

ESSENTIALS OF DATA SCIENCE

Theory Activity No. 1

Name – Rushikesh Ganesh Shinde

PRN – 202401040040

Roll NO – CS2-11

Div – CS2

❖ 20 problem statements for **Kaggle Text Classification Dataset**
using Numpy and Pandas.

1) Numpy Question and Answers :-

```
Untitled0.ipynb
File Edit View Insert Runtime Tools Help
Q Commands + Code + Text
RUSHIKESH GANESH SHINDE CS2-11 ROLL NO 11 PRN- 202401040040

**PANDAS AND NUMPY CODE**

[24] # Import necessary libraries
import pandas as pd
import numpy as np
df = pd.read_csv('/train_40k.csv')

# 1. What is the average helpfulness ratio using NumPy?
df[['HelpfulVotes', 'TotalVotes']] = df['Helpfulness'].str.split('/', expand=True).astype(int)
df['HelpfulnessRatio'] = df['HelpfulVotes'] / df['TotalVotes'].replace(0, np.nan)
average_ratio = np.nanmean(df['HelpfulnessRatio'].values)

print("Average Helpfulness Ratio:", average_ratio)
Average Helpfulness Ratio: 0.803883678903371

[27] # 2. Find the median review score using NumPy
np.median(df['Score'].values)
np.float64(5.0)

[28] # 3. Get the standard deviation of scores
np.std(df['Score'].values)
```

```
Untitled0.ipynb
File Edit View Insert Runtime Tools Help
Q Commands + Code + Text
np.float64(1.3570742313429285)

[29] # 4. Convert Score column to a NumPy array and display its shape
score_array = df['Score'].values
score_array.shape
(40000,)

[30] # 5. How many reviews have helpful votes greater than 2?
np.sum(df['HelpfulVotes'].values > 2)
np.int64(10650)

[31] # 6. Create a boolean mask for reviews with Score >= 4
mask = df['Score'].values >= 4
mask[:5]
array([False,  True,  True,  True,  True])

[32] # 7. Use NumPy to find the indices of reviews with Score = 1
np.where(df['Score'].values == 1)[0][:5]
array([ 11,  74,  91, 138, 181])
```

```
Q Commands + Code + Text
[33] # 8. Normalize the 'Score' column (min-max normalization)
normalized_score = (df['Score'] - df['Score'].min()) / (df['Score'].max() - df['Score'].min())
normalized_score.head()
Score
0    0.5
1    1.0
2    1.0
3    1.0
4    1.0
dtype: float64

[34] # 9. Count how many reviews have TotalVotes == 0 using NumPy
np.sum(df['TotalVotes'].values == 0)
np.int64(15149)

# 10. What is the max and min Score using NumPy?
np.max(df['Score'].values), np.min(df['Score'].values)
(np.float64(5.0), np.float64(1.0))
```

2) Pandas Question and Answers :-

```
Untitled0.ipynb
File Edit View Insert Runtime Tools Help
Q Commands + Code + Text
[14] # Import necessary libraries
import pandas as pd
import numpy as np
df = pd.read_csv('/train_40k.csv')

import numpy as np

# 1. How many missing values are there in each column?
df.isnull().sum()
```

	0
productId	0
Title	16
userId	0
Helpfulness	0
Score	0
Time	0
Text	0
Cat1	0
Cat2	0
Cat3	0

```
Untitled0.ipynb
File Edit View Insert Runtime Tools Help
Q Commands + Code + Text
[12] # 2. What are the top 5 most reviewed products?
df[['productId']].value_counts().head()
```

	count
productId	
B000FSFNUE	146
B00003TL7P	94
B0002DK2DU	85
B000GLRREU	83
B0009V1YR8	79

dtype: int64

```
# 3. What is the average score of all products?
df['Score'].mean()
np.float64(4.070179)
```

```
# 4. How many unique users are in the dataset?
df['userId'].nunique()
36298
```

0s completed at 11:12 PM

```
Untitled0.ipynb
File Edit View Insert Runtime Tools Help
Q Commands + Code + Text
[17] # 5. What are the most common categories in Cat1?
df['Cat1'].value_counts().head()
```

	count
Cat1	
toys games	10266
health personal care	9772
beauty	5846
baby products	5637
pet supplies	4862

dtype: int64

```
# 6. Convert 'Helpfulness' column into two numeric columns: 'HelpfulVotes' and 'TotalVotes'
df[['HelpfulVotes', 'TotalVotes']] = df['Helpfulness'].str.split('/', expand=True).astype(int)
df[['HelpfulVotes', 'TotalVotes']].head()
```

	HelpfulVotes	TotalVotes
0	0	0
1	0	0
2	0	0
3	0	0

0s completed at 11:12 PM

Untitled0.ipynb

File Edit View Insert Runtime Tools Help

Q Commands + Code + Text

RAM Disk

```
[19] # 7. What is the proportion of helpful votes?
df['helpfulnessRatio'] = df['HelpfulVotes'] / df['TotalVotes'].replace(0, np.nan)
df['helpfulnessRatio'].mean()

np.float64(0.803803678903371)
```

```
[20] # 8. How many reviews have a score of 5?
(df['Score'] == 5).sum()

np.int64(23362)
```

```
# 9. Get all rows where 'Title' is missing.
df[df['Title'].isnull()]
```

5202	B0009S3K1E	NaN	A2PWSZJ1Q21NYQ	0/1	5.0	1131408000	I bought this for my wife last christmas for l...	beauty	fragrance	unknown	0	1	0.000000	
7340	B0001Y6DUG	NaN	AW18N7F16XTR4	0/0	5.0	1163030400	This is one of my favorite lip balms ever! If...	beauty	makeup	lips	0	0	NaN	
9374	B00028MIKK	NaN	A1H7CKOA2DQJU0	2/2	5.0	1178409600	This tea is really, really good -- it's very s...	grocery	gourmet food	beverages	tea	2	2	1.000000

0s completed at 11:12 PM

Untitled0.ipynb

File Edit View Insert Runtime Tools Help

Q Commands + Code + Text

RAM Disk

12117	B000KGRNMM	NaN	A2EVLVSF7IDC6	99/100	5.0	1199232000	take care of my Mother...	personal care	supplies	equipment	daily living aids	99	100	0.990000
13089	B000GUFFQS	NaN	A1P0IHMP5WVOZ	1/1	5.0	1204934400	My 93 year old father in law recently fell (u...	health personal care	medical supplies	equipment	beds accessories	1	1	1.000000
14881	B000KGRNMM	NaN	A3V0HBE74ZUFKB	52/52	5.0	1218672000	I am delighted with my nail care kit. I am in ...	health personal care	medical supplies	equipment	daily living aids	52	52	1.000000
18850	B000634EEO	NaN	A3A6B6RYANZYJN	6/6	1.0	1244937600	We bought this yesterday at Petco. Our betta, ...	pet supplies	fish aquatic	aquarium starter kits	6	6	1.000000	
19395	B000GUFFQS	NaN	A1W0PU8J06LJU	3/3	5.0	1249603200	I put this on my grandmother's bed. It was ver...	health personal care	medical supplies	equipment	beds accessories	3	3	1.000000
20193	B000GUFFQS	NaN	A3URJ91T1WXQHE	1/1	5.0	1255392000	My father, who had a stroke several years ago...	health personal care	medical supplies	equipment	beds accessories	1	1	1.000000
24905	B000GUFFQS	NaN	A3IEU3BT51AUKF	0/0	5.0	1284336000	My mother was issued a hospital bed to prevent...	health personal care	medical supplies	equipment	beds accessories	0	0	NaN
25759	B000PC4KX6	NaN	A2S8ECPKOH229L	1/1	4.0	1288828800	I bought these after they were used on my son ...	health personal care	medical supplies	equipment	health monitors	1	1	1.000000
26912	B000GUFFQS	NaN	A2FXS0OQKM3JQD	0/0	4.0	1294012800	Purchased it for older neighbor and she loves ...	health personal care	medical supplies	equipment	beds accessories	0	0	NaN

0s completed at 11:12 PM

Untitled0.ipynb

File Edit View Insert Runtime Tools Help

Q Commands + Code + Text

RAM Disk

26912	B000GUFFQS	NaN	A2FXS0OQKM3JQD	0/0	4.0	1294012800	Purchased it for older neighbor and she loves ...	health personal care	medical supplies	equipment	beds accessories	0	0	NaN
31253	B000PJNRC	NaN	A1J9ELJUBARM1	1/1	5.0	1312588800	Handy prepackaged alcohol wipes a great bulk p...	health personal care	health care	first aid	1	1	1.000000	
32483	B000GUFFQS	NaN	A2TH88LHAR9KXC	2/2	5.0	1317772800	I ordered this bedcane for my husband who is 6...	health personal care	medical supplies	equipment	beds accessories	2	2	1.000000
39455	B000GUFFQS	NaN	A1H5TCWJB5CV77	0/0	5.0	1342224000	I purchased another bedrail that was too low f...	health personal care	medical supplies	equipment	beds accessories	0	0	NaN

```
# 10. Convert 'Time' column from UNIX timestamp to readable date
df['ReviewDate'] = pd.to_datetime(df['Time'], unit='s', errors='coerce')
df[['Time', 'ReviewDate']].head()
```

	Time	ReviewDate
0	-1	1969-12-31 23:59:59
1	860630400	1997-04-10 00:00:00
2	883008000	1997-12-25 00:00:00
3	897696000	1998-06-13 00:00:00
4	911865600	1998-11-24 00:00:00