```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
df = pd.read_csv('/content/drive/MyDrive/Dataset ML LAB 205122021/train.csv')
df.shape
```

```
(404290, 6)
```

```
df.sample(10)
```

	id	qid1	qid2	question1	question2	is_duplicate
399275	399275	532524	35705	FOX against Clinton?	Is Fox News biased against Hillary Clinton?	1
40216	40216	72783	72784	What is dating like in Germany?	What is German dating culture like?	1
41743	41743	75321	75322	What is the reaction between magnesium and hyd...	What type of reaction is created by mixing hyd...	1
348886	348886	225768	477497	How can we stop thinking negatively about others?	What are the technique to stop negative thinki...	1
217343	217343	34543	5298	How can I manage my anger?	How do I control my emotions and anger?	1
403375	403375	89437	20491	What is better to use: MIT OCW or Khan Academy?	Is it better to learn calculus from Khan acade...	1
331137	331137	407476	416384	What is the difference between DRAM, SRAM and ...	What are CISC and RISC architecture? How do th...	0
17712	17712	27923	33614	If I delete WhatsApp, will the messages sent t...	If I delete my WhatsApp account, will that can...	0

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 404290 entries, 0 to 404289
Data columns (total 6 columns):
 #   Column        Non-Null Count   Dtype 
---  ------        --------------   ----- 
 0   id            404290 non-null  int64 
 1   qid1          404290 non-null  int64 
 2   qid2          404290 non-null  int64 
 3   question1     404289 non-null  object
 4   question2     404288 non-null  object
 5   is_duplicate  404290 non-null  int64 
dtypes: int64(4), object(2)
memory usage: 18.5+ MB
```

```
# missing values
df.isnull().sum()
```

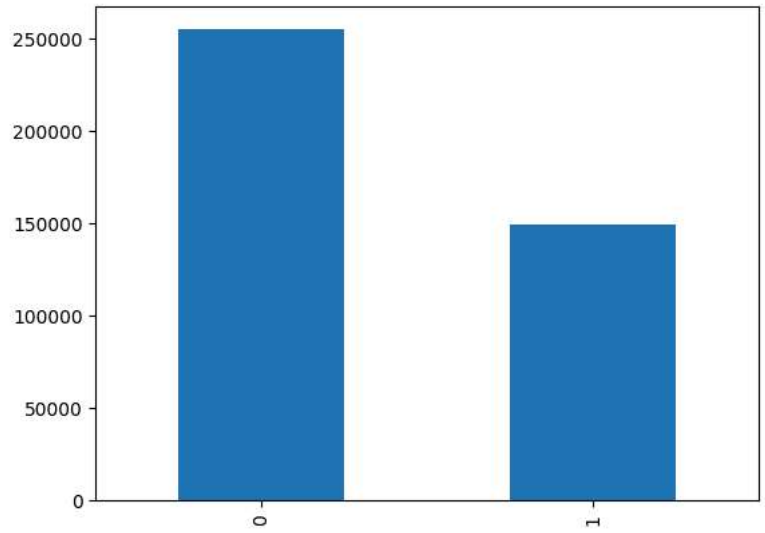
```
id              0
qid1            0
qid2            0
question1       1
question2       2
is_duplicate    0
dtype: int64
```

```
# duplicate rows
df.duplicated().sum()
```

```
0
```

```
# Distribution of duplicate and non-duplicate questions

print(df['is_duplicate'].value_counts())
print((df['is_duplicate'].value_counts()/df['is_duplicate'].count())*100)
df['is_duplicate'].value_counts().plot(kind='bar')
```

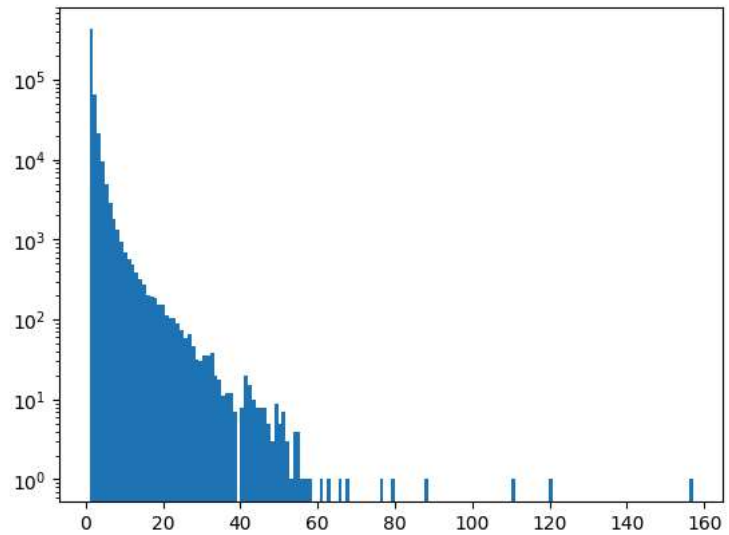
```
0    255027
1    149263
Name: is_duplicate, dtype: int64
0    63.080215
1    36.919785
Name: is_duplicate, dtype: float64
<Axes: >
```




```
# Repeated questions

qid = pd.Series(df['qid1'].tolist() + df['qid2'].tolist())

print('Number of unique questions',np.unique(qid).shape[0])
x = qid.value_counts()>1
print('Number of questions getting repeated',x[x].shape[0])
```

```
Number of unique questions 537933
Number of questions getting repeated 111780
```

```
# Repeated questions histogram

plt.hist(qid.value_counts().values,bins=160)
plt.yscale('log')
plt.show()
```



```
Start coding or generate with AI.
```