```
!kaggle datasets download -d ravishah1/us-jobs-on-and-off-site
```

```
!ls -lh *.zip
```

```
-rw-r--r-- 1 root root 1.1G Jan 12 21:42 us-jobs-on-and-off-site.zip
```

```
!unzip -q -o -j us-jobs-on-and-off-site.zip -d .
!rm -f us-jobs-on-and-off-site.zip
!ls -lh *.parquet
```

```
-rw-r--r-- 1 root root 1.1G Jan 12 21:44 jobs_train.parquet
```

```
import pandas as pd
df = pd.read_parquet("jobs_train.parquet")
print(f"Shape: {df.shape[0]:,} rows × {df.shape[1]} columns")
print(f"Memory: {df.memory_usage(deep=True).sum()/1024**2:.1f} MB")
df.info()
```

```
Shape: 1,048,575 rows × 22 columns
Memory: 1,024.0 MB
<class'pandas.core.frame.DataFrame'>
RangeIndex: 1,048,575 entries, 0 to 1,048,574
Data columns (total 22):
 #   Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
 0   job_id                 1,048,575 non-null   int64
 1   title                  1,048,575 non-null   object
 2   location               1,048,575 non-null   object
 3   department             1,048,575 non-null   object
 4   salary_range           1,048,575 non-null   object
 5   company_profile        1,048,575 non-null   object
 6   description            1,048,575 non-null   object
 7   requirements           1,048,575 non-null   object
 8   benefits               1,048,575 non-null   object
 9   telecommuting          1,048,575 non-null   int64
 10  has_company_logo       1,048,575 non-null   int64
 11  has_questions          1,048,575 non-null   int64
 12  employment_type        1,048,575 non-null   object
 13  required_experience    1,048,575 non-null   object
 14  required_education     1,048,575 non-null   object
 15  industry               1,048,575 non-null   object
 16  function               1,048,575 non-null   object
 17  fraudulent             1,048,575 non-null   int64
 18  text                   1,048,575 non-null   object
 19  jobid                  1,048,575 non-null   int64
 20  place_id               1,048,575 non-null   int64
 21  country_code           1,048,575 non-null   object
dtypes: int64(9), object(13)
```

```
df.head(3)
```

	job_id	title	location	department	salary_range	company_profile	description	requirements	benefits	telecommuting	has_company_logo	has_questions	employment_type
0	1	Marketing Intern	US, NY, New York	Marketing		We're Food52, and we've been voted one of the 'Best Places to Work' for a groundbreaking,	We are seeking a Marketing Intern to join our team! A creative, detail-oriented, highly motivated intern,	Experience with content management systems a major plus (any CMS is fine), Ability to juggle multiple projects at once, Ability to		0	1	0	Other
1	2	Customer Service Cloud Video Production	NZ, Auckland,	Success		90 Seconds, the worldwide Cloud Video Production service.	Overview\n\n\tWe're looking for a Customer Service Cloud Video Production Specialist to join our team. This role is a	What we expect from you:\n\nYou have hands-on experience producing high quality video content, and	What you can expect from us:\n\n\tWe offer a competitive salary, and a range of benefits including	0	1	0	Other
2	3	Commissioning Machine Learning Intern	US, NY, Wilton	Data		Vetro Sciences\nVetro Sciences is a company focused on developing novel therapeutics for cancer patients.	Our client, located in Houston, is using machine learning and other advanced analytics methods to optimize clinical trial site	Implement pre-processing and feature extraction modules for handling raw clinical	Our Lively and Vibrant Culture\n\n\tOur culture is alive and well. Our	0	1	0	Other
4	4	Account Executive - Washington DC	US, DC, Washington	Sales		Gen Digital\n\n\tGen Digital (NASDAQ: GEN) is a global company with a portfolio of cybersafety brands	THE COMPANY: ESRI – ENVIRONMENTAL SYSTEMS RESEARCH INSTITUTE\n\n\tEDUCATION: Bachelor's or Master's in GIS,	EDUCATION: Bachelor's or Master's in GIS, GEOGRAPHY, or related field\n\n\tYEARS OF WORK EXPERIENCE: Three	Full Benefits Offered	0	1	0	Full-time
5	5	Bill Review Manager	US, FL, Jacksonville	Health		SynCentrix\n\n\tSynCentrix is a Stamford-based, \n\n\tJOB TITLE: Associate Bill Review Manager\n\n\tJOB	QUALIFICATIONS:\n\n\tBachelor's degree in Computer Science, Information	QUALIFICATIONS:\n\n\tBachelor's degree in Computer Science, Information	Full Benefits Offered	0	1	0	Full-time

```
df.describe()
```

	job_id	telecommuting	has_company_logo	has_questions	salary_range	employment_type
count	1,048,575.0	1,048,575.0	1,048,575.0	1,048,575.0	1,048,575.0	1,048,575.0
mean	524,288.0	0.0	0.8	0.5	0.0	0.0
std	302,754.4	0.1	0.4	0.5	0.0	0.0
min	1.0	0.0	0.0	0.0	0.0	0.0
25%	262,144.0	0.0	0.0	0.0	0.0	0.0
50%	524,288.0	0.0	1.0	1.0	0.0	0.0
75%	786,431.0	0.0	1.0	1.0	0.0	0.0
max	1,048,575.0	1.0	1.0	1.0	0.0	0.0

```
df.shape, df.dtypes
```

```
(1,048,575, 22)
```

```
df.dtypes[df.dtypes.isin(['object', 'int64', 'bool'])]
```

```
job_id                 int64
title                 object
location              object
department            object
salary_range          object
company_profile       object
description           object
requirements          object
benefits              object
telecommuting          int64
has_company_logo       int64
has_questions          int64
employment_type       object
required_experience   object
required_education    object
industry              object
function              object
fraudulent             int64
text                  object
jobid                  int64
place_id               int64
country_code          object
dtype: int64
```

```
df.sample(3)[['telecommuting', 'has_company_logo', 'has_questions', 'salary_range', 'employment_type']]
```

	telecommuting	has_company_logo	has_questions	salary_range	employment_type
0	0	1	0	0	Other
1	0	1	0	0	Other
2	0	1	0	0	Other

```
# Fraud and Real distribution
ax = df['fraudulent'].value_counts().sort_index().plot(kind='bar')
ax.set_xlabel('fraudulent'); ax.set_ylabel('Count')
ax.bar_label(ax.containers[0])
```

```
/usr/local/lib/python3.12/dist-packages/matplotlib/cbar.py:1462: FutureWarning: Passing the following variables as a keyword a
Mpl3.9.0 deprecation. In the future, this will raise an error. Pass the value of the parameter to the relevant keyword arg(s) instead or use a
  cbar = fig.colorbar(sm, ax=ax, **kw)
/usr/local/lib/python3.12/dist-packages/matplotlib/axes/_axes.py:2.9.0 deprecation. In the future, this will raise an error. Pass the value of the parameter to the relevant keyword arg(s) instead or use a
  ax.bar_label(p, fmt=fmt, **kwargs)
```

```
df['fraudulent'] = df['fraudulent'].astype(int)
df['fraudulent'].value_counts()
```

```
fraudulent
0    1,047,716
1        859
Name: count, dtype: int64
```

```
# Aggregation job postings by countries
df = df[df['location'].\
        ...\
        ...\
        ...
```