**Análisis de Datos 2024**

**Practico 1: Obtención, exploración y pretratamiento de datos**



**Integrantes**: Patrick Hobbins, Nicolas Silva, Nicolas Sere, Felipe Ham, Marcos Montenegro



INDICE

1. Carga de Datos

El primer paso fue cargar el dataset “permisos\_construccion.csv” para comenzar a trabajar con él. Este archivo tiene información detallada sobre los permisos de construcción en San Francisco (Estados Unidos), como números de permiso, tipos de permiso, fechas de creación, direcciones, entre otros. A partir de aquí, se comenzó la exploración y pretratamiento de los datos.

2. Resumen de las Variables

Cada columna en el dataset representa una variable que contiene información específica sobre los permisos de construcción.

Algunas de las variables importantes son:

* **Permit Number:** Número único de identificación del permiso.
* Street Number Suffix: los datos vacíos pueden ser porque no se conocen o no apliquen para esta dirección. Esto se dedujo porque hay muchos vacíos.
* Completed Date: los faltantes de esta columna deberían corresponder a las construcciones que aún no finalizaron.
* Structural Notification: hay una gran cantidad de datos faltantes, lo que significa que la gran mayoría de los permisos no tendrían que cumplir o tener en cuenta alguna notificación estructural.
* **Permit Type:** Código que indica el tipo de permiso emitido.
* **Permit Creation Date:** Fecha en que se creó el permiso.
* **Street Number, Street Name, Street Suffix:** Información relacionada con la dirección donde se realiza la construcción.
* **Completed Date:** Fecha en la que se completó el proyecto (si es aplicable).
* **Block, Lot:** Información sobre la parcela donde se lleva a cabo la construcción.

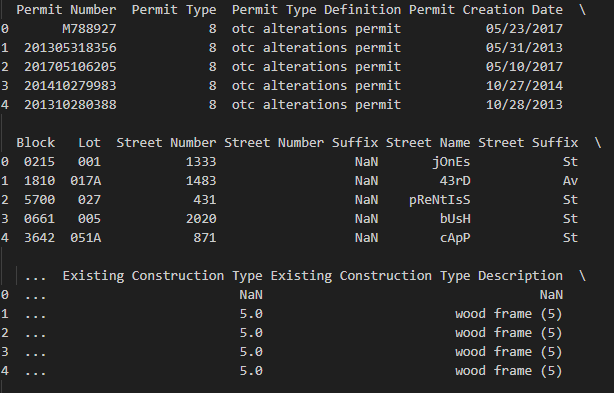
3. Análisis Exploratorio de los Datos

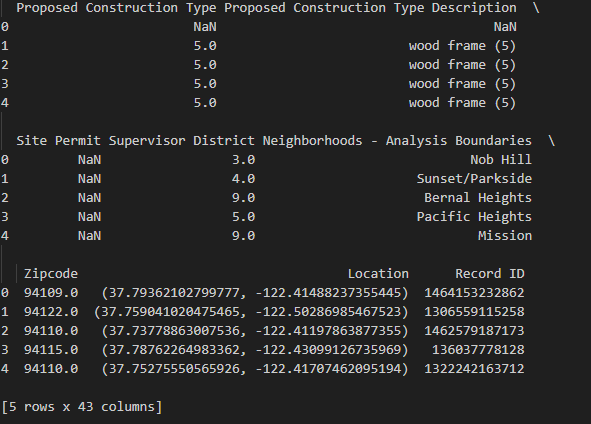
1. **Cantidad de filas y columnas**

El *dataset* está formado por 198910 filas y 43 columnas. Por lo tanto, hay casi 200000 registros divididos en 43 variables.

1. **Observación de las primeras 5 filas**

Luego se obtuvieron las primeras 5 filas para tener una primera vista de la información que tiene cada variable. Además, permite identificar problemas o inconsistencias desde el inicio.





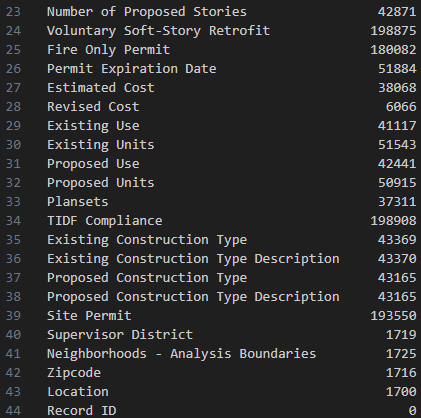
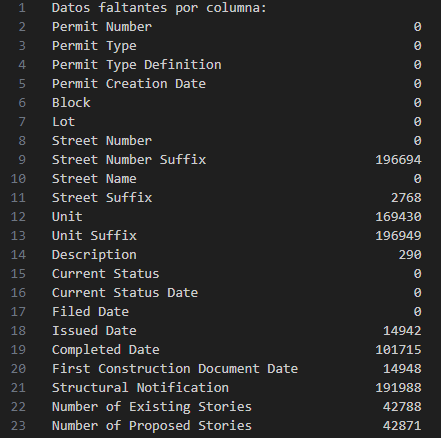
1. **Existencia de datos faltantes y duplicados.**

Durante la exploración, se identificaron datos faltantes en varias columnas:

Las columnas como ‘*Street Number Suffix’, ‘Unit’, y ‘Voluntary Soft-Story Retrofit’* presentan un alto porcentaje de datos faltantes, de los cuales se explicarán posibles motivos en la siguiente parte.

No se encontraron filas duplicadas. Por lo tanto, se puede deducir que cada registro es único.

La cantidad de datos faltantes de cada variable son los siguientes:



1. **Posibles motivos para los datos faltantes.**

Existen varios motivos por los cuales hay datos faltantes, así que a continuación se analizaran los motivos más importantes, en otras palabras, los motivos que abundan más y son más comunes.

Hay varios índices con el mismo motivo, y es que no aplican para este caso, como Street Number Suffix, Street Suffix, Unit Suffix, o Unit.

Mas adelante hay índices como Completed Date, que significa que los faltantes de esta columna deberían corresponder a las construcciones que aún no finalizaron. Luego se encuentran los índices como Structural Notification, lo que quiere decir que hay una gran cantidad de datos faltantes, significando que la gran mayoría de los permisos no tendrían que cumplir o tener en cuenta alguna notificación estructural, o Voluntary Soft-Story Retrofit, que quiere decir que solo unos pocos permisos cuentan con la protección contra terremotos. También hay índices como Fire Only Permit, que quiere decir que son pocos los permisos relacionados a la prevención de incendios. Por último, se encuentra el TIDF Compliance, significando que hay un desconocimiento de este requerimiento legal (por ser nuevo).

1. **Evaluación de valores únicos en variables discretas**

Se evaluaron las variables categóricas. Son aquellas que pueden tomar un conjunto limitado de valores, como el ‘tipo de permiso’ o el ‘sufijo de la calle’. Esto ayuda a comprender la diversidad de valores que hay en cada variable y si algunos valores son más comunes que otros.

1. **Cuantificación de valores únicos y creación de histogramas**

Se cuantificaron los valores únicos de cada variable categórica y se generaron histogramas para poder ver la distribución de estos valores.

Los histogramas son útiles para identificar patrones en los datos, como si algunos tipos de permisos son mucho más comunes que otros.

1. **Evaluación de datos inconsistentes**

4. Limpieza de Datos

5. Normalización de Datos

6. Comparación de inconsistencias