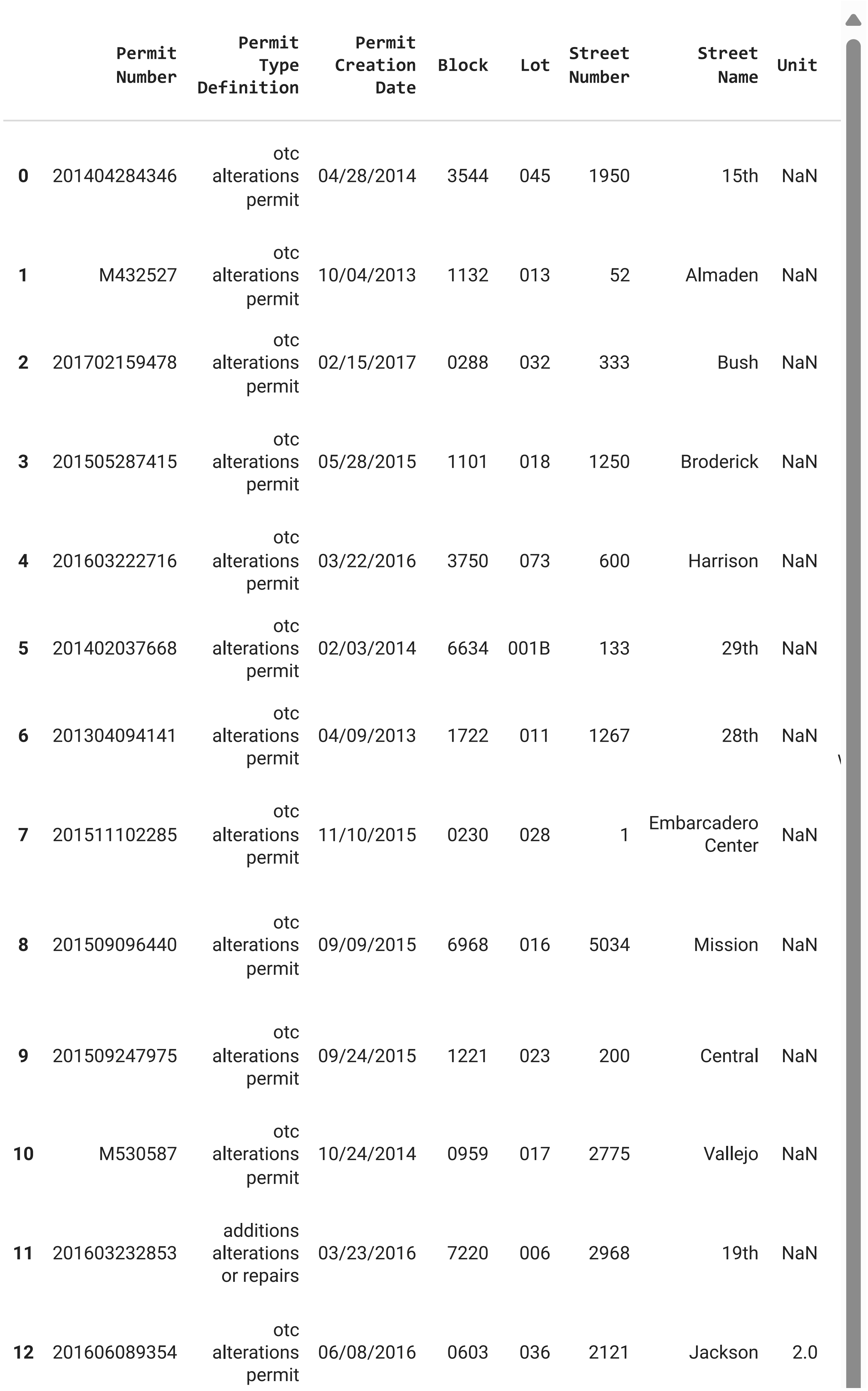
Program 3 : Use regular expressions in Python to clean and preprocess text data effectively. Use the `re` module for pattern matching and text manipulation.

import pandas and Read data

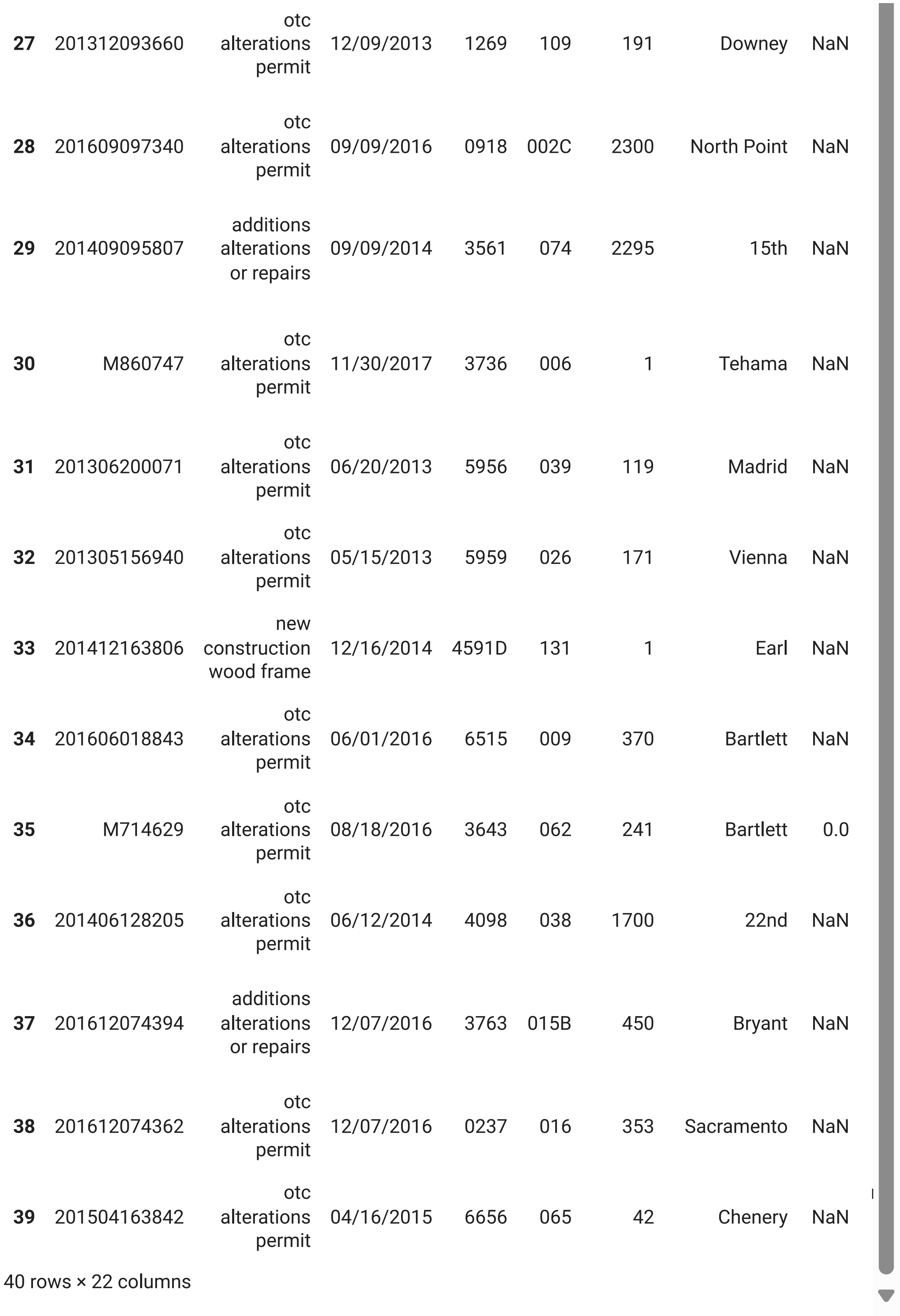
```
import pandas as pd

df = pd.read_csv('build_data.csv')
df.head(40)
```



	Permit Number	Permit Type Definition	Permit Creation Date	Block	Lot	Street Number	Street Name	Unit
0	201404284346	otc alterations permit	04/28/2014	3544	045	1950	15th	NaN
1	M432527	otc alterations permit	10/04/2013	1132	013	52	Almaden	NaN
2	201702159478	otc alterations permit	02/15/2017	0288	032	333	Bush	NaN
3	201505287415	otc alterations permit	05/28/2015	1101	018	1250	Broderick	NaN
4	201603222716	otc alterations permit	03/22/2016	3750	073	600	Harrison	NaN
5	201402037668	otc alterations permit	02/03/2014	6634	001B	133	29th	NaN
6	201304094141	otc alterations permit	04/09/2013	1722	011	1267	28th	NaN
7	201511102285	otc alterations permit	11/10/2015	0230	028	1	Embarcadero Center	NaN
8	201509096440	otc alterations permit	09/09/2015	6968	016	5034	Mission	NaN
9	201509247975	otc alterations permit	09/24/2015	1221	023	200	Central	NaN
10	M530587	otc alterations permit	10/24/2014	0959	017	2775	Vallejo	NaN
11	201603232853	additions alterations or repairs	03/23/2016	7220	006	2968	19th	NaN
12	201606089354	otc alterations permit	06/08/2016	0603	036	2121	Jackson	2.0

13	201408133771	otc alterations permit	08/13/2014	5339	012	1665	Quesada	NaN
14	201611283589	otc alterations permit	11/28/2016	0998	001	3201	Washington	NaN
15	M413307	otc alterations permit	08/01/2013	1237	010	922	Haight	NaN
16	201712075788	demolitions	12/07/2017	2799	042	49	Hopkins	NaN
17	201410169091	otc alterations permit	10/16/2014	1523	032	460	22nd	0.0
18	201402219041	additions alterations or repairs	02/21/2014	3705	042	865	Market	NaN
19	201408284928	otc alterations permit	08/28/2014	3641	077	1269	South Van Ness	0.0
20	201308154362	otc alterations permit	08/15/2013	3780	006	888	Brannan	NaN
21	201607223145	otc alterations permit	07/22/2016	5871	030	282	Trumbull	NaN
22	201501064993	otc alterations permit	01/06/2015	4272	033	3027	25th	NaN
23	201704063407	additions alterations or repairs	04/06/2017	5297	027	1103	Phelps	NaN
24	201706209881	otc alterations permit	06/20/2017	1530	011A	469	14th	NaN
25	M627367	otc alterations permit	10/06/2015	1389	017	286	30th	NaN
26	M800847	otc alterations permit	06/22/2017	3544	067	2075	Market	NaN



27	201312093660	otc alterations permit	12/09/2013	1269	109	191	Downey	NaN
28	201609097340	otc alterations permit	09/09/2016	0918	002C	2300	North Point	NaN
29	201409095807	additions alterations or repairs	09/09/2014	3561	074	2295	15th	NaN
30	M860747	otc alterations permit	11/30/2017	3736	006	1	Tehama	NaN
31	201306200071	otc alterations permit	06/20/2013	5956	039	119	Madrid	NaN
32	201305156940	otc alterations permit	05/15/2013	5959	026	171	Vienna	NaN
33	201412163806	new construction wood frame	12/16/2014	4591D	131	1	Earl	NaN
34	201606018843	otc alterations permit	06/01/2016	6515	009	370	Bartlett	NaN
35	M714629	otc alterations permit	08/18/2016	3643	062	241	Bartlett	0.0
36	201406128205	otc alterations permit	06/12/2014	4098	038	1700	22nd	NaN
37	201612074394	additions alterations or repairs	12/07/2016	3763	015B	450	Bryant	NaN
38	201612074362	otc alterations permit	12/07/2016	0237	016	353	Sacramento	NaN
39	201504163842	otc alterations permit	04/16/2015	6656	065	42	Chenery	NaN

40 rows × 22 columns

find the data types in each fields. find the data type of block column

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2676 entries, 0 to 2675
Data columns (total 22 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   Permit Number               2676 non-null   object 
 1   Permit Type Definition      2676 non-null   object 
 2   Permit Creation Date        2676 non-null   object 
 3   Block                       2676 non-null   object 
 4   Lot                         2676 non-null   object 
 5   Street Number               2676 non-null   int64  
 6   Street Name                 2676 non-null   object 
 7   Unit                        365 non-null    float64
 8   Description                 2674 non-null   object 
 9   Current Status              2676 non-null   object 
 10  Structural Notification     82 non-null     object 
 11  Number of Existing Stories  2078 non-null   float64
 12  Number of Proposed Stories  2075 non-null   float64
 13  Estimated Cost              2138 non-null   float64
 14  Revised Cost                2601 non-null   float64
 15  Existing Use                2104 non-null   object 
 16  Existing Units              1963 non-null   float64
 17  Supervisor District         2654 non-null   float64
 18  Proposed Use                2086 non-null   object 
 19  Proposed Units              1974 non-null   float64
 20  Plansets                    2147 non-null   float64
 21  Zipcode                     2654 non-null   float64
dtypes: float64(10), int64(1), object(11)
memory usage: 460.1+ KB
```

Try to convert object datatype to int. You will get an error

```
df['Block'] = df['Block'].astype(int)
```

Import regular expressions package regex. Change the field to regular expression.

Syntax for re.sub() is re.sub(pattern,replacement,string[, count,flags])

```
import re

block_list = []
for string in df['Block']:
  #use the re.sub function to remove all the non umeric characters of the string.
  normal_string = re.sub('[^0-9]',"",string)
  #store it in updated list called block_list that you have initialised
  block_list.append(normal_string)

#assign the values of updated list called block_list to Block column
df['Block'] = block_list
```

Again try to change the code

```
df['Block'] =df['Block'].astype(int)
```

Again , check if the field is changed to int type

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2676 entries, 0 to 2675
Data columns (total 22 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   Permit Number               2676 non-null   object 
 1   Permit Type Definition      2676 non-null   object 
 2   Permit Creation Date        2676 non-null   object 
 3   Block                       2676 non-null   int64  
 4   Lot                         2676 non-null   object 
 5   Street Number               2676 non-null   int64  
 6   Street Name                 2676 non-null   object 
 7   Unit                        365 non-null    float64
 8   Description                 2674 non-null   object 
 9   Current Status              2676 non-null   object 
 10  Structural Notification     82 non-null     object 
 11  Number of Existing Stories  2078 non-null   float64
 12  Number of Proposed Stories  2075 non-null   float64
 13  Estimated Cost              2138 non-null   float64
 14  Revised Cost                2601 non-null   float64
 15  Existing Use                2104 non-null   object 
 16  Existing Units              1963 non-null   float64
 17  Supervisor District         2654 non-null   float64
 18  Proposed Use                2086 non-null   object 
 19  Proposed Units              1974 non-null   float64
 20  Plansets                    2147 non-null   float64
 21  Zipcode                     2654 non-null   float64
dtypes: float64(10), int64(2), object(10)
memory usage: 460.1+ KB
```