

Guía 4 – Anteproyecto

El equipo ha determinado que son necesarias las siguientes fuentes de datos para cumplir con los requisitos y el alcance planteado en anteriores entregas:

Nombre de la Fuente	Fuente	Granularidad	Filas	Observaciones
SECOPII - Multas y Sanciones	https://www.datos.gov.co/Gastos-Gubernamentales/SECOPII-Multas-y-Sanciones/it5q-hg94	Proyecto-Proveedor	0.5K	Fuente externa (Datos Abiertos). Accesible mediante descarga de csv o API
SECOP II - Contratos Electrónicos	https://www.datos.gov.co/Gastos-Gubernamentales/SECOP-II-Contratos-Electronicos/jbjy-vk9h	Contrato-Proveedor	2M	Fuente externa (Datos Abiertos). Accesible mediante descarga de csv o API
SECOP II - Procesos de Contratación	https://www.datos.gov.co/Gastos-Gubernamentales/SECOP-II-Procesos-de-Contratacion/p6dx-8zbt	Proyecto-Proveedor	2M	Fuente externa (Datos Abiertos). Accesible mediante descarga de csv o API

En términos generales, las fuentes primarias de datos son accesibles desde cualquier red pública y pueden ser consumidas directamente por medio de un API que permite filtrar en tiempo de ejecución las filas que se desean obtener de cada fuente o se puede descargar por medio de archivos csv cada fuente y posterior a ese cargue iniciar, se puede gestionar su actualización diaria o mensual según se desee.

Para cada una de las fuentes identificadas, daremos respuesta a los 3 puntos solicitados en la guía:

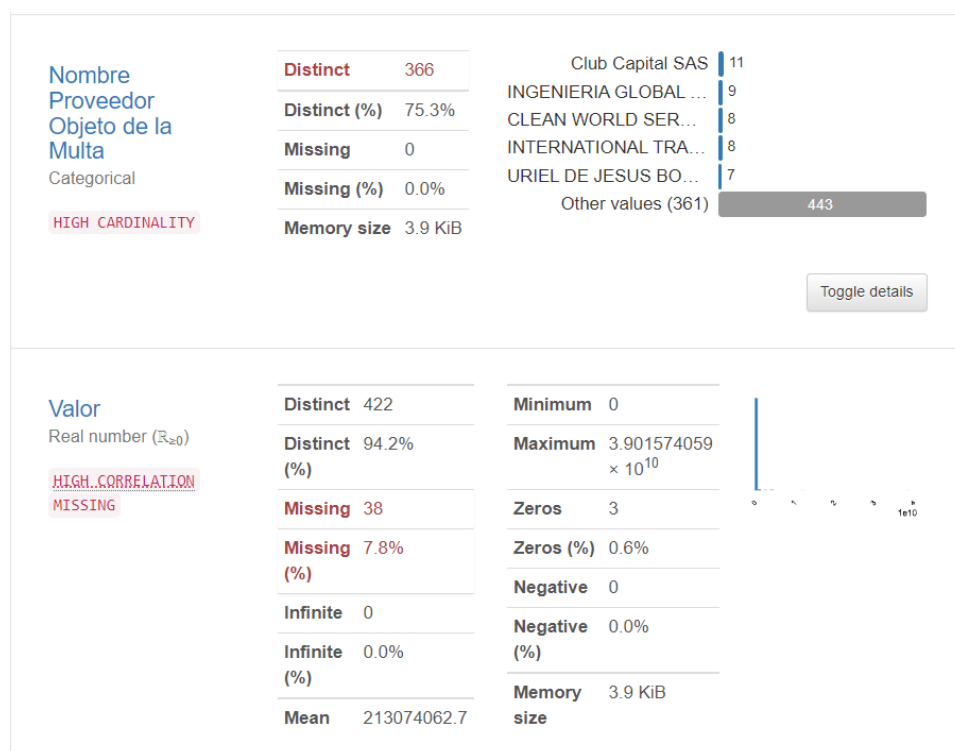
1. Características de Calidad de Datos

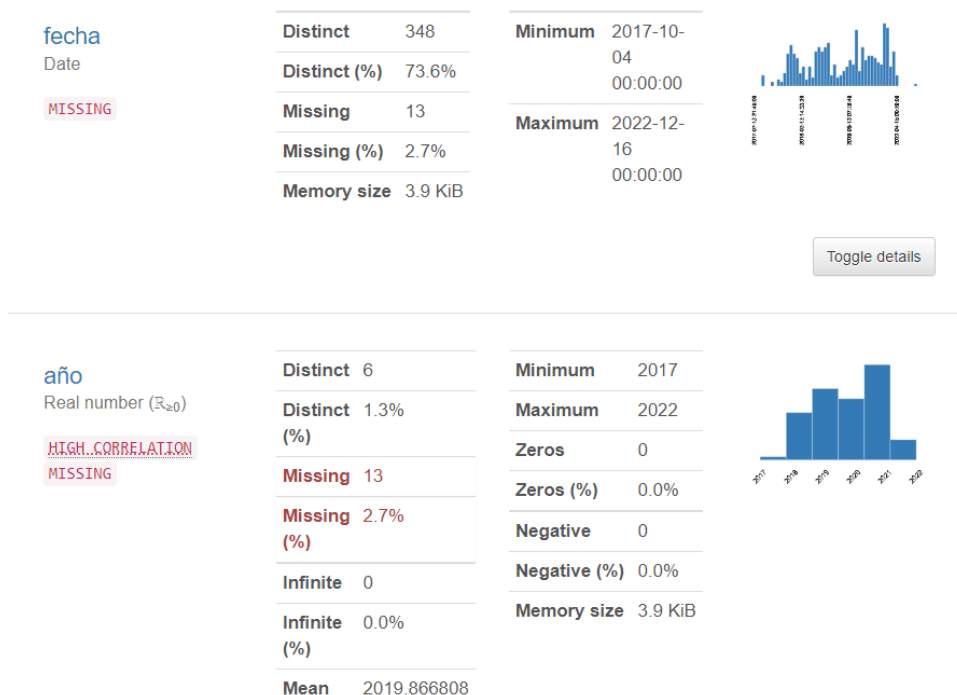
- SECOP II – Multas y sanciones

- Completitud: La base contiene un 8.4% de valores nulos, los cuales se concentran en solamente 9 de las 19 columnas que la componen. En términos generales, es una fuente de datos muy buena. Los datos faltantes corresponden al ID del Contrato, el valor de la multa, el valor pagado, la fecha del evento, si aplico garantías, la descripción de la multa y la fecha de la multa. Desafortunadamente, algunas de ellas son muy relevantes para el dominio del problema como lo son el valor de la multa y la fecha, por lo que se deben evaluar alternativas para manejar dichos faltantes.

Overview		Alerts 36	Reproduction
Dataset statistics		Variable types	
Number of variables	19	Categorical	11
Number of observations	486	Numeric	6
Missing cells	772	Boolean	1
Missing cells (%)	8.4%	Date Time	1
Duplicate rows	0		
Duplicate rows (%)	0.0%		
Total size in memory	72.3 KiB		
Average record size in memory	152.3 B		

- Consistencia: La fuente posee una alta coherencia. Se respeta la definición de los tipos de datos de acuerdo a la documentación oficial que la soporta en el sitio web de datos abiertos.
- Claridad: Los resultados obtenidos por técnicas estadísticas son claros. Por ejemplo, para los datos numéricos, es posible obtener su distribución por cuartiles. Para las variables categóricas también factible obtener las cifras asociadas con cada una:



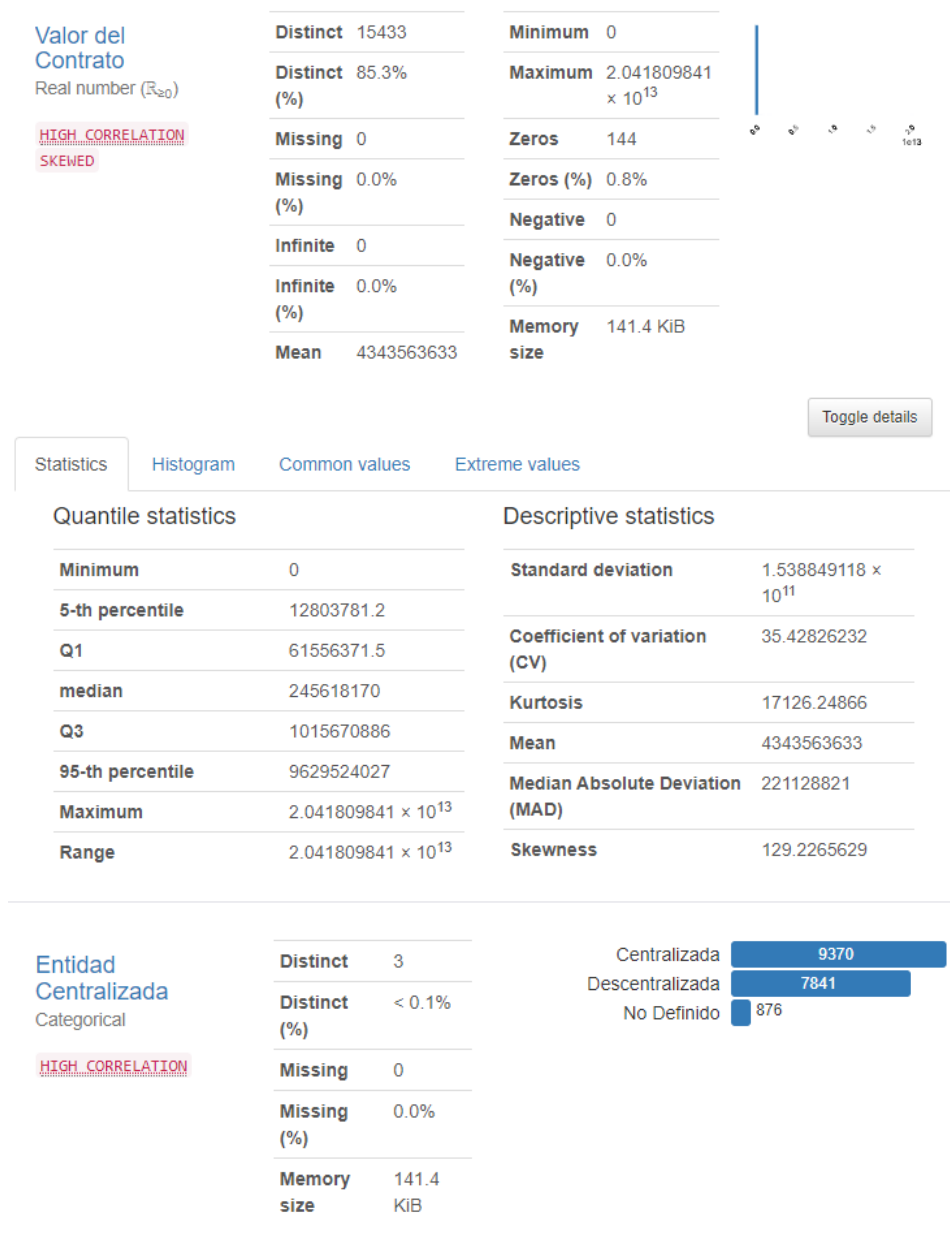


- Formato: Hasta el momento, el análisis realizado en los datos ha certificado formato valido para todos los campos de la fuente. Las fechas se encuentran en formatos válidos y los campos que hacen referencia a valores numéricos se encuentran dentro de las escalas esperadas (millones para cifras monetarias y miles para conteo de días).
- SECOP II - Contratos Electrónicos
 - Completitud: La base contiene un 8.4% de valores nulos, los cuales se concentran en solamente 8 de las 70 columnas que la componen. En términos generales, es una fuente de datos muy buena. Los datos faltantes corresponden a fechas (fecha de última actualización, fecha de liquidación etc.). Desafortunadamente algunas de ellas son muy relevantes para el dominio del problema, por lo que se deben evaluar alternativas para manejar dichos faltantes. (Más adelante se aclara el motivo por el cual se listan acá solo 18 mil observaciones en lugar de los 2M mencionados anteriormente).

Overview	Alerts 114	Reproduction
----------	------------	--------------

Dataset statistics		Variable types	
Number of variables	70	Numeric	19
Number of observations	18087	Categorical	50
Missing cells	55371	Unsupported	1
Missing cells (%)	4.4%		
Duplicate rows	0		
Duplicate rows (%)	0.0%		
Total size in memory	9.7 MiB		
Average record size in memory	560.0 B		

- Consistencia: La fuente posee una alta coherencia. Se respeta la definición de los tipos de datos de acuerdo a la documentación oficial que la soporta en el sitio web de datos abiertos.
- Claridad: Los resultados obtenidos por técnicas estadísticas son claros. Por ejemplo, para los datos numéricos, es posible obtener su distribución por cuartiles. Para las variables categóricas también factible obtener las cifras asociadas con cada una:



- Formato: Hasta el momento, el análisis realizado en los datos ha certificado formato valido para todos los campos de la fuente. Las fechas se encuentran en formatos válidos y los campos que hacen referencia a valores numéricos se encuentran dentro de las escalas esperadas (millones para cifras monetarias y miles para conteo de días).
- SECOPII - Procesos de Contratación
 - Completitud: La base tiene un total de 5.6% de celdas vacías. Tiene un nivel de completitud bastante alto para poder ser utilizada en nuestros modelos.

Overview

Overview

Alerts 81

Reproduction

Dataset statistics


Number of variables	57
Number of observations	34896
Missing cells	110912
Missing cells (%)	5.6%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	15.2 MiB
Average record size in memory	456.0 B

Variable types

Categorical	44
Numeric	12
Unsupported	1

- Consistencia: Los datos contenidos en cada una de las variables/columnas respetan la estructura definida para el tipo de dato especificado en el sitio web.
- Claridad: Los resultados obtenidos por técnicas estadísticas son claros. Por ejemplo, para los datos tipo texto se sacan las categorías principales, para los datos numéricos se obtienen información de mínimos, máximos, media, y cuartiles.

Variables

entidad Categorical HIGH_CARDINALITY	<table><tr><td>Distinct</td><td>1871</td></tr><tr><td>Distinct (%)</td><td>5.4%</td></tr><tr><td>Missing</td><td>0</td></tr><tr><td>Missing (%)</td><td>0.0%</td></tr><tr><td>Memory size</td><td>272.8 KiB</td></tr></table> <div><div>INSTITUTO NACIONAL DE VIAS3694</div><div>UNIDAD ADMINISTRATIVA ESPECIAL ...1405</div><div>INSTITUTO DE INFRAESTRUCTURA ...1273</div><div>GOBERNACION DE CALDAS748</div><div>MUNICIPIO DE MANIZALES429</div><div>Other values (1866)27347</div></div> <div>Toggle details</div>	Distinct	1871	Distinct (%)	5.4%	Missing	0	Missing (%)	0.0%	Memory size	272.8 KiB																		
Distinct	1871																												
Distinct (%)	5.4%																												
Missing	0																												
Missing (%)	0.0%																												
Memory size	272.8 KiB																												
nit_entidad Real number (R ₃₂)	<table><tr><td>Distinct</td><td>1684</td></tr><tr><td>Distinct (%)</td><td>4.8%</td></tr><tr><td>Missing</td><td>0</td></tr><tr><td>Missing (%)</td><td>0.0%</td></tr><tr><td>Infinite</td><td>0</td></tr><tr><td>Infinite (%)</td><td>0.0%</td></tr><tr><td>Mean</td><td>1192481206</td></tr></table> <table><tr><td>Minimum</td><td>800000118</td></tr><tr><td>Maximum</td><td>9015641901</td></tr><tr><td>Zeros</td><td>0</td></tr><tr><td>Zeros (%)</td><td>0.0%</td></tr><tr><td>Negative</td><td>0</td></tr><tr><td>Negative (%)</td><td>0.0%</td></tr><tr><td>Memory size</td><td>272.8 KiB</td></tr></table> <div></div> <div>Toggle details</div>	Distinct	1684	Distinct (%)	4.8%	Missing	0	Missing (%)	0.0%	Infinite	0	Infinite (%)	0.0%	Mean	1192481206	Minimum	800000118	Maximum	9015641901	Zeros	0	Zeros (%)	0.0%	Negative	0	Negative (%)	0.0%	Memory size	272.8 KiB
Distinct	1684																												
Distinct (%)	4.8%																												
Missing	0																												
Missing (%)	0.0%																												
Infinite	0																												
Infinite (%)	0.0%																												
Mean	1192481206																												
Minimum	800000118																												
Maximum	9015641901																												
Zeros	0																												
Zeros (%)	0.0%																												
Negative	0																												
Negative (%)	0.0%																												
Memory size	272.8 KiB																												
departamento_entidad Categorical HIGH_CORRELATION	<table><tr><td>Distinct</td><td>33</td></tr><tr><td>Distinct (%)</td><td>0.1%</td></tr><tr><td>Missing</td><td>0</td></tr><tr><td>Missing (%)</td><td>0.0%</td></tr><tr><td>Memory size</td><td>272.8 KiB</td></tr></table> <div><div>No Definido12248</div><div>Distrito Capital de Bogotá8681</div><div>Valle del Cauca2196</div><div>Antioquia1980</div><div>Santander1908</div><div>Other values (28)7883</div></div> <div>Toggle details</div>	Distinct	33	Distinct (%)	0.1%	Missing	0	Missing (%)	0.0%	Memory size	272.8 KiB																		
Distinct	33																												
Distinct (%)	0.1%																												
Missing	0																												
Missing (%)	0.0%																												
Memory size	272.8 KiB																												

- Formato: Los datos respetan la estructura de datos y los formatos de cada una de las variables. Los campos numéricos, de fecha y de texto son extraídos, leídos e interpretados correctamente.

2. Técnicas de Limpieza de datos

- SECOP II – Multas y sanciones

Comenzamos a realizar especial énfasis en los campos con valores nulos, asociados a 9 columnas del dataset:

ID Contrato: identificamos que a pesar de que esta variable podía ser la llave para conectar varias fuentes de datos se iba a utilizar otros campos como llaves para conectar diferentes fuentes de información. Se concluyó lo anterior ya que había fuentes de datos que no contaban con este campo de ID Contrato y además teníamos bastantes nulos por lo que no resultaba ser una buena llave. Por lo pronto, se decide no realizar imputación ni eliminación de este campo.

Valor: Identificamos que los valores nulos de esta columna podrían corresponder a aquellos proyectos que finalmente no fueron multados. Por lo tanto, se decide no realizar imputación ni eliminación de este campo.

Valor Pagado: Resulta coherente que, si el campo del valor total de la multa tiene nulos, el campo de valor pagado de la multa también puede contener nulos. Además, también es posible que a pesar de que un proyecto haya sido multado, todavía no haya pagado nada correspondiente. Por lo tanto, se decide no realizar imputación ni eliminación de este campo.

Fecha Evento: llama la atención que existan contratos activos sin este valor. Por lo pronto, se decide no realizar imputación ni eliminación de este campo. De igual forma son pocos registros que no tienen una fecha del evento asociada (solo 13 contratos). Sin embargo, se tiene pendiente un análisis más a fondo con la ayuda de los SH sobre las 13 filas asociadas a contratos activos sin fecha del evento.

Aplico garantías: Llama la atención de que no se tenga conocimiento de si un contrato aplica garantías o no. Sin embargo, puede corresponder a que hay contratos específicos que no necesitan aplicar garantías por lo que podría tener este campo vacío. Sin embargo, tenemos pendiente aclarar estas dudas con los SH sobre las 12 filas que tienen este campo vacío. Por lo tanto, se decide no realizar imputación ni eliminación de este campo.

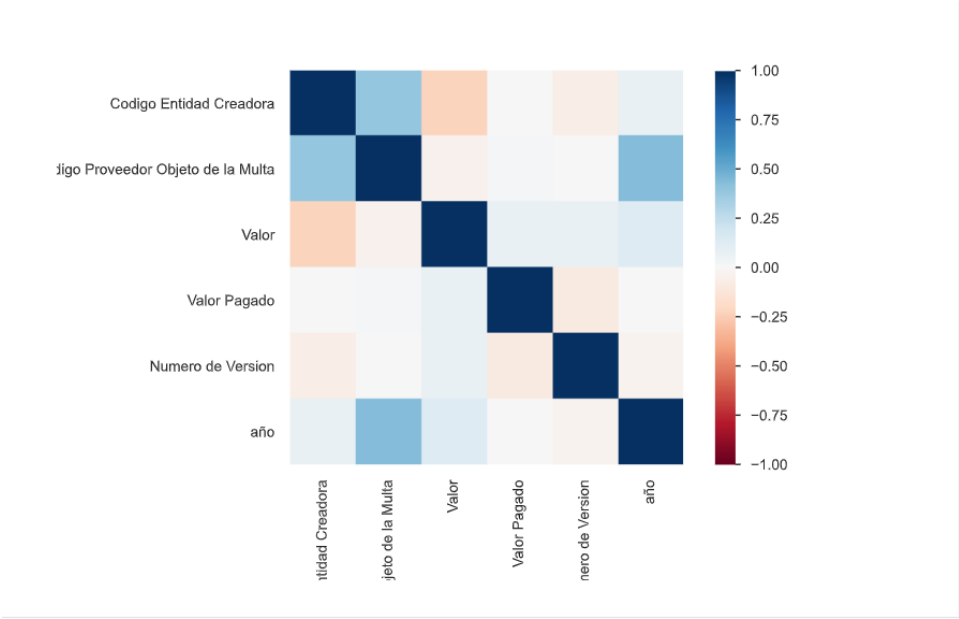
Numero de Acto: A pesar de que es un campo que podría darnos detalles sobre el tipo de multa o infracción que incurrió un contrato, para los objetivos de nuestro proyecto esta variable no resulta importante. Lo anterior es dado a que nosotros solo queremos predecir si un proyecto será multado o no (sin importar el tipo de resolución que violó). Por lo tanto, se decide no realizar imputación ni eliminación de este campo.

Descripción Otro Tipo de Sanción: A pesar de que es un campo que podría darnos detalles sobre el tipo de multa o infracción que incurrió un contrato, para los objetivos de nuestro proyecto esta variable no resulta importante. Lo anterior es dado a que nosotros solo queremos predecir si un proyecto será multado o no (sin importar el tipo de infracción). Por lo tanto, se decide no realizar imputación ni eliminación de este campo.

Fecha: Resulta un campo al que se debe indagar más información del porque se pueden tener valores nulos. Sin embargo, no aporta mucha información valiosa para el objetivo de nuestro proyecto ya que predeciremos si un contrato es exitoso o no con base en otras características. Por el momento, se decide no realizar imputaciones ni eliminaciones de filas de este campo.

Año: Es coherente que este campo presente valores nulos dado que la variable Fecha también contiene valores nulos. Dichas filas que tienen valores nulos corresponden 100% a las filas con valores nulos en el campo de Fecha. Por lo tanto, una vez se aclaren las dudas con los SH del porque podríamos tener fechas vacías, se decidirá eliminar o imputar de la misma manera la variable Año. Pero, por el momento, se decide no realizar imputaciones ni eliminaciones hasta no tener mayor claridad.

Por último, no se identifican correlaciones significativas entre los datos que permita sospechar la eliminación de variables.



- SECOP II - Contratos Electrónicos

Debido al alcance planteado en el proyecto, en donde se busca analizar específicamente los contratos asociados a proyectos de infraestructura, el primer paso en la limpieza realizada en esta fuente fue la identificación de los contratos que estuvieran asociados únicamente a dicho tipo de proyectos:

Tipo de Contrato	
Acuerdo Marco de Precios	1110
Acuerdo de cooperación	19
Arrendamiento de inmuebles	17955
Arrendamiento de muebles	688
Asociación Público Privada	512
Comisión	139
Comodato	4926
Compraventa	54505
Concesión	90
Consultoría	4650
DecreeLaw092/2017	54948
Emprestito	283
Interventoría	5269
Negocio fiduciario	101
No Especificado	1727
Obra	18087
Otro	78284
Prestación de servicios	1682019
Seguros	3178
Servicios financieros	400
Suministros	56271
Venta inmuebles	37
Venta muebles	166
Name: Tipo de Contrato, dtype: int64	

Así, de los más de 2M de registros de esta fuente de datos, solo 18K están asociados a Tipos de Contrato de Obras públicas, por lo cual, procedimos a descartar todos los demás.

Luego, comenzamos a realizar especial énfasis en los campos con valores nulos, asociados a 8 columnas del dataset:

Fecha de firma: Al realizar el análisis correspondiente por Estado del contrato, se puede llegar a la conclusión de que es normal que haya contratos de ciertos estados que no hayan sido firmados aún. No es necesario realizar imputaciones o eliminaciones.

```
In [61]: df4 = df['Fecha de Firma'].isnull().groupby([df['Estado Contrato']]).sum().astype(int).reset_index(name='count')
print(df4)
```

	Estado Contrato	count
0	Activo	0
1	Borrador	1381
2	Cerrado	0
3	En aprobación	291
4	En ejecución	0
5	Modificado	0
6	Suspendido	0
7	cedido	0
8	enviado Proveedor	419
9	terminado	0

Fecha Inicio de Contrato: llama la atención que existan contratos activos sin este valor. Los demás pueden ser explicados por el tipo de contrato. Por lo pronto, se decide no realizar imputación ni eliminación de este campo. Se tiene pendiente un análisis más a fondo con la ayuda de los SH sobre las 923 filas asociadas a contratos activos sin fecha de Inicio de contrato.

```
In [62]: df4 = df['Fecha de Inicio del Contrato'].isnull().groupby([df['Estado Contrato']]).sum().astype(int).reset_index(name='count')
print(df4)
```

	Estado Contrato	count
0	Activo	923
1	Borrador	1278
2	Cerrado	0
3	En aprobación	83
4	En ejecución	0
5	Modificado	0
6	Suspendido	0
7	cedido	0
8	enviado Proveedor	137
9	terminado	0

Fecha de Fin de Contrato: Los faltantes están asociados a Borradores. Comienza a ser tentador eliminar todos los contratos borrador, pero aún no se toma dicha decisión. No se realiza imputación.

```
In [63]: df4 = df['Fecha de Fin del Contrato'].isnull().groupby([df['Estado Contrato']]).sum().astype(int).reset_index(name='count')
print(df4)
```

	Estado Contrato	count
0	Activo	0
1	Borrador	782
2	Cerrado	0
3	En aprobación	0
4	En ejecución	0
5	Modificado	0
6	Suspendido	0
7	cedido	0
8	enviado Proveedor	0
9	terminado	0

Fecha de Inicio y Fin de Ejecución: Se concentra una buena parte de los faltantes en estas columnas. No es claro el motivo por el cual pueda estar la fuente en este estado, ya que incluso contratos cerrados y terminados tienen dichos faltantes. Imputar valores en fechas de contrato no es nada fácil, puesto que no se conoce a ciencia cierta cual puede ser un valor adecuado. Por lo pronto, se decide no imputar ni eliminar estas filas (de eliminar, nos quedaríamos sin gran parte de la información).


```
In [64]: df4 = df['Fecha de Inicio de Ejecucion'].isnull().groupby([df['Estado Contrato']]).sum().astype(int).reset_index(name='count')
print(df4)
```

Estado Contrato	count
0 Activo	1389
1 Borrador	1338
2 Cerrado	178
3 En aprobación	211
4 En ejecución	3721
5 Modificado	5273
6 Suspendido	353
7 cedido	22
8 enviado Proveedor	312
9 terminado	1417

```
In [65]: df4 = df['Fecha de Fin de Ejecucion'].isnull().groupby([df['Estado Contrato']]).sum().astype(int).reset_index(name='count')
print(df4)
```

Estado Contrato	count
0 Activo	1382
1 Borrador	1314
2 Cerrado	178
3 En aprobación	288
4 En ejecución	3721
5 Modificado	5257
6 Suspendido	353
7 cedido	22
8 enviado Proveedor	262
9 terminado	1486

Fecha Inicio y Fin Liquidación: Columnas candidatas a ser excluidas del dataset, ya que casi la totalidad de las filas (alrededor de un 61%) no tienen valores. Por lo pronto, se mantienen en el dataset para ser incluidas en el análisis de características principales. Tampoco se realiza imputación. Este campo tiene cierta coherencia con el campo flag “Liquidación” ya que se esperaría que únicamente las filas marcadas como “No” tuvieran los nulos, sin embargo, 9 registros que están marcados como “Si” tienen nulos en estas filas.

```
In [66]: df4 = df['Fecha Inicio Liquidacion'].isnull().groupby([df['Estado Contrato']]).sum().astype(int).reset_index(name='count')
print(df4)
```

Estado Contrato	count
0 Activo	1235
1 Borrador	1326
2 Cerrado	214
3 En aprobación	222
4 En ejecución	2338
5 Modificado	3820
6 Suspendido	256
7 cedido	15
8 enviado Proveedor	325
9 terminado	1131

```
In [67]: df4 = df['Fecha Fin Liquidacion'].isnull().groupby([df['Estado Contrato']]).sum().astype(int).reset_index(name='count')
print(df4)
```

Estado Contrato	count
0 Activo	1235
1 Borrador	1326
2 Cerrado	214
3 En aprobación	222
4 En ejecución	2338
5 Modificado	3821
6 Suspendido	256
7 cedido	15
8 enviado Proveedor	325
9 terminado	1131

```
In [79]: df4 = df['Fecha Inicio Liquidacion'].isnull().groupby([df['Liquidación']]).sum().astype(int).reset_index(name='count')
print(df4)
```

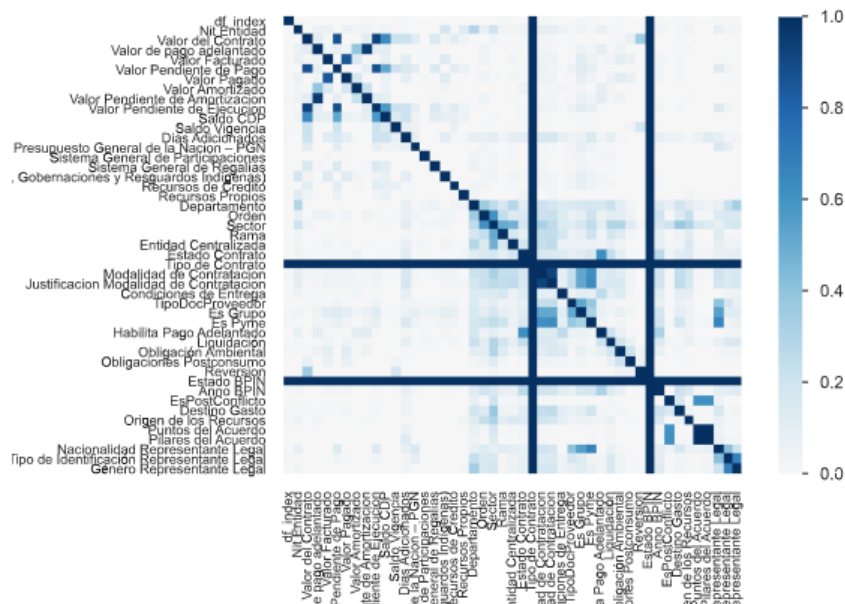
Liquidación	count
0 No	10873
1 Si	9

Sobre valores asociados a rangos o valores específicos, se tienen los siguientes comentarios:

- En muchos campos, la fuente cuenta con el valor “No Definido” cuando carece de otro valor. Ejemplos: Documento Proveedor o Tipo Documento Proveedor.
- Los valores monetarios tienen valores esperados para su mínimo, media y máximo, teniendo en cuenta que estamos analizando únicamente obras de infraestructura.

	Nit Entidad	Valor del Contrato	Valor de pago adelantado	Valor Facturado	Valor Pendiente de Pago	Valor Pagado	Valor Amortizado	Valor Pendiente de Amortización	Valor Pendiente de Ejecución	Saldo CDP
count	1.808700e+04	1.808700e+04	1.808700e+04	1.808700e+04	1.808700e+04	1.808700e+04	1.808700e+04	1.808700e+04	1.808700e+04	1.808700e+04
mean	1.184762e+09	4.343564e+09	3.845038e+07	1.876713e+08	2.920087e+09	1.316329e+08	1.991770e+06	3.645861e+07	2.988094e+09	4.585635e+09
std	1.568508e+09	1.538849e+11	6.666249e+08	1.179148e+09	2.005723e+10	9.263977e+08	5.368137e+07	6.635474e+08	2.006106e+10	1.052116e+11
min	8.000029e+08	0.000000e+00	0.000000e+00	0.000000e+00	-3.000000e+04	0.000000e+00	0.000000e+00	-1.188033e+09	-1.413358e+08	0.000000e+00
25%	8.002158e+08	6.155637e+07	0.000000e+00	0.000000e+00	3.467065e+07	0.000000e+00	0.000000e+00	0.000000e+00	4.964553e+07	3.911849e+07
50%	8.908011e+08	2.456182e+08	0.000000e+00	0.000000e+00	1.794345e+08	0.000000e+00	0.000000e+00	0.000000e+00	2.108806e+08	2.089380e+08
75%	8.999991e+08	1.015671e+09	0.000000e+00	0.000000e+00	9.000000e+08	0.000000e+00	0.000000e+00	0.000000e+00	9.727776e+08	1.104169e+09
max	9.015500e+09	2.041810e+13	5.266143e+10	4.226820e+10	6.300077e+11	3.823300e+10	4.240182e+09	5.266143e+10	6.300077e+11	1.321777e+13

Finalmente, existen correlaciones naturales entre los datos, especialmente cuando se refiere a Valores de contratos que por su propia naturaleza crean la correlación, como el valor pagado y el valor pendiente de pago (relación inversa). Sin embargo, no se decide tampoco remover campos por alta correlación ya que como veremos más adelante, se utilizará un método de selección de variables que es resistente a variables relacionadas.



- SECOP II - Procesos de Contratación

Al igual que con la base de Contratos electrónicos, se busca analizar específicamente los contratos asociados a proyectos de infraestructura, por ello se filtró en la extracción de datos para que solo tomara los tipos de contrato de obra:

```
results = client.get("p6dx-8zbt", where = 'tipo_de_contrato = "Obra"')
```

Con este filtro pasamos de tener una base de más de 2 millones de registros a tener alrededor de 34 mil registros, es decir, se está tomando 1.58% de datos de la base total. Los demás datos no serán tenidos en cuenta.

Ahora, se procede a realizar una inspección o análisis sobre datos faltantes, se revisa para cada variable la cantidad de valores nulos. La siguiente tabla muestra la cantidad de valores nulos para cada una de las variables que tienen campos vacíos.

No	VARIABLE	CANTIDAD	NULOS	% Nulos
1	Fecha_de_Publicacion_Fase_Planeacion_Precalificacion_	34651	1,58%	
2	Fecha_de_Publicacion_Fase_Seleccion_Precalificacion_	34651	1,58%	
3	Fecha_de_Publicacion_Manifestacion_de_Interes_	29682	1,36%	
4	Fecha_de_Publicacion_Fase_Borrador_	25738	1,18%	
5	Fecha_Adjudicacion	20882	0,96%	
6	Fecha_de_Publicacion_Fase_Seleccion_	13936	0,64%	
7	Fecha_de_Apertura_de_Respuesta	6553	0,30%	
8	Fecha_de_Apertura_Efectiva	6325	0,29%	
9	Fecha_de_Recepcion_de_Respuestas	5841	0,27%	
10	Descripci_n_del_Procedimiento	547	0,03%	
11	Nombre_del_Procedimiento	253	0,01%	
12	Fecha_de_Publicacion_del_Proceso	54	0,00%	
13	Fecha_de_Ultima_Publicaci_n	54	0,00%	
14	Nombre del Adjudicador	1	0,00%	

En total de las 57 variables, 14 tienen datos nulos. Sin embargo, al ver el % de nulos de cada variable podemos observar que este es inferior al 1,6%. Esto habla muy bien de la información, pues el nivel de completitud es bastante alto.

Como se puede observar, las primeras 9 variables son las que poseen la mayor cantidad de nulos y sus campos de tipo fecha al igual que las variables 12 y 13. Estos tipos de datos no se realizará imputación de datos puesto que imputar valores en fechas no es nada fácil, puesto que no se conoce a ciencia cierta cual puede ser un valor adecuado.

Los campos 10 Descripción del procedimiento, 11 Nombre del procedimiento y 14 nombre del adjudicador son campos de tipo texto, los cuales tampoco se imputarán datos, pues en el caso de la variable 10 no aporta información relevante, de hecho, se puede prescindir de esta variable.

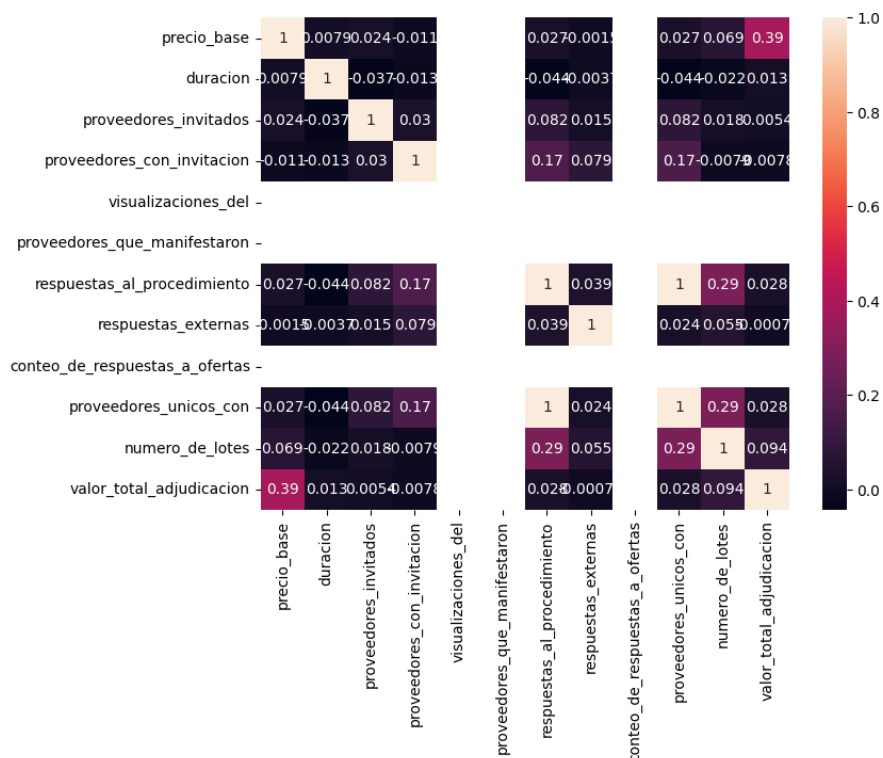
A continuación, se muestran las estadísticas descriptivas sobre los campos numéricos:

	precio_base	duracion	proveedores_invitados	proveedores_con_invitacion	visualizaciones_del	proveedores_que_manifestaron	respuestas_al_procedimiento	respuestas_externas	conteo_de_respuestas_a_ofertas	proveedores_unicos_con	numero_de_lotes	valor_total_adjudicacion
count	3.489600e+04	34896.000000	34896.000000	34896.000000	34896.0	34896.0	34896.000000	34896.000000	34896.0	34896.000000	34896.000000	3.489600e+04
mean	6.385117e+09	36.643799	85.413343	5.251230	0.0	0.0	4.944779	0.013440	0.0	4.931224	0.416925	4.186227e+09
std	1.210529e+11	88.050536	135.351213	22.685905	0.0	0.0	15.006337	0.233754	0.0	14.998770	2.237351	1.072132e+11
min	-3.971665e+08	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.000000	0.000000	0.000000e+00
25%	9.300000e+07	3.000000	10.000000	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.000000	0.000000	0.000000e+00
50%	3.000000e+08	7.000000	51.000000	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.000000	0.000000	0.000000e+00
75%	1.000000e+09	35.000000	121.000000	0.000000	0.0	0.0	3.000000	0.000000	0.0	3.000000	0.000000	1.178246e+08
max	1.321229e+13	3600.000000	5180.000000	389.000000	0.0	0.0	254.000000	17.000000	0.0	254.000000	38.000000	1.328415e+13

Observaciones:

- Los valores monetarios (precio base y valor adjudicación) tienen valores esperados para su mínimo, media y máximo, teniendo en cuenta que estamos analizando únicamente obras de infraestructura.
- El campo duración es coherente con respecto a su mínimo, máximo y media.
- Los campos ‘visualizaciones_del’, ‘proveedores_que_manifestaron’ y ‘conteo_de_respuesta_a_ofertas’ vienen en su totalidad con 0.

Ahora, se muestran las correlaciones de los datos. Como se puede observar en la siguiente figura, entre más negro el color de mapa, menor es la correlación, las casillas blancas es porque no encuentra relación entre las variables. En general la correlación entre las variables numéricas es muy baja. Por tanto, todas las variables son candidatas a ser usadas en el modelo. No habrá eliminación de variables por este concepto



3. Identificación de técnicas para un primer entendimiento de los datos

Dentro de los procedimientos de análisis de datos realizados hasta ahora se tiene filtrado y limpieza de datos, imputación de datos, explicación de estadísticas básicas, y elección de bases adicionales que aportan al entendimiento de la pregunta de negocio, tal y como se describió en los puntos 1 y 2.

En las técnicas subsiguientes tenemos por implementar:

a. Cruce de Bases de Datos

Actualmente contamos con las bases del SECOP II procesos de contratación, contratos electrónicos y multas y sanciones. Hemos realizado el cruce de dichas bases mediante algunas llaves identificadas como los id de los contratos. Sin embargo, algunos de los cruces resultantes están generando algunos duplicados de datos. Para ellos, hemos realizado algunos cruces temporales que nos permitan relacionar las 3 bases de datos.

Sin embargo, seguimos estudiando los datos a fondo para tener un mejor entendimiento de las bases y así poder identificar las llaves adecuadas que nos permitan relacionarlas correctamente sin generar duplicados que podrían afectar el comportamiento de los modelos.

b. Elección de la Variable de respuesta

Para responder a la pregunta de negocio y de analytics, donde lo que se quiere es poder predecir el éxito de un proyecto, tomaremos como variable de respuesta Y si el algunos de los proyectos han tenido o no multas y sanciones.

Como estaos tomando información de los tipos de contratos de obras, estamos cruzándolos con la base de multas. Pocos contratos (80) han tenido multas. Dado que son pocos los contratos que tienes esta característica, debemos pensar en un rebalanceo al momento de aplicar alguna técnica de ML.

c. Elección de Features o variables importantes

Para la selección de variables, se aplicará en un primer paso una función que permita encontrar aquellas variables que podrían explicar el comportamiento de la variable de respuesta. Para ello utilizaremos *mutual_info_regresion*, la cual mide la dependencia entre variables y se dejan aquellas que tienen una menor dependencia entre variables.

d. Reducción de la Dimensionalidad

Hecha la elección de las variables, se pretende también aplicar algunos métodos de reducción de dimensionalidad como Análisis de componentes Principales (PCA) o Aproximación y Proyección de Colector Uniforme (UMAP) donde se dejan aquellas variables o componentes que expliquen al menos un 95% de la varianza, para luego si aplicar los modelos de machine learning (ML).

e. Elección de los algoritmos a aplicar

Hasta el momento solo se ha analizado la información del SECOP II en relación a los proyectos de obras públicas. Aunque ya se tiene la pregunta de negocio, no se han definido los modelos de ML a utilizar. Dado que el objetivo base es poder identificar aquellos proyectos que resultarían ser exitosos o no, se utilizarán algoritmos de clasificación como arboles de decisión, Naive Bayes, regresión logística, Support Vector Machine, Random Forest, redes neuronales entre otros.

4. Propuesta de arquitectura

Para el desarrollo del proyecto se elaboró una arquitectura tal como se muestra en la siguiente imagen;

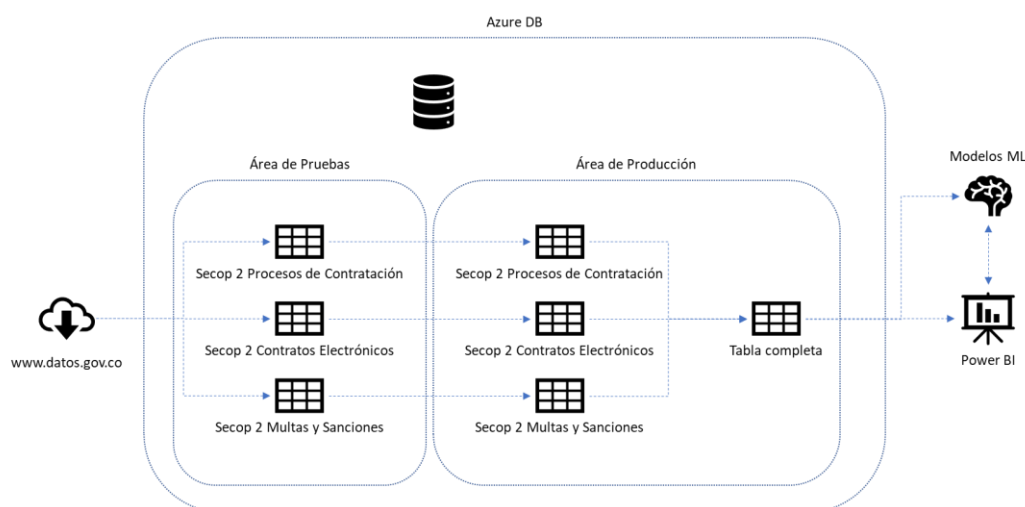


Ilustración 1 - Arquitectura de Software

De lo anterior se propone tener una base de datos desplegada sobre Azure, el cual es un servicio de Azure DB, y dentro de este servicio desplegar dos áreas; una de pruebas y una de producción.

El área de pruebas permite almacenar dentro de nuestro servicio los datos provenientes de la fuente, manteniendo la estructura y la forma en la que vienen dichos datos.

Por otro lado, el área de producción lo que hace es almacenar los datos a los cuales ya se les ha realizado un respectivo trabajo de limpieza y transformación necesaria para poder dar respuesta al objetivo del proyecto. Finalmente, dentro de esta misma área tenemos una unión entre las diferentes fuentes limpias, el cuál es el insumo a utilizar para desplegar las etapas de visualización y de ML en nuestro proyecto.

5. Anexos

El equipo ha creado el siguiente repositorio público en el cual ha consolidado todos los entregables documentales, así como los técnicos, de tal forma que puedan ser explorados por cualquier lector:
https://github.com/jsantacruz/ProyectoDeGradoMIAD_o3b