ESSENTIALS OF DATA SCIENCE THEORY ACTIVITY NO.1

Name: Amayara Gani

Division: CS2

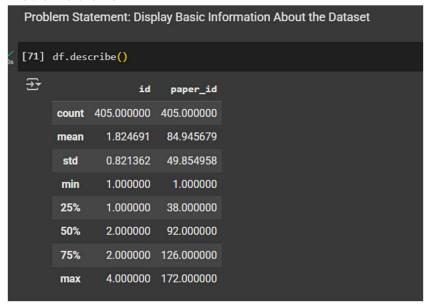
Roll No: CS2-61

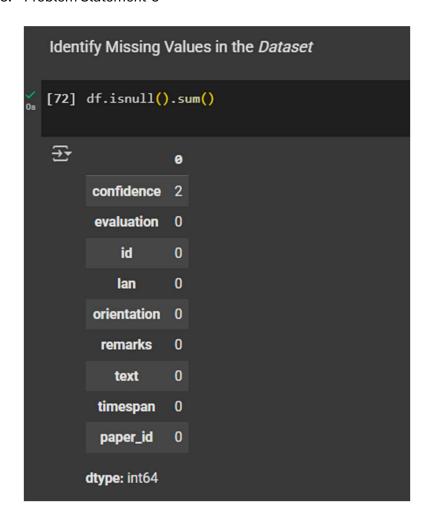
PRN: 202401040273

```
Dataset Link: https://archive.ics.uci.edu/dataset/410/paper+reviews

Import Libraries

[93] import numpy as np
import pandas as pd
```



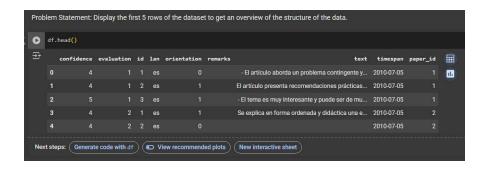


```
Problem Statement: Find shape of the given dataset

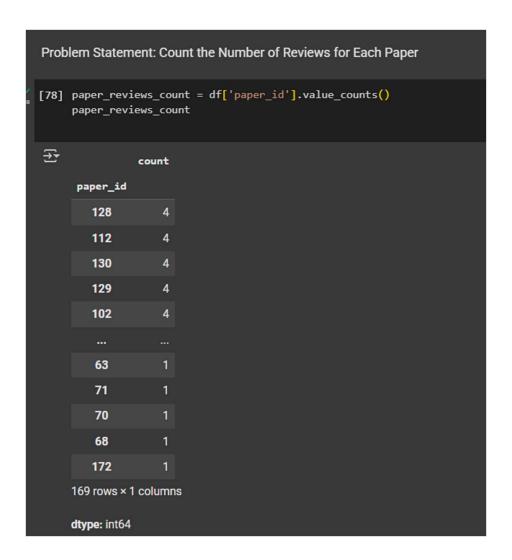
[73] df.shape

(405, 9)
```

5. Problem Statement-5



```
Problem Statement: Find the papers that received the most reviews.
[77] paper_reviews_count = df['paper_id'].value_counts()
     most_reviewed_papers = paper_reviews_count.head(10)
     print(most_reviewed_papers)
→ paper_id
     128
           4
     112
            4
     130
           4
     129
     102
     80
           4
     92
           4
     50
           4
     55
           4
     40
     Name: count, dtype: int64
```



```
Problem Statement: Filter reviews where the evaluation is very positive (evaluation = 2).

[79] positive_reviews = df[df['evaluation'] == 2]
    print(positive_reviews[['paper_id', 'text', 'evaluation']].head())

The statement: Filter reviews where the evaluation'] == 2]
    print(positive_reviews[['paper_id', 'text', 'evaluation']].head())

The statement: Filter reviews where the evaluation is very positive (evaluation = 2).
```

```
Problem Statement: Count the number of reviews written in Spanish (lan == 'es') and English (lan == 'en')

[80] language_distribution = df['lan'].value_counts()
print(language_distribution)

2 lan
es 388
en 17
Name: count, dtype: int64
```

12. Problem Statement-12

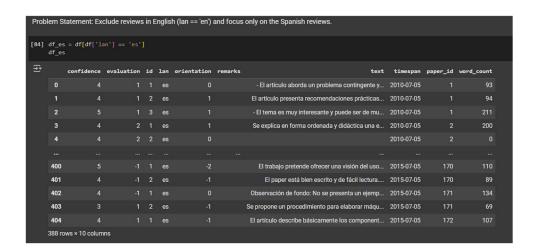
```
Problem Statement: Filter out reviews that have fewer than 10 words.

[81] df['word_count'] = df['text'].apply(lambda x: len(x.split()))
filtered_reviews = df[df['word_count'] >= 10]
print(filtered_reviews[['paper_id', 'text', 'word_count']].head())

paper_id text word_count
0 1 - El artículo aborda un problema contingente y... 93
1 El artículo presenta recomendaciones prácticas... 94
2 1 - El tema es muy interesante y puede ser de mu... 211
3 2 Se explica en forma ordenada y didáctica una e... 200
5 2 Los autores describen una metodología para des... 299
```



15. Problem Statement-15



```
Problem Statement: Display the last 5 reviews where the evaluation score is -2 (very negative).

[85] print(df[df['evaluation'] == -2].tail())

Empty DataFrame
Columns: [confidence, evaluation, id, lan, orientation, remarks, text, timespan, paper_id, word_count]
Index: []
```

```
Problem Statement: Check if there are any duplicate rows in the dataset.

[86] print(df.duplicated().sum())

3. 0
```

```
Problem Statement: Sort the dataset by the 'timespan' column in descending order.
[87] print(df.sort_values(by='timespan', ascending=False))
                          confidence evaluation id lan orientation \
                                                                                                              evaluation id land of the second seco
                            346
                            347
                             348
                            404
                             364 El artículo aborda una temática de gran releva...
                             347 Si se considera un track para mostrar proyecto...
                            348 Desde mi punto de vista, es un trabajo regular...
                           83
84
                            85
                          text timespan
404 El artículo describe básicamente los component... 2015-07-05
364 Los datos abiertos de Chile Este trabajo anali... 2015-07-05
346 el trabajo presenta una posible solución a una... 2015-07-05
347 El artículo presenta el desarrollo de una apli... 2015-07-05
                                                                                                                                                                                                                                                                                                                                                    timespan paper_id \
                                                                                                                                                                                                                                                                                                                                                                                                                                                  172
```

```
348 Es una buena aplicación, que usa los recursos ... 2015-07-05
                                                                                144
82 Es un excelente trabajo de investigación de ap... 2010-07-05
                                                                                 31
83 Este es un trabajo que muestra la experiencia ... 2010-07-05
                                                                                 31
              Me ha gustado mucho Lo dejaría tal cual. 2010-07-05
84
                                                                                 32

    La ponencia es muy completa es un tema importa... 2010-07-05
    El artículo aborda un problema contingente y... 2010-07-05

                                                                                 32
     word_count
404
            107
364
            476
346
            217
            182
347
348
            203
82
             66
83
            106
84
             8
85
             21
0
             93
[405 rows x 10 columns]
```

```
Problem Statement: Get the count of unique values in the 'lan' column (language column).

[88] print(df['lan'].value_counts())

The search of the count of unique values in the 'lan' column (language column).

[88] print(df['lan'].value_counts())

The search of the count of unique values in the 'lan' column (language column).
```