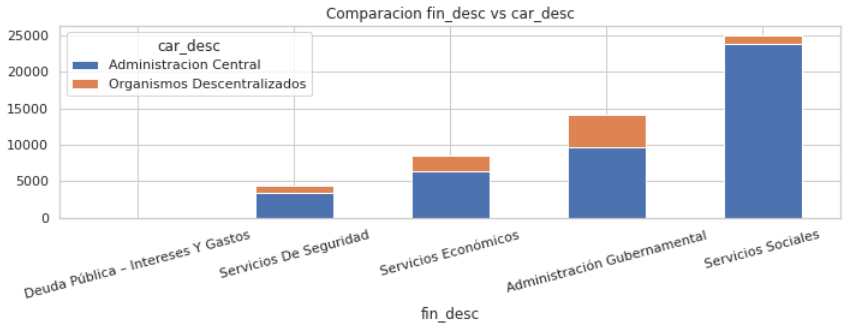


Figura 14: Gráfico de barras apiladas comparando fin_desc vs car_desc

5.2. Categórica vs Numérica

En esta sección se analizó el gasto devengado en el presupuesto por cada comuna de la ciudad en la figura 15. Es interesante ver que se mantiene una relación entre la cantidad de presupuestos asignados y el total del gasto, es decir, no es común que haya pocas asignaciones de montos muy altos, de manera tal que se mantienen la Comuna 1 como la de mayor presupuesto, seguida de la 4 y la 3 en ese orden, al igual que en el gráfico 8 de frecuencia en la sección de análisis univariado.

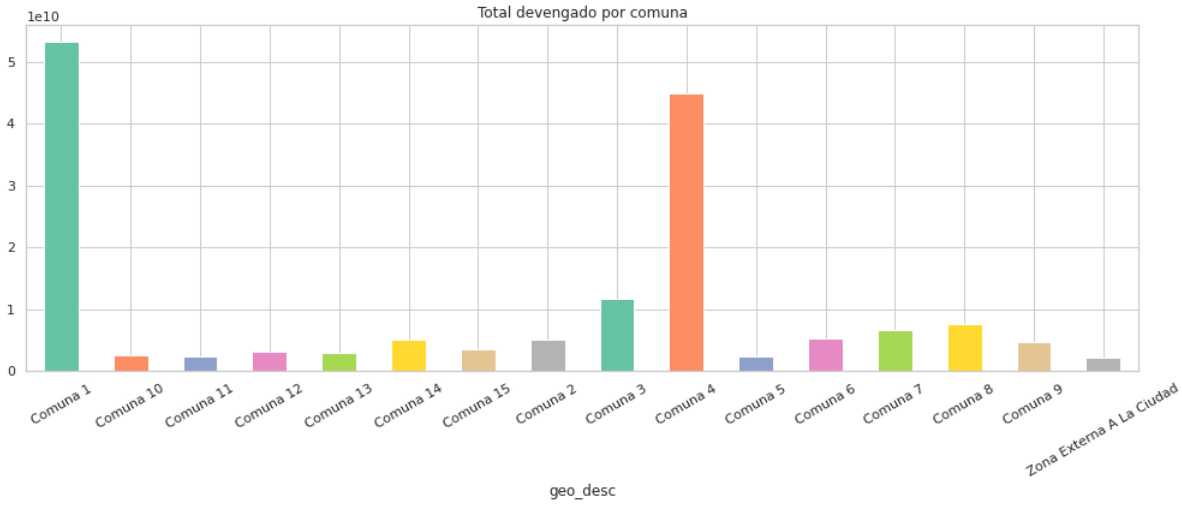
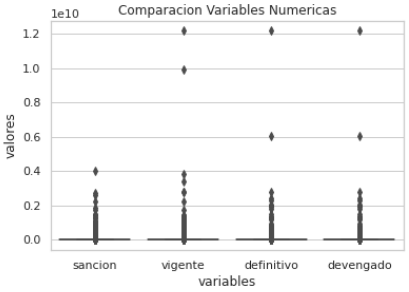


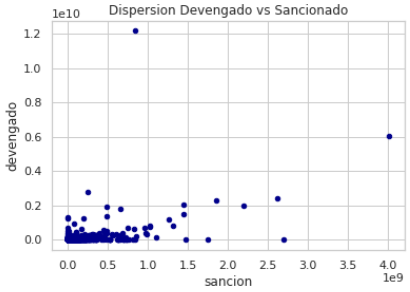
Figura 15: Gráfico de barras comparando geo_desc y el total devengado por comuna

5.3. Numérico vs Numérico

Se compararon las variables numéricas utilizando un box plot para comparar las distribuciones de cada una para observar con facilidad la similitud existente entre estas, así como sus percentiles, outliers y medias. Al analizar la figura 16 se puede ver en los outliers la marcada diferencia entre el presupuesto inicial o sancionado en comparación con las otras categorías posteriores.



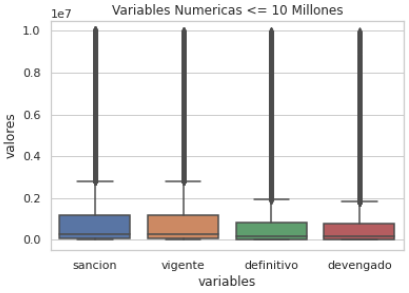
(a) Variables > 10 Millones



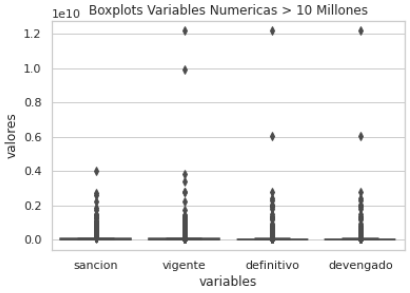
(b) Scatterplot Sancionado vs Devengado

Figura 16: Comparación de variable numéricas

Además, en la figura 16(b) se puede ver una scatterplot comparando lo sancionado y devengado para ver que por lo general es mayor lo devengado que lo sancionado originalmente.



(a) Variables $\leq$ 10 Millones



(b) Variables > 10 Millones

Figura 17: Comparaciones entre boxplots variables numéricas.

Por otra parte, para visualizar mejor de manera comparativa las distribuciones de las variables en la figura 17 se dividieron los boxplots en aquellos cuando todas las variables numéricas eran menores o iguales a 10 millones y diferentes de cero en el gráfico (a) y todos los demás casos en el gráfico (b) excluyendo también los ceros.

Cuadro 12: Correlación de las variables numéricas

	sanción	vigente	definitivo	devengado
sanción	1.000000	0.635581	0.513513	0.505972
vigente	0.635581	1.000000	0.933877	0.931916
definitivo	0.513513	0.933877	1.000000	0.998214
devengado	0.505972	0.931916	0.998214	1.000000

Finalmente, en el cuadro 12 se puede apreciar la correlación de las variables numéricas y en la figura 18 se puede ver la matriz de correlación donde se evidencia la distancia o diferencia entre la

variable sanción y las demás. También se puede observar como devengado y definitivo tienen una fuerte correlación, es decir que entre estas no hay muchos cambios de presupuesto.

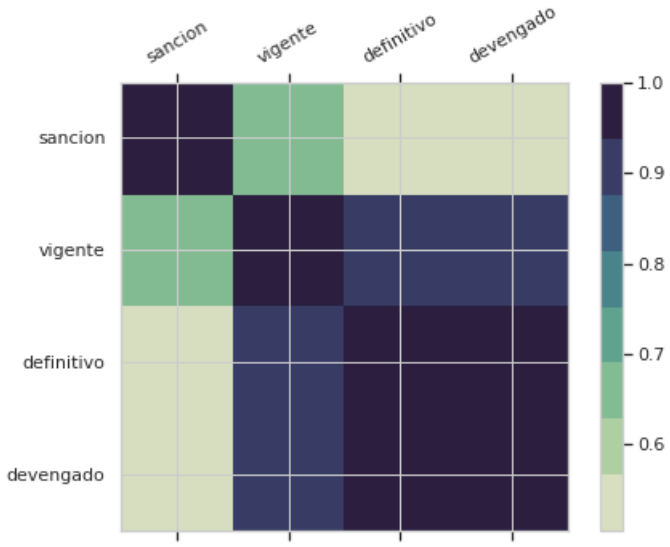


Figura 18: Box Plot comparativo entre las variables numéricas

6. Conclusiones

El dataset del presupuesto de la Ciudad de Buenos Aires se encuentra correctamente estructurado y categorizado en sus 52,053 filas y 44 columnas con información consistente y completa. Al visualizar cada variable mediante gráficos o tablas se pudo apreciar cualidades interesantes del gasto y la distribución del presupuesto, así como la correlación entre diferentes variables.

Durante el presente trabajo se se encontraron varios aspectos que llaman la atención como la disparidad del gasto por comuna donde se evidencia aun la centralización del poder Ejecutivo y Administrativo de la ciudad en diferentes zonas geográficas, esta diferencia presupuestaria seguramente sería aun mas marcada si evaluáramos el presupuesto de la nación. para evitar estas concentraciones de poder económico y político otras países han optado por separar su capital de su ciudad más grande, como es el caso de Australia con Canberra como capital sin embargo la ciudad más poblada es Sydney. Una diversificación y distribución mas equitativa del presupuesto de CABA permitiría hacer una ciudad mas inclusiva y beneficiosa para todos los bonaerenses. Por otra parte se pudo apreciar que los mayores gasto van a los servicios sociales como la educación y salud, pero también que la ciudad ha invertido poco en finanzas o deudas lo cual puede ser un síntoma de la inflación, sin embargo, esta clase de conclusiones son mas adecuadas para ser tomadas por un experto del modelo de negocio.

Ahora bien, este dataset 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 ya que se encontraron errores e inconsistencias con el `.csv` del dataset. A la vez aún es necesario hacer una limpieza de datos para mejorar variables que son casi idénticas pero difieren en *typos* o espaciado como el caso de Actividad y Obra.

Por último, al hacer el análisis bivariado también saltó un hecho relevante o que llama la atención que es la gran diferencia entre lo sancionado y lo devengado, o en si la poco correlación entre lo sancionado y las otras variables numéricas, es decir muy pocas veces realmente se cumple

el presupuesto inicial y en algunos casos este presupuesto aumentó hasta en un orden de magnitud con respecto al original. Ahora bien, para finalizar se recomienda añadir la fecha de los cambios en vigente, definitivo y devengado a fin de mantener un mejor control sobre el histórico de las transacciones con respecto a la sanción inicial y si estas variaciones están relacionadas con la inflación del año o fluctuaciones en los precios de divisas.

7. Anexo (código fuente)

El manejo del dataset así como el análisis y la visualización de los datos fue realizado utilizando Python sobre un Kaggle notebook tomando la data del sitio [3].

Se deja más abajo el link a dicho notebook, junto con la correspondencia a cada ejercicio y una breve explicación sobre el razonamiento del código.

LINK AQUÍ [4]

Además, se deja el link a los spreedshets con las diferencias encontradas entre variables casi idénticas:

- **eg_desc vs act_desc.**[5]
- **ob_desc vs act_desc.**[6]
- **par_desc vs spar_desc**[7]

[3] https://data.buenosaires.gob.ar/dataset/presupuesto-ejecutado/archivo/juqdkmgo-16623-resource

[4] https://www.kaggle.com/duerunner/aproximaci-n-a-la-calidad-de-datos/

[5] https://www.kaggle.com/duerunner/presupuesto-de-la-ciudad-de-buenos-aires-2018?select=comparacion-eg-vs-ent.csvg

[6] https://www.kaggle.com/duerunner/presupuesto-de-la-ciudad-de-buenos-aires-2018?select=comparacion+ob_desc+vs+act_desc.csv

[7] https://www.kaggle.com/duerunner/presupuesto-de-la-ciudad-de-buenos-aires-2018?select=comparacion+par+vs+spar.csv