1) Veamos los datos

```
# modulo que usaremos
import pandas as pd
import numpy as np

# Leer en todos nuestros datos
sf_permits = pd.read_csv("../content/Building_Permits.csv")

# set seed for reproducibility
np.random.seed(0)
```

```
<ipython-input-1-d569fab7a32b>:6: DtypeWarning: Columns (22,32) have mixed types. Specify dtype option on import or set low_memory=False.
  sf_permits = pd.read_csv("../content/Building_Permits.csv")
```

```
sf_permits.head()
```

	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	Site Permit	Supervisor District	Neighborhoods - Analysis Boundaries	Zipcode	Location	Record ID
0	201505065519	4	sign - erect	05/06/2015	0326	023	140	NaN	Ellis	St	...	3.0	constr type 3	NaN	NaN	NaN	3.0	Tenderloin	94102.0	(37.785719256680785, -122.40852313194863)	1.380611e+12
1	201604195146	4	sign - erect	04/19/2016	0306	007	440	NaN	Geary	St	...	3.0	constr type 3	NaN	NaN	NaN	3.0	Tenderloin	94102.0	(37.78733980600732, -122.41063199757738)	1.420164e+12
2	201605278609	3	additions alterations or repairs	05/27/2016	0595	203	1647	NaN	Pacific	Av	...	1.0	constr type 1	1.0	constr type 1	NaN	3.0	Russian Hill	94109.0	(37.7946573324287, -122.42232562979227)	1.424857e+12
3	201611072166	8	otc alterations permit	11/07/2016	0156	011	1230	NaN	Pacific	Av	...	5.0	wood frame (5)	5.0	wood frame (5)	NaN	3.0	Nob Hill	94109.0	(37.79595867909168, -122.41557405519474)	1.443574e+12
4	201611283529	6	demolitions	11/28/2016	0342	001	950	NaN	Market	St	...	3.0	constr type 3	NaN	NaN	NaN	6.0	Tenderloin	94102.0	(37.78315261897309, -122.40950883997789)	1.445482e+11

5 rows × 43 columns

¿El conjunto de datos tiene valores faltantes? La respuesta es si, a continuación mostramos la cantidad de datos faltantes que existen en este set de datos.

```
missing_values_count = sf_permits.isnull().sum()
missing_values_count[0:43]
```

	0
Permit Number	0
Permit Type	0
Permit Type Definition	0
Permit Creation Date	0
Block	0
Lot	0
Street Number	0
Street Number Suffix	69365
Street Name	0
Street Suffix	846
Unit	60212
Unit Suffix	69403
Description	113
Current Status	1
Current Status Date	1
Filed Date	1
Issued Date	3023
Completed Date	28924
First Construction Document Date	3032
Structural Notification	67794
Number of Existing Stories	14348
Number of Proposed Stories	14719
Voluntary Soft-Story Retrofit	70115
Fire Only Permit	63667
Permit Expiration Date	15396
Estimated Cost	12735
Revised Cost	1122
Existing Use	13846
Existing Units	17061
Proposed Use	14525
Proposed Units	17128
Plansets	12487
TIDF Compliance	70131
Existing Construction Type	14545
Existing Construction Type Description	14545
Proposed Construction Type	14838
Proposed Construction Type Description	14838
Site Permit	68533
Supervisor District	642

Neighborhoods - Analysis Boundaries	642
Zipcode	643
Location	637
Record ID	1

dtype: int64

```
total_cells = np.product(sf_permits.shape)
total_missing = missing_values_count.sum()


print(total_cells)
print("Total de datos faltantes: ",total_missing)
```

```
3015676
Total de datos faltantes:  769859
```

2) ¿Qué porcentaje de los valores en el conjunto de datos faltan?

La respuesta es 25.5 %

```
percent_missing = (total_missing/total_cells)*100
print(percent_missing)
```

```
25.52857137172561
```

3) Averiguar por qué faltan los datos

¿Cuáles, si es que faltan, por qué no existen? Los datos de **Street Number Suffix** porque la mayoría de las casas no existe un sufijo de su numero, entonces este atributo no existe.

¿Cuales si es que faltan por qué no se registraron? **Zipcode**, hay filas en las que la mayoría de información de los atributos es nula así que es dificil saber el zipcode de las direcciones.

```
subset_sf_permits = sf_permits.loc[:,'Street Number Suffix':'Zipcode'].head()
subset_sf_permits
```

	Street Number Suffix	Street Name	Street Suffix	Unit	Unit Suffix	Description	Current Status	Current Status Date	Filed Date	Issued Date	...	Plansets	TIDF Compliance	Existing Construction Type	Existing Construction Type Description	Proposed Construction Type	Proposed Construction Type Description	Site Permit	Supervisor District	Neighborhoods - Analysis Boundaries	Zipcode
0	NaN	Ellis	St	NaN	NaN	ground fl facade: to erect illuminated, electr...	expired	12/21/2017	05/06/2015	11/09/2015	...	2.0	NaN	3.0	constr type 3	NaN	NaN	NaN	3.0	Tenderloin	94102.0
1	NaN	Geary	St	0.0	NaN	remove (e) awning and associated signs.	issued	08/03/2017	04/19/2016	08/03/2017	...	2.0	NaN	3.0	constr type 3	NaN	NaN	NaN	3.0	Tenderloin	94102.0
2	NaN	Pacific	Av	NaN	NaN	installation of separating wall	withdrawn	09/26/2017	05/27/2016	NaN	...	2.0	NaN	1.0	constr type 1	1.0	constr type 1	NaN	3.0	Russian Hill	94109.0
3	NaN	Pacific	Av	0.0	NaN	repair dryrot & stucco at front of bldg.	complete	07/24/2017	11/07/2016	07/18/2017	...	2.0	NaN	5.0	wood frame (5)	5.0	wood frame (5)	NaN	3.0	Nob Hill	94109.0
4	NaN	Market	St	NaN	NaN	demolish retail/office/commercial 3-story buil...	issued	12/01/2017	11/28/2016	12/01/2017	...	2.0	NaN	3.0	constr type 3	NaN	NaN	NaN	6.0	Tenderloin	94102.0

```
display(sf_permits)
```

	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	Site Permit	Supervisor District	Neighborhoods - Analysis Boundaries	Zipcode	Location	Record ID
0	201505065519	4	sign - erect	05/06/2015	0326	023	140	NaN	Ellis	St	...	3.0	constr type 3	NaN	NaN	NaN	3.0	Tenderloin	94102.0	(37.785719256680785, -122.40852313194863)	1.380611e+12
1	201604195146	4	sign - erect	04/19/2016	0306	007	440	NaN	Geary	St	...	3.0	constr type 3	NaN	NaN	NaN	3.0	Tenderloin	94102.0	(37.78733980600732, -122.41063199757738)	1.420164e+12
2	201605278609	3	additions alterations or repairs	05/27/2016	0595	203	1647	NaN	Pacific	Av	...	1.0	constr type 1	1.0	constr type 1	NaN	3.0	Russian Hill	94109.0	(37.7946573324287, -122.42232562979227)	1.424857e+12
3	201611072166	8	otc alterations permit	11/07/2016	0156	011	1230	NaN	Pacific	Av	...	5.0	wood frame (5)	5.0	wood frame (5)	NaN	3.0	Nob Hill	94109.0	(37.79595867909168, -122.41557405519474)	1.443574e+12
4	201611283529	6	demolitions	11/28/2016	0342	001	950	NaN	Market	St	...	3.0	constr type 3	NaN	NaN	NaN	6.0	Tenderloin	94102.0	(37.78315261897309, -122.40950883997789)	1.445482e+11
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
70127	201412123647	8	otc alterations permit	12/12/2014	4591C	008	451	NaN	Galvez	Av	...	5.0	wood frame (5)	5.0	wood frame (5)	NaN	10.0	Bayview Hunters Point	94124.0	(37.72859693061181, -122.36704632757674)	1.365062e+12
70128	201412123648	8	otc alterations permit	12/12/2014	1218	014	1056	NaN	Page	St	...	5.0	wood frame (5)	5.0	wood frame (5)	NaN	5.0	Haight Ashbury	94117.0	(37.77235620560034, -122.43828353332393)	1.365063e+12
70129	201412123648	8	otc alterations permit	12/12/2014	1218	014	1058	NaN	Page	St	...	5.0	wood frame (5)	5.0	wood frame (5)	NaN	5.0	Haight Ashbury	94117.0	(37.77235620560034, -122.43828353332393)	1.365064e+12
70130	201412123646	8	otc alterations permit	12/12/2014	6531	030	1442	NaN	Valencia	St	...	5.0	wood frame (5)	5.0	wood frame (5)	NaN	8.0	Mission	94110.0	(37.74973543181233, -122.42073191683869)	1.365065e+12
70131	201412123646	8	otc alterations permit	12/12/2014	6531	030	1446	NaN	Valencia	St	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

70132 rows × 43 columns