Reemplazar valores con expresiones regulares en un DataFrame de Pandas

Cargar los datos en Pandas

In [1]:

```python
import pandas as pd

df = pd.read_csv(f'..\data/internshala_dataset_raw.csv')
df.head()
```

Out[1]:

	internship	company_name	skills	perks	location	duration	stipend	applicants	ifSkillsorPerksMissingUseThis
0	Software Testing	Times Internet	Software Testing	Certificate, 5 days a week	Noida	6 Months	8000 /month	119 applicants	Software Testing, Certificate\n5 days a week
1	Technical Operations - Networking And Monitoring	Paytm Payments Bank	Java, SQL, Unix, Oracle, MS SQL Server, Hibern...	Certificate, Letter of recommendation, 5 days ...	Noida	6 Months	10000 /month	194 applicants	Java\nSQL\nUnix\nOracle\nMS SQL Server\nHibern...
2	Software Project Management	IIT Bombay	English Proficiency (Spoken), English Proficie...	Certificate, Letter of recommendation, Flexibl...	Work From Home	6 Months	1000-2000 /month	113 applicants	English Proficiency (Spoken)\nEnglish Proficie...
3	Web Development	IIT Bombay	HTML, CSS, Flask, Python, Django	Certificate, Letter of recommendation, Flexibl...	Work From Home	6 Months	1000-2000 /month	183 applicants	HTML\nCSS\nFlask\nPython\nDjango, Certificate\...
4	Front End Development	IIT Bombay	HTML, CSS, JavaScript, ReactJS, Redux	Certificate, Letter of recommendation, Flexibl...	Work From Home	6 Months	1000-2000 /month	205 applicants	HTML\nCSS\nJavaScript\nReactJS\nRedux, Certifi...

Reemplazar los valores de cadena en una columna

1. Reemplazar el valor de una sola cadena con espacios

In [2]:

```python
df['applicants'].str.replace(r'\sapplicants', '', regex=True) # r=raw -> literal
```

Out[2]:

```
0                         119
1                         194
2                         113
3                         183
4                         205
                ...          
2562                       30
2563    Be an early applicant
2564    Be an early applicant
2565    Be an early applicant
2566    Be an early applicant
Name: applicants, Length: 2567, dtype: object
```

2. Reemplazar el valor de varias cadenas con espacios

In [3]:

```python
df['applicants'].str.replace(r'(\sapplicants|Be an early applicant)', '', regex=True)
```

Out[3]:

```
0       119
1       194
2       113
3       183
4       205
       ... 
2562     30
2563       
2564       
2565       
2566       
Name: applicants, Length: 2567, dtype: object
```

3. Reemplazar varios valores de cadena con un cero

In [4]:

```python
df['applicants'].replace([r'(\d+) applicants', 'Be an early applicant'],[r'\1',0], regex=True)
```

Out[4]:

```
0       119
1       194
2       113
3       183
4       205
       ... 
2562     30
2563      0
2564      0
2565      0
2566      0
Name: applicants, Length: 2567, dtype: object
```

Reemplazar caracteres especiales con espacios

In [5]:

```
df['internship'].str.replace(r'[^0-9a-zA-Z:,\s]+', '', regex=True)
```

Out[5]:

```
0                                       Software Testing
1       Technical Operations  Networking And Monitoring
2                             Software Project Management
3                                         Web Development
4                                   Front End Development
                              ...
2562                                Internet Of Things IoT
2563                                      Summer Research
2564                   Academic Research Computer Science
2565                                     Computer Science
2566                                     Computer Science
Name: internship, Length: 2567, dtype: object
```

Reemplazar números o caracteres que no son dígitos

1. Reemplazar todos los símbolos no numéricos y mapea en caso de que falten

In [6]:

```
df['applicants'].str.replace(r'\D+', '', regex=True).replace({'':0}).astype('int')
```

Out[6]:

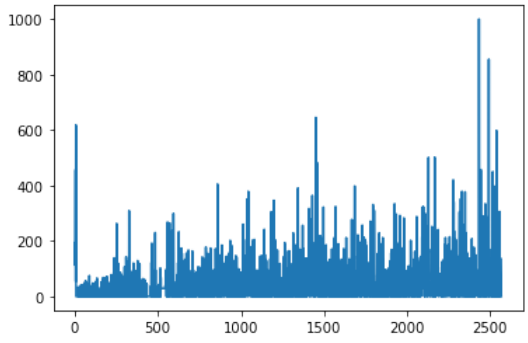
```
0       119
1       194
2       113
3       183
4       205
       ...
2562     30
2563      0
2564      0
2565      0
2566      0
Name: applicants, Length: 2567, dtype: int32
```

In [7]:

```
df['applicants'].str.replace(r'\D+', '', regex=True).replace({'':0}).astype('int').plot()
```

Out[7]:

```
<AxesSubplot:>
```



2. Reemplazar todos los números de la columna

In [8]:

```
df['applicants'].replace(to_replace=r"\d+", value=r" ", regex=True)
```

Out[8]:

```
0                  applicants
1                  applicants
2                  applicants
3                  applicants
4                  applicants
                ...
2562               applicants
2563    Be an early applicant
2564    Be an early applicant
2565    Be an early applicant
2566    Be an early applicant
Name: applicants, Length: 2567, dtype: object
```

Reemplazar valores numéricos en cuatro columnas del DataFrame

In [9]:

```
cols = ['stipend', 'applicants', 'internship', 'duration']
df[cols].replace(to_replace=r"\d+", value=r" ", regex=True)
```

Out[9]:

```
           stipend            applicants                                  internship  duration
   0        /month            applicants                            Software Testing    Months
   1        /month            applicants  Technical Operations - Networking And Monitoring    Months
   2      - /month            applicants                 Software Project Management    Months
   3      - /month            applicants                             Web Development    Months
   4      - /month            applicants                      Front End Development    Months
 ...           ...                   ...                                         ...       ...
2562      - /month            applicants                     Internet Of Things (IoT)    Months
2563        /month  Be an early applicant                             Summer Research    Months
2564  Not provided  Be an early applicant          Academic Research (Computer Science)    Months
2565        /month  Be an early applicant                            Computer Science    Months
2566        /month  Be an early applicant                            Computer Science    Months

2567 rows × 4 columns
```