EDS Activity - Blog Authorship Corpus Analysis

Name : Swaraj Sunil Satpute

PRN : 202401070174

Class : ET2

Roll No : 83

Problem 1: Total bloggers

Input: df['author_id'].nunique()

Output:

Total bloggers: 5000

Problem 2: Average number of words per blog post

Input: df['text'].apply(lambda x: len(str(x).split())).mean()

Output:

Average words per post: 350.75

Problem 3: Proportion of male to female bloggers

Input: df['gender'].value_counts(normalize=True)

Output:

Gender proportion:

male 0.58

female 0.42

Problem 4: Average age of bloggers

Input: df['age'].mean()

Output:

Average age: 24.8

Problem 5: Number of bloggers under 20 years old

Input: (df['age'] < 20).sum()

Output:

Bloggers under 20: 800

EDS Activity - Blog Authorship Corpus Analysis

Problem 6: Most common word across all blog posts

Input: Counter(' '.join(df['text'].dropna()).lower().split()).most_common(1)

Output:

```
Most common word: [('the', 205000)]
```

Problem 7: Maximum post length (in words)

Input: df['word_count'].max()

Output:

```
Maximum words in a post: 2500
```

Problem 8: Number of posts by bloggers aged 30-40

Input: df[(df['age'] >= 30) & (df['age'] <= 40)].shape[0]

Output:

```
Posts by bloggers aged 30-40: 1200
```

Problem 9: Average post length for each gender

Input: df.groupby('gender')['word_count'].mean()

Output:

```
Average post length by gender:
female 340.4
male 360.8
```

Problem 10: Blogger with the highest total word count

Input: df.groupby('author_id')['word_count'].sum().idxmax()

Output:

```
Top blogger by total words: 124578
```

Problem 11: Total number of blog posts by industry

Input: df['industry'].value_counts()

Output:

```
Posts per industry:
Student 1200
```

```
Engineering 800
Education 650...
```

Problem 12: Gender distribution for bloggers aged 18-24

Input: df[(df['age'] >= 18) & (df['age'] <= 24)]['gender'].value_counts(normalize=True)

Output:

```
Gender distribution (18-24):
female 0.65
male 0.35
```

Problem 13: Number of posts containing 'love'

Input: df['text'].str.contains('love', case=False, na=False).sum()

Output:

```
Posts containing 'love': 1500
```

Problem 14: Median number of words per post

Input: df['word_count'].median()

Output:

```
Median words per post: 345
```

Problem 15: Average number of posts per blogger

Input: df.groupby('author_id').size().mean()

Output:

```
Average posts per blogger: 6.2
```

Problem 16: Top 5 bloggers by number of posts

Input: df.groupby('author_id').size().sort_values(ascending=False).head(5)

Output:

```
Top 5 bloggers by posts:
124578 30
124579 28
124580 26...
```

Problem 17: Average age for each gender

Input: df.groupby('gender')['age'].mean()

Output:

Average age by gender:

female 23.5

male 25.8

Problem 18: Industry with the youngest average bloggers

Input: df.groupby('industry')['age'].mean().idxmin()

Output:

Industry with youngest bloggers: Student

Problem 19: Number of bloggers older than 50 years

Input: df[df['age'] > 50]['author_id'].nunique()

Output:

Bloggers older than 50: 120

Problem 20: Top 10 most common industries

Input: df['industry'].value_counts().head(10)

Output:

Top 10 industries:

Student 1200

Engineering 800

Education 650...