

EDS Activity – 1

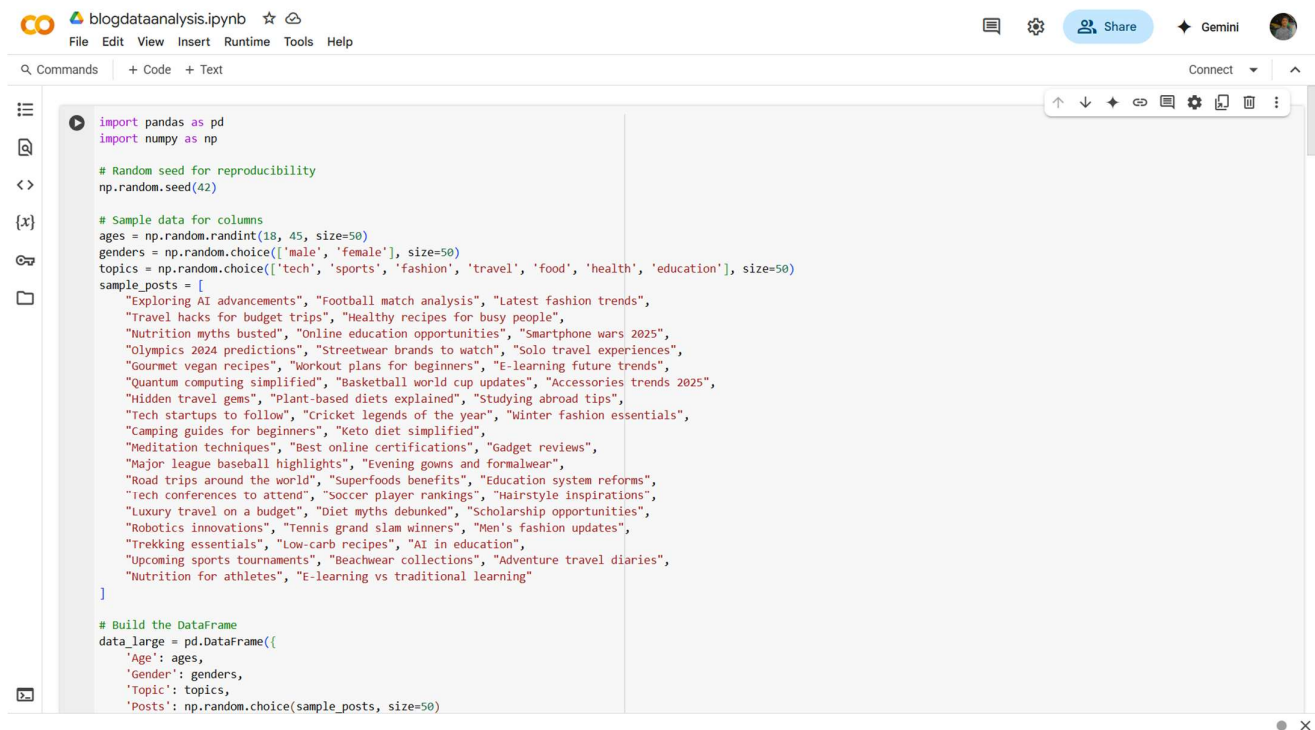
Name: - Dhiraj Ashok Jadhav

Div.: - CS8

Roll No.: - CS8-42

PRN: - 202401120053

Dataset: - The Blog Authorship Corpus



The screenshot shows a Jupyter Notebook interface with a code editor and a sidebar. The code in the editor is as follows:

```
import pandas as pd
import numpy as np

# Random seed for reproducibility
np.random.seed(42)

# Sample data for columns
ages = np.random.randint(18, 45, size=50)
genders = np.random.choice(['male', 'female'], size=50)
topics = np.random.choice(['tech', 'sports', 'fashion', 'travel', 'food', 'health', 'education'], size=50)
sample_posts = [
    "Exploring AI advancements", "Football match analysis", "Latest fashion trends",
    "Travel hacks for budget trips", "Healthy recipes for busy people",
    "Nutrition myths busted", "Online education opportunities", "Smartphone wars 2025",
    "Olympics 2024 predictions", "Streetwear brands to watch", "Solo travel experiences",
    "Gourmet vegan recipes", "Workout plans for beginners", "E-learning future trends",
    "Quantum computing simplified", "Basketball world cup updates", "Accessories trends 2025",
    "Hidden travel gems", "Plant-based diets explained", "Studying abroad tips",
    "Tech startups to follow", "Cricket legends of the year", "Winter fashion essentials",
    "Camping guides for beginners", "Keto diet simplified",
    "Meditation techniques", "Best online certifications", "Gadget reviews",
    "Major league baseball highlights", "Evening gowns and formalwear",
    "Road trips around the world", "Superfoods benefits", "Education system reforms",
    "Tech conferences to attend", "Soccer player rankings", "Hairstyle inspirations",
    "Luxury travel on a budget", "Diet myths debunked", "Scholarship opportunities",
    "Robotics innovations", "Tennis grand slam winners", "Men's fashion updates",
    "Trekking essentials", "Low-carb recipes", "AI in education",
    "Upcoming sports tournaments", "Beachwear collections", "Adventure travel diaries",
    "Nutrition for athletes", "E-learning vs traditional learning"
]

# Build the DataFrame
data_large = pd.DataFrame({
    'Age': ages,
    'Gender': genders,
    'Topic': topics,
    'Posts': np.random.choice(sample_posts, size=50)
})
```

blogdataanalysis.ipynb

File Edit View Insert Runtime Tools Help

Commands Code Text

Build the DataFrame

data_large = pd.DataFrame({
 'Age': ages,
 'Gender': genders,
 'Topic': topics,
 'Posts': np.random.choice(sample_posts, size=50)
})

print(data_large)

	Age	Gender	Topic	Posts
0	24	male	education	Robotics innovations
1	37	female	travel	Cricket legends of the year
2	32	female	tech	Best online certifications
3	28	male	health	Soccer player rankings
4	25	female	food	Exploring AI advancements
5	38	female	food	Soccer player rankings
6	24	female	sports	Luxury travel on a budget
7	43	female	education	Beachwear collections
8	36	male	food	E-learning future trends
9	40	female	sports	Latest fashion trends
10	28	male	tech	Exploring AI advancements
11	28	female	travel	Healthy recipes for busy people
12	41	female	travel	Meditation techniques
13	38	female	travel	E-learning future trends
14	21	male	food	Scholarship opportunities
15	25	female	tech	Best online certifications
16	41	male	food	Olympics 2024 predictions
17	20	female	education	Quantum computing simplified
18	39	male	food	Quantum computing simplified
19	38	female	tech	Meditation techniques
20	19	male	tech	Men's fashion updates
21	41	male	education	Workout plans for beginners
22	29	female	tech	Superfoods benefits
23	23	male	tech	Scholarship opportunities
24	19	female	travel	Nutrition for athletes
25	38	female	education	Superfoods benefits
26	18	female	fashion	Travel hacks for budget trips
27	29	female	fashion	Evening gowns and formalwear
28	43	female	tech	Luxury travel on a budget

blogdataanalysis.ipynb

File Edit View Insert Runtime Tools Help

Commands Code Text

27 29 female fashion Evening gowns and formalwear

28 43 female tech Luxury travel on a budget

29 39 female fashion Winter fashion essentials

30 29 female fashion Scholarship opportunities

31 42 female tech AI in education

32 34 female fashion Quantum computing simplified

33 44 female food Trekking essentials

34 44 female sports Major league baseball highlights

35 27 male education Hairstyle inspirations

36 33 male sports Workout plans for beginners

37 32 female tech Superfoods benefits

38 32 female travel Online education opportunities

39 36 female education Cricket legends of the year

40 29 female tech Gadget reviews

41 40 female travel Football match analysis

42 37 female sports Men's fashion updates

43 42 female tech AI in education

44 20 female education Nutrition myths busted

45 22 male education Gadget reviews

46 36 female health Gadget reviews

47 24 male food Low-carb recipes

48 38 female fashion Low-carb recipes

49 26 female travel Studying abroad tips

[] # 1) Find the average age of all bloggers.

average_age = np.mean(df['Age'])
print("Average Age:", average_age)

Average Age: 30.5

[] #2) Count number of male and female bloggers.

gender_count = df['Gender'].value_counts()
print(gender_count)

Gender
female 5

blogdataanalysis.ipynb

File Edit View Insert Runtime Tools Help

Q Commands + Code + Text

Connect

Share Gemini

Gender
female 5
male 5
Name: count, dtype: int64

[] # 3) Find the most popular blog topic.

popular_topic = df['Topic'].mode()[0]
print("Most Popular Topic:", popular_topic)

Most Popular Topic: fashion

[] # 4: Blogger with maximum age.

oldest_blogger = df[df['Age'] == df['Age'].max()]
print(oldest_blogger)

Age Gender Topic Posts
5 40 male sports Cricket world cup 2024 preview

[] # 5: Blogger with minimum age.

youngest_blogger = df[df['Age'] == df['Age'].min()]
print(youngest_blogger)

Age Gender Topic Posts
4 22 female food Top 10 vegan restaurants 2025

[] # 6: Number of unique blogging topics.

unique_topics = df['Topic'].nunique()
print("Unique Topics:", unique_topics)

blogdataanalysis.ipynb

File Edit View Insert Runtime Tools Help

Q Commands + Code + Text

Connect

Share Gemini

Unique Topics: 5

[] # 7: Bloggers who are older than 30.

older_than_30 = df[df['Age'] > 30]
print(older_than_30)

Age Gender Topic Posts
1 31 male sports Champions league final highlights
3 35 male travel Backpacking through Europe guide
5 40 male sports Cricket world cup 2024 preview
7 36 female fashion Sustainable fashion brands to watch
9 33 female travel Solo travel safety tips

[] # 8: Bloggers who wrote about "food".

food_bloggers = df[df['Topic'] == 'food']
print(food_bloggers)

Age Gender Topic Posts
4 22 female food Top 10 vegan restaurants 2025
8 28 male food Delicious quick breakfast recipes

[] # 9: Total number of words across all posts.

total_words = df['Posts'].apply(lambda x: len(x.split())).sum()
print("Total Words:", total_words)

Total Words: 44

[] # 10: Which post has the maximum number of words?

longest_post_idx = df['Posts'].apply(lambda x: len(x.split())).idxmax()

blogdataanalysis.ipynb

File Edit View Insert Runtime Tools Help

Share Gemini

Commands Code Text Connect

[] # 10: Which post has the maximum number of words?

longest_post_idx = df['Posts'].apply(lambda x: len(x.split())).idxmax()
print(df.loc[longest_post_idx])

Age 24
Gender female
Topic tech
Posts AI innovations changing the world
Name: 0, dtype: object

[] # 11: Calculate the mean age for each blogging topic.

mean_age_topic = df.groupby('Topic')['Age'].mean()
print(mean_age_topic)

Topic
fashion 31.5
food 25.0
sports 35.5
tech 26.5
travel 34.0
Name: Age, dtype: float64

[] # 12: How many bloggers have age between 25 and 35?

age_between = df[(df['Age'] >= 25) & (df['Age'] <= 35)].shape[0]
print("Bloggers aged 25-35:", age_between)

Bloggers aged 25-35: 6

[] # 13: What percentage of bloggers are writing about 'tech'?

blogdataanalysis.ipynb

File Edit View Insert Runtime Tools Help

Share Gemini

Commands Code Text Connect

[] # 13: What percentage of bloggers are writing about 'tech'?

tech_percentage = (df['Topic'].value_counts(normalize=True)['tech']) * 100
print("Tech bloggers %:", tech_percentage)

Tech bloggers %: 20.0

[] # 14: Create new column "Post_Length" (word counts).

df['Post_Length'] = df['Posts'].apply(lambda x: len(x.split()))
print(df[['Posts', 'Post_Length']])

Posts Post_Length
0 AI innovations changing the world 5
1 Champions league final highlights 4
2 Streetwear trends for summer 4
3 Backpacking through Europe guide 4
4 Top 10 vegan restaurants 2025 5
5 Cricket world cup 2024 preview 5
6 New smartphones launch comparison 4
7 Sustainable fashion brands to watch 5
8 Delicious quick breakfast recipes 4
9 Solo travel safety tips 4

[] # 15: Find topic with bloggers having highest average post length.

topic_max_post_len = df.groupby('Topic')['Post_Length'].mean().idxmax()
print("Topic with longest posts:", topic_max_post_len)

Topic with longest posts: fashion

[] # 16: Sort bloggers by Age ascending.

sorted_asc = df.sort_values(by='Age')

blogdataanalysis.ipynb

File Edit View Insert Runtime Tools Help

Q Commands + Code + Text

Connect

16: Sort bloggers by Age ascending.

```
sorted_asc = df.sort_values(by='Age')
print(sorted_asc[['Age', 'Gender', 'Topic']])
```

	Age	Gender	Topic
4	22	female	food
0	24	female	tech
2	27	female	fashion
8	28	male	food
6	29	male	tech
1	31	male	sports
9	33	female	travel
3	35	male	travel
7	36	female	fashion
5	40	male	sports

17: Find the median age of bloggers.

```
median_age = np.median(df['Age'])
print("Median Age:", median_age)
```

Median Age: 30.0

18: Group bloggers by gender and find average Post_Length.

```
avg_post_gender = df.groupby('Gender')['Post_Length'].mean()
print(avg_post_gender)
```

Gender	
female	4.6
male	4.2

Name: Post_Length, dtype: float64

19: Bloggers writing about travel.

```
travel_bloggers = df[df['Topic'] == 'travel']
print(travel_bloggers)
```

blogdataanalysis.ipynb

File Edit View Insert Runtime Tools Help

Q Commands + Code + Text

Connect

19: Bloggers writing about travel.

```
travel_bloggers = df[df['Topic'] == 'travel']
print(travel_bloggers)
```

	Age	Gender	Topic	Posts	Post_Length
3	35	male	travel	Backpacking through Europe guide	4
9	33	female	travel	Solo travel safety tips	4

20: Bloggers with posts having less than 6 words.

```
travel_bloggers = df[df['Topic'] == 'travel']
print(travel_bloggers)
```

	Age	Gender	Topic	Posts	Post_Length
3	35	male	travel	Backpacking through Europe guide	4
9	33	female	travel	Solo travel safety tips	4

21: Find the proportion of bloggers aged above 30.

```
above_30 = (df['Age'] > 30).sum()
proportion_above_30 = above_30 / df.shape[0]
print("Proportion of bloggers aged above 30:", proportion_above_30)
```

Proportion of bloggers aged above 30: 0.5

22: Find the proportion of bloggers aged above 30.

```
above_30 = (df['Age'] > 30).sum()
proportion_above_30 = above_30 / df.shape[0]
print("Proportion of bloggers aged above 30:", proportion_above_30)
```

Proportion of bloggers aged above 30: 0.5