Facultad de Ingeniería

Análisis de datos

Camila Berreta, María Paz Díaz,
Bruno Ramos, Martin Sena, Rodrigo Sotelo

Práctico 1

Docente: Juan Carlos Pellegrini

Montevideo, 31/08/2024

Índice:

Objetivo

Analizar y manipular un set de datos, que contiene información acerca de los permisos de construcción emitidos en la ciudad de San Francisco, Estados Unidos, para posteriormente realizar conclusiones sobre las inconsistencias encontradas en el set original.

Explicación de las variables del set de datos

- *Permit Number* - Número de Permiso (int): Número asignado cuando se quiere un permiso.
- *Permit Type* - Tipo de Permiso (int): Tipo de permiso representado numéricamente.
- *Permit Type Definition* - Definición del Tipo de Permiso (str): Descripción del tipo de permiso, por ejemplo: nueva construcción, modificaciones.
- *Permit Creation Date* - Fecha de Creación del Permiso (date): Fecha de cuando se creó el permiso, debe ser igual o posterior a la fecha de presentación.
- *Block* - Manzana (str): Relacionado a la dirección.
- *Lot* - Lote (str): Relacionado a la dirección.
- *Street Number* - Número de Puerta (str): Relacionado a la dirección.
- *Street Number Suffix* - Sufijo del Número de Puerta (str): Relacionado a la dirección.
- *Street Name* - Calle (str): Relacionado a la dirección.
- *Street Suffix* - Sufijo de la Calle (str): Relacionado a la dirección.
- *Unit* - Unidad (int): Unidad de un edificio.
- *Unit Suffix* - Sufijo de la Unidad (str): Sufijo de la unidad, si la hay.
- *Description* - Descripción (str): Detalles sobre el propósito del permiso. Por ejemplo: renovación en el baño, etc.
- *Current Status* - Estado Actual (str): Estado actual de la solicitud del permiso.
- *Current Status Date* - Fecha del Estado Actual (date): Fecha en la que ingresó el estado actual.

- *Filed Date* - Fecha de Presentación (str): Fecha en la que se presentó la solicitud del permiso.
- *Issued Date* - Fecha de Emisión (str): Fecha en la que se emitió el permiso.
- *Completed Date* - Fecha de Finalización (str): Fecha en la que se terminó el proyecto, se aplica si el Estado Actual = "completado".
- *First Construction Document Date* - Fecha del Primer Documento de Construcción (date): Fecha de cuando se documentó la construcción por primera vez.
- *Structural Notification* - Notificación Estructural (str): Notificación estructural para cumplir con algún requisito legal.
- *Number of Existing Stories* - Número de Pisos Existentes (int): Cantidad de pisos existentes en un edificio, no aplica a cierto tipo de edificios.
- *Number of Proposed Stories* - Número de Pisos Propuestos (int): Cantidad de pisos propuestos para la construcción o alteraciones.
- *Voluntary Soft-Story Retrofit* - Modificación Voluntaria de Piso Blando (str): Modificación voluntaria de pisos blandos para cumplir con regulaciones sísmicas.
- *Fire Only Permit* - Permiso Solo de Incendios (str): Permiso relacionado con la prevención de riesgos de incendios.
- *Permit Expiration Date* - Fecha de Vencimiento del Permiso (date): Fecha de vencimiento del permiso emitido.
- *Estimated Cost* - Costo Estimado (float): Estimación inicial del costo del proyecto.
- *Revised Cost* - Costo Revisado (float): Estimación revisada del costo del proyecto.
- *Existing Use* - Uso Existente (str): Uso existente del edificio.
- *Existing Units* - Unidades Existentes (int): Cantidad de unidades existentes.
- *Proposed Use* - Uso Propuesto (str): Uso propuesto del edificio.
- *Proposed Units* - Unidades Propuestas (int): Cantidad de unidades propuestas.
- *Plansets* - Conjuntos de Planos (int): Representación del plano que indica la intención general del diseño de la fundación.

- *TIDF Compliance* - Cumplimiento de TIDF (str): Nuevo requerimiento legal.
- *Existing Construction Type* - Tipo de Construcción Existente (int): Tipo de construcción existente, representado numéricamente.
- *Existing Construction Type Description* - Descripción del Tipo de Construcción Existente (str): Descripción del tipo de construcción, por ejemplo, madera u otros tipos de construcción.
- *Proposed Construction Type* - Tipo de Construcción Propuesto (int): Tipo de construcción propuesto, representado numéricamente.
- *Proposed Construction Type Description* - Descripción del Tipo de Construcción Propuesto (str): Descripción del tipo de construcción propuesto.
- *Site Permit* - Permiso del Sitio (str): Permiso para el sitio.
- *Supervisor District* - Distrito del Supervisor (str): Distrito al cual pertenece el edificio.
- *Neighborhoods - Analysis Boundaries* - Barrios (str): Barrio al cual pertenece el edificio.
- *Zipcode* - Código Postal (int): Código postal donde se encuentra el edificio.
- *Location* - Ubicación (float): Ubicación con las coordenadas de la latitud y la longitud.
- *Record ID* - ID de Registro (int): Un identificador que no es útil para este análisis.

Análisis exploratorio de los datos

A. Cantidad de filas y columnas

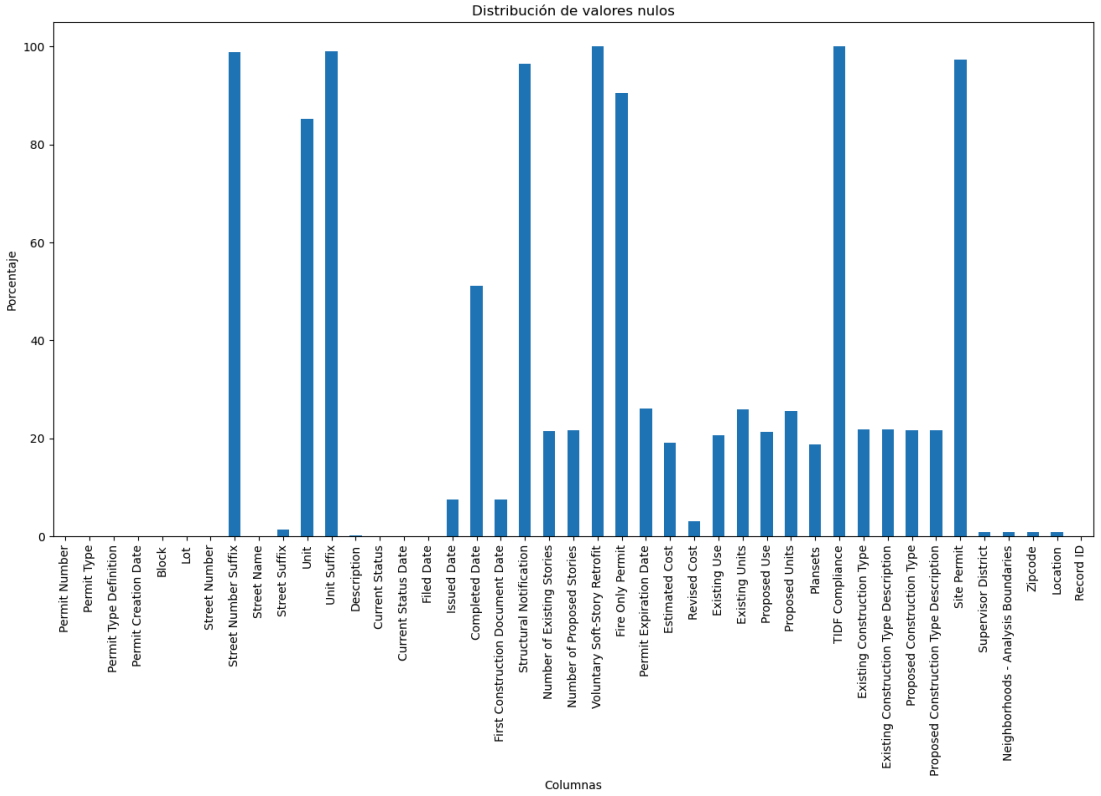
La cantidad de filas es 198910 y la cantidad de columnas es 43.

B. Primeras filas (ejemplo)

	Permit Number	Permit Type	Permit Type Definition	Permit Creation Date	Block	Lot	Street Number	Street Number Suffix	Street Name	Street Suffix	...	Existing Construction Type	Existing Construction Type Description	Proposed Construction Type	Proposed Construction Type Description	F
0	M788927	8	otc alterations permit	05/23/2017	0215	001	1333	NaN	jOnEs	St	...	NaN	NaN	NaN	NaN	
1	201305318356	8	otc alterations permit	05/31/2013	1810	017A	1483	NaN	43rD	Av	...	5.0	wood frame (5)	5.0	wood frame (5)	
2	201705106205	8	otc alterations permit	05/10/2017	5700	027	431	NaN	pReNtIsS	St	...	5.0	wood frame (5)	5.0	wood frame (5)	
3	201410279983	8	otc alterations permit	10/27/2014	0661	005	2020	NaN	bUsH	St	...	5.0	wood frame (5)	5.0	wood frame (5)	
4	201310280388	8	otc alterations permit	10/28/2013	3642	051A	871	NaN	cApP	St	...	5.0	wood frame (5)	5.0	wood frame (5)	

C. Evaluar la existencia de datos faltantes y duplicados

El histograma a continuación muestra los porcentajes de datos faltantes para cada columna.



Para calcular los valores duplicados se tomaron en cuenta cada una de las filas en su totalidad. Luego del análisis no se encontraron filas duplicadas.

D. Para los datos faltantes, evaluar posibles motivos de esto en cada caso

Street Number Suffix (98.8%): Es razonable que dé este valor debido a que en la ciudad de San Francisco, más del 80% de las propiedades no tienen un sufijo en el número de puerta[1].

*Street Suffix (*1.39%*):* Al igual que con los sufijos de números de puerta, no todas las calles tienen un sufijo. Los sufijos de las calles pueden ser menos comunes y solo aparecer en ciertos tipos de direcciones o áreas.

Unit (85.1%)*:* Este campo generalmente se refiere a un número de unidades dentro de una estructura mayor, por ejemplo, un edificio de apartamentos. Si el proyecto es una casa o similar, este campo puede quedar nulo.

[1] https://data.sfgov.org/Geographic-Locations-and-Boundaries/Street-Names/6d9h-4u5v/data_preview

Unit Suffix (99.0%)*:* El mismo se utiliza para especificar un sufijo adicional que pueda ser necesario para identificar con precisión una unidad, como, por ejemplo, "A", "B", "C", etc. Si no hay un sufijo asociado a la unidad, este campo puede quedar nulo. También, podría ser nulo si el proyecto no involucra unidades que requieren un sufijo adicional para su identificación.

*Issued Date (*7.5%*), Completed Date* (51.1%)*, First Construction Document Date* (7.5%)*:* Las fechas pueden faltar por varias razones. Por ejemplo, si el permiso aún está en proceso o no se ha completado el trabajo, las fechas de finalización o los documentos de construcción iniciales pueden no estar disponibles aún.

Structural Notification (96.5%)*:* Asumimos que el tipo de notificación es por algún elemento erróneo que hay en la construcción. Estos elementos erróneos pueden deberse a legislaciones relacionadas con construcciones anti-terremotos que hay en San Francisco dada la falla de San Andrés.

Number of Existing Stories (21.5%)*:* En la metadata aclara que este campo no aplica para todas las construcciones, por lo tanto, es lógico que falten datos.

Number of Proposed Stories (21.5%)*:* No se proponen nuevos pisos en el proyecto. El permiso puede estar relacionado con una remodelación interna o una actualización que no afecta la estructura del edificio.

Voluntary Soft-Story Retrofit (99.9%)*:* Este campo se refiere a un tipo específico de refuerzo estructural. Si el proyecto no implica este tipo de refuerzo, el campo quedará nulo.

Fire-Only Permit (90.5%): Aplica a permisos específicos relacionados con sistemas de seguridad contra incendios. Si el proyecto no incluye instalaciones de este tipo, no se necesita este permiso y el campo estará nulo.

Permit Expiration Date (26.1%): La fecha de expiración puede no estar definida aún si el permiso está en proceso de aprobación o si se trata eventualmente de un permiso permanente.

Estimated Cost (19.1%): El costo estimado puede no estar disponible en las primeras etapas de planificación o si el proyecto está aún en revisión.

Revised Cost (3.0%): Si no ha habido revisiones o ajustes al costo original, este campo puede permanecer nulo.

Existing Use (20.7%): Este campo puede estar nulo si el uso actual no se ha documentado aún o si el proyecto involucra un terreno vacío o una estructura sin uso definido.

Existing Units (25.9%): En caso de que no existan unidades específicas en el sitio actual, este campo puede quedar nulo.

Proposed Use (21.3%): Si el uso propuesto aún no está decidido o es el mismo que el uso existente, este campo puede estar vacío.

Proposed Units (25.6%): Si no se proponen nuevas unidades, el mismo puede quedar nulo.

Plansets (18.7%): Si los planos aún no han sido presentados o no son requeridos en esta etapa, el campo estará vacío.

TIDF Compliance (99.9%): Si el proyecto no está sujeto a este tipo de regulación o si el cumplimiento aún no ha sido determinado, este campo puede ser nulo.

Existing Construction Type, Existing Construction Type Description, Proposed Construction Type, *Proposed Construction Type Description (~21.7%):* Estos campos pueden estar nulos si aún no se ha detallado la información que llevan.

Site Permit (97.3%): Si no se ha solicitado o emitido un permiso específico para el sitio, este campo puede estar vacío.

Supervisor District (0.86%): El campo podría estar nulo si la información geográfica no se encuentra registrada o si el proyecto no está vinculado a un distrito específico.

Neighborhoods - Analysis Boundaries (0.86%): Puede estar vacío si el análisis del barrio no se ha realizado o no es relevante para el proyecto.

Zipcode (0.86%): Si el código postal no ha sido ingresado o si la ubicación es incierta, puede estar vacío.

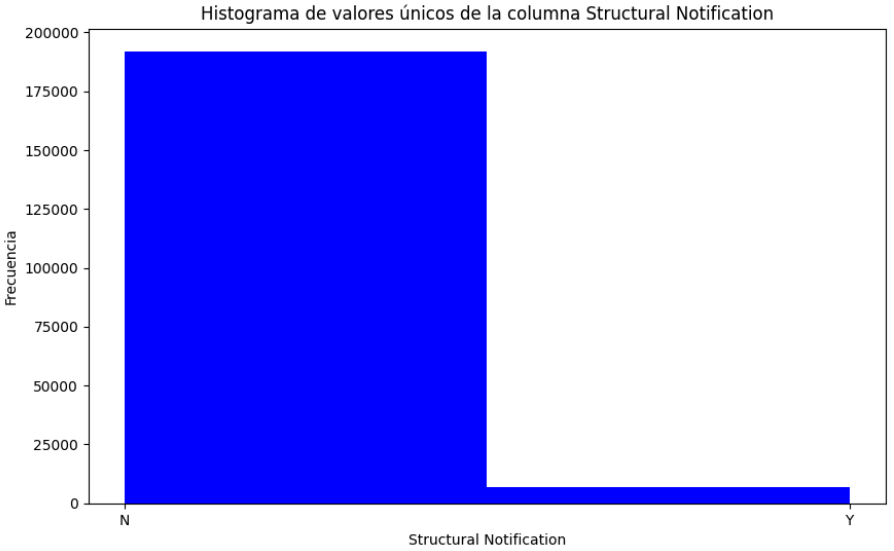
Location (0.85%): Este campo puede estar nulo si la ubicación exacta del proyecto aún no se ha determinado o ingresado.

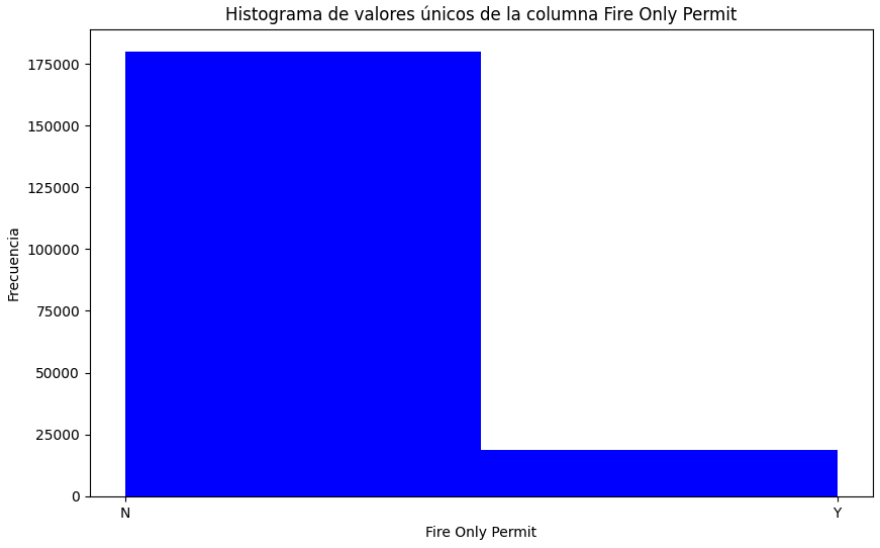
E. Variables discretas

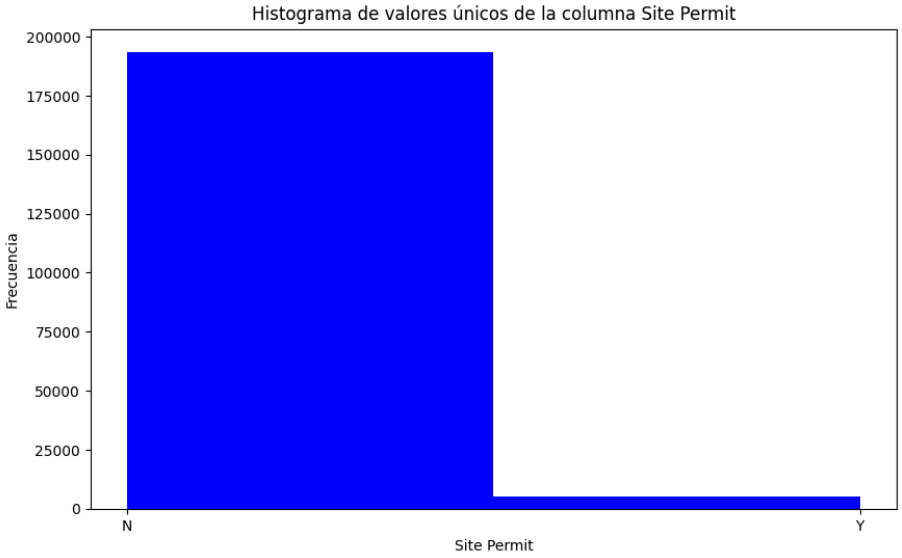
Lo primero que hicimos fue ver cuáles columnas tenían menos de 10 valores únicos. A partir de ese resultado, tomamos aquellas columnas cuyos valores eran “etiquetas”. Las columnas que tomamos y evaluamos son: *“Structural Notification”, “Voluntary Soft-Story Retrofit”, “Fire Only Permit”, “Existing Construction Type Description”, “Proposed Construction Type Description”, “TIDF Compliance” y “Site Permit”.*

F. Valores de las variables discretas

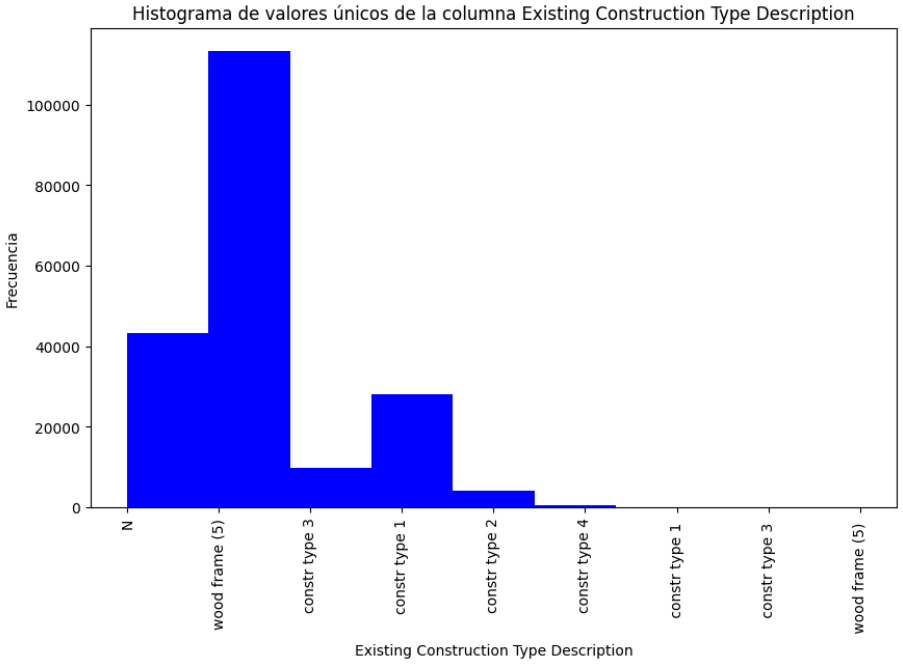
La columna *“Structural Notification”* tiene dos valores únicos que son *“N”* e *“Y”*. Asimismo sucede con las columnas *“Voluntary Soft-Story Retrofit”, “Fire Only Permit”* y *“Site Permit”.* A continuación, podemos ver los histogramas correspondientes.



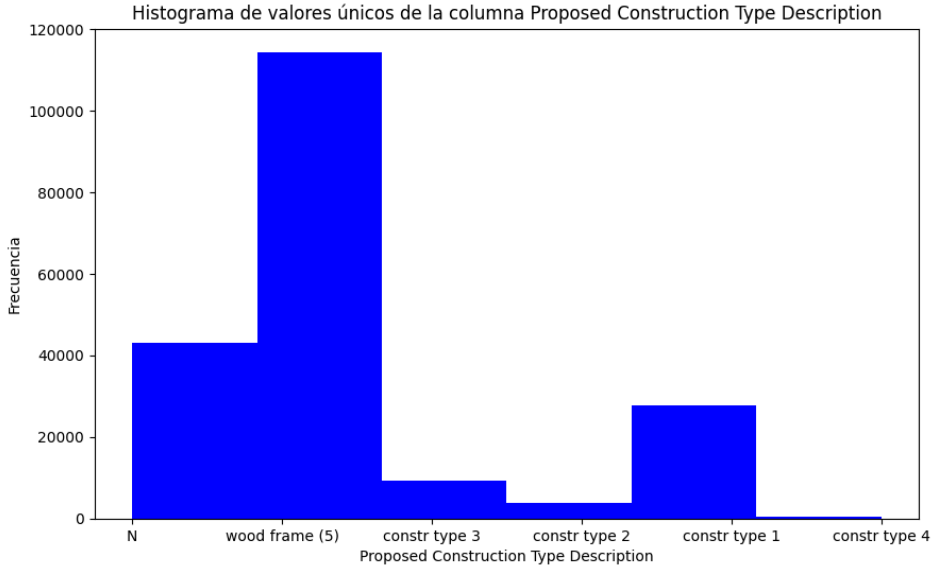




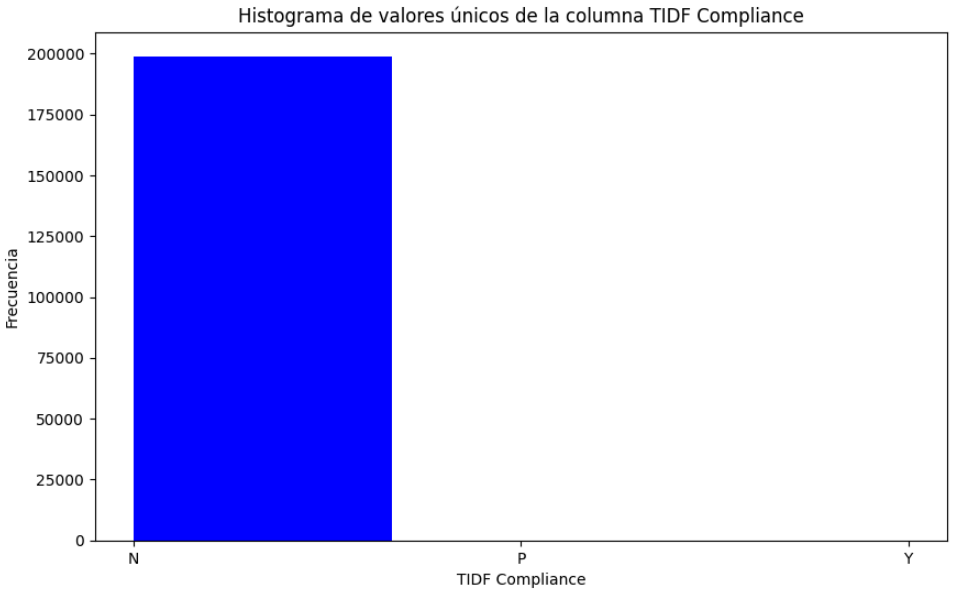
Por otra parte, la columna *“Existing Construction Type Description”* tiene 9 valores distintos que son: *“N”, “wood frame (5)”, “constr type 3”, “constr type 1”, “constr type 2”, “constr type 4”, “constr type 1”, “constr type 3”* y *“wood frame (5)”.* Consideramos que, si bien es una cantidad mayor de valores únicos comparado a las columnas anteriores, los valores son acotados.



De la misma forma sucede con *“Proposed Construction Type Description”* que tiene 6 valores únicos que son *“N”, “wood frame (5)”, “constr type 3”, “constr type 2”, “constr type 1”* y *“constr type 4”.* La siguiente imágen muestra el histograma de los valores únicos de esta columna.



Finalmente, la columna *“TIDF Compliance”* tiene 3 valores únicos que son *“N”, “P”* e *“Y”.*



G. Inconsistencias en los datos:

Se encontraron dos inconsistencias principales en los datos, ambas relacionadas con pares de columnas: en cada par, una columna contiene un valor representativo de una definición, mientras que la otra contiene su descripción correspondiente.

La primera inconsistencia es en las columnas *‘Permit Type’* y *‘Permit Type Definition’*. Aquí se muestra todos los posibles valores que se encuentran en el dataset para este par de columnas:

	Permit Type	Permit Type Definition
0	8	otc alterations permit
10	2	new construction wood frame
16	4	sign - erect
20	3	additions alterations or repairs
189	6	demolitions
263	1	new construction
2189	7	wall or painted sign
3092	5	grade or quarry or fill or excavate
4336	8	otc alterations permit
19458	1	new construction
21234	8	otc alterations permit #
77460	1	new construction #
85100	3	additions alterations or repairs

Aquí se puede ver que se encuentran múltiples definiciones para el *'Permit Type*" tipo 1: "*new construction", "new construction #*" y "*new construction "* que contiene espacios al final que lo convierten en un valor distinto. Lo mismo sucede con los valores "*otc alterations permit*" y *"additions alterations or repairs*". Cada valor numérico de la columna "*Permit Type*" debería corresponder con un único valor de "*Permit Type Definition*".

Algo similar sucede con las columnas "*Existing Construction Type*" y "*Existing Construction Type Description*".

	Existing Construction Type	Existing Construction Type Description
0	NaN	NaN
1	5.0	wood frame (5)
8	3.0	constr type 3
14	1.0	constr type 1
91	2.0	constr type 2
437	4.0	constr type 4
60250	1.0	constr type 1
63044	3.0	constr type 3
69554	-99999.0	wood frame (5)
71803	-99999.0	NaN
79768	99999.0	wood frame (5)
85100	5.0	wood frame (5)

En este caso, para los tipos "*1.0*", "*3.0*" y "*5.0*" existe más de una definición a causa de espacios o caracteres vacíos. A su vez, existen valores numéricos como "*-99999.0*" y "*99999.0*" que parecen ser errores de ingreso, no solo por su valor tan

extremo, sino también porque suceden en tres filas únicamente cada uno y se corresponden con una definición existente.

Limpieza de los datos

A. Toma de decisiones

Con respecto a los datos faltantes, debajo se listan todos los datos faltantes, la decisión y el porqué de esta última.

Dato	Decisión	Explicación
Street Number Suffix	Eliminar columnas	La mayoría de las casas de San Francisco no tienen un sufijo numérico.
Unit Suffix		
Structural Notification		
Voluntary Soft-Story Retrofit		
TIDF Compliance		
Site Permit		
Street Suffix	Colocar la moda	Consideramos colocarle el sufijo "st" a las calles que no tienen sufijo, y considerarlas como calles al igual que las demás.
Issued Date	Eliminar aquellas con valores faltantes	Como el porcentaje es relativamente bajo, creemos que no afectará en el análisis.
Completed Date		
First Construction Document Date		
Number of Existing Stories	Dejarlo nulo	Algunas localidades pueden ser casas y no requerir una cantidad de pisos y, en ese caso, tampoco será necesario un número de pisos propuestos.
Number of Proposed Stories		

Fire-Only Permit	Colocar nueva categoría "N"	Como los valores que no son nulos tienen el valor "Y" (Yes), es correcto colocarles a los que no tienen el valor "N" (No), dado que no tiene el permiso Fire-Only.
Permit Expiration Date	Colocar nueva categoría "indefinida"	Existen construcciones que no tienen definida la fecha de expiración del permiso.
Estimated Cost	Dejar los valores nulos	Podría ser que el costo estimado no se haya fijado todavía y, por ende, tampoco que se haya revisado el costo.
Revised Cost		
Unit	Dejar los valores nulos	No se conocen las unidades residenciales que hay en el lugar, y como el campo es numérico, el nulo representaría el desconocimiento de la cantidad.
Existing Units		
Proposed Use	Colocar nueva categoría "No definido"	Existen proyectos en etapas tempranas que no sepan bien el uso que se les va a dar y/o no conocen el uso que tienen ahora.
Existing Use		
Proposed Units	Dejar los valores nulos	Refieren al número de unidades residenciales que se colocarán en la construcción, podría ser que el proyecto esté en una etapa temprana y no se haya decidido el número exacto.
Plansets	Colocar valor 0.0	Refiere al número de la revisión de los planos, asumimos que los valores que sean nulos no se les han hecho revisión o son versiones preliminares, por ende les colocamos el valor 0.0.
Existing Construction Type	Dejar los valores nulos	No se conoce el tipo de construcción existente y/o propuesta para la construcción.
Existing Construction Type Description		
Proposed Construction Type		

Proposed Construction Type Description		
Supervisor District		
Neighborhoods - Analysis Boundaries		
Zipcode		
Location		

Con respecto a las inconsistencias se tomaron las siguientes decisiones:

En la primer inconsistencia, se borraron los caracteres de “#” y espacios para que coincidan las definiciones con la real. A continuación se muestran los valores posibles que tomaron las columnas luego de las alteraciones:

	Permit Type	Permit Type Definition
263	1	new construction
10	2	new construction wood frame
20	3	additions alterations or repairs
16	4	sign - erect
3092	5	grade or quarry or fill or excavate
189	6	demolitions
2189	7	wall or painted sign
0	8	otc alterations permit

En la segunda inconsistencia, se hizo lo mismo para los valores que correspondía. Para las anomalías en las que el valor es “-99999.0” y “99999.0” se cambió para que coincida con la definición que tenían. Se considera que el valor numérico es un error y la definición indica que valor numérico debería tener. Los valores posibles luego de las alteraciones son:

	Existing Construction Type	Existing Construction Type Description
14	1.0	constr type 1
91	2.0	constr type 2
8	3.0	constr type 3
437	4.0	constr type 4
1	5.0	wood frame (5)
0	NaN	NaN

Normalización de los datos

A. Normalización de variables continuas.

Se normalizaron las variables cuantitativas del dataset: "*Estimated Cost*", "*Revised Cost", "Number of Existing Stories"* y *"Number of Proposed Stories".*

B. Ventajas y desventajas de la normalización.

La normalización de datos es el proceso de ajustar las escalas de las variables numéricas en un conjunto de datos a un rango estandarizado. El objetivo es que ninguna variable influya desproporcionadamente en el análisis debido a su escala. Esto tiene algunas ventajas, por ejemplo:

a. Ciertos algoritmos que utilizan estos datos funcionan mejor cuando todas las variables están en la misma escala.

b. Aquellos algoritmos que básicamente dependen del cálculo del gradiente, por ejemplo, redes neuronales, pueden converger más rápidamente.

c. Reduce los problemas de precisión numérica que pueden surgir al trabajar con valores extremadamente grandes o pequeños.

Por otro lado, dentro de las desventajas se tiene que:

a. Si el rango original de datos es muy grande, se pueden generar resultados que sean menos interpretables.

b. Algunos métodos de normalización son muy sensibles a los valores atípicos, los cuales pueden distorsionar significativamente la media y la desviación estándar. Nuestro caso es un poco más robusto ya que

no utilizamos normalización sino que escalamos a un rango específico (-1 y 1).

Reporte de datos

Se utilizó la librería *ydata-profiling* para realizar un reporte de datos automático para hacer una comparativa con las inconsistencias encontradas y un análisis más profundo de los datos.

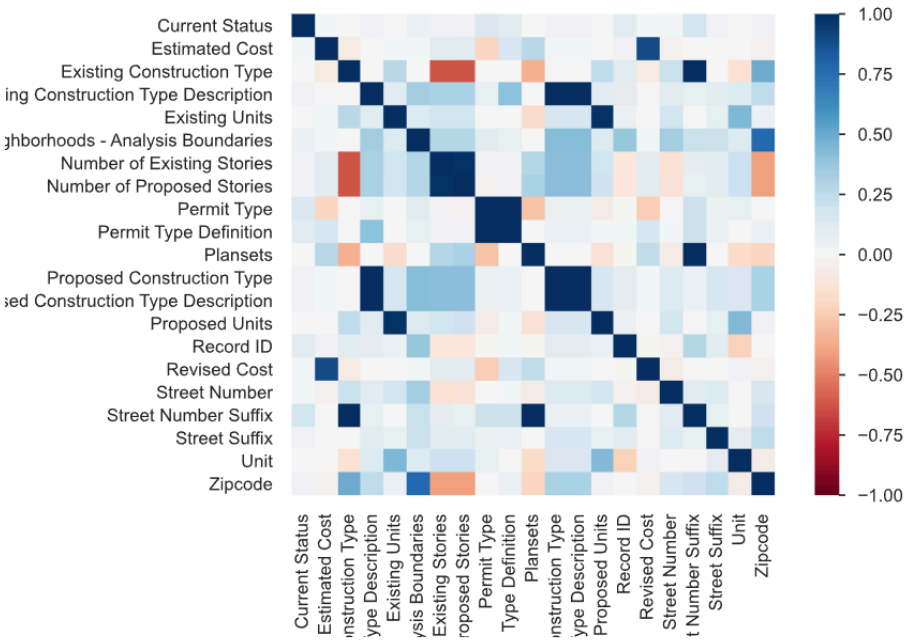
Para la columna '*Permit Type Definition*', por ejemplo, podemos ver los valores similares pero distintos que tenía el dataset originalmente y como estos son menos del 1% de los casos, lo cual indica claramente una anomalía y un error.

Value	Count	Frequency (%)
otc alterations permit	178836	89.9%
additions alterations or repairs	14663	7.4%
sign - erect	2892	1.5%
new construction wood frame	950	0.5%
demolitions	600	0.3%
wall or painted sign	511	0.3%
new construction	347	0.2%
grade or quarry or fill or excavate	91	< 0.1%
otc alterations permit	8	< 0.1%
otc alterations permit #	8	< 0.1%
Other values (3)	4	< 0.1%

Lo mismo sucede para la columna '*Existing Construction Type Description':*

Value	Count	Frequency (%)
wood frame (5)	113350	57.0%
constr type 1	28072	14.1%
constr type 3	9663	4.9%
constr type 2	4068	2.0%
constr type 4	381	0.2%
wood frame (5)	4	< 0.1%
constr type 1	1	< 0.1%
constr type 3	1	< 0.1%
(Missing)	43370	21.8%

Entre las gráficas proporcionadas por el reporte, una de la cual podemos inferir información es la de correlación entre variables. La correlación entre dos variables apunta a la proporcionalidad entre ellos: si es cercana a 1, cuando una variable crece, la otra también; si es cercana a 0, no tienen relación; si es cercana a -1, cuando una variable crece, la otra decrece.



Podemos ver correlaciones esperadas como "*Number of Existing Stories"* y *"Number of Proposed Stories"*, ya que son columnas estrechamente relacionadas. Pero también podemos extraer información que no es evidente. Podemos ver que las columnas "*Number of Existing Stories*" y "*Number of Proposed Stories*" tienen correlación negativa con "*Existing Construction Type*", lo cual indica algún tipo de jerarquía entre los tipos de construcción con la cantidad de pisos; cuanto mayor es el tipo, menor son la cantidad de pisos de la construcción.

Conclusiones

En este análisis de los permisos de construcción en San Francisco, identificamos y corregimos dos tipos principales de inconsistencias en los datos diferencias en las definiciones de tipos de permisos y construcciones, que fueron ajustadas eliminando caracteres especiales y espacios adicionales. También detectamos errores numéricos que se corrigieron para alinearlos con las descripciones correspondientes.

En cuanto a los datos faltantes, tomamos decisiones específicas para cada columna según su relevancia y aplicabilidad, eliminando aquellas con altos porcentajes de datos ausentes y conservando otras importantes.

Finalmente, normalizamos las variables continuas para asegurar un análisis más preciso y mejorar el rendimiento de futuros modelos. Con estas mejoras, el dataset ahora es más coherente y útil para analizar patrones de permisos de construcción en la ciudad.